MATRIX Book Series 4

David R. Wood *Editor-in-Chief*
Jan de Gier
Cheryl E. Praeger
Terence Tao *Editors*

2019-20 MATRIX Annals

MATRI

Springer

MATRIX Book Series 4

MATRIX is Australia's international and residential mathematical research institute. It facilitates new collaborations and mathematical advances through intensive residential research programs, each lasting 1–4 weeks.

More information about this series at https://link.springer.com/bookseries/15890

David R. Wood
Editor-in-Chief

Jan de Gier · Cheryl E. Praeger · Terence Tao
Editors

2019-20 MATRIX Annals

ISSN 2523-3041 ISSN 2523-305X (electronic)
MATRIX Book Series
ISBN 978-3-030-62496-5 ISBN 978-3-030-62497-2 (eBook)
https://doi.org/10.1007/978-3-030-62497-2

Mathematics Subject Classification: 00-XX

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

MATRIX is Australia's international and residential mathematical research institute. It was established in 2015 and launched in 2016 as a joint partnership between Monash University and The University of Melbourne, with seed funding from the ARC Centre of Excellence for Mathematical and Statistical Frontiers. In 2020, The Australian National University joined MATRIX in a three-way partnership. The purpose of MATRIX is to facilitate new collaborations and mathematical advances through intensive residential research programs, which are currently held in Creswick, a small town nestled in the beautiful forests of the Macedon Ranges, 130km west of Melbourne.

This book is a scientific record of the ten programs held at MATRIX in 2019 and two programs held in January 2020:

- *Topology of Manifolds: Interactions Between High and Low Dimensions*
  Guest editors: Diarmuid Crowley, Stefan Friedl, Stephan Tillmann
- *Australian-German Workshop on Differential Geometry in the Large*
- *Aperiodic Order meets Number Theory*
  Guest editors: Michael Baake and Uwe Grimm
- *Ergodic Theory, Diophantine Approximation and Related Topics*
  Guest editor: Mumtaz Hussain
- *Influencing Public Health Policy with Data-informed Mathematical Models of Infectious Diseases*
  Guest editor: Jennifer Flegg
- *International Workshop on Spatial Statistics*
  Guest editor: Pavel Krupskiy
- *Mathematics of Physiological Rhythms*
  Guest editor: Maia Angelova
- *Conservation Laws, Interfaces and Mixing*
  Guest editors: Snezhana I. Abarzhi, Alexander Nepomnyashchy, Anthony J. Roberts, Joseph Klewicki
- *Structural Graph Theory Downunder*
  Guest editor: Anita Liebenau
- *Tropical Geometry and Mirror Symmetry*
  Guest editor: Mandy Cheung
- *Early Career Researchers Workshop on Geometric Analysis and PDEs*
  Guest editor: Paul Bryan
- *Harmonic Analysis and Dispersive PDEs: Problems and Progress*
  Guest editor: Kenji Nakanishi

The MATRIX Scientific Committee selected these programs based on scientific excellence and the participation rate of high-profile international participants. This committee consists of: David Wood (Monash Uni., Chair), Ben Andrews (Australian National Uni.), Santiago Badia (Monash Uni.), Peter Bouwknegt (Australian National Uni.), Peter Bühlmann (ETH Zurich), Alison Etheridge (Uni. Oxford), Jan de Gier (Uni. Melbourne), Cecilia González Tokman (Uni. Queensland), Frances Kuo

(UNSW Sydney), Joshua Ross (Uni. Adelaide), Terence Tao (Uni. California, Los Angeles), Ole Warnaar (Uni. Queensland), and Geordie Williamson (Uni. Sydney).

These programs involved organisers from a variety of Australian universities, including Adelaide, Deakin, LaTrobe, Macquarie, Monash, Melbourne, Newcastle, UNSW, Sydney, Western Australia, along with international organisers and participants.

Each program lasted 1–4 weeks, and included ample unstructured time to encourage collaborative research. Some of the longer programs had an embedded conference or lecture series. All participants were encouraged to submit articles to the MATRIX Annals.

The articles were grouped into refereed contributions and other contributions. Refereed articles contain original results or reviews on a topic related to the MATRIX program. The other contributions are typically lecture notes or short articles based on talks or activities at MATRIX. A guest editor organised appropriate refereeing and ensured the scientific quality of submitted articles arising from each program. The Editors (Jan de Gier, Cheryl E. Praeger, Terence Tao and myself) finally evaluated and approved the papers.

Many thanks to the authors and to the guest editors for their wonderful work.

MATRIX is hosting 12 programs in 2021, with more to come beyond that; see `www.matrix-inst.org.au`. Our goal is to facilitate collaboration between researchers in universities and industry, and increase the international impact of Australian research in the mathematical sciences.

David R. Wood
MATRIX Annals Editor-in-Chief

Topology of Manifolds: Interactions Between High and Low Dimensions

7 – 18 January 2019

Organisers

Jonathan Bowden
Uni. Regensburg

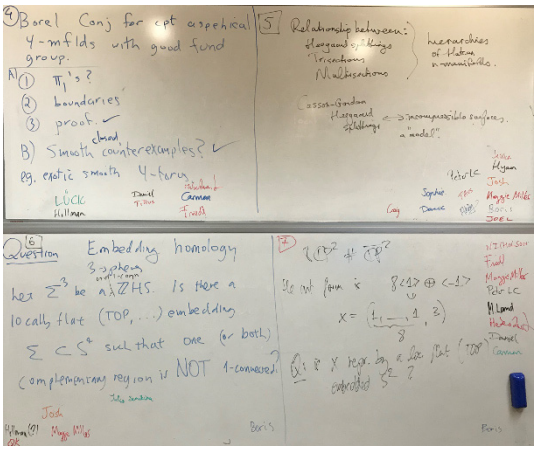
Diarmuid Crowley
Uni. Melbourne

Stefan Friedl
Uni. Regensburg

Stephan Tillmann
Uni. Sydney

Jim Davis
Indiana Uni.

Carmen Rovi
Uni. Heidelberg



Participants

Jonathan Bowden (Monash Uni.), Diarmuid Crowley (Uni. Melbourne), Jim Davis (Indiana Uni.), Stefan Friedl (Uni. Regensburg), Carmen Rovi (Indiana Uni.), Stephan Tillmann (Uni. Sydney), Wolfgang Lück (Uni. Bonn, Germany), Andras Stipsicz (Hungarian Acad. Sci.), Bea Bleile (Uni. New England), Jessica Purcell (Monash Uni.), Jae Choon Cha (POSTECH), Abby Thompson (Uni. California, Davis), Ana Lecuona (Uni. Glasgow), Daniel Kasprowski (Uni. Bonn), Imi Bokor, Jonathan Hillman (Uni. Sydney), Craig Hodgson (Uni. Melbourne), Hyam Rubinstein (Uni. Melbourne), Markus Land (Uni. Regensburg), Boris Lishak (Uni. Sydney), Fabian Hebestreit (Uni. Bonn), Ruth Kellerhals (Uni. Fribourg), Joel Hass (Uni. California, Davis), Julia Semikina (Uni. Bonn), Christoph Winges (Uni. Bonn), Fabian Henneke (Uni. Bonn), Fabio Gironella (Alfred Renyi Inst.), Josh Howie (Monash Uni.), Anthony Conway (Uni. Durham), Sylvain Cappell (Courant Inst.), Irving Dai (Princeton Uni.), Jen Hom (Georgia Tech.), Qayum Khan (Saint Louis Uni.), Peter Lambert-Cole (Georgia Tech.), Adam Levine (Duke Uni.), Duncan McCoy (Uni. Texas Austin), Maggie Miller (Princeton Uni.), Kent Orr (Indiana Uni.), Lisa Piccirillo (Uni. Texas Austin), Linh Truong (Columbia Uni.), Min Hoon Kim (Korea Inst. Advanced Study), Csaba Nagy (Uni. Melbourne), Johanna Meumertzheim (Uni. Regensburg), Dominic Tate (Uni. Sydney), Huijun Yang (Henan Uni.), Johnny Nicholson (Uni. College London), Sophie Ham (Monash Uni.), Kevin Yin (Courant Inst.)

This workshop explored connections between the study of manifolds in high and low dimensions, via the comparison of phenomena and methods across dimensions and via analysing higher dimensional spaces in terms of lower-dimensional subspaces.

Low-dimensional spaces ($n \leqslant 4$) appear naturally in physics, for example as the dimensions of space and space-time, and exhibit unique phenomena. Higher dimensional spaces ($n > 4$) arise as the parameters spaces of complex systems. The areas of low-dimensional topology and high-dimensional topology have developed rather independently since the days of Milnor and Smale, reflecting the differing nature of problems in dimensions three and four and in higher dimensions. In dimension three Thurston's geometrisation program led to the possibility of a complete classification of 3–manifolds. Dimension four is marked by the failure of the Whitney trick and is intermediate between high and low dimensions. In dimensions five and higher, surgery theory and smoothing theory provide powerful tools for analysing manifolds.

The workshop was organised around three key elements, listed here in the order in which they made their first entrance during the two-week program:

- **Lecture Series** designed to provide bridges between the different areas represented at the workshop,
- **Problem Sessions and Working Groups** designed to stimulate interaction and collaboration between researchers from different areas, and
- **Research Talks** addressed at a wide audience.

Lecture Series

Three lecture series were given in the mornings of the first week of the workshop, and were supported by discussion and exercise sessions in the afternoons.

Surgery: high-dimensional methods in low dimensions
by Diarmuid Crowley, Jim Davis and Kent Orr

This lecture series gave an introduction to topological 4–manifolds, normal maps and the surgery obstruction, reviewing the work of Wall and Cappell-Shaneson and the stable s/h cobordism theorem, as well as Kreck's surgery machine for classification. This led to the stable classification of $2q$–manifolds and in particular 4–manifolds. Further topics included the Q–form conjecture, application of the surgery machine in low dimensions and an overview of the current state of knowledge concerning topological concordance of classical knots.

The (stable) Cannon Conjecture
by Wolfgang Lück

Starting with an introduction of 3–manifold theory and properties of hyperbolic groups, this lecture series centred around the statement of Cannon's conjecture that a torsionfree hyperbolic group has the 2–sphere as its boundary if and only if it is the fundamental group of a closed hyperbolic 3–manifold. After a discussion of

topological rigidity and L^2–invariants, the lectures culminated in a sketch of the recent proof by Ferry, Lück and Weinberger of the Stable Cannon Conjecture.

Invariants of knots from Heegaard Floer homology
by András Stipsicz

The third lecture series moved from the theory of Heegaard diagrams of 3–manifolds to the definitions and properties of Floer homology theory and Knot Floer homology theory. With these tools in hand, the invariant Υ_K of a knot K invented by Ozsváth, Stipsicz, and Szabó was defined, and numerous applications of this concordance invariant were given.

Problem Sessions and Working Groups

Key elements to our workshop were the organised problem sessions and working groups. We had invited participants working in different areas of topology, and many of them had never met before, let alone glanced at each other's work. Before the workshop, we encouraged participants to submit difficult problems that they feel cannot be tackled from one viewpoint alone, or which aim to translate methods or insights from one area to another.

During the lunch break before our first Open Problem Session on the Monday afternoon of the first week, we asked participants to write their problems on the boards in the main lecture hall. During the session, they then had five minutes (or thereabouts) to explain their problem and answer questions. The session concluded with each participant writing their name next to every problem they were interested in. An example of this is shown in the image above.

The organisers then looked at which problems made sense to run concurrently and allocated time and space for groups to meet and work on a subset of the problems. A deciding factor was to create diverse groups, bringing together researchers from different areas and career stages. The groups would meet each day of the workshop, and we also had regular sessions with all participants in which the working groups reported on progress, asked for input and received feedback. This gave the opportunity to shift focus (for instance, after declaring victory or defeat on a problem), to move to other problems that were initially posed or to formulate new ones.

MATRIX house, which allowed participants to wander from one working group to another, provided an ideal environment for this flexible and collaborative approach.

The problems and progress reports were collected on the online platform Manifold Atlas, `http://www.map.mpim-bonn.mpg.de/`, where we expect to keep track of these and related problems.

Several new participants joined in the second week of the workshop, which therefore included another Open Problem Session to expand and continue the work done in the first week. The papers and the problem list published in this book, as well as additional publications, and the progress reported in the Manifold Atlas are testament to the fruitful interactions at the workshop and indicate that there is scope for deeper synergy between these areas.

Research Talks

The second week featured research talks by invited speakers in the mornings, ranging from graduate students to seasoned and established experts, and covering all aspects of this program.

Bea Bleile (Uni. New England)
Homotopy Types of Poincare Duality Complexes
Sylvain Cappell (NYU Courant Institute)
Using Atiyah-Bott classes to produce polynomial invariants of 3–manifolds
Jae Choon Cha (Postech)
Freely slicing good boundary links with a homotopically trivial plus property
Jen Hom (Georgia Tech)
An infinite-rank summand of the homology cobordism group
Qayum Khan (Saint Louis Uni.)
Stable existence of incompressible 3–manifolds in 4–manifolds
Daniel Kasprowski (Uni. Bonn)
$\mathbb{C}P^2$*–stable diffeomorphism of 4–manifolds*
Peter Lambert-Cole (Georgia Tech.)
Bridge trisections and the Thom conjecture
Markus Land (Uni. Regensburg)
A vanishing theorem for tautological classes of aspherical manifolds
Ana Lecuona (Uni. Glasgow)
Torus knots and rational homology balls
Adam S. Levine (Duke Uni.)
Simply-connected, spineless 4–manifolds
Maggie Miller (Princeton Uni.)
Extending fibrations from knot complements to ribbon disk complements
Csaba Nagy (Uni. Melbourne)
The Q–form conjecture for some 1–connected manifolds
Lisa Piccirillo (Uni. Texas Austin)
The Conway knot is not slice
Jessica Purcell (Monash Uni.)
Combinatorial criteria to determine whether a state surface is a fiber
Hyam Rubinstein (Uni. Melbourne)
Multisections of PL manifolds
Abigail Thompson (Uni. California Davis)
Trisections and surgery questions on links in 3–manifolds
Christoph Winges (Uni. Bonn)
Mapping class groups of high-dimensional, aspherical manifolds

Conclusion

The papers collected in this volume give evidence that this workshop did indeed achieve its aim of stimulating new work through the interaction of topologists working in different subfields that do not usually meet. The organisers expect further work to be submitted elsewhere to materialise in the near future.

We are writing this document in August 2020, a time when most conferences and workshops planned for the current year have been cancelled or been moved to on-line formats. The energetic discussions in Creswick often lasted over many hours spent at blackboards with intermittent walks through the bush. They resulted in knowledge transfer and progress in research that would have otherwise not seemed possible. Such intensive interaction is difficult to accomplished via on-line solutions with the currently available technology. We hope that what now feels like a distant past will become a (virtual?) reality in the not so distant future.

Diarmuid Crowley, Stefan Friedl, Stephan Tillmann
Guest editors

Australian-German Workshop on Differential Geometry in the Large

4 – 15 February 2019

Organisers

Owen Dearricott
Uni. Melbourne

Diarmuid Crowley
Uni. Melbourne

Thomas Leistner
Uni. Adelaide

Yuri Nikolayevsky
LaTrobe Uni.

Wilderich Tuschmann
Karlsruhe Uni.

Katrin Wendland
Freiburg Uni.

Participants

Diarmuid Crowley (Uni. Melbourne), Owen Dearricott (Uni. Melbourne), Thomas Leistner (Uni. Adelaide), Wilderich Tuschmann (Karlsruhe Institute of Technology), Yuri Nikolayevsky (La Trobe Uni.), Ben Andrews (Australian National Uni.), Burkhard Wilking (Uni. Muenster), Christoph Böhm (Uni. Muenster), Claude LeBrun (Stony Brook Uni.), Thomas Farrell (Tsinghua Uni.), Frances Kirwan (Oxford Uni.), Fuquan Fang (Capital Normal Uni.), Guofang Wei (Uni. California Santa Barbara), Neil Trudinger (Australian National Uni.), Peter Petersen (Uni. California Los Angeles), Robert Bryant (Duke Uni.), Rod Gover (Uni. Auckland), Ramiro Lafuente (Uni. Queensland), Karsten Grove (Uni. Notre Dame), Sebastian Goette (Uni. Freiburg), Lashi Bandara (Uni. Potsdam), Katharina Neusser (Masaryk Uni.), Artem Pulemotov (Uni. Queensland), Jesse Gell-Redman (Uni. Melbourne), Lee Kennard (Syracuse Uni.), Haotian Wu (Uni. Sydney), Paul Bryan (Macquarie Uni.), Julian Scheuer (Uni. Freiburg), Krishnan Shankar (Uni. Oklahoma), Xianzhe Dai (Uni. California Santa Barbara), Fernando Galaz-Garcia (Karlsruhe Institute of Technology), Valentina Wheeler (Uni. Wollongong), Julie Clutterbuck (Monash Uni.), Martin Kerin (Uni. Muenster), Fred Wilhelm (Uni. California Riverside), Catherine Searle (Wichita State Uni.), Mathew Langford (Uni. Tennessee, Knoxville), Uwe Semmelmann (Uni. Stuttgart), Joseph Wolf (Uni. California Berkeley), Tracy Payne (Idaho State Uni.), Boris Vertman (Uni. Oldenburg), Pedro Solarzano (UNAM-CONACYT Oaxaca), Jim Davis (Indiana Uni.), Lorenz Schwachhoefer (TU Dortmund), Stephan Klaus (MFO Oberwolfach/Uni. Mainz), Klaus Kröncke (Uni. Hamburg), Matthias Ludewig (Uni. Adelaide), Vicente Cortes (Uni. Hamburg), Vladimir

Matveev (Uni. Jena), Charles Boyer (Uni. New Mexico), Vincent Pencastaing (Uni. Luxembourg), Fernando Cortes Kuehnast (TU Berlin), William Campbell Wylie (Syracuse Uni.), Franziska Beitz (WWU Münster), James McCoy (Uni. Newcastle), Megan Kerr (Wellesley College), Adam Moreno (Uni. Notre Dame), Anusha Krishnan (Uni. Pennsylvania), Curtis Porter (North Carolina State Uni.), Romina Arroyo (Uni. Queensland), Gerd Schmalz (Uni. New England), Nan Li (City Uni. New York), Zheting Dong (Oregon State Uni.), Changwei Xiong (Australian National Uni.), Xianfeng Wang (Australian National Uni.), Yuhan Wu (Uni. Wollongong), Brett Parker (Monash Uni.), Jian He (Monash Uni.)

The first week of this program took the form of an international conference with several prominent keynote speakers. These included Ben Andrews and Neil Trudinger from Australia; Rod Gover from New Zealand; Christoph Böhm and Burkhard Wilking from Germany; Robert Bryant, Karsten Grove, Claude LeBrun, Peter Petersen and Guofang Wei from the United States; Dame Frances Kirwan from the United Kingdom; and Tom Farrell and Fuquan Fang from China. Additional contributed talks were delivered in topics across differential geometry and geometric analysis.

In the second week, the meeting was less formal with specialised talks in parallel sessions in the mornings and free time for discussion and research in the remainder of the day. The parallel sessions were organised around themes which included:

- geometric evolutions equations and curvature flow,
- structures on manifolds and mathematical physics,
- higher invariants and positive scalar curvature, and
- recent developments in non-negative sectional curvature.

The organisers wish to thank MATRIX for hosting the event; and the Australian Mathematical Sciences Institute, the Australian Mathematical Society, DFG national priority research scheme “Geometry at Infinity, SPP2026”, the National Science Foundation, the University of Melbourne International Research and Research Training Fund, La Trobe University, and the Ian Potter Foundation for their financial support.

A separate proceedings volume for this meeting will be published as “Differential Geometry in the Large”, London Mathematical Society Lecture Note Series (463), Cambridge University Press.

Owen Dearricott, Diarmuid Crowley
for the organisers


AUSTRALIAN-GERMAN
WORKSHOP ON
DIFFERENTIAL
GEOMETRY IN THE LARGE
CRESWICK
MATRIX
MATRIX

Aperiodic Order meets Number Theory

25 February – 1 March 2019

Organisers

Michael Baake
Bielefeld Uni.

Michael Coons
Uni. Newcastle

Uwe Grimm
Open U.

John A. G. Roberts
UNSW Sydney

Reem Yassawi
Uni. Claude Bernard Lyon 1

Participants

Shigeki Akiyama (Uni. Tsukuba, Japan), Michael Baake (Bielefeld Uni.), Valérie Berthé (Uni. Paris Diderot), Yann Bugeaud (Uni. Strasbourg), Álvaro Bustos (Uni. Chile), Michael Coons (Uni. Newcastle), María-Isabel Cortez (Uni. Santiago Chile), Karma Dajani (Utrecht Uni.), David Damanik (Rice Uni.), Robbert Fokkink (TU Delft), Franz Gähler (Bielefeld Uni.), Amy Glen (Murdoch Uni.), Uwe Grimm (Open Uni.), Mumtaz Hussain (LaTrobe Uni.), Jeffrey C. Lagarias (Uni. Michigan), Dong-il Lee (Seoul Women's Uni.), Jeong-Yup Lee (Kwandong Uni.), Mariusz Lemańczyk (Nicolaus Copernicus Uni.), Manuel J. C. Loquias (Uni. Philippines), Michael Mampusti (Uni. Wollongong), Neil Mañibo (Bielefeld Uni.), Robert V. Moody (Uni. Victoria, Canada), John A. G. Roberts (UNSW Sydney), Tanja Schindler (Australian National Uni.), Bernd Sing (Uni. West Indies), Nicolae Strungaru (MacEwan Uni.), Venta Terauds (Uni. Tasmania), Franco Vivaldi (Queen Mary Uni. London), Peter Zeiner (Xiamen Uni. Malaysia)

This workshop benefited from the participation of a diverse group of 29 mathematicians ranging from world-experts and rising stars to eager new doctoral students. Our common thread was a desire to understand the connections between aperiodic order and number theory and to consider the further development of those connections.

During the week of our workshop, we had about four talks a day, two of which formed a pair of shorter talks on a coordinated theme. They covered topics from harmonic analysis, dynamical systems, ergodic theory, discrete geometry, number theory, topological dynamics, spectral theory, algebra and invariants. Most topics had connections to number theory, which occurred on various levels. At present, the majority of connections are of the form that known results from elementary, algebraic and analytic number theory are helping to answer questions in aperiodic order.

However, there is an increasing activity on open problems in number theory such as the Möbius disjointness conjecture or connections to the Riemann hypothesis.

Even though aperiodic order at present is profiting more from number theory than the other way round, it became clear that there is an increasing potential for the reverse direction. This view was strengthened by conversations with number theorists in attendance including Yann Bugeaud, Jeffrey Lagarias and Michael Coons. Each of these number theorists has interests in integer sequences and, in particular, the statistical properties of base expansions of integers. Questions in this area are in a unique position to be considered in the context of aperiodic order, and it is our hope that results in aperiodic order can lead to new number theoretic results. What is interesting is that this connection is not new — indeed it goes back to a near collaboration between the famous American mathematician Norbert Wiener and the famous German-Australian number theorist Kurt Mahler.

In 1926, Norbert Wiener received a Guggenheim fellowship to work with Max Born in Göttingen and then to travel on to work with Niels Bohr in Copenhagen. In that year, Born's assistant was Werner Heisenberg, who would follow Wiener to Copenhagen and develop what would later become his famous uncertainty principle. It is in this setting that, while in Göttingen, Wiener was given an (unpaid) assistant — the young Kurt Mahler! Collectively, Wiener and Mahler produced a two-part series of papers entitled, "The spectrum of an array and its application to the study of the translational properties of a simple class of arithmetical functions." As Wiener introduces his part, he writes

> "The purpose of the present paper is to extend the spectrum theory already developed by the author in a series of papers to the harmonic analysis of functions only defined for a denumerable set of arguments — *arrays*, as we shall call them — and **the application of this theory to the study of certain power series admitting the unit circle as an essential boundary**." (Boldness added by author.)

Concerning the actual contribution, given a sequence A, Wiener describes a method to construct a monotone non-decreasing function $A(x)$, which he calls the *spectral function of* A. By a result of Fréchet, $A(x)$ may contain three possible additive parts: a monotone step function, a function which is the integral of its derivative, and a continuous function which has almost everywhere a zero derivative. In modern terminology, what Wiener is describing is how one can associate a measure to the sequence A. The three possible parts of the measure are then described by the Lebesgue decomposition theorem: *Any regular Borel measure* μ *on* $\mathbb{R}^d$ *has a unique decomposition* $\mu = \mu_{\mathrm{pp}} + \mu_{\mathrm{ac}} + \mu_{\mathrm{sc}}$ *where* $\mu_{\mathrm{pp}} \perp \mu_{\mathrm{ac}} \perp \mu_{\mathrm{sc}} \perp \mu_{\mathrm{pp}}$ *and also* $|\mu| = |\mu_{\mathrm{pp}}| + |\mu_{\mathrm{ac}}| + |\mu_{\mathrm{sc}}|$. Here, μ_{pp} is a pure point measure corresponding to the monotone step function, μ_{ac} is an absolutely continuous measure corresponding to the function that is the integral of its derivative, and μ_{sc} is a singular continuous measure corresponding to the continuous function which has almost everywhere a zero derivative. Wiener provided two examples giving pure point measures and absolutely continuous measures, respectively, and an 'almost all' result for examples having a singular continuous measure. As it turns out, periodic sequences give pure point measures. Wiener's example giving an absolutely continuous measure is reminiscent of the sequence of digits of Champernowne's number. Mahler's contribution

is to find a piece of hay in the haystack — an example of a sequence whose associated measure is singular continuous. His example, the Thue–Morse sequence, is paradigmatic and started an area of transcendence theory now called Mahler's method. The Thue–Morse sequence $\{t(n)\}_{n\geqslant 0}$ is defined by $t(0) = 1$, $t(1) = -1$, $t(2n) = t(n)$ and $t(2n+1) = -t(n)$. This sequence is now ubiquitous in the areas of theoretical computer science and symbolic dynamics.

Two areas emerged, then diverged, from these two related papers. Therein lies what the participants of this conference intend to do: *to bring back together these areas and to use the results of aperiodic order to address fundamental questions in number theory, such as those concerning power series that have the unit circle as a natural boundary*; that is, to address Wiener's original purpose in studying the harmonic analysis of functions on countable sets!

The first paper arising from this program discusses the origin and structure of the field of aperiodic order. The other 18 papers are extended abstracts of the presented talks.

The Guest Editors would like to thank Michael Coons who co-authored this summary.

Michael Baake and Uwe Grimm
Guest editors

Ergodic Theory, Diophantine Approximation and Related Topics

17 – 28 June 2019

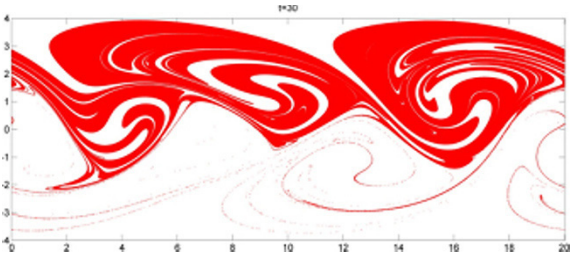
Organisers

Dzmitry Badziahin
Uni. Sydney

Alexander Fish
Uni. Sydney

Mumtaz Hussain
La Trobe Uni.

Bao-Wei Wang
Huazhong Uni.



Participants

Jinpeng An (Peking Uni.), Dzmitry Badziahin (Uni. Sydney), Michael Bjorklund (Chalmers Uni), Michael Coons (Uni. Newcastle), Alexander Fish (Uni. Sydney), Alexander Gorodnik (Uni. Zürich), Mumtaz Hussain (LaTrobe Uni.), Dmitry Kleinbock (Brandeis Uni.), Bing Li (South China Uni. Tech.), Nikolay Moshchevitin (Moscow State Uni.), Johannes Schleischitz (Middle East Cyprus Uni.), Lovy Singhal (Peking Uni.), Sanju Velani (Uni. York), Oleg German (Moscow State Uni.), Sam Chow (Uni. Oxford), Changhao Chen (UNSW Sydney), Ayreena Bakhtawar (LaTrobe Uni.)

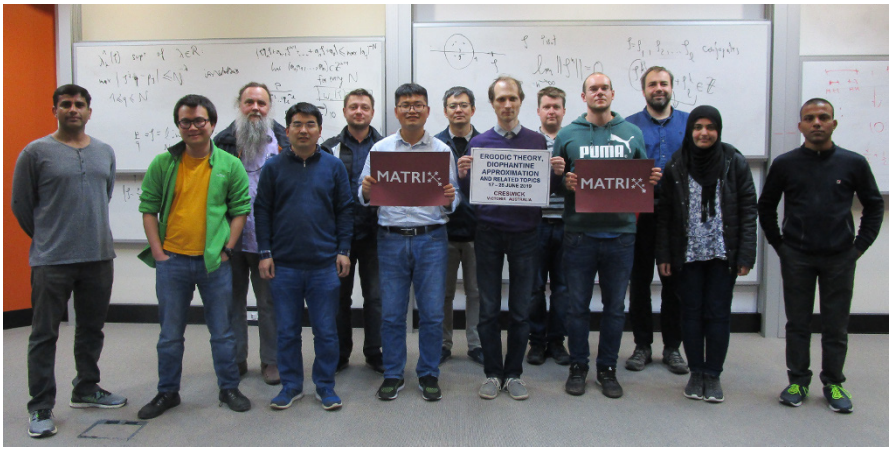
This two-week research workshop was a continuation of the conference "Dynamics and number theory" held at the University of Sydney (11–14 June 2019). The workshop was on interconnected topics in Ergodic Theory and Analytical Number Theory with the focus on Diophantine Approximation. Progress on cutting edge problems in these fields were presented and discussed in a free flowing manner. The focus was on methods and techniques that could lead to the resolution of some long standing open problems such as the Littlewood Conjecture (1930), Wirsing's problem (1961), and Generalised Baker-Schmidt Problem (1970), etc. During the workshop we learned that Koukoulopoulos and Maynard had resolved the Duffin-Schaeffer Conjecture; the official announcement followed a few days later.

During the workshop several new collaborations emerged (as detailed below) and progress was made on several long standing open problems such as the Wirsing Problem. There were two expository talks every day followed by several hours of research collaboration time. The covered topics included,

- **Diophantine exponents:** This topic was specifically discussed by Badziahin, Moshchevitin, German, Schleischitz, and Chow. In particular, Badziahin and Schleischitz improved bounds on Wirsing's problem during this workshop (https://arxiv.org/abs/1912.09013). This paper is now published in Transactions of the American Mathematical Society (https://doi.org/10.1090/tran/8245).

- **Diophantine approximations on fractal sets:** Schleischitz and Singhal discussed various problems concerning Diophantine approximation on fractal sets.
- **Metric recurrence and shrinking target problems**: Hussain and Li worked on this problem and made some progress in establishing the metrical theory for shrinking target and recurrence problems for dynamical systems satisfying some natural conditions. The systems include the continued fractions, beta dynamical systems, and homogeneous self-similar sets.
- **Singular vectors on manifolds and fractals:** Kleinbock and Moshchevitin worked on proving the existence of totally irrational vectors and linear forms with large uniform Diophantine exponents; see https://arxiv.org/abs/1912.13070.
- **Generalised Baker-Schmidt problem on manifolds:** Badziahin, Hussain, and Schleschitz discussed Diophantine approximation problems on manifolds especially the generalised Baker-Schmidt problem. In particular, Hussain and Schleischitz made progress in settling this problem for all non-degenerate co-dimensional two manifolds not only for the Euclidean setting but also for p-adics.
- **Uniform Diophantine approximation:** Hussain and Kleinbock discussed improvements to Dirichlet's theorem. An article is in preparation on this topic.
- **Multiplicative Diophantine approximation:** Gorodnik, Badziahin, Fish, Chow, Moshchevitin, and German discussed problems within the theory of multiplicative Diophantine approximation such as the well-known Littlewood conjecture (1930). In particular, Chow presented his results using Bohr sets and German by using the parametric geometry of numbers.
- **Central limit theorems and Diophantine approximation:** Bjorklund and Gorodnik discussed this topic.
- **Hitting probabilities and shrinking targets:** Li and Velani initiated a collaboration on hitting probabilities within the shrinking target settings of dynamical systems.

Mumtaz Hussain
Guest editor



Influencing Public Health Policy with Data-informed Mathematical Models of Infectious Diseases

1 – 12 July 2019

Organisers

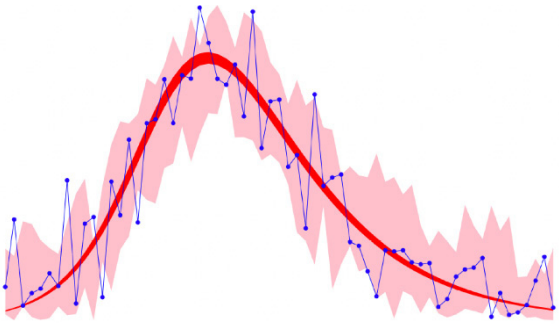
Jennifer Flegg
Uni. Melbourne

James McCaw
Uni. Melbourne

Joshua Ross
Uni. Adelaide

Thomas House
Uni. Manchester

Ben Cooper
Mahidol, Uni. Oxford



Participants

Jennifer Flegg (Melbourne), James McCaw (Melbourne), Joshua Ross (Adelaide), Thomas House (Manchester), Jonathan Keith (Monash), Lisa White (Oxford), Nick Golding (Melbourne), Deborah Cromer (UNSW), Andrew Black (Adelaide), Ada Yan (Imperial College), Jason Whyte (Melbourne), Sai Thein Than Tun (Oxford), Carla Ewels (JCU), Alex Zarebski (Oxford), James Walker (Adelaide), Amani Alahmadi (Monash), Sarah Belet (Monash), Nurul Anwar (Melbourne), Freya Shearer (Melbourne), Pavithra Jayasundara (UNSW), Hom Nath Dhungana (UTS), Zari Dzalilov (Federation), Rob Moss (Melbourne)

With the ever-growing emphasis on the importance of sound evidence in healthcare decision-making and policy, the power of data-informed mathematical models to provide much needed insight is substantial. In order for conclusions drawn from a mathematical model to be reliable, it is essential for unknown model parameters to be estimated from data in a statistically sound manner and to account for uncertainty in the parameter values. Our MATRIX workshop brought together local and international experts in this area to discuss the use of existing statistical methods and showcase new methods for parameter estimation in models of infectious diseases.

During the program there were three groups, each working on a focus problem:

- using prior knowledge to improve inference and forecasting of infectious disease transmission;
- integrating multiple data sources in infectious disease modelling;
- fitting complex models: identifying the problems and the solutions.

The first week of the program saw the introduction of the three focus problems, collaborative time on the focus problems, a software demonstration (GRETA) and

scientific talks. The second week of the program was focussed around more collaborative time but also saw a software demonstration (SHINY) and more scientific talks.

The third focus group, which focused on issues of parameter identifiability, soon found that there was significant theory underlying the topic. One of the focus group leaders, Dr Jason Whyte, has put together a review for this book which is entitled "Model structures and structural identifiability: What? Why? How?". This paper provides an overview of the importance of structural global identifiability in dynamical systems models, details some essential theory and distinctions, and demonstrates these by some key examples.

Jennifer Flegg
Guest editor

International Workshop on Spatial Statistics

15 – 19 July 2019

Organisers

Tingjin Chu
Uni. Melbourne

Davide Ferrari
Free Uni. Bozen-Bolzano

Wei Lin
Peking Uni.

Aihua Xia
Uni. Melbourne

Sanming Zhou
Uni. Melbourne

Lawrence Reeves
Uni. Melbourne

Participants

Tingjin Chu (Uni. Melbourne), Pavel Krupskiy (Uni. Melbourne), Guoqi Qian (Uni. Melbourne), Noel Cressie (Uni. Wollongong), Jun Zhu (Uni. Wisconsin-Madison), Erin Peterson (Queensland Uni. Technology), Le Chang (The Australian National Uni.), Andriy Olenko (La Trobe Uni.), Yan Wang (RMIT Uni.), Wei Lin (Peking University, China), Wenlin Dai (Renmin Uni. China), Xianbin Cao (Uni. Melbourne), Illia Donhauzer (La Trobe Uni.), Ravindi Nanayakkara (La Trobe Uni.), Nan Zheng (Monash Uni.), Jianzhong Qi (Uni. Melbourne), Jiabao Li (Leon), (Uni. Melbourne), Benjamin Hines (Uni. Melbourne), Manling Qian (Uni. Melbourne), Paul Keeler (Uni. Melbourne), Alan Pearse (Queensland Uni. Technology)

The program brought together internationally recognised experts in spatial statistics as well as early career researchers to discuss statistical modelling and applications. Scholars from fields outside of statistics also attended the workshop, and interdisciplinary research topics were discussed. The focus of this workshop was to explore new methodology and develop computational tools to make challenging problems in spatial statistics more tractable.

During the program, the following topics were discussed in small groups:

- New methodology for spatial data with complex dependence structures including models for non-Gaussian spatial processes, spherical data and categorical/proportional spatial multivariate data.
- Models and inference for high-dimensional data with temporal dependence dynamics and spatial covariates with applications in climatology, geology and environmental science. Outlier detection for multivariate functional data.

- Spatial data reconstruction tools and construction of spatial maps for sparse data. Uncertainty quantification for data coming from different sources.

Young researchers had opportunity to interact with senior academics, and several discussion groups were organized to discuss challenging problems in spatial statistics and to create new collaboration opportunities. This lead to the paper in this volume by B. Hines, Y. Kuleshov and G. Qian "Spatial modelling of linear regression coefficients for gauge measurements against satellite estimates," which studies the problem of predicting rainfall in remote areas of Australia using satellite estimates.

Pavel Krupskiy
Guest editor

Mathematics of Physiological Rhythms

9 – 13 September 2019

Organisers

Maia Angelova
Deakin Uni.

James Sneyd
Uni. Auckland

Aneta Stefanovska
Lancaster Uni.

Plamen Ivanov
Uni. Boston, Uni. Harvard

Participants

Maia Angelova (Deakin Uni.), Aneta Stefanovska (Lancaster Uni.), Plamen Ch. Ivanov (Boston Uni.), Anne Skeldon (Uni. Surrey), Krasimira Tsaneva-Atanassova (Uni. Exeter), Adelle Coster (UNSW Sydney), Andrew Phillips (Monash Uni.), David Liley (Uni. Melbourne), Ruben Fossion (National Autonomous Uni. Mexico), Chandan Karmakar (Deakin Uni.), Ye Zhu (Deakin Uni.), Sutharshan Rajasegarar (Deakin Uni.), Christopher Stephens (National Autonomous Uni. Mexico), Sergiy Shelyag (Deakin Uni.), Shitanshu Kusmakar (Deakin Uni.), Jyotheesh Gaddam (Deakin Uni.), Mohammad Abdul Motin (Uni. Melbourne), Emerson Keenan (Uni. Melbourne), Shreyasi Datta (Uni. Melbourne), Md Ahsan Habib (Deakin Uni.), Jason Whyte (Uni. Melbourne), Tania Pencheva (Bulgarian Acad. Sci.), Anuroop Gaddam (Deakin Uni.)

This research retreat was devoted to novel dynamical system methodologies underpinning the modelling of complex physiological systems, and focused on four main topics: Network Physiology, Brain, Diabetes, and Sleep.

The aim was to unite and combine current trends in dynamical systems and time series analysis for solving problems in physiology which are governed by repeating processes. Examples are cardio-dynamics, sleep processes, glucose-insulin regulation and diabetes, and many others. The invited participants were experts in mathematics, physics and computer sciences working in applications of dynamical systems and time series in physiology, biology and medicine. The program explored the state-of the-art research underlying the mathematics of periodic and periodic-like processes in human physiology.

The program was attended by 20 participants funded by the MATRIX Institute, and three participants funded by other institutions. The participants were involved in four discussion and collaborative sessions each afternoon led by one of the plenary speakers. Each session was devoted to one of the main topics of the program. The participants were from five countries: Australia, USA, UK, Mexico and Bulgaria.

Women were well represented, four were plenary speakers, one invited speaker and one a PhD student. Furthermore, two out of the four organisers were women. The participants included world leading researchers in the field, early career researchers, postdocs and PhD students. They were experts in mathematical physiology, mathematical biology, differential equations, functional analysis, time series, fractals, statistical mechanics and phase transitions. A number of participants were also experts in data mining and machine learning, which would facilitate the use of such methods for parameter estimation.

The retreat focused on models based on deterministic and stochastic differential equations and delay differential equations, dynamical system approach to time series, statistical mechanics, phase transitions and mean field approaches. The mathematical models of regulation processes are often informed by data driven models, derived from spectral analysis and signal processing. Furthermore, as the large number of physiological parameters are difficult to measure, machine learning and statistical approaches were exploited to evaluate parameters. The models are based on real data measured from humans (ECG, EEG, actigraphy, eye movements), and complexity for building models from such data was discussed. The program addressed the aims of MATRIX by focusing on new mathematical models governing regulation and control processes in human physiology.

The program had one keynote talk and one invited talk each morning. The afternoons were spent on directed discussions around current trends and coordinated collaborative work. The first and the second day were devoted to Network Physiology and Diabetes. Plamen Ivanov (Boston) gave a fascinating lecture on Network Physiology. During the afternoons there was a session on Open Problems in Network Physiology led by Plamen Ivanov. On Tuesday morning Aneta Stefanovska (Lancaster) continued the theme on Network Physiology, and facilitated a 3 hour workshop on the new time series software tool, MODA, developed in Lancaster. This workshop was very useful for the PhD students attending the program. Another plenary session was focused on diabetes, where Adelle Coster gave a plenary talk and led a discussion session in the afternoon on Open Problems in Diabetes research. The topic on Wednesday was Brain. Krasimira Tsaneva-Atanassova (Exeter) gave the first plenary talk, followed by the talk given by David Liley. David led the discussion session before lunch on fitting complex mathematical models with a large number of parameters. The plenary talks on Thursday were on Sleep, given by Maia Angelova (Deakin) and Andrew Phillips (Monash). The afternoon was focused on collaborative work on Sleep. The closing session on Friday by plenary speaker Christopher Stephens (UNAM), a renowned expert in Data Mining in Healthcare, was a part of Network Physiology topic. The two invited talks given by Tania Pencheva (BAS) and Ruben Fossion (UNAM), and another two short “ignit” talks, were presented by Sutharshan Rajasegarar (Deakin) and Anuroop Gaddam (Deakin). A working group on Sleep was formed to work on models of insomnia; this group met daily during all days of the program. The group is currently active, submitting jointly co-authored papers and preparing an ARC grant proposal. On Wednesday afternoon a walk around Creswick wss organised to facilitate networking and the forming of new research links. This was particularly useful for PhD

students, early- and mid-career researchers, as it allowed them to talk to the leaders in the field in a relaxed atmosphere.

The aim and objectives of the program were completed. The participants expressed their gratitude to The MATRIX Institute for providing excellent conditions that enabled new research collaborations. The program was very useful and contributed to our long term goals to develop Deakin University as a Hub for Mathematical and AI modelling translated to health, physiology, wellbeing and health care. In addition to the papers appearing in this book, a number of papers arising from the program will appear in a special issue of *Frontiers of Physiology*. The organisers and participants gratefully acknowledge funding from the MATRIX Institute and Deakin University, School of IT.

Maia Angelova
Guest editor

Conservation Laws, Interfaces and Mixing

4 – 8 November 2019

Organisers

Snezhana I. Abarzhi
Uni. Western Australia

Neville Fowkes
Uni. Western Australia

Alik Nepomnyashchy
Technion, Israel

Anthony J. Roberts
Uni. Adelaide

Yvonne Stokes
Uni. Adelaide

Participants

Snezhana I. Abarzhi (Uni. Western Australia), Yasuhide Fukumoto (Kyushu Uni.), Ashleigh Hutchinson (Uni. Witwatersrand), Alexander Klimenko (Uni. Queensland), Joseph Klewicki (Uni. Melbourne), Alexander Nepomnyashchy (Technion), Xiaolin Li (State Uni. New York, Stony Brook), Tony Roberts (Uni. Adelaide), Mako Sato (Osaka City Uni.), Helen Wang (Zeta Global Inc.), Kurt Williams (Uni. Western Australia), Paulo de Almeida (Altron Bytes Systems Integration), Cameron Wright (Uni. Western Australia), Tanmay Agrawal (Uni. Melbourne), Saleh Tanveer (Ohio State Uni.), Ash Khan (RMIT Uni.)

Interfacial transport and mixing are non-equilibrium processes coupling kinetic and macroscopic scales. They occur in molecules, fluids, plasmas and materials over celestial events. Examples include supernovae and fusion, planetary convection and reactive fluids, wetting and adhesion, turbulence and mixing, nano-fabrication and bio-technology. Addressing the societal challenges posed by alternative energy sources, efficient use of non-renewable resources, and purification of water requires a better understanding of non-equilibrium interfacial transport and mixing.

The dynamics of interfacial transport and mixing often involve sharp changes of vector and scalar fields, and may also include strong accelerations and shocks, radiation transport and chemical reactions, diffusion of species and electric charges, among other effects. Interfacial transport and mixing are inhomogeneous, anisotropic, non-local, and statistically unsteady. At macroscopic scales, their spectral and invariant properties differ substantially from those of canonical turbulence. At atomistic and meso-scales, the non-equilibrium dynamics depart dramatically from the standard scenario given by Gibbs ensemble averages and the quasi-static Boltzmann equation. At the same time, non-equilibrium transport may lead to self-organization and order, thus offering new opportunities for diagnostics and control. Capturing

the properties of interfaces and mixing enables: the accurate description of conservation properties, the solution of boundary value problems, better understanding of Eulerian and Lagrangian dynamics, and the development of methods for control of non-equilibrium transport in nature and technology.

Significant success was recently achieved in the understanding of interfacial transport and mixing in terms of of theoretical analysis, large-scale numerical simulations, and data analysis. This success opened new opportunities for the study of the fundamentals of non-equilibrium dynamics across the scales, for developing a unified description of particles and fields on the basis of the synergy of theory and numerical data, and for applying the fundamentals of non-equilibrium transport to address the contemporary challenges of modern science, technology and society.

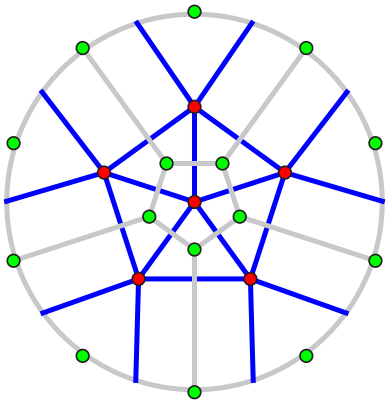
This program built upon recent achievements in understanding interfacial transport and mixing using theoretical analysis, large-scale numerical simulations, and data analysis. The focus was on conservation laws and boundary value problems. The program brought together researchers from applied mathematics, applied analysis, dynamical and complex systems, stochastic processes and data analysis, dynamics of fluid and plasmas, industrial mathematics and materials science. The program motivated discussions of rigorous mathematical problems, theoretical approaches and state-of-the-art numerical simulations along with advanced data analysis techniques. The program explored the state-of-the-art in the areas of interfaces and non-equilibrium transport, and charted new research directions in this field.

The participants included leading experts and researchers at all career stages from Australia and from abroad.

Snezhana Abarzhi, Alexander Nepomnyashchy, Anthony Roberts, Joseph Klewicki
Guest editors

Structural Graph Theory Downunder

25 November – 1 December 2019



Organisers

David Wood
Monash Uni.

Anita Liebenau
UNSW Sydney

Alex Scott
Uni. Oxford

Participants

Maria Chudnovsky (Princeton Uni.), Zdeněk Dvořák (Charles Uni.), Kevin Hendrey (IBS Korea), Tony Huynh (Monash Uni.), Gwenaël Joret (ULB Belgium), Nina Kamčev (Monash Uni.), Ringi Kim (KAIST Korea), Tereza Klimošová (Charles Uni.), Anita Liebenau (UNSW Sydney), Chun-Hung Liu (Texas A&M Uni.), Natasha Morrison (Uni. Cambridge), Marcin Pilipczuk (Uni. Warsaw), Bruce Reed (McGill Uni., Montréal), Alex Scott (Uni. Oxford), Paul Seymour (Princeton Uni.), Maya Stein (Uni. Chile), Jane Tan (Uni. Oxford), David Wood (Monash Uni.), Liana Yepremyan (London School Econs.), Yelena Yuditsky (Ben-Gurion Uni.), Xuding Zhu (Zhejiang Uni.)

This program consisted of a 1-week intensive research workshop, where mathematicians from across the globe came together to work on open problems in structural graph theory. The program featured a mix of early-career, mid-career and senior researchers; a mix of women and men; and a mix of people from Australia, Europe, North America, South America, Israel, China, and Korea. The goal was to create an environment where mathematicians at all career stages worked side-by-side. This goal was certainly achieved. Many participants commented on how conducive the MATRIX House was for doing collaborative research.

The majority of the time was allocated to collaborative research. In addition, there were six research talks about recent significant results:

- Xuding Zhu (Zhejiang Uni.) surveyed recent developments on Hedetniemi's Conjecture and the Poljak-Rödl function, including Shitov's recent breakthrough;
- Liana Yepremyan (London School Econs.) presented a proof of the size-Ramsey number of graphs of bounded degree and bounded treewidth;
- Tereza Klimošová (Charles Uni.) talked about edge-partitioning 3-edge-connected graphs;
- Chun-Hung Liu (Texas A&M Uni.) talked about clustered graph colouring, in particular, clustered variants of Hajós' Conjecture;

- Paul Seymour (Princeton Uni.) discussed recent results on the structure of graphs excluding certain graphs as induced subgraphs;
- Maria Chudnovsky (Princeton Uni.) described a polynomial-time algorithm for finding a maximum independent set in a graph with no hole of length at least five; and
- Zdeněk Dvořák (Charles Uni.) talked about the interplay between bounded expansion classes and sub-linear separators.

Prior to the more formal talks, every participant gave a 5-minute talk introducing their research interests and an open problem that they would like to work on during the workshop. People then naturally formed groups working on problems of common interest. These topics included: Hadwiger's Conjecture, Hedetniemi's Conjecture, induced subgraphs, graph product structure theory, and centred colouring. On each of these topics, significant progress was made during the program.

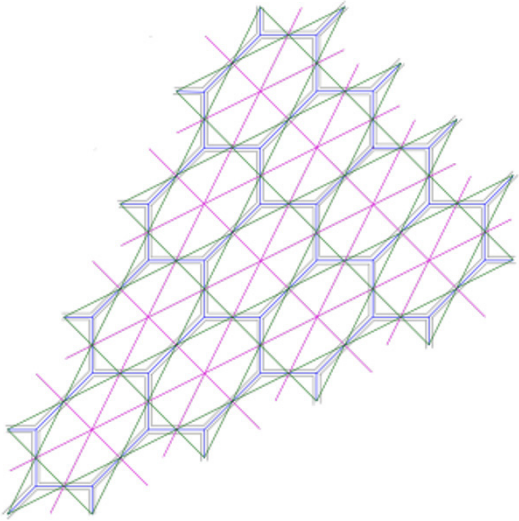
This led to four papers in the MATRIX Annals. Xuding Zhu's lecture notes describe Shitov's proof in detail. Maria Chudnovsky and Paul Seymour present work completed at the workshop on the clique-stable set separation property. Zdeněk Dvořák, Tony Huynh, Gwenaël Joret, Chun-Hung Liu and David Wood survey recent results on graph product structure theory, including some new work done at the workshop, as well discussing many open problems. Tony Huynh, Bruce Reed, David Wood and Liana Yepremyan describe results on the tree- and path-chromatic number. They also present a tree-chromatic version of Hadwiger's Conjecture and give evidence that this conjecture may be more tractable than the original conjecture.

All in all, the program was a great success. The participants were keen that another program in structural graph theory be held at MATRIX soon.

Anita Liebenau
Guest editor

Tropical Geometry and Mirror Symmetry

9 – 20 December 2019



Organisers

Nick Sheridan
Uni. Edinburgh

Brett Parker
Monash Uni.

Paul Norbury
Uni. Melbourne

Jian He
Monash Uni.

Kristin Shaw
Uni. Oslo

Participants

Brett Parker (Monash Uni.), Paul Norbury (Uni. Melbourne), Nick Sheridan (Uni. Edinburgh), Jian He (Monash Uni.), Kristin Shaw (Uni. Oslo), Renato Vianna (Federal Uni. Rio de Janeiro), Siu-Cheong Lau (Boston Uni.), Cheol Hyun Cho (Seoul National Uni.), Helge Ruddat (Johannes Gutenberg Uni. Mainz), Mandy Cheung (Harvard Uni.), Pierrick Bousseau (ETH Zürich), Jeff Hicks (Uni. Cambridge), Ilia Zharkov (Kansas State Uni.), Xiao Zheng (Boston Uni.), Mehdi Tavakol (Uni. Melbourne), Ziming Nikolas Ma (Chinese Uni. Hong Kong), Johanes Rau (Uni. Tübingen), Lucia Lopez de Medrano (Uni. Nacional Autonoma de Mexico), Masahiro Futaki (Chiba Uni.), Ellena Moscovski (Monash Uni.), Wee Chaimanowong (Uni. Melbourne), Michael Swaddle (Uni. Melbourne), Urs Fuchs (Monash Uni.), Daniel Mathews (Monash Uni.), Norm Do (Monash Uni.)

Mirror symmetry studies the relationship between algebraic and symplectic geometry. The duality passes through the adiabatic limit of the geometries—the tropical geometry. In recent years, significant advances were established in each of these areas. Thus there is a need for the communities to understand the intrinsic connections between mirror symmetry and tropical geometry by using these newly developed technical tools.

This two-week program brought together researchers in algebraic, symplectic, and tropical geometry. The workshop started with a series of introductory lectures to invite junior participants, in particular graduate students, to become familiar with the subjects. In tandem, there were around two research talks each day on various topics. One of the goals of the program was to encourage communications between different groups in mirror symmetry. Hence there was plenty of time allocated each

day for informal discussion. The program created fertile advances in mirror symmetry, and new interdisciplinary collaborations resulted.

Three articles arose from the program. In ‘Observations on disks with tropical Lagrangian boundary’, Jeffs Hicks studies Lagrangian submanifolds which are built as the lifts of tropical curves. In ‘Compactifying torus fibrations over integral affine manifolds with singularities’, Helge Ruddat and Ilia Zharkov announce a construction in which they build a space X which is a torus fibration over a given integral affine manifold. Discussions during the program led Man-Wai Cheung and Renato Vianna to explore the correspondence of the compactifications of cluster varieties from the algebro geometric and symplectic perspectives. Their discoveries are presented in the paper, ‘Algebraic and symplectic viewpoint on compactifications of two-dimensional cluster varieties of finite type’.

Mandy Cheung
Guest editor

Early Career Researchers Workshop on Geometric Analysis and PDEs

13 – 24 January 2020

Organisers

Paul Bryan
Macquarie Uni.

Jiakun Liu
Uni. Wollongong

Mariel Sáez
Pontificia Universidad
Católica de Chile

Haotian Wu
Uni. Sydney

Participants

Paul Bryan (Macquarie Uni.), Jiakun Liu (Uni. Wollongong), Mariel Sáez (Pontificia Uni. Católica Chile), Haotian Wu (Uni. Sydney), Julian Scheuer (Columbia Uni.), Guoyi Xu (Tsinghua Uni.), Valentina Wheeler (Uni. Wollongong), Lu Wang (UW-Madison/IAS), Davi Maximo (Uni. Pennsylvania), Otis Chodosh (Princeton/IAS), Artem Pulemetov (Uni. Queensland), Xianfeng Wang (ANU/Nankai Uni.), Katarzyna (Kasia), Mazowiecka (UCLouvain), Yong Wei (Australian National Uni.), Ben Sharp (Uni. Leeds), Shibing Chen (Uni. Science & Tech. China), Azahara de la Torre (Albert-Ludwigs-Uni. Freiburg), Mircea Petrache (P. Uni. Catolica de Chile.), Alessandra Pluda (Uni. Pisa), Brett Kotschwar (Arizona State Uni.), Mathew (Mat), Langford (Uni. Tennessee), Theodora Bourni (Uni. Tennessee), Ben Lambert (Uni. College London), Changwei Xiong (Australian National Uni.), Kui Wang (Soochow Uni.), Kwok-Kun Kwong (Uni. Wollongong), Medet Nursultanov (Chalmers/Uni. Sydney), Yuhan Wu (Uni. Wollongong), Qiang Guang (Australian National Uni.), Julie Clutterbuck (Monash Uni.), Peter Olanipekun (Monash Uni.), Lachlann O'Donnell (Uni. Wollongong)

The principal aim of this program was to draw together early career researchers working in the fields of calculus of variations, optimal transport, fully nonlinear PDEs and geometric flows.

The first week consisted of a series of mini-courses on contemporary areas of study in geometric PDE. It was a very interactive week with considerable discussion. The lecturers themselves gained immeasurably from the expert comments and feedback. Researchers in geometric flows benefited enormously from Guoyi Xu's lectures on isometric embedding, whilst much of the audience—especially those not well versed in flows—learnt the breadth of applications of inverse curvature flows with a very complete and concise set of lectures by Julian Scheuer, whose notes

may be found in this volume. Valentina Wheeler made the brave choice to begin a course on free boundary mean curvature flow by lecturing to an expert audience on Huisken's original paper on the mean curvature flow. A lively and robust discussion on the optimal way to prove this seminal result followed with many experts learning something new! All in all the first week generated a large amount of discussion and initiated several collaborations. The lectures were:

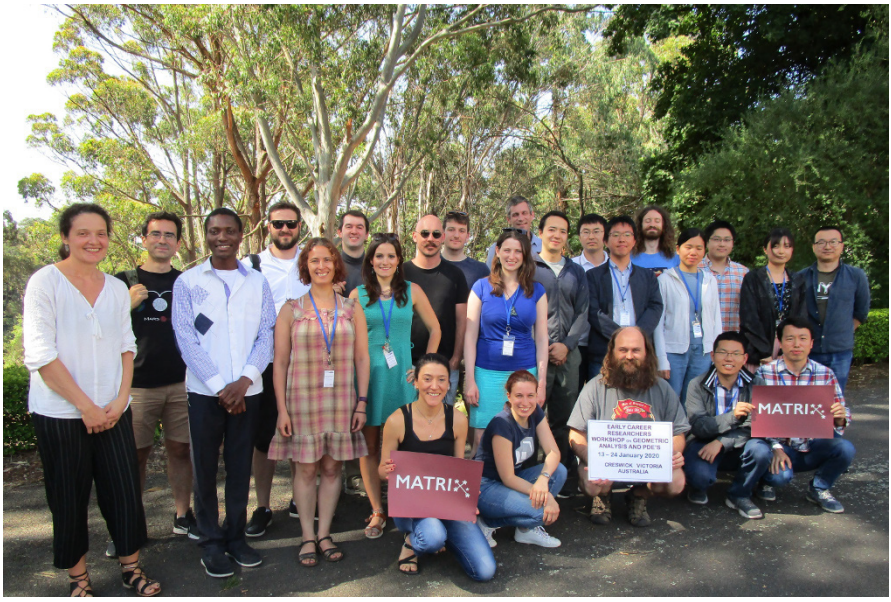
- *Isometric Embeddings*, Guoyi Xu
- *Extrinsic curvature flows and applications*, Julian Scheuer
- *Mean Curvature Flow with free boundary*, Valentina Wheeler

The second week comprised ample time for research collaboration, during which the coffee machine at MATRIX house was pushed to its limits. Embedded within this week was a research seminar where we heard about a diverse range of topics including intrinsic curvature flows such as Ricci flow, hypersurface flows, minimal surfaces, isoperimetric estimates, harmonic maps, prescribed curvature problems and eigenvalue problems.

The submissions to this volume are:

- *Extrinsic curvature flows and applications*, Julian Scheuer
- *Short time existence for higher order curvature flows with and without boundary conditions*, Yuhan Wu

Paul Bryan
Guest Editor



Harmonic Analysis and Dispersive PDEs: Problems and Progress

3 – 7 February 2020

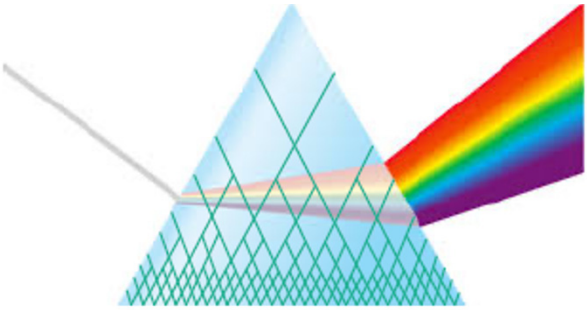
Organisers

Zihua Guo
Monash Uni.

Ji Li
Macquarie Uni.

Kenji Nakanishi
Kyoto Uni.

Wenhui Shi
Monash Uni.



Participants

Adam Sikora (Macquarie Uni.), Andrew Hassel (Australian National Uni.), Brett D. Wick (Washington Uni. Saint Louis), Chuhee Cho (Seoul National Uni.), Chunyan Huang (Central Uni. Finance Economics), Jacob Shapiro (Australian National Uni.), Jan Rozendaal (Australian National Uni.), Ji Li (Macquarie Uni.), Kenji Nakanishi (Kyoto Uni.), Lifeng Zhao (Uni. Science & Tech. China), Lixin Yan (Sun Yat-sen Uni.), Loredana Lanzani (Syracuse Uni.), Luca Fanelli (Uni. Roma "La Sapienza"), Melissa Tacy (Uni. Otago), Pierre Portal (Australian National Uni.), Sanghyuk Lee (Seoul National Uni.), Satoshi Masaki (Osaka Uni.), Sebastian Herr (Bielefield Uni.), Soonsik Kwon (KAIST), Timothy Candy (Uni. Otago), Wenhui Shi (Monash Uni.), Xing Cheng (Hohai Uni.), Xuan Duong (Macquarie Uni.), Yoshio Tsutsumi (Kyoto Uni.), Zihua Guo (Monash Uni.), Stephen Deng (Monash Uni.)

This one-week program focused on very recent and on-going progress in harmonic analysis and its applications to dispersive PDEs. With 25 participants from various countries over the world, the workshop had three talks in the morning, and the whole afternoon for discussions, both in a relaxing and stimulating mood.

The topics include: multilinear restriction estimates; the Strichartz estimate with dispersive potential and orthonormal initial data, micro-local and semi-classical analysis for dispersive equations and resolvents, with variable coefficients or with randomness; L^p-theory and the Hardy spaces for dispersive equations, quasimodes, and general integral transforms; large-data global existence and scattering with variational characters for nonlinear dispersive equations; local well-posedness, (modified) scattering, blow-up, stability and instability of solutions.

"A note on bilinear wave-Schrödinger interactions", by Timothy Candy, focuses on the interaction between the wave and the Schrödinger equations, in terms of bilinear restriction estimate, which was recently extended by the same author to the full mixed exponents for more general dispersion relations. A counter-example is given to show necessity of some geometric conditions beyond the transversality of

the characteristic surfaces. Moreover, a transference principle of the bilinear estimate is established for the space of ℓ^2-bounded oscillation in time.

"A note on the scattering for 3D quantum Zakharov system with non-radial data in L^2", by Chunyan Huang, treats the scattering problem for the quantum Zakharov system, which has additional bi-Laplacian both in the wave and Schrödinger equations, compared with the standard Zakharov system. Using the normal form transform and the improved Strichartz estimate in the L^2 spherical average, the author establishes small-data scattering result for L^2 initial data with angular regularity in three space dimensions. This improves the preceding results in several respects.

"Hankel transforms and weak dispersion", by Federico Cacciafesta and Luca Fanelli, is a concise survey of recent developments on the linear dispersive estimates for scaling-critical perturbations: the Dirac and the fractional Schrödinger equations with the Coulomb potential and with the Aharonov-Bohm field. Specifically, it focuses on the analysis based on some explicit representations using the Hankel transforms and special functions, taking advantage of scaling invariance.

"A priori bounds for the kinetic DNLS", by Nobu Kishimoto and Yoshio Tsutsumi, deals with a nonlinear Schrödinger-type equation, with both local and non-local derivative nonlinear terms in one dimensional torus. The non-local part models Landau damping with a nonlinear dissipation. The authors derive a priori upper bound on the energy norm and a lower bound on the L^2 norm. Combined with local analysis using multilinear and nonlinear estimates, it yields global existence of solutions for small data in the energy space.

Kenji Nakanishi
Guest editor

Contents

Part I

Refereed Articles

Chapter 1

Topology of Manifolds: Interactions Between High and Low Dimensions

Connected sum decompositions of high-dimensional manifolds

Imre Bokor, Diarmuid Crowley, Stefan Friedl, Fabian Hebestreit, Daniel Kasprowski, Markus Land and Johnny Nicholson

Abstract The classical Kneser-Milnor theorem says that every closed oriented connected 3-dimensional manifold admits a unique connected sum decomposition into manifolds that cannot be decomposed any further. We discuss to what degree such decompositions exist in higher dimensions and we show that in many settings uniqueness fails in higher dimensions.

The original version of this chapter was revised. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-62497-2_65

Imre Bokor
13 Holmes Avenue, Armidale, NSW 2350, Australia, e-mail: ibokor@bigpond.net.au

Diarmuid Crowley
School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia
e-mail: dcrowley@unimelb.edu.au

Stefan Friedl
Fakultät für Mathematik, Universität Regensburg, Germany e-mail: sfriedl@gmail.com

Fabian Hebestreit
Rheinische Friedrich-Wilhelms-Universität Bonn, Mathematisches Institut, Endenicher Allee 60, 53115 Bonn, Germany, e-mail: f.hebestreit@math.uni-bonn.de

Daniel Kasprowski
Rheinische Friedrich-Wilhelms-Universität Bonn, Mathematisches Institut, Endenicher Allee 60, 53115 Bonn, Germany, e-mail: kasprowski@uni-bonn.de

Markus Land
Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark e-mail: markus.land@math.ku.dk

Johnny Nicholson
Department of Mathematics, UCL, Gower Street, London, WC1E 6BT, United Kingdom
e-mail: j.k.nicholson@ucl.ac.uk

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_1

1 Introduction

Consider the set $\mathscr{M}_n^{\mathrm{Cat}}$ of n-dimensional, oriented Cat-isomorphism classes of Cat-manifolds, where $\mathrm{Cat} = \mathrm{Top}, \mathrm{PL}$ or Diff; unless explicitly stated, all manifolds are assumed non-empty, closed, connected and oriented. $\mathscr{M}_n^{\mathrm{Cat}}$ forms a monoid under connected sum (see Section 2) as do its subsets

$$\mathscr{M}_n^{\mathrm{Cat,sc}} = \{M \in \mathscr{M}_n^{\mathrm{Cat}} \,|\, M \text{ is simply connected}\}$$

and

$$\mathscr{M}_n^{\mathrm{Cat,hc}} = \{M \in \mathscr{M}_n^{\mathrm{Cat}} \,|\, M \text{ is highly connected}\};$$

here an n-manifold M is called *highly connected* if $\pi_i(M) = 0$ for $i = 0, \ldots, \lfloor \frac{n}{2} \rfloor - 1$. Recall that the theorems of Radó [56] and Moise [48, 49] show $\mathscr{M}_n^{\mathrm{Top}} = \mathscr{M}_n^{\mathrm{PL}} = \mathscr{M}_n^{\mathrm{Diff}}$ for $n \leq 3$, and by Cerf's work [10, p. IX] that $\mathscr{M}_n^{\mathrm{PL}} = \mathscr{M}_n^{\mathrm{Diff}}$ for $n \leq 6$. These monoids are countable. For PL and Diff this follows from the fact that triangulations exist. For Top this follows from work of Cheeger and Kister [11].

In this paper we want to study the question whether or not these monoids are unique factorisation monoids. First we need to make clear what we mean by a unique factorisation monoid.

Definition 1.

1. Let $\mathscr{M}$ be an abelian monoid (written multiplicatively). We say $m \in M$ is *prime* if m is not a unit and if it divides a product only if it divides one of the factors.
2. Given a monoid $\mathscr{M}$ we denote by $\mathscr{M}^*$ the units of $\mathscr{M}$. We write $\overline{\mathscr{M}} := \mathscr{M} / \mathscr{M}^*$.
3. Let $\mathscr{M}$ be an abelian monoid. We denote by $\mathscr{P}(\mathscr{M})$ the set of prime elements in $\overline{\mathscr{M}}$. We say $\mathscr{M}$ is a *unique factorisation monoid* if the canonical monoid morphism $\mathbb{N}^{\mathscr{P}(\mathscr{M})} \to \overline{\mathscr{M}}$ is an isomorphism.

In dimensions 1 and 2 we of course have $\mathscr{M}_1^{\mathrm{Cat}} = \overline{\mathscr{M}}_1^{\mathrm{Cat}} = \{[S^1]\}$ and $\mathscr{M}_2^{\mathrm{Cat}} = \overline{\mathscr{M}}_2^{\mathrm{Cat}} \cong \mathbb{N}$ via the genus. In particular these monoids are unique factorisation monoids. In dimension 3 there is the celebrated *prime decomposition theorem* which was stated and proved, in rather different language, by Kneser [35] and Milnor [46]. See also [26, Chapter 3] for a proof.

Theorem 1 (Kneser-Milnor).

1. The monoid $\mathscr{M}_3^{\mathrm{Cat}}$ has no non-trivial units.
2. The monoid $\mathscr{M}_3^{\mathrm{Cat}} = \overline{\mathscr{M}}_3^{\mathrm{Cat}}$ is a unique factorisation monoid.

The purpose of the present note is to study to what degree these statements hold in higher dimensions.

First, note that all units of $\mathscr{M}_n^{\mathrm{Cat}}$ are homotopy spheres, as we deduce from an elementary complexity argument in Proposition 3 below. It now follows from various incarnations of the Poincaré conjecture that $\mathscr{M}_n^{\mathrm{Top}}$ never has non-trivial units, and neither does $\mathscr{M}_n^{\mathrm{PL}}$, except potentially if $n = 4$. In the smooth category the current

status is the following: The only odd dimensions in which $\mathscr{M}_n^{\mathrm{Diff}}$ has no non-trivial units are $1,3,5$ and 61, see [27, 74]. In even dimensions greater than 5 Milnor and Kervaire construct an isomorphism $\Theta_n \cong \mathbb{S}_n/J_n$ from the group of smooth homotopy n-spheres Θ_n to the cokernel of the stable J-homomorphism $\pi_n(\mathrm{SO}) \to \mathbb{S}_n$, where $\mathbb{S}$ is the sphere spectrum. In dimensions below 140, the only even dimension where J_n is surjective are $2,4,6,12$ and 56, see [3]. Wang and Xu recently conjectured that $1,2,3$, possibly 4, and $5,6,12,56,61$ are the only dimensions without exotic spheres [74].

Let us also mention that in dimension 4 neither is it known whether every homotopy sphere is a unit nor whether S^4 is the only unit in $\mathscr{M}_4^{\mathrm{Diff}}$ (and of course these questions combine into the smooth 4-dimensional Poincaré conjecture).

Before we continue we introduce the following definitions.

Definition 2. Let $\mathscr{M}$ be an abelian monoid with neutral element e.

1. Two elements $m,n \in \mathscr{M}$ are called *associated* if there is a unit $u \in \mathscr{M}^*$, such that $m = u \cdot n$.
2. An element m is called *irreducible*, if it is not a unit and if all its divisors are associated to either e or m.
3. An element a is *cancellable*, if $ab = ac$ implies $b = c$ for all elements $b,c \in \mathscr{M}$.

We make four remarks regarding these definitions.

1. We warn the reader that for the monoids $\mathscr{M}_3^{\mathrm{Cat}}$ our usage of "irreducible" does *not* conform with standard use in 3-dimensional topology. More precisely, in our language $S^1 \times S^2 \in \mathscr{M}_3^{\mathrm{Cat}}$ is irreducible, whereas in the usual language used in 3-dimensional topology, see [26, p. 28], the manifold $S^1 \times S^2$ is not irreducible. Fortunately [26, Lemma 3.13] says that this is the only 3-dimensional manifold for which the two definitions of irreducibility diverge (in fact by the prime decomposition theorem our notion of irreducible 3-dimensional manifold coincides with the usual use of the term prime 3-manifold).
2. Let $\mathscr{M}$ be an abelian monoid. If $\mathscr{M}$ is a unique factorisation monoid, then every element in $\overline{\mathscr{M}}$ is cancellable.
3. Another warning worthy of utterance is that in general, given a monoid $\mathscr{M}$, neither all irreducible elements are prime, nor does a prime element need to be irreducible, unless it is also cancellable.
4. Finally, note, that if $\mathscr{M}$ is a unique factorisation monoid, this does not necessarily imply that every element in $\mathscr{M}$ is cancellable: A good example is given by non-zero integers under multiplication modulo the relation that $x \sim -x$ if $|x| \geq 2$. In this case $\overline{\mathscr{M}} \cong \mathbb{N}_{\geq 1}$ under multiplication, so $\mathscr{M}$ is a unique factorisation monoid by prime decomposition. On the other hand we have $2 \cdot (-1) = 2 \cdot 1$ and $-1 \neq 1$. Thus we see that 2 is not cancellable.

Using a fairly simple complexity argument we obtain the following result (Corollary 1 below).

Proposition 1. *Every element in $\mathscr{M}_n^{\mathrm{Cat}}$ admits a connected sum decomposition into a homotopy sphere and irreducible manifolds.*

Unless $n = 4$ and Cat = Diff or PL, the homotopy sphere can of course, by the resolution of the Poincaré Conjecture, be absorbed into one of the irreducible factors.

As an example of the failure of cancellation and unique factorisation consider the manifolds $\mathbb{C}\mathrm{P}^2\#\overline{\mathbb{C}\mathrm{P}^2}$ and $S^2 \times S^2$. The intersection forms show that these manifolds are not homotopy equivalent, but it is well-known that $(S^2 \times S^2)\#\mathbb{C}\mathrm{P}^2$ and $\mathbb{C}\mathrm{P}^2\#\overline{\mathbb{C}\mathrm{P}^2}\#\mathbb{C}\mathrm{P}^2$ are diffeomorphic [18]. This implies easily that $\mathbb{C}\mathrm{P}^2$ is not cancellable in $\overline{\mathscr{M}}_4^{\mathrm{Cat}}$ and thus that $\mathscr{M}_4^{\mathrm{Cat}}$ is not a unique factorisation monoid.

The following is the main result of Section 5:

Theorem 2. *For $n \geq 4$ the manifold $S^2 \times S^{n-2}$ is not cancellable in any of the monoids $\overline{\mathscr{M}}_n^{\mathrm{Cat}}$ and thus none of the monoids $\mathscr{M}_n^{\mathrm{Cat}}$ is a unique factorisation monoid in that range.*

The proof we provide crucially involves manifolds with non-trivial fundamental groups. This leaves open the possibility that the submonoid $\overline{\mathscr{M}}_n^{\mathrm{Cat,sc}}$ consisting of simply connected Cat-manifolds is better behaved. However, we show, for most dimensions, in Section 6 that this is still not the case:

Theorem 3. *For $n \geq 17$, the manifold $S^5 \times S^{n-5}$ is not cancellable in any of the monoids $\overline{\mathscr{M}}_n^{\mathrm{Cat,sc}}$ and thus $\mathscr{M}_n^{\mathrm{Cat,sc}}$ is not a unique factorisation monoid in that range.*

The bound $n \geq 17$ is by no means intrinsic for finding non-cancellative elements in $\overline{\mathscr{M}}_n^{\mathrm{Cat,sc}}$; we already gave the example of $\mathbb{C}\mathrm{P}^2 \in \overline{\mathscr{M}}_4^{\mathrm{Cat}}$, and indeed by Wall's classification [67] the element $S^{2n} \times S^{2n}$ is non-cancellable in $\overline{\mathscr{M}}_{4n}^{\mathrm{Cat,hc}}$, the monoid of highly connected $4n$-manifolds, once $n > 1$.

Interestingly, in some cases the monoids $\mathscr{M}_n^{\mathrm{Diff,hc}}$ are actually unique factorisation monoids. More precisely, by [67] and [62, Corollary 1.3] we have the following theorem.

Theorem 4 (Smale, Wall). *For $k \equiv 3, 5, 7 \bmod 8$, and $k \neq 15, 31, 63$, half the rank of H_k gives an isomorphism $\overline{\mathscr{M}}_{2k}^{\mathrm{Diff,hc}} \cong \mathbb{N}$. In particular, $\mathscr{M}_{2k}^{\mathrm{Diff,hc}}$ is a unique factorisation monoid in these cases.*

As a final remark consider for $n \neq 4$ the exact sequence

$$0 \to \Theta_n \to \mathscr{M}_n^{\mathrm{Diff}} \to \overline{\mathscr{M}}_n^{\mathrm{Diff}} \to 0$$

of abelian monoids. In general this sequence does not admit a retraction $\mathscr{M}_n^{\mathrm{Diff}} \to \Theta_n$: It is well-known that there are smooth manifolds M for which there exists a non-trivial homotopy sphere Σ such that $M\#\Sigma \cong M$, i.e. where the inertia group of M is non-trivial, [77, Theorem 1], [57, Theorem 1.1]. For instance one can consider $M = \mathbb{HP}^2$. The group of homotopy 8-spheres is isomorphic to $\mathbb{Z}/2$ and equals the inertia group of $\mathbb{HP}^2$. Applying a potential retraction of the above sequence to the equation $[M\#\Sigma] = [M]$ gives a contradiction as Θ_n is a group. Potentially, the above sequence might admit a splitting, but we will not investigate this further.

Remark 1. The topic of this paper is related to the notion of knot factorization. More precisely, Schubert [59] showed in 1949 that the monoid of oriented knots in S^3, where the operation is given by connected sum, is a unique factorisation monoid. It was shown by Kearton [31, 32] and Bayer [2] that the higher-dimensional analogue does *not* hold.

Remark 2. The question of whether a given n-manifold M is reducible is in general a hard problem and the answer can depend on the category. When $n \geq 3$ and M is a j-fold connected sum $M = M_1 \# \ldots \# M_j$, then $\pi_1(M) \cong \pi_1(M_1) * \cdots * \pi_1(M_j)$ is the free product of the fundamental groups of the summands. The converse of this statement goes by the name of the Kneser Conjecture. When $n = 3$, the Kneser Conjecture was proved by Stallings in his PhD thesis, see also [26, Theorem 7.1]. In higher dimensions, results of Cappell showed that the Kneser Conjecture fails [7, 8] and when $n = 4$, Kreck, Lück and Teichner [37] showed that the Kneser Conjecture fails in both $\mathscr{M}_4^{\mathrm{Diff}}$ and $\mathscr{M}_4^{\mathrm{Top}}$ and even gave an example of an irreducible smooth 4-manifold which is topologically reducible.

Organisation

The paper is organized as follows. In Section 2 we study the behaviour of complexity functions under the connected sum operation and use the results to provide the proof of Proposition 1 and the characterisation of the units. In Section 3 we extract some results from Wall's classification of highly-connected manifolds. In Section 4 we recall Wall's thickening operation which makes it possible to associate manifolds to CW-complexes. In Section 5 we show the existence of interesting pairs of 2-dimensional CW-complexes which allow us to prove Theorem 9 and 10. Furthermore in Section 6 we recall the construction of interesting pairs of 8-dimensional simply connected CW-complexes which leads to the proof of Theorem 3. In Section 7 we discuss the existence of prime manifolds in various monoids, in particular in we show that the Wu manifold $W = \mathrm{SU}(3)/\mathrm{SO}(3)$ is prime in $\mathscr{M}_5^{\mathrm{Cat,sc}}$. Finally in Section 8 we list some open problems.

2 The connected sum operation

We recall the definition of the connected sum. Let $n \in \mathbb{N}$ and let $M, N \in \mathscr{M}_n^{\mathrm{Cat}}$ be two Cat-manifolds. Given an orientation-preserving Cat-embedding $\varphi\colon \overline{B}^n \to M$ and given an orientation-reversing Cat-embedding $\psi\colon \overline{B}^n \to N$ we define the *connected sum* of M and N as

$$M \# N := \big(M \setminus \varphi\big(B^n\big)\big) \sqcup \big(N \setminus \psi\big(B^n\big)\big) \,/\, \varphi(P) = \psi(P) \text{ for all } P \in S^{n-1},$$

given a smooth structure (for Cat = Diff) by rounding the corners, and a piecewise linear one (for Cat = PL) by choosing appropriate triangulations that make the images of φ and ψ sub-complexes.

For Cat = Diff or PL the fact that the isomorphism type of the connected sum of two manifolds does not depend on the choice of embedding is a standard fact in (differential) topology, see e.g. [73, Theorem 2.7.4] and [58, Disc Theorem 3.34]. The analogous statement also holds for Cat = Top, but the proof (for $n > 3$) is significantly harder. It follows in a relatively straightforward way from the "annulus theorem" that was proved in 1969 by Kirby [34] for $n \neq 4$ and in 1982 by Quinn [54, 14] for $n = 4$. We refer to [15] for more details.

It is well-known that many invariants are well-behaved under the connected sum operation. Before we state the corresponding lemma that summarizes the relevant results we introduce our notation for intersection forms and linking forms. Given a $2k$-dimensional manifold W we denote by $Q_W\colon H_k(W;\mathbb{Z}) \times H_k(W;\mathbb{Z}) \to \mathbb{Z}$ the intersection form of W. Furthermore, given a $(2k+1)$-dimensional manifold W we denote by $\mathrm{lk}_W\colon \operatorname{Tors} H_k(W;\mathbb{Z}) \times \operatorname{Tors} H_k(W;\mathbb{Z}) \to \mathbb{Q}/\mathbb{Z}$ the linking form of W.

Lemma 1. *If $M_1, \dots, M_k$ are n-dimensional manifolds, then the following statements hold:*

1. *If $n \geq 3$, then $\pi_1(M_1 \# \dots \# M_k) \cong \pi_1(M_1) * \dots * \pi_1(M_k)$,*
2. *Let R be a ring. Then the cohomology ring $H^*(M_1 \# \dots \# M_k; R)$ is a quotient of a subring of the product $H^*(M_1;R) \times \dots \times H^*(M_k;R)$: First, consider the subring of this product where the elements in degree zero are in the image of the diagonal map $R \to R^k$ (recall that all M_i are connected). Then divide out the ideal generated by $(\mu_i - \mu_j)$ for $1 \leq i, j \leq k$, where μ_i is the cohomological fundamental class of M_i,*
3. *if n is even, then $Q_{M_1 \# \dots \# M_k} \cong Q_{M_1} \oplus \dots \oplus Q_{M_k}$,*
4. *if n is odd, then $\mathrm{lk}_{M_1 \# \dots \# M_k} \cong \mathrm{lk}_{M_1} \oplus \dots \oplus \mathrm{lk}_{M_k}$.*

Proof. By induction, it suffices to treat the case $k = 2$. (1) is an elementary application of the Seifert–van Kampen theorem. (2) follows from the cofibre sequence

$$S^{n-1} \to M_1 \# M_2 \to M_1 \vee M_2 \to S^n :$$

Recall that the subring described in the statement is precisely the cohomology of $M_1 \vee M_2$. It is then easy to see that the map $M_1 \vee M_2 \to S^n$ induces an injection on $H^n(-;R)$ with image precisely $\mu_1 - \mu_2$. Statements (3) and (4) follow from (2) and the fact that the described isomorphism of rings is compatible with the Bockstein operator. □

One can extract a crude complexity invariant from the above data, as follows. Given a finitely generated abelian group A we denote by $\operatorname{rank}(A) = \dim_{\mathbb{Q}}(A \otimes \mathbb{Q})$ its rank and we define

$$t(A) \;:=\; \ln(\#\text{torsion subgroup of } A).$$

Furthermore, given a finitely generated group we denote by $d(G) \in \mathbb{N}$ the minimal number of elements in a generating set for G.

Let M be an n-dimensional topological manifold. Then M is a retract of a finite CW-complex [5, p. 538], and thus has the homotopy of a CW complex (e.g. [19, Proposition A.11]) and its fundamental group and the (co-) homology groups of M are finitely presented. Thus we can define the *complexity of* M as

$$c(M) := d(\pi_1(M)) + \operatorname{rank}\left(\bigoplus_{i=1}^{n-1} H_i(M;\mathbb{Z})\right) + t\left(\bigoplus_{i=1}^{n-1} H_i(M;\mathbb{Z})\right) \in \mathbb{R}_{\geq 0}.$$

The following proposition summarizes two key properties of $c(M)$.

Proposition 2.

1. *For $n \geq 3$ the complexity gives a homomorphism $\mathscr{M}_n^{\mathrm{Cat}} \to \mathbb{R}_{\geq 0}$.*
2. *The kernel of c consists entirely of homotopy spheres.*

For the proof, recall the Grushko-Neumann Theorem which is proved in most text books on combinatorial group theory, e.g. [42, Corollary IV.1.9].

Theorem 5. (Grushko-Neumann Theorem) *Given any two finitely generated groups A and B we have*

$$d(A * B) = d(A) + d(B).$$

Proof (Proof of Proposition 2). The first statement follows from Lemma 1 and the Grushko-Neumann Theorem 5. The second statement follows since by the Hurewicz Theorem we have $\pi_n(M) \cong \mathbb{Z}$ if $c(M) = 0$, and a generator of $\pi_n(M)$ is represented by a map $S^n \to M$ which is a homotopy equivalence by Whitehead's theorem (and the observation above, that M has the homotopy type of a CW-complex). □

Proposition 3.

1. *All units of $\mathscr{M}_n^{\mathrm{Cat}}$ are homotopy spheres.*
2. *The converse to* (1) *holds if $n \neq 4$ or if $n = 4$ and* $\mathrm{Cat} = \mathrm{Top}$.
3. *The neutral element is the only unit in $\mathscr{M}_n^{\mathrm{Top}}$ and $\mathscr{M}_n^{\mathrm{PL}}$.*

By the work of Smale, Newman, Milnor and Kervaire the groups of homotopy spheres are of course relatively well understood for $n \geq 5$.

Proof.

1. Since units are mapped to units under homomorphisms it follows from Proposition 2 (2) that units are homotopy spheres.
2. For $n \leq 2$ the classification of n-manifolds clearly implies the converse to (1). In dimensions 3 and 4 the desired result follows straight from the Poincaré conjecture proved by Perelman and Freedman. Now let $n \geq 5$. Given an n-dimensional homotopy sphere Σ and an n-disc $D \subseteq \Sigma^n$, $(\Sigma \setminus \mathrm{int}(\mathrm{D})) \times I$ with an open $n+1$-disc removed away from the boundary is an h-cobordism between $\Sigma \# \overline{\Sigma}$ and S^n. For $n \neq 5$ the h-cobordism theorem then implies that Σ is indeed a unit. For $n = 5$, every homotopy sphere is h-cobordant to the standard sphere, see e.g. [36, Chapter X (6.3)], and hence diffeomorphic to S^5 by the h-cobordism theorem.

3. This result follows from the resolution of the Poincaré Conjecture. □

The following corollary implies in particular Proposition 1.

Corollary 1. *Unless $n = 4$ and* Cat $=$ Diff *or* PL*, the monoid $\mathscr{M}_n^{\mathrm{Cat}}$ does not admit infinite divisor chains, and therefore every manifold admits a decomposition into irreducible manifolds.*

Proof. An infinite divisor chain M_i gives rise to a descending sequence of natural numbers under c, which becomes stationary after index i say. But then for $j \geq i$ the elements witnessing that M_j is a summand of M_i have vanishing complexity and thus units by Proposition 3, so M_j is associated to M_i for all $j \geq i$. □

Similar arguments also allow us to identify some irreducible elements of $\mathscr{M}_n^{\mathrm{Cat}}$.

Corollary 2. *The manifolds $\mathbb{R}\mathrm{P}^{2n-1}, \mathbb{C}\mathrm{P}^n, \mathbb{H}\mathrm{P}^n, \mathbb{O}\mathrm{P}^2$ and $S^n \times S^{k-n}$ are irreducible in the monoid $\mathscr{M}_m^{\mathrm{Cat}}$ for any choice of* Cat *and appropriate dimension m, except possibly for $\mathscr{M}_4^{\mathrm{Diff}}$.*

Proof. This follows immediately from Lemma 1 (1) and (2). □

3 Wall's work on highly connected manifolds

The possibility of prime factorisations in $\overline{\mathscr{M}}_{2k}^{\mathrm{PL,hc}}$ and $\overline{\mathscr{M}}_{2k}^{\mathrm{Diff,hc}}$ was studied by Wall in [67, Problem 2A]. He classified such smooth manifolds in terms of their intersection form and an additional invariant $\alpha\colon H_k(M) \to \pi_k \mathrm{BSO}(k)$, which is given by representing an element in $H_k(M)$ by an embedded sphere and taking its normal bundle. In the case of piecewise linear manifolds Wall restricts attention to manifolds that can be smoothed away from a point; for even k, the map α is then well-defined (i.e. independent of the chosen smoothing) by [67, Lemma 2 & Formula (13)] and the injectivity of the stable J-homomorphism (which was not known at the time). For k odd, one furthermore has to invest the injectivity of the unstable J-homomorphism $\pi_k(\mathrm{BSO}(k)) \to \pi_{2k-1}(S^k)$.

Theorem 6 (Wall). *Unique factorisation in $\overline{\mathscr{M}}_{2k}^{\mathrm{PL,hc}}$ holds only for $k = 1, 3$ and possibly $k = 7$. In addition to these cases unique factorisation in $\overline{\mathscr{M}}_{2k}^{\mathrm{Diff,hc}}$ holds exactly for $k \equiv 3, 5, 7 \mod 8$ with $k \neq 15, 31$ and possibly $k \neq 63$ (if there exists a Kervaire sphere in dimension* 126*).*

In all cases that unique factorisation holds, the monoid in question is actually isomorphic to $\mathbb{N}$ via half the rank of the middle homology group, except possibly $\overline{\mathscr{M}}_{14}^{\mathrm{PL,hc}}$. In fact there does not seem to be a full description of $\overline{\mathscr{M}}_n^{\mathrm{PL,hc}}$ (or $\overline{\mathscr{M}}_n^{\mathrm{Top,hc}}$) in the literature. Let us remark, that the work of Kirby-Siebenmann implies $\mathscr{M}_n^{\mathrm{PL,hc}} \cong \mathscr{M}_n^{\mathrm{Top,hc}}$, once $n \geq 10$, as the obstruction to finding a PL-structure on

a topological manifold M is located in $H^4(M;\mathbb{Z}/2)$, with $H^3(M;\mathbb{Z}/2)$ acting transitively on isotopy classes of PL-structures. Wall's argument also shows that unique factorisation fails in $\overline{\mathscr{M}}_8^{\mathrm{Top,hc}}$.

Proof. It follows from Wall's work that for $k \geq 4$ even, the monoids $\overline{\mathscr{M}}_{2k}^{\mathrm{PL,hc}}$ and $\overline{\mathscr{M}}_{2k}^{\mathrm{Diff,hc}}$ never admit unique factorisations; this can be seen by picking an even positive definite unimodular form, and realizing it by a $(k-1)$-connected $2k$-dimensional manifold M with α-invariant whose values lie in $\ker(\pi_k \mathrm{BSO}(2k) \to \pi_k \mathrm{BSO})$; such α is uniquely determined by the intersection form by [67, Lemma 2] and the computation of $\pi_k \mathrm{BSO}(2k)$ on [67, p. 171]. By [67, Proposition 5], in this case a smooth realizing manifold exists whenever the signature is divisible by a certain index. Then $M\# -M \cong m(S^k \times S^k)\#\Sigma$ for some homotopy sphere Σ, where m is the rank of $H_k(M)$. But $S^k \times S^k$ cannot divide M or $-M$. Indeed, this follows from Lemma 1 (3), the fact that the intersection form of $S^k \times S^k$ is indefinite and the fact that the intersection forms of $\pm M$ are definite. See Proposition 9 below for a stronger statement in the case $k = 2$.

For odd values of k there are several cases to be distinguished. To start, for $k = 1,3$ and $\mathrm{Cat} = \mathrm{PL}$ or $k = 1,3,7$ and $\mathrm{Cat} = \mathrm{Diff}$ the monoid $\overline{\mathscr{M}}_{2k}^{\mathrm{Cat,hc}}$ is isomorphic to $\mathbb{N}$ via half the rank of the middle homology by [67, Lemma 5].

For other odd values of $k \neq 1,3,7$ unique decomposition in $\mathscr{M}_{2k}^{\mathrm{PL,hc}}$ never holds. This can be seen via the Arf-Kervaire invariant; this is the Arf invariant of a certain quadratic refinement of the intersection form. By the work of Jones and Rees and Stong [30, 63] any highly connected manifold of even dimension not $2,4,8$ or 16 possesses a canonical such refinement. We proceed by taking a manifold M which is smoothable away from a point with non-trivial Arf-Kervaire invariant and then decompose $M\#M$ into manifolds with vanishing Kervaire invariant and intersection form hyperbolic of rank 2; this is possible by [67, Lemmata 5 and 9] and the fact that the Arf-Kervaire invariant is additive.

For $\mathscr{M}_{2k}^{\mathrm{Diff,hc}}$ the argument above works equally well if there exists a smooth $2k$-manifold with Kervaire invariant one (which also implies the existence of an irreducible one by Corollary 1). This famously is the case if and only if $k = 1,3,7,15,31$ and possibly $k = 63$ [27], which rules out unique factorisation in these dimensions.

For $k = 3,5,7 \mod 8$ with $k \neq 3,7,15,31,63$ the monoid $\overline{\mathscr{M}}_{2k}^{\mathrm{Diff,hc}}$ is in fact isomorphic to $\mathbb{N}$ via half the rank of the middle homology by [67, Lemma 5] (with the case $\overline{\mathscr{M}}_{126}^{\mathrm{Diff,hc}}$ being open). In contrast, for $k \equiv 1 \mod 8$ the failure of unique decomposability can be seen by considering the composite homomorphism

$$S\alpha \colon H_k(M) \xrightarrow{\alpha} \pi_k \mathrm{BSO}(k) \longrightarrow \pi_k \mathrm{BSO} = \mathbb{Z}/2 :$$

Wall says that a manifold is of type 0 if $S\alpha$ is non-trivial and of type 1 if $S\alpha$ is trivial, see [67, p. 173, Case 5]. Note that the type is not additive but the connected sum of two manifolds has type 1 if and only if both manifolds have type 1. Hence by [67, Theorem 3] we can pick any two irreducible manifolds W_0, W_1 such the invariants from [67, Lemma 5] agree except that the type of W_i is i. Then $W_0\#W_1 \cong W_0\#W_0$ and unique decomposition fails. □

Remark 3. Given the first part of the proof above one might wonder whether for any highly connected manifold M, whose intersection form is even, $M\#\overline{M}$ is homeomorphic to $\#^k(S^n \times S^n)$. This is in fact not correct as the following example shows. Let M be the total space of an S^4-bundle over S^4 with non-trivial first Pontryagin class and trivial Euler class. It is then easy to see that the intersection form of M is even. However, since the rational Pontryagin classes are homeomorphism invariants we find that $M\#\overline{M}$ is not homeomorphic to $S^4 \times S^4 \# S^4 \times S^4$.

Cancellation in $\mathscr{M}_{2k}^{\mathrm{PL,hc}}$ and $\mathscr{M}_{2k}^{\mathrm{Diff,hc}}$ was also studied by Wall in [67, Problem 2C].

Theorem 7 (Wall). *Cancellation in $\overline{\mathscr{M}}_{2k}^{\mathrm{Diff,hc}}$ holds if and only if either $k = 1$ or $k \equiv 3,5,7 \mod 8$. In $\mathscr{M}_{2k}^{\mathrm{PL,hc}}$ the "only if" part still holds.*

Proof. Wall's classification directly shows that $S^k \times S^k$ is not cancellable in either $\mathscr{M}_{2k}^{\mathrm{PL,hc}}$ and $\mathscr{M}_{2k}^{\mathrm{Diff,hc}}$ if k is even by an argument similar to the one above. For $k = 2$, the failure of cancellation follows from the fact that $(S^2 \times S^2)\#\mathbb{C}\mathrm{P}^2$ and $\mathbb{C}\mathrm{P}^2\#\overline{\mathbb{C}\mathrm{P}}^2\#\mathbb{C}\mathrm{P}^2$ are diffeomorphic.

By [67, Lemma 5] and the classification of almost closed $n-1$-connected $2n$-manifolds by their n-types (see [67, page 170]), for $k \equiv 3,5,7 \mod 8$ the monoid $\mathscr{M}_{2k}^{\mathrm{Diff,hc}}$ embeds into $\mathbb{Z} \times \mathbb{Z}/2$ via the rank and the Arf-Kervaire invariant. If $k \equiv 1 \mod 8$, $k \neq 1$, then the examples from the previous proof exhibit the failure of cancellation. □

A similar analysis of $\overline{\mathscr{M}}_{2k+1}^{\mathrm{Cat,hc}}$ can be carried out using [72], but we refrain from spelling this out here.

4 Thickenings of finite CW-complexes

In this section we will see that one can associate to a finite CW-complex a smooth manifold which is unique in an appropriate sense. This procedure allows us to translate information about CW-complexes to manifolds. We will use this procedure in the proofs of Theorems 9, 10 and 3 in the following sections.

Convention 1 *By a* finite complex *we mean a finite connected CW-complex and by a* finite n-complex *we mean a finite connected n-dimensional CW-complex.*

4.1 Thickenings of finite CW-complexes

In this section we will summarize the theory of smooth thickenings of CW-complexes as developed in [70]. As is explained in [70] there is also an analogous theory of PL-thickenings.

We start out with the following definition, which is an adaptation of the definition on [70, p. 74] for our purposes.

Definition 3. Let X be a finite complex.

1. A *k-thickening of X* is a pair (M,ϕ) where M is an oriented, smooth, compact k-dimensional manifold with trivial tangent bundle, which has the property that the inclusion induced map $\pi_1(\partial M) \to \pi_1(M)$ is an isomorphism and where $\phi\colon X \to M$ is a simple homotopy equivalence.
2. Two k-thickenings (M,ϕ) and (N,ψ) of X are called *equivalent* if there exists an orientation-preserving diffeomorphism $f\colon M \to N$ such that $f\circ\phi$ is homotopic to ψ.

Note that ϕ does not have to be an embedding in the definition above.

Theorem 8. *Let X be a finite n-complex.*

1. If $k \geq 2n$, then there exists a k-thickening of X.
2. If $k \geq 2n+1$ and $k \geq 6$, then all k-thickenings of X are equivalent.

Proof. The theorem follows from [70, p. 76] using that every map from a path-connected space to $\mathbb{R}^k$ is 1-connected. □

Definition 4. Let $k \geq 2n+1$ and $k \geq 6$. Let X be a finite n-complex. We denote by $N^k(X)$ the oriented diffeomorphism type of the k-dimensional thickening of X. In our notation we will not distinguish between $N^k(X)$ and any representative thereof.

For convenience we state the following example.

Lemma 2. *If $k \geq 2n+1$ and $k \geq 6$, then $N^k(S^n) = S^n \times B^{k-n}$.*

Proposition 4. *Let X and Y be finite complexes. We suppose that $k \geq 2\dim(X)+1$, $k \geq 2\dim(Y)+1$ and $k \geq 6$. If X and Y are simple homotopy equivalent, then there exists an orientation-preserving diffeomorphism from $N^k(X)$ to $N^k(Y)$.*

Proof. Let $f\colon X \to Y$ be a simple homotopy equivalence. Let (M,ϕ) be a k-thickening for X and let (N,ψ) be a k-thickening for Y. Note that $(N,\psi\circ f)$ is a k-thickening for X. It follows from Theorem 8, and our dimension restrictions on k, that $N = N^k(X)$ is diffeomorphic to $M = N^k(Y)$. □

Lemma 3. *Let X and Y be finite complexes. If $k \geq 2\dim(X)+1$, $k \geq 2\dim(Y)+1$ and $k \geq 6$, then*

$$N^k(X \vee Y) \;=\; N^k(X) \#_b N^k(Y),$$

where "$\#_b$" denotes the boundary connected sum.

Proof. Let (M,ϕ) and (N,ψ) be k-thickenings of X and Y, respectively. After a simple homotopy we can and will assume that the the image of the wedge points under ϕ and ψ lies on the boundary of M and N. Note that $\phi\vee\psi\colon X\vee Y \to M\vee N$ is a simple homotopy equivalence and note that the inclusion $M\vee N \to M\#_b N$ is a simple homotopy equivalence. Thus we see that the map $\phi\vee\psi\colon X\vee Y \to M\#_b N$ is a simple homotopy equivalence. It follows almost immediately from Theorem 8 that $N^k(X\vee Y) = N^k(X)\#_b N^k(Y)$. □

4.2 Boundaries of thickenings of finite complexes

Definition 5. Let X be a finite complex and let $k \geq 2\dim(X)$ and $k \geq 5$. We write $M^k(X) := \partial N^{k+1}(X)$. Recall that $N^{k+1}(X)$ is an oriented manifold and we equip $M^k(X)$ with the corresponding orientation.

The following lemma is an immediate consequence of Lemma 2.

Lemma 4. *If $k \geq 2n$ and $k \geq 5$, then $M^k(S^n) = S^n \times S^{k-n}$.*

In the following proposition we summarize a few properties of $M^k(X)$.

Proposition 5. *For $n \in \mathbb{N}$ let X and Y be finite n-complexes. Furthermore let $k \in \mathbb{N}$ with $k \geq 2n$ and $k \geq 5$.*

1. *$M^k(X)$ is a closed oriented k-dimensional manifold,*
2. *if X and Y are simple homotopy equivalent, then there exists an orientation-preserving diffeomorphism from $M^k(X)$ to $M^k(Y)$,*
3. *$M^k(X \vee Y) = M^k(X) \# M^k(Y)$.*

If we have in fact $k \geq 2n+1$, then the following also holds:

4. *if $M^k(X)$ and $M^k(Y)$ are homotopy equivalent, then X and Y are homotopy equivalent.*

Proof. The first statement follows immediately from the definitions, the second from Proposition 4 and the third is a straightforward consequence of Lemma 3. The fourth statement is proved in [38, Proposition II.1]. □

Let $n \in \mathbb{N}$. Furthermore let $k \in \mathbb{N}$ with $k \geq 2n$ and $k \geq 5$. By Proposition 5 we obtain a well-defined map

$$M^k \colon \{\text{finite } n\text{-complexes}\}/\simeq_s \to \mathscr{M}_k^{\mathrm{Diff}}$$

where $\simeq_s$ denotes simple homotopy equivalence.

We conclude this section with the following corollary, which we will make use of in the proofs of Theorems 9, 10 and 3 respectively.

Corollary 3. *Let $n \in \mathbb{N}$. Furthermore let $k \in \mathbb{N}$ with $k \geq 2n$ and $k \geq 5$. Suppose that X and Y are finite complexes of dimension $\leq n$. We suppose that $X \not\simeq Y$ and $X \vee S^n \simeq_s Y \vee S^n$. Then $M^k(X) \not\simeq M^k(Y)$, but there is an orientation preserving diffeomorphism between $M^k(X)\#(S^n \times S^{k-n})$ and $M^k(Y)\#(S^n \times S^{k-n})$.*

Proof. This corollary follows immediately from the four statements of Proposition 5. □

4.3 5-dimensional thickenings

Now let X be a finite 2-complex. By Theorem 8 there exists a 5-thickening of X. We can no longer conclude from Theorem 8 that the thickening is well-defined up to diffeomorphism. But in fact the following weaker statement holds:

Proposition 6. *Let X be a finite 2-complex. If (M,ϕ) and (N,ψ) are 5-dimensional thickenings for X, then ∂M and ∂N are s-cobordant.*

This statement is implicit in [70], see also [38, p. 15]. The same way that we deduced Proposition 4 from Theorem 8 we can also deduce the following proposition from Proposition 6.

Proposition 7. *Let X and Y be finite 2-complexes. If X and Y are simple homotopy equivalent, then given any 5-dimensional thickenings A of X and B of Y the boundaries ∂A and ∂B are s-cobordant.*

5 Finite 2-complexes, group presentations and the D2-problem

The goal of this section is to prove Theorem 2 from the introduction and to give a survey of the various constructions that can be used to construct examples of non-cancellation in $\mathscr{M}_n^{\mathrm{Cat}}$. We will use:

Lemma 5. *Let $n \in \mathbb{N}$. Suppose there exist n-dimensional smooth manifolds M and N which are not homotopy equivalent but such that there is $r \geq 1$, an n-dimensional smooth manifold W and an orientation preserving diffeomorphism between $M\#r\cdot W$ and $N\#r\cdot W$. Then for every* Cat $=$ Top, PL *and* Diff *the following two statements hold:*

1. *The element W is not cancellable in $\overline{\mathscr{M}}_n^{\mathrm{Cat}}$.*
2. *The monoid $\mathscr{M}_n^{\mathrm{Cat}}$ is not a unique factorisation monoid.*

Proof. By hypothesis we know that $[M_1] \neq [M_2] \in \mathscr{M}_n^{\mathrm{Top}}$. By Proposition 3 we know that $\mathscr{M}_n^{\mathrm{Top}} = \overline{\mathscr{M}}_n^{\mathrm{Top}}$. In particular $[M_1] \neq [M_2] \in \overline{\mathscr{M}}_n^{\mathrm{Top}}$ and thus $[M_1] \neq [M_2] \in \overline{\mathscr{M}}_n^{\mathrm{Cat}}$. Furthermore we know that $[M_1] + r\cdot[W] = [M_2] + r\cdot[W] \in \overline{\mathscr{M}}_n^{\mathrm{Cat}}$. By induction we see that $[W]$ is not cancellable in $\overline{\mathscr{M}}_n^{\mathrm{Cat}}$. This implies that $\overline{\mathscr{M}}_n^{\mathrm{Cat}}$ is not isomorphic to some $\mathbb{N}^P$, i.e. $\overline{\mathscr{M}}_n^{\mathrm{Cat}}$ is not a unique factorisation monoid, hence $\mathscr{M}_n^{\mathrm{Cat}}$ is not a unique factorisation monoid. □

We will exploit this lemma with the following result:

Theorem 9. *Let $n \in \mathbb{N}_{\geq 5}$. Then there exist n-dimensional smooth manifolds M and N which are not homotopy equivalent but such that there is an orientation preserving diffeomorphism between $M\#(S^2 \times S^{n-2})$ and $N\#(S^2 \times S^{n-2})$.*

A slightly weaker result is also available in dimension 4:

Theorem 10. *There exist 4-dimensional smooth manifolds M and N which are not homotopy equivalent but such that there is an orientation preserving diffeomorphism between $M\#r\cdot(S^2\times S^2)$ and $N\#r\cdot(S^2\times S^2)$ for some $r\geq 1$.*

Taken together these results immediately imply Theorem 2 from the introduction.

5.1 Proof of Theorem 9

In this section we will provide the proof for Theorem 9. The key idea for finding suitable manifolds is to use Corollary 3. Thus our goal is to find finite 2-complexes X and Y which are not homotopy equivalent but such that $X\vee S^2$ and $Y\vee S^2$ are simple homotopy equivalent.

We begin our discussion with the following well-known construction of a finite 2-complex $X_{\mathscr{P}}$ with $\pi_1(X_{\mathscr{P}})=G$ from a group presentation

$$\mathscr{P}=\langle x_1,\cdots,x_s\mid r_1,\cdots,r_t\rangle$$

of a finitely presented group G. This is known as the Cayley complex $X_{\mathscr{P}}$ of the presentation $\mathscr{P}$ and has 1-skeleton a wedge of s circles, one circle for each generator x_i, with its 2-cells attached along the paths given by each relation r_i expressed as a word in the generators.

The following can be found in [6], [24, Theorem B]:

Theorem 11. *If X and Y are finite 2-complexes with $\pi_1(X)\cong\pi_1(Y)$ finite and $\chi(X)=\chi(Y)$, then $X\vee S^2\simeq_s Y\vee S^2$.*

Recall that, if $\mathscr{P}$ is a presentation of G with s generators and t relations, then the deficiency $\mathrm{def}(\mathscr{P})$ of $\mathscr{P}$ is $s-t$. The Euler characteristic of a presentation complex can be completely understood in terms of the deficiency:

Lemma 6. *If $\mathscr{P}$ is a group presentation, then $\chi(X_{\mathscr{P}})=1-\mathrm{def}(\mathscr{P})$.*

The task is therefore to find a finite group G with presentations $\mathscr{P}_1$ and $\mathscr{P}_2$ such that $X_{\mathscr{P}_1}\not\simeq X_{\mathscr{P}_2}$ and $\mathrm{def}(\mathscr{P}_1)=\mathrm{def}(\mathscr{P}_2)$. That $X_{\mathscr{P}_1}\vee S^2\simeq_s X_{\mathscr{P}_2}\vee S^2$ would then follow automatically from Theorem 11.

The first examples of such presentations were found by Metzler in [45] in the case $\pi_1(X)=(\mathbb{Z}/p)^s$ for $s\geq 3$ odd and $p\equiv 1 \mod 4$ prime. See [28, p. 297] for a convenient reference.

Theorem 12. *For $s\geq 3$ odd, $p\equiv 1 \mod 4$ prime and $p\nmid q$, consider presentations*

$$\mathscr{P}_q=\langle x_1,\ldots,x_s\mid x_i^p=1,[x_1^q,x_2]=1,[x_i,x_j]=1,1\leq i<j\leq s,(i,j)\neq(1,2)\rangle$$

for the group $(\mathbb{Z}/p)^s$. Then $X_{\mathscr{P}_q}\not\simeq X_{\mathscr{P}_{q'}}$ if $q(q')^{-1}$ is not a square mod p.

Remark 4. The smallest case for which this is satisfied is the case $p=5$, $s=3$, $q=1$ and $q'=2$, corresponding to the group $(\mathbb{Z}/5)^3$.

To prove these complexes are not homotopy equivalent, Metzler defined the bias invariant. This is a homotopy invariant defined for all finite 2-complexes and which was later shown in [61], [6] to be a complete invariant for finite 2-complexes with finite abelian fundamental group which led to a full (simple) homotopy classification in these cases.

We are now ready to prove Theorem 9.

Proof (Proof of Theorem 9). Let $s \geq 3$ be odd, $p \equiv 1 \mod 4$ be prime, and choose $q, q' \geq 1$ such that $p \nmid q$, $p \nmid q'$ and such that $q(q')^{-1}$ is not a square mod p. Let $\mathscr{P}_q$ and $\mathscr{P}'_q$ be the presentations for $(\mathbb{Z}/p)^s$ constructed above. Then $X_{\mathscr{P}_q} \not\simeq X_{\mathscr{P}'_q}$ by Theorem 12.

Since $\mathrm{def}(\mathscr{P}_q) = \mathrm{def}(\mathscr{P}'_q)$, we have $\chi(X_{\mathscr{P}_q}) = \chi(X_{\mathscr{P}'_q})$ by Lemma 6 and so

$$X_{\mathscr{P}_q} \vee S^2 \simeq_s X_{\mathscr{P}'_q} \vee S^2$$

are simply homotopy equivalent, by Theorem 11. Since $n \geq 5$ this fulfills the conditions of Corollary 3. Thus we see that $M = M^n(X_{\mathscr{P}_q})$ and $N = M^n(X_{\mathscr{P}'_q})$ have the desired properties. □

5.2 Proof of Theorem 10

Theorem 10. *There exist 4-dimensional smooth manifolds M and N which are not homotopy equivalent but such that there is an orientation preserving diffeomorphism between $M\#r\cdot(S^2\times S^2)$ and $N\#r\cdot(S^2\times S^2)$ for some $r \geq 1$.*

Proof. We use the same notation as in the proof of Theorem 9. By Theorem 8 there exist 5-dimensional thickenings A for $X_{\mathscr{P}_q}$ and B for $X_{\mathscr{P}'_q}$. We write $M = \partial A$ and $N = \partial B$.

As in Lemma 3 we see that $A\#_b(S^2\times\overline{B}^3)$ is a thickening of $X_{\mathscr{P}_q}\vee S^2$ and we see that $B\#_b(S^2\times\overline{B}^3)$ is a thickening of $X_{\mathscr{P}'_q}\vee S^2$. As in the proof of Theorem 9 we note that $X_{\mathscr{P}_q}\vee S^2$ is simple homotopy equivalent to $X_{\mathscr{P}'_q}\vee S^2$. Thus we obtain from Proposition 7 that $M\#(S^2\times S^2)$ and $N\#(S^2\times S^2)$ are s-cobordant. It follows from Wall [68, Theorem 3] (see also [60, p. 149] and [55, Theorem 1.1]) that these manifolds are diffeomorphic (via an orientation presentation diffeomorphism) after stabilisation by sufficiently many copies of $S^2\times S^2$, i.e.

$$M\#\underbrace{(S^2\times S^2)\#r(S^2\times S^2)}_{=(r+1)\cdot(S^2\times S^2)} \cong_{\mathrm{Diff}} N\#\underbrace{(S^2\times S^2)\#r(S^2\times S^2)}_{=(r+1)\cdot(S^2\times S^2)}$$

for some $r \geq 0$.

We still need to show that M and N are not homotopy equivalent. Since we are now dealing with the case $k = 4$ we cannot appeal to Proposition 5. But it is shown in [38, Theorem III.3] (see also [25, Proposition 4.3]) that M and N are indeed not homotopy equivalent. □

5.3 The D2 problem

We will now discuss a link to the work of C. T. C. Wall on the structure of finite complexes as it places the examples above into the framework of a more general conjecture.

Wall asked [69] whether or not a Dn complex, i.e. a finite complex X such that $H_i(X;M) = 0$ and $H^i(X;M) = 0$ for all $i \geq n+1$ and all finitely generated left $\mathbb{Z}[\pi_1(X)]$-modules M, is necessarily homotopy equivalent to a finite n-complex. This was shown to be true in the case $n > 2$ [71, Corollary 5.1] and in the case $n = 1$ [64], [65]. The case $n = 2$ remains a major open problem and is known as Wall's D2-problem.

Question 1. **(D2 Problem)** Let X be a D2 complex. Is X homotopy equivalent to a finite 2-complex?

We say that a group G has the *D2-property* if the D2-problem is true for all D2 complexes X with $\pi_1(X) = G$. This is relevant to the present discussion due to the following equivalent formulation.

Define an *algebraic 2-complex* $E = (F_*, \partial_*)$ over $\mathbb{Z}[G]$ to be a chain complex consisting of an exact sequence

$$F_2 \xrightarrow{\partial_2} F_1 \xrightarrow{\partial_1} F_0 \xrightarrow{\partial_0} \mathbb{Z} \longrightarrow 0$$

where $\mathbb{Z}$ is the $\mathbb{Z}[G]$-module with trivial G action and where the F_i are stably free $\mathbb{Z}[G]$-modules, i.e. $F_i \oplus \mathbb{Z}[G]^r \cong \mathbb{Z}[G]^s$ for some $r, s \geq 0$.

For example, if X is a finite 2-complex with a choice of polarisation $\pi_1(X) \cong G$, the chain complex $C_*(\widetilde{X})$ of the universal cover is a chain complex over $\mathbb{Z}[G]$ under the deck transformation action of G on $\widetilde{X}$. Since the action is free, $C_i(\widetilde{X})$ is free for all $i \geq 0$ and so $C_*(\widetilde{X})$ is an algebraic 2-complex over $\mathbb{Z}[G]$. We say an algebraic 2-complex over $\mathbb{Z}[G]$ is *geometrically realisable* if it is chain homotopy equivalent to $C_*(\widetilde{X})$ for some finite 2-complex X.

The following correspondence is established in [53, Theorem 1.1]:

Theorem 13. *If G is a finitely presented group, then there is a one-to-one correspondence between polarised D2 complexes X with $\pi_1(X) \cong G$ up to polarised homotopy and algebraic 2-complexes over $\mathbb{Z}[G]$ up to chain homotopy given by $X \mapsto C_*(\widetilde{X})$.*

In particular, G has the D2-property if and only if every algebraic 2-complex over $\mathbb{Z}[G]$ is geometrically realisable, as was already shown in [29] and [44].

One can thus search for further examples of finite 2-complexes X and Y for which $X \not\simeq Y$ and $X \vee S^2 \simeq_s Y \vee S^2$ by studying the chain homotopy types of algebraic 2-complexes over $\mathbb{Z}[G]$ for G having the D2-property.

A class of groups G for which it is feasible to classify algebraic 2-complexes over $\mathbb{Z}[G]$ up to chain homotopy are those with *n-periodic cohomology*, i.e. for which the Tate cohomology groups satisfy $\widehat{H}^i(G;\mathbb{Z}) = \widehat{H}^{i+n}(G;\mathbb{Z})$ for all $i \in \mathbb{Z}$. Let $m_{\mathbb{H}}(G)$ denote the number of copies of $\mathbb{H}$ in the Wedderburn decomposition of $\mathbb{R}[G]$ for a finite group G, i.e. the number of one-dimensional quaternionic representations. The following is a consequence of combining Theorem 11 with a special case of [52, Theorem A], which is proven as an application of a recent cancellation result for projective $\mathbb{Z}[G]$ modules [51].

Theorem 14. *If G has 4-periodic cohomology and $m_{\mathbb{H}}(G) \geq 3$. If G has the D2 property, then there exists finite 2-complexes X and Y with $\pi_1(X) \cong \pi_1(Y) \cong G$ for which $X \not\simeq Y$ and $X \vee S^2 \simeq_s Y \vee S^2$.*

Remark 5. More generally, [52, Theorem A] gives non-cancellation examples for finite n-complexes for all even $n > 2$ without any assumption on the D2 property.

Examples of groups with 4-periodic cohomology and $m_{\mathbb{H}}(G) \geq 3$ include the generalised quaternion groups

$$Q_{4n} = \langle x, y \mid x^n = y^2, yxy^{-1} = x^{-1} \rangle$$

for $n \geq 6$ and the groups $Q(2^n a; b, c)$ which appear in Milnor's list [47] for $n = 3$ or $n \geq 5$, and a, b, c odd coprime with $c \neq 1$ [53, Theorem 5.10].

It was shown in [53, Theorem 7.7] that Q_{28} has the D2-property (contrary to a previous conjecture [4]), and so Q_{28} gives an example where the hypotheses of Theorem 14 hold. In fact, the examples predicted by Theorem 14 were determined explicitly in [43]:

Proposition 8. *Consider the following presentations for Q_{28}:*

$$\mathscr{P}_1 = \langle x, y \mid x^7 = y^2, yxy^{-1} = x^{-1} \rangle, \quad \mathscr{P}_2 = \langle x, y \mid x^7 = y^2, y^{-1}xyx^2 = x^3y^{-1}x^2y \rangle.$$

Then $X_{\mathscr{P}_1} \not\simeq X_{\mathscr{P}_2}$ and $X_{\mathscr{P}_1} \vee S^2 \simeq_s X_{\mathscr{P}_2} \vee S^2$.

By Corollary 3 this shows that, for $k \geq 5$, we have that $M^k(X_{\mathscr{P}_1}) \not\simeq M^k(X_{\mathscr{P}_2})$ and

$$M^k(X_{\mathscr{P}_1}) \# (S^2 \times S^{k-2}) \cong M^k(X_{\mathscr{P}_2}) \# (S^2 \times S^{k-2}).$$

This gives an alternate way to prove Theorem 9, as well as giving an example whose fundamental group is non-abelian.

We conclude this section by remarking that, whilst Theorem 14 gives a reasonable place to look to find further non-cancellation examples, there is currently no known method to show that such examples exist without an explicit construction. Indeed, the presentations found in Proposition 8 are used in the proof of [53, Theorem 7.7].

6 Simply-connected complexes

Theorem 3 follows from the next theorem together with Lemma 5.

Theorem 15. *Let $k \geq 17$. There exist simply connected k-dimensional smooth manifolds M and N which are not homotopy equivalent but such that there is an orientation preserving diffeomorphism between $M\#(S^5 \times S^{k-5})$ and $N\#(S^5 \times S^{k-5})$.*

Remark 6. The bound $k \geq 17$ is an artifact of our method, and we expect similar examples to exist in a much lower range of dimensions. The strict analogue of Theorem 15 cannot, however, hold in dimension 4: It follows from Donaldson's Theorem [13, Theorem A], the classification of indefinite intersection forms and Freedman's' Theorem [16, Theorem 1.5] that any two 4-dimensional simply connected *smooth* manifolds that become diffeomorphic after the connected sum with $r \cdot (S^2 \times S^2)$ where $r \geq 1$, are already homeomorphic.

The key idea is once again to use Theorem 3. But this time we will use mapping cones to produce useful CW-complexes. We introduce the following notation.

Notation 1 *Let $\alpha\colon S^{m-1} \to S^n$ be a map. We denote its mapping cone by C_α. Note that C_α has a CW-structure with three cells, one in dimension 0, one in dimension n and one in dimension m.*

The following theorem is a practical machine for constructing interesting CW-complexes, see [22, Theorem 3.1 & Corollary 3.3]. Note that [22, 50] contain many other examples of CW-complexes exhibiting similar phenomena.

Theorem 16. *Let $m,n \in \mathbb{N}_{\geq 3}$ and let $[\alpha],[\beta] \in \pi_{m-1}(S^n)$ be elements of finite order. If $[\alpha]$ is in the image of the suspension homomorphism $\pi_{m-2}(S^{n-1}) \to \pi_{m-1}(S^n)$, then the following two statements hold:*

1. *$C_\alpha \simeq C_\beta$ if and only if $[\beta] = \pm[\alpha] \in \pi_{m-1}(S^n)$.*
2. *If $[\alpha]$ and $[\beta]$ generate the same subgroup of $\pi_{m-1}(S^n)$, then $C_\alpha \vee S^m \simeq C_\beta \vee S^m$.*

Proof. We prove (1) first and may assume that $m \geq n+2$, else the statements become easy (if $m-1=n$) or trivial (if $m-1<n$). The "if" part is obvious. To see the "only if" we consider a homotopy equivalence $f\colon C_\alpha \simeq C_\beta$. Such a homotopy equivalence induces an isomorphism on π_n, so that we find that there is a homotopy commutative diagram

$$\begin{array}{ccc} S^n & \longrightarrow & C_\alpha \\ \downarrow{\scriptstyle \pm 1} & & \downarrow{\scriptstyle f} \\ S^n & \longrightarrow & C_\beta \end{array}$$

in which the horizontal maps are the canonical maps. We denote by F_α and F_β the homotopy fibres of these horizontal maps. Since the composites $S^{m-1} \to S^n \to C$ ($C = C_\alpha, C_\beta$) are null homotopic, we obtain a canonical homotopy commutative diagram

$$\begin{array}{ccccccc} S^{m-1} & \longrightarrow & F_\alpha & \longrightarrow & S^n & \longrightarrow & C_\alpha \\ \vdots\pm 1 \downarrow & & \downarrow \bar{f} & & \downarrow \pm 1 & & \downarrow f \\ S^{m-1} & \longrightarrow & F_\beta & \longrightarrow & S^n & \longrightarrow & C_\beta \end{array}$$

where $\bar{f}$ is the induced homotopy equivalence of homotopy fibres. The relative Hurewicz theorem for the maps $S^n \to C_\alpha$ and $S^n \to C_\beta$, together with $m \geq n+2$, implies that the two maps $S^{m-1} \to F_\alpha$ and $S^{m-1} \to F_\beta$ induce isomorphisms on π_{m-1} so one obtains a dashed arrow making the diagrams commute up to homotopy. We deduce that there is a homotopy commutative diagram

$$\begin{array}{ccc} S^{m-1} & \xrightarrow{\alpha} & S^n \\ \downarrow \pm 1 & & \downarrow \pm 1 \\ S^{m-1} & \xrightarrow{\beta} & S^n \end{array}$$

Now we use that α is a suspension, so that post composition of α with a degree -1 map is just $-\alpha$ in $\pi_{m-1}(S^n)$.

To see (2) we consider the space $C_{\alpha,\beta} = C_\alpha \cup_\beta D^m$. Here, we view β as the composite $S^{m-1} \to S^n \to C_\alpha$. We note that $C_{\alpha,\beta} \cong C_{\beta,\alpha}$. By assumption β is contained in the subgroup generated by α. This implies the composite $S^{m-1} \to S^n \to C_\alpha$ is null homotopic so that $C_{\alpha,\beta} \simeq C_\alpha \vee S^m$. We thus obtain

$$C_\alpha \vee S^m \simeq C_{\alpha,\beta} \simeq C_{\beta,\alpha} \simeq C_\beta \vee S^m$$

where the very last equivalence follows from the assumption that α is also contained in the subgroup generated by β. □

Proof (Proof of Theorem 15). It is well known that the $\pi_6(S^3) \cong \mathbb{Z}/12$ and that suspension homomorphism $\pi_6(S^3) \to \pi_7(S^4)$ is injective with image $t(\pi_7(S^4))$; see [66, Proposition 5.6, Lemma 13.5]. Let μ be a generator of $t(\pi_7(S^4))$. We then consider the elements $\alpha = \mu$ and $\beta = 5 \cdot \mu$ in $t(\pi_7(S^4))$.

It follows from Theorem 16 that $C_\alpha \not\simeq C_\beta$ and that $C_\alpha \vee S^8 \simeq C_\beta \vee S^8$. Since these CW-complexes are simply connected, we have in fact $C_\alpha \vee S^8 \simeq_s C_\beta \vee S^8$. The theorem is now an immediate consequence of Corollary 3 applied to the 8-dimensional CW-complexes $X = C_\alpha$, $Y = C_\beta$, $n = 4$ and the given $k \geq 17$. □

7 Prime manifolds

Let $\mathscr{M}_n^{\mathrm{Cat,sc}}$ denote the submonoid of $\mathscr{M}_n^{\mathrm{Cat}}$ of simply connected manifolds. The question we want to ask in the present section is whether there exist prime manifolds in higher dimensions at all. While we do not know the answer, we will show that the Wu-manifold is prime among simply connected 5-folds.

As a warm-up recall that on the one hand the manifolds $\mathbb{C}\mathrm{P}^n, \overline{\mathbb{C}\mathrm{P}^n}$ and $S^n \times S^n$ are all irreducible by Corollary 2, except possibly for $n = 2$ and Cat = Diff or PL. On the other, as mentioned before, $(S^2 \times S^2)\#\mathbb{C}\mathrm{P}^2$ and $\mathbb{C}\mathrm{P}^2\#\overline{\mathbb{C}\mathrm{P}}^2\#\mathbb{C}\mathrm{P}^2$ are well-known to be diffeomorphic, see e.g. [18, p. 151] for details. These two observations imply immediately that none of $S^2 \times S^2, \mathbb{C}\mathrm{P}^2$ or $\overline{\mathbb{C}\mathrm{P}^2}$ are prime in $\mathscr{M}_4^{\mathrm{Cat,sc}}$ or $\mathscr{M}_4^{\mathrm{Cat}}$. In higher dimension we recorded similar behaviour for $S^{2k} \times S^{2k}$ in the first lines of the proof of Theorem 6.

Corollary 4. *Let $n \in \mathbb{N}_{\geq 2}$ be even. Then $S^n \times S^n$ is not prime in $\mathscr{M}_{2n}^{\mathrm{Cat,sc}}$ or $\mathscr{M}_{2n}^{\mathrm{Cat}}$.*

By contrast, for some odd n Theorem 6 also implies that $S^n \times S^n$ is prime in $\mathscr{M}_{2n}^{\mathrm{Diff,hc}}$. We do not know whether this extends to simply-connected manifolds, i.e. we do no know whether for those odd $S^n \times S^n$ is prime in $\mathscr{M}_{2n}^{\mathrm{Diff,sc}}$.

Proposition 9. *The monoid $\mathscr{M}_4^{\mathrm{Cat,sc}}$ has no prime elements. In particular, no simply-connected manifold is prime in $\mathscr{M}_4^{\mathrm{Cat}}$.*

Proof. By Freedman's classification, two simply connected, topological 4-manifolds are homeomorphic if and only if they have isomorphic intersection forms and the same Kirby-Siebenmann invariant. Hence for any such manifolds M there exist $m, m', n, n' \in \mathbb{N}, \varepsilon \in \{0,1\}$ such that $M\#m\mathbb{C}\mathrm{P}^2\#m'\overline{\mathbb{C}\mathrm{P}^2}\#\varepsilon * \mathbb{C}\mathrm{P}^2$ and $n\mathbb{C}\mathrm{P}^2\#n'\overline{\mathbb{C}\mathrm{P}^2}$ are homeomorphic. Similarly, if M is assumed smooth (or piecewise linear), it follows from the above and work of Wall [68, Theorem 3] (see also [60, p. 149] and [55, Theorem 1.1]) that there exist m, m', n, n' and $k \in \mathbb{N}$ such that $M\#m\mathbb{C}\mathrm{P}^2\#m'\overline{\mathbb{C}\mathrm{P}^2}\#k(S^2 \times S^2)$ and $n\mathbb{C}\mathrm{P}^2\#n'\overline{\mathbb{C}\mathrm{P}^2}\#k(S^2 \times S^2)$ are diffeomorphic. Using the fact that $(S^2 \times S^2)\#\mathbb{C}\mathrm{P}^2$ and $\mathbb{C}\mathrm{P}^2\#\overline{\mathbb{C}\mathrm{P}}^2\#\mathbb{C}\mathrm{P}^2$ are diffeomorphic we can arrange, at the cost of increasing m, m', n, n' that $k = 0$. In either case, if M is prime it follows that M is either $\mathbb{C}\mathrm{P}^2$ or $\overline{\mathbb{C}\mathrm{P}^2}$, since the latter manifolds are irreducible. But we observed above that they are not prime. □

Turning to dimension 5, recall that the Wu manifold $\mathrm{SU}(3)/\mathrm{SO}(3)$ is a simply connected, non-spin 5-manifold with $H_2(W;\mathbb{Z}) = \mathbb{Z}/2$.

Proposition 10. *The Wu-manifold $W = \mathrm{SU}(3)/\mathrm{SO}(3)$ is prime in $\mathscr{M}_5^{\mathrm{Cat,sc}}$.*

We will use Barden's classification of smooth simply connected 5-manifolds [1, 12]. There are two invariants which are important to us:

1. $H_2(M;\mathbb{Z})$ with its torsion subgroup $TH_2(M;\mathbb{Z})$. The group $TH_2(M;\mathbb{Z})$ is always isomorphic to $A \oplus A \oplus C$ where C is either trivial or cyclic of order 2 and A is some finite abelian group.
2. The height $h(M) \in \mathbb{N}_0 \cup \{\infty\}$: If M is spin, one sets $h(M) = 0$. If M is non-spin, $w_2 \colon H_2(M;\mathbb{Z}) \to \mathbb{Z}/2$ is a surjection. It is an algebraic fact that for any surjection $w \colon H \to \mathbb{Z}/2$ where H is a finitely generated abelian group, there exists an isomorphism $H \cong H' \oplus \mathbb{Z}/2^\ell$ such that w corresponds to the composite $H' \oplus \mathbb{Z}/2^\ell \to \mathbb{Z}/2^\ell \to \mathbb{Z}/2$. Here, ℓ is allowed to be ∞, where we (ab)use the notation

that $\mathbb{Z}/2^\infty = \mathbb{Z}$. The number ℓ is then defined to be the height $h(M)$ of M. Note that here we follow the wording of [1], an equivalent definition of the height is given in [12].

Barden's classification says that the map $\mathscr{M}_5^{\mathrm{Diff,sc}} \to \mathrm{Ab} \times (\mathbb{N}_0 \cup \{\infty\})$ sending a manifold M to the pair $(H_2(M;\mathbb{Z}), h(M))$ is injective, and that the following two statements are equivalent:

1. a pair (B,k) lies in the image,
2. the torsion subgroup TB is of the form $A \oplus A \oplus \mathbb{Z}/2$ if $k = 1$ and it is of the form $A \oplus A$ otherwise.

Moreover, the above map restricts to a bijection between spin manifolds and the pairs $(B,0)$ where $TB \cong A \oplus A$.

Lemma 7. *1.* $h(M\#N) = \begin{cases} h(M) & \text{if } h(N) = 0, \\ h(N) & \text{if } h(M) = 0, \\ \min(h(M), h(N)) & \text{if } h(M) \neq 0 \neq h(N) \end{cases}$

2. $h(M\#N) = 1$ *if and only if* $h(M) = 1$ *or* $h(N) = 1$,

3. The Wu manifold W *divides* M *if and only if* $h(M) = 1$,

Proof. To see (1), we observe that $M\#N$ is spin if and only if both M and N are spin. Furthermore, it is clear from the above definition of the height that if M is spin, then $h(M\#N) = h(N)$. To see the case where both M and N are not spin, it suffices to argue that if $\ell \leq k$, and we consider the map $\mathbb{Z}/2^k \oplus \mathbb{Z}/2^\ell \to \mathbb{Z}/2$ which is the sum of the canonical projections, then there is an automorphism of $\mathbb{Z}/2^k \oplus \mathbb{Z}/2^\ell$ such that this map corresponds to the map $\mathbb{Z}/2^k \oplus \mathbb{Z}/'^\ell \to \mathbb{Z}/2^\ell \to \mathbb{Z}/2$: The automorphism is given by sending $(1,0)$ to $(1,1)$ and $(0,1)$ to $(0,1)$. Statement (2) is then an immediate consequence of (1). To see (3) we first assume that W divides M, i.e. that M is diffeomorphic to $W\#L$. We find that $h(W\#L) = 1$ by (1). Conversely, suppose that $h(M) = 1$. By Barden's classification, we know that the *torsion* subgroup of $H_2(M;\mathbb{Z})$ is of the form $A \oplus A \oplus \mathbb{Z}/2$, in particular $H_2(M;\mathbb{Z}) \cong \mathbb{Z}^n \oplus A \oplus A \oplus \mathbb{Z}/2$ for some $n \geq 0$. Again, by the classification, there exists a spin manifold L with $H_2(L;\mathbb{Z}) \cong \mathbb{Z}^n \oplus A \oplus A$. We find that $W\#L$ and M have isomorphic $H_2(-;\mathbb{Z})$ and both height 1, so they are diffeomorphic, and thus W divides M. □

Proof (Proof of Proposition 10). First we consider $\mathscr{M}_5^{\mathrm{PL,sc}} = \mathscr{M}_5^{\mathrm{Diff,sc}}$. If W divides $M\#N$, then $h(M\#N) = 1$, we may then without loss of generality assume that $h(M) = 1$, so that W divides M.

Every simply connected topological 5-manifold admits a smooth structure, since the Kirby-Siebenmann invariant lies in $H^4(M;\mathbb{Z}/2) \cong H_1(M;\mathbb{Z}/2) = 0$. As the invariants in Barden's classification are homotopy invariants, it follows that $\mathscr{M}_5^{\mathrm{Top,sc}} = \mathscr{M}_5^{\mathrm{Diff,sc}}$. In particular, the Wu manifold is also prime in $\mathscr{M}_5^{\mathrm{Top,sc}}$. □

8 Questions and problems

We conclude this paper with a few questions and challenges.

Question 2.

1. Let $n \geq 4$. Does there exist a non-trivial cancellable element in any of the monoids $\mathscr{M}_n^{\mathrm{Top}}$, $\mathscr{M}_n^{\mathrm{PL}}$, $\mathscr{M}_n^{\mathrm{Diff}}$?
2. Let $n \geq 6$. Does there exist a non-trivial cancellable element in any of the monoids $\mathscr{M}_n^{\mathrm{Top,sc}}$, $\mathscr{M}_n^{\mathrm{PL,sc}}$, $\mathscr{M}_n^{\mathrm{Diff,sc}}$?

Question 3.

1. Let $n \geq 4$. Does there exist a prime element in any of the monoids $\mathscr{M}_n^{\mathrm{Top}}$, $\mathscr{M}_n^{\mathrm{PL}}$, $\mathscr{M}_n^{\mathrm{Diff}}$?
2. Let $n \geq 6$. Does there exist a prime element in any of the monoids $\mathscr{M}_n^{\mathrm{Top,sc}}$, $\mathscr{M}_n^{\mathrm{PL,sc}}$?

In light of Theorem 3 and the remark on page 8 we also raise the following related question.

Question 4. For which $n \in 5, \ldots, 16$ is $\mathscr{M}_n^{\mathrm{Cat,sc}}$ a unique factorisation monoid?

The following question arises naturally from Proposition 10.

Question 5. Is the Wu manifold prime in $\mathscr{M}_5^{\mathrm{Top}}$ or $\mathscr{M}_5^{\mathrm{Diff}}$?

Throughout the paper we worked mostly with simply connected and highly connected manifolds. It is reasonable to ask what is happening at the end of the spectrum, namely when we restrict ourselves to aspherical manifolds.

Question 6. In any of the three categories Top, PL and Diff, is the monoid generated by aspherical manifolds a unique decomposition factorization monoid for $n \geq 4$?

As a partial answer to Question 6, we would like to thank the referee for pointing out that in the topological category unique decomposition factorization is implied by the Borel conjecture. To see this, we first note that the fundamental group of an aspherical manifold can not be a non-trivial free product of groups.

Lemma 8. *Let M be aspherical closed n-manifold and let $\pi_1(M) \cong G * H$. Then either G or H is trivial.*

Proof. As M is aspherical the assumption that $\pi_1(M) \cong G * H$ implies that M is homotopy equivalent to $B(G*H) \simeq BG \vee BH$. Note that for all $k \in \mathbb{N}$ we have

$$\widetilde{H}_k(BG \vee BH; \mathbb{Z}/2) \cong \widetilde{H}_k(BG; \mathbb{Z}/2) \oplus \widetilde{H}_k(BH; \mathbb{Z}/2).$$

Hence $\mathbb{Z}/2 \cong H_n(M; \mathbb{Z}/2) \cong H_n(BG \vee BH; \mathbb{Z}/2) \cong H_n(BG; \mathbb{Z}/2) \oplus H_n(BH; \mathbb{Z}/2)$. Now we suppose without loss of generality that $H_n(BH; \mathbb{Z}/2)$ is trivial. Consider the cover $\widehat{M}$ of M corresponding to the canonical map $\pi_1(M) \cong G * H \to H$. Then $\widehat{M}$ is homotopy equivalent to $\bigvee_{i=1}^{|H|} BG$ and hence $H_n(\widehat{M}; \mathbb{Z}/2) \cong \bigoplus_{i=1}^{|H|} \mathbb{Z}/2$. As $\widehat{M}$ is a manifold, this implies $|H| = 1$ and thus H is trivial. □

Recall that by the Grushko–Neumann theorem [42, Corollay IV.1.9] together with the Kurosh isomorphism theorem [40, Isomorphiesatz][41, p. 27] we obtain the following lemma.

Lemma 9. *Every non-trivial, finitely generated group G can be decomposed as a free product*

$$G \cong A_1 * \ldots * A_r * F_k,$$

where F_k is a free group of rank k, each of the groups A_i is non-trivial, freely indecomposable and not infinite cyclic; moreover, for a given G, the numbers r and k are uniquely determined and the groups $A_1, \ldots, A_r$ are unique up to reordering and conjugation in G. That is, if $G \cong B_1 * \ldots B_s * F_l$ is another such decomposition then $r = s$, $k = l$, and there exists a permutation $\sigma \in S_r$ such that for each $i = 1, \ldots, r$ the subgroups A_i and $B_{\sigma(i)}$ are conjugate in G.

Proposition 11. *Assume that the Borel conjecture is true. Then in* Top *the monoid generated by aspherical manifolds is a unique decomposition factorization monoid for $n \geq 4$.*

Proof. Let $N := N_1 \# \ldots \# N_k$ and $M := M_1 \# \ldots \# M_l$ with N_i and M_j aspherical for all i and j. Crushing all connecting spheres to points, we obtain projection maps $p_N \colon N \to N_1 \vee \ldots \vee N_k$ and $p_M \colon M \to M_1 \vee \ldots \vee M_l$. Suppose there is an orientation preserving homeomorphism $f \colon N \xrightarrow{\cong} M$. The maps p_N and p_M induce isomorphisms on the fundamental groups by Lemma 1.(1). We consider the isomorphism $\varphi := p_{M*} \circ f_* \circ p_{N*}^{-1} \colon \pi_1(N_1) * \cdots * \pi_1(N_k) \to \pi_1(M_1) * \cdots * \pi_1(M_l)$. Note that $\pi_1(N_j)$ is never infinite cyclic since the dimension of N_j is larger than one. By Lemma 8 and Lemma 9, $k = l$ and for each j there is an i with $\pi_1(N_j) \cong \pi_1(M_i)$. Moreover, since $\varphi(\pi_1(N_j))$ is conjugate to $\pi_1(M_i)$ in $\pi_1(M)$, we see that the map $\pi_1(N_j) \to \pi_1(N_1) * \cdots * N_k \xrightarrow{\varphi} \pi_1(M_1) * \cdots * \pi_1(M_k) \to \pi_1(M_i)$ is an isomorphism. Hence it induces a homotopy equivalence $N_j \to M_i$. The fundamental class of N is mapped to the fundamental classes $([N_j])_j \in \bigoplus_{r=1}^{k} H_n(N_r; \mathbb{Z}) \cong H_n(N_1 \vee \ldots \vee N_k; \mathbb{Z})$. Hence the homotopy equivalence $N_j \to M_i$ is orientation preserving. Assuming the Borel conjecture, it follows that N_j is orientation preserving homeomorphic to M_i. Thus the decomposition is unique. □

Remark 7. Finally, in the smooth and *PL* categories in dimensions $n \geq 5$, we also thank the referee for suggesting that the existence of exotic tori might lead to the failure of unique factorisation in the monoid generated by aspherical manifolds: We think that this is an attractive approach to attacking Question 6.

Acknowledgements Most of the work on this paper happened while the authors attended the workshop "Topology of Manifolds: Interactions Between High and Low Dimensions" that took place January 7th-18th at the Mathematical Research Institute MATRIX in Creswick, Australia. We are very grateful to MATRIX for providing an excellent research environment.

SF and ML were supported by the SFB 1085 "Higher Invariants" at the University of Regensburg funded by the DFG, FH and DK were funded by the Deutsche Forschungsgemeinschaft

(DFG, German Research Foundation) under Germany's Excellence Strategy - GZ 2047/1, Projekt-ID 390685813. JN was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/N509577/1.

We wish to thank Mark Powell, Manuel Krannich, and Patrick Orson for helpful conversations. We also wish to thank the referee for several useful comments. In particular the proof of Proposition 11 was suggested to us by the referee.

References

1. D. Barden: Simply connected five-manifolds. Ann. of Math. **82**, 365–385 (1965)
2. E. Bayer: Factorisation is not unique for higher dimensional knots. Comment. Math. Helv. **55**, 583-592 (1980)
3. M. Behrens, M. Hill, M. Hopkins and M. Mahowald: Detecting exotic spheres in low dimensions using coker J, J. Lond. Math. Soc. (2) **101**, 1173–1218 (2020)/
4. F. R. Beyl and N. Waller: A stably free nonfree module and its relevance for homotopy classification, case Q_{28}. Algebr. Geom. Topol. **5**, 899–910 (2005)
5. G. Bredon: Topology and geometry, Graduate Texts in Mathematics 139. Springer-Verlag (1993)
6. W. H. Browning. Homotopy types of certain finite CW-complexes with finite fundamental group, PhD Thesis, Cornell University (1979)
7. S. E. Cappell: On connected sums of manifolds, Topology **13**, 395–400 (1974)
8. S. E. Cappell: A spitting theorem for manifolds, Inventiones Math. **33**, 69–170 (1976)
9. S. E. Cappell and J. L. Shaneson: On four-dimensional surgery and applications, Comment. Math. Helv. **46**, 500–528 (1971)
10. J. Cerf: Sur les difféomorphismes de la sphère de dimension trois $(\Gamma_4 = 0)$, Lecture Notes in Mathematics 53, Springer Verlag (1968)
11. J. Cheeger and J. M. Kister: Counting topological manifolds, Topology **9**, 149–151 (1970)
12. D. Crowley: 5-manifolds: 1-connected, Bulletin of the Manifold Atlas, 49-55 (2011)
13. S. K. Donaldson: An application of gauge theory to four-dimensional topology, J. Diff. Geom. **18**, 279–315 (1983)
14. R. D. Edwards: The solution of the 4-dimensional Annulus conjecture (after Frank Quinn), in "Four-manifold Theory", Gordon and Kirby ed., Contemporary Math. **35**, 211–264 (1984)
15. S. Friedl, M. Nagel, P. Orson and M. Powell. A survey of the foundations of 4-manifold theory, Preprint (2019), arXiv:1910.07372.
16. M. Freedman: The topology of four-dimensional manifolds, J. Diff. Geom. **17**, 357–453 (1982)
17. H. Freudenthal: Über die Klassen der Sphärenabbildungen I. Große Dimensionen, Compos. Math. **5**, 299–314 (1938)
18. R. Gompf and A. Stipsicz: 4-manifolds and Kirby calculus, Graduate Studies in Mathematics 20, AMS (1999)
19. A. Hatcher: Algebraic Topology, Cambridge University Press (2001)
20. P. J. Hilton: On the homotopy groups of the union of spheres, J. London Math. Soc., Second Series, **30**, 154–172 (1955)
21. P. J. Hilton: Homotopy Theory and Duality, Gordon and Breach, New York, (1965)
22. P. J. Hilton: On the Grothendieck group of compact polyhedra, Fund. Math. **61**, 199–214 (1967)
23. I. Hambleton and M. Kreck: On the classification of topological 4-manifolds with finite fundamental group, Math. Annalen. **280**, 85–104 (1988)
24. I. Hambleton and M. Kreck: Cancellation of lattices and finite two-complexes, J. reine angew. Math **442**, 91–109 (1993).

25. I. Hambleton and M. Kreck: Cancellation results for 2-complexes and 4-manifolds and some applications, Two-dimensional homotopy and combinatorial group theory, 281-308, London Math. Soc. Lecture Note Ser., 197, Cambridge Univ. Press, Cambridge (1993).
26. J. Hempel: 3-manifolds, Annals of Mathematics Studies 86. Princeton, New Jersey: Princeton University Press and University of Tokyo Press. XII (1976)
27. M. Hill, M. Hopkins and D. Ravenel: On the nonexistence of elements of Kervaire invariant one, Ann. of Math. (2) **184**, 1 - 262 (2016)
28. C. Hog-Angeloni, W. Metzler and A. J. Sieradski: Two-Dimensional Homotopy and Combinatorial Group Theory, London Math. Soc. Lecture Note Ser. 197, Cambridge University Press (1993)
29. F. E. A. Johnson: Stable Modules and the D(2)-Problem, LMS Lecture Notes Series 301 (2003)
30. J. Jones and E. Rees: Kervaire's invariant for framed manifolds, Proc. Symp. Pure Math. **32**, 141–147 (1978)
31. C. Kearton: Factorisation is not unique for 3-knots, Indiana Univ. Math. J. **28**, 451-452 (1979)
32. C. Kearton: The factorisation of knots, Low-dimensional topology, Proc. Conf., Bangor/Engl. 1979, Vol. 1, Lond. Math. Soc. Lect. Note Ser. **48**, 71-80 (1982)
33. M. Kervaire and J. Milnor: Groups of Homotopy Spheres: I, Ann. Math. **77**, 504–537 (1963)
34. R. Kirby: Stable homeomorphisms and the annulus conjecture, Ann. Math. **89**, 575–582 (1969)
35. H. Kneser: Geschlossene Flächen in dreidimensionalen Mannigfaltigkeiten, Jber. Deutsch. Math.-Verein. **38**, 248–260 (1929)
36. A. Kosinski: Differential Manifolds, Dover Publications, 1993.
37. M. Kreck, W. Lück and P. Teichner, Counterexamples to the Kneser conjecture in dimension four, Comment. Math. Helvetici **70**, 423–433 (1995)
38. M. Kreck and J. A. Schafer: Classification and stable classification of manifolds: some examples, Comment. Math. Helv. **59**, 12–38 (1984)
39. M. Kreck: Surgery and duality, Ann. Math. (2) **149**, 707–754 (1999)
40. A. G. Kurosh: Die Untergruppen der freien Produkte von beliebigen Gruppen, Math. Annalen **109**, 647–660 (1934)
41. A. G. Kurosh: The theory of groups, Volume 2, Chelsea Publishing Co. (1960)
42. R. Lyndon and P. Schupp: Combinatorial group theory, Springer Verlag (1977)
43. W. Mannan and T. Popiel: An exotic presentation of Q_{28}, arXiv:1901.10786 (2019)
44. W. H. Mannan: Realizing algebraic 2-complexes by CW-complexes, Math. Proc. Cam. Phil. Soc. **146** (3), 671–673 (2009)
45. W. Metzler: Über den Homotopietyp zweidimensionaler CW-Komplexe und Elementartransformationen bei Darstellungen von Gruppen durch Erzeugende und definierende Relationen, J. reine angew. Math. **285**, 7–23 (1976)
46. J. Milnor: A unique factorisation theorem for 3-manifolds, Amer. J. Math. **84**, 1-7 (1962)
47. J. Milnor: Groups which act on S^n without fixed points, Amer. J. Math. **79** (3), 623–630 (1957)
48. E. Moise: Affine structures in 3-manifolds V. The triangulation theorem and Hauptvermutung, Ann. of Math. **56**, 96–114 (1952)
49. E. Moise: Geometric Topology in Dimensions 2 and 3, Graduate Texts in Mathematics 47, Springer, New York-Heidelberg, 1977.
50. E. A. Molnar: Relation between cancellation and localization for complexes with two cells, J. Pure Appl. Algebra **3**, 141–158 (1973)
51. J. Nicholson: A cancellation theorem for modules over integral group rings, Math. Proc. Cambridge Philos. Soc., to appear, arXiv:1807.00307 (2018)
52. J. Nicholson: Cancellation for (G,n)-complexes and the Swan finiteness obstruction, arXiv:2005.01664 (2020)
53. J. Nicholson: On CW-complexes over groups with periodic cohomology, arXiv:1905.12018 (2019)
54. F. Quinn: Ends of maps. III: Dimensions 4 and 5, J. Differ. Geom. **17**, 503–521 (1982)

55. F. Quinn: The stable topology of 4-manifolds, Top. Appl. **15**, 71–77 (1983)
56. T. Radó: Über den Begriff der Riemannschen Fläche, Acta Szeged **2**, 101–121 (1926)
57. K. Ramesh: Inertia groups and smooth structures of $(n-1)$-connected $2n$-manifolds, Osaka J. Math. **53**, 303–319 (2016)
58. C. P. Rourke and B. J. Sanderson: Introduction to piecewise-linear topology, Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag (1972)
59. H. Schubert: Die eindeutige Zerlegbarkeit eines Knotens in Primknoten, S.-B. Heidelberger Akad. Wiss. Math.-Nat. Kl. 1949 no. 3, 57–104 (1949)
60. A. Scorpan: The wild world of 4-manifolds, Providence, RI: American Mathematical Society (2005)
61. A. J. Sieradski: A semigroup of simple homotopy types, Math. Zeit. **153**, 135–148 (1977)
62. S. Smale: On the structure of manifolds, Amer. J. Math. **84**, 387–399 (1962).
63. R. E. Stong: Determination of $H^*(BO\langle k\rangle, \mathbb{Z}/2)$ and $H^*(BU\langle k\rangle, \mathbb{Z}/2)$, Trans. Amer. Math. Soc. **107**, 526-544 (1963)
64. J. R. Stallings: On torsion free groups with infinitely many ends, Ann. Math. **88**, 312–334 (1968)
65. R. G. Swan: Groups of cohomological dimension one, J. of Algebra. **12**, 585–610 (1969)
66. H. Toda: Composition methods in homotopy groups of spheres, Princeton University Press, 1962.
67. C. T. C. Wall: Classification of $(n-1)$-connected 2n-manifolds, Ann. Math. **75**, 163–189 (1962)
68. C. T. C. Wall: On simply-connected 4-manifolds, J. London Math. Soc. **39**, 141–149 (1964)
69. C. T. C. Wall: Finiteness conditions for CW Complexes, Ann. of Math. **81**, 56–69 (1965)
70. C. T. C. Wall: Classification problems in Differential Topology - IV Thickenings, Topology **5**, 73–94 (1966)
71. C. T. C. Wall: Finiteness conditions for CW complexes, II. Proc. Roy. Soc. Ser. A **295**, 129–139 (1966)
72. C. T. C. Wall: Classification problems in differential topology. VI. Classification of $(s-1)$-connected $(2s+1)$-manifolds. Topology **6**, 273-296 (1967)
73. C. T. C. Wall: Differential topology, Cambridge Studies in Advanced Mathematics 156. Cambridge University Press (2016)
74. G. Wang and Z. Xu, *The triviality of the 61-stem in the stable homotopy groups of spheres*. Ann. Math. (2) **186**, 501 - 580 (2017)
75. G. W. Whitehead: Elements of Homotopy Theory, Graduate Texts in Mathematics 61, Springer-Verlag, New York, (1978)
76. J. H. C. Whitehead: Manifolds with transverse fields in euclidean space, Ann. Math. (2) **73**, 154-212 (1961)
77. D. L. Wilkens: On the inertia group of certain manifolds, J. London Math. Soc. **9**, 537–548 (1974/75)

The Levine-Tristram signature: a survey

Anthony Conway

Abstract The Levine-Tristram signature associates to each oriented link L in S^3 a function $\sigma_L\colon S^1 \to \mathbb{Z}$. This invariant can be defined in a variety of ways, and its numerous applications include the study of unlinking numbers and link concordance. In this survey, we recall the three and four dimensional definitions of σ_L, list its main properties and applications, and give comprehensive references for the proofs of these statements.

1 Introduction

Given an oriented link $L \subset S^3$, the Levine-Tristram signature is a function $\sigma_L\colon S^1 \to \mathbb{Z}$ whose study goes back to the sixties [94, 57]. The main goal of this survey article is to collect the various definitions of σ_L, while a secondary aim is to list its properties. Although some elementary arguments are outlined in the text, we provide detailed external references for most of the proofs. Briefly, we will discuss the definition in terms of Seifert matrices, various 4-dimensional interpretations as well as connections to pairings on the Alexander module. The next paragraphs give the flavor of some of these constructions.

Most knot theory textbooks that cover the Levine-Tristram signature introduce it using Seifert matrices [61, 49, 46, 67]. Indeed, as we review in Section 2, the Levine-Tristram signature at $\omega \in S^1$ can be defined using any Seifert matrix A for L by setting

$$\sigma_L(\omega) = \operatorname{sign}(1-\omega)A + (1-\overline{\omega})A^T.$$

In the same section, we collect the numerous properties of σ_L: after listing its behavior under mirror images, orientation reversals and satellite operations, we review ap-

Anthony Conway
Department of Mathematical Sciences, Durham University, DH1 3LE, United Kingdom
e-mail: anthonyyconway@gmail.com

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_2

plications to unlinking numbers, link concordance and discuss various incarnations of the Murasugi-Tristram inequality [78, 94].

The signature admits several 4-dimensional interpretations: either using covers of D^4 branched along surfaces cobounding L [96], or applying twisted signatures, or as invariants of the zero framed surgery along L. Before discussing these constructions in detail in Section 3, let us briefly sketch one of them. Given a locally flat compact connected oriented surface $F \subset D^4$ with boundary L, we set $W_F := D^4 \setminus \nu F$ and consider the coefficient system $\pi_1(W_F) \to H_1(W_F) \cong \mathbb{Z} \to \mathbb{C}$ which maps the meridian of F to ω. This gives rise to a twisted intersection form $\lambda_{\mathbb{C}_\omega}(W_F)$ on the twisted homology $\mathbb{C}$-vector space $H_2(W_F; \mathbb{C}_\omega)$ whose signature coincides with the Levine-Tristram signature:

$$\sigma_L(\omega) = \operatorname{sign} \lambda_{\mathbb{C}_\omega}(W_F).$$

Section 4 is concerned with methods of extracting $\sigma_K(\omega)$ from pairings on the Alexander module $H_1(X_K; \mathbb{Z}[t^{\pm 1}])$ of a knot K (here we write $X_K := S^3 \setminus \nu K$ for the exterior of K) [76, 52]. Briefly, the signature σ_K can be extracted by considering the primary decomposition of $H_1(X_K; \mathbb{R}[t^{\pm 1}])$ and by studying the Milnor pairing or the Blanchfield pairing

$$H_1(X_K; \mathbb{Z}[t^{\pm 1}]) \times H_1(X_K; \mathbb{Z}[t^{\pm 1}]) \to \mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}].$$

In fact, as we discuss in Section 5, the signature can also be understood as a signed count of $\mathrm{SU}(2)$ representations of $\pi_1(X_K)$ with fixed meridional traces [64, 45], or in terms of the Meyer cocycle and the Burau representation [34]. Summarizing, σ_L admits a wealth of definitions, which never seemed to have been collected in a single article.

We conclude this introduction with two remarks. Firstly, note that we mention neither the Gordon-Litherland pairing [40] nor the multivariable signature [15]. Secondly, we stress that even though σ_L was defined 50 years ago, it continues to be actively studied nowadays. We mention some recent examples: results involving concordance properties of positive knots can be found in [3]; the behavior of σ_L under splicing is now understood [23]; the relation between the jumps of σ_L and the zeroes of Δ_L has been clarified [38, 63]; a diagrammatic interpretation of σ_L (inspired by quantum topology) is conjectured in [87]; there is a characterization of the functions that arise as knot signatures [70]; new lower bounds on unknotting numbers have been obtained via σ_K [71]; there is a complete description of the $\omega \in S^1$ at which σ_L is a concordance invariant [80]; and σ_L is invariant under *topological* concordance [84].

This survey is organized is as follows. In Section 2, we review the Seifert matrix definition of σ_L and list its properties. In Section 3, we outline and compare the various four dimensional interpretations of σ_L. In Section 4, we give an overview of the definitions using the Milnor and Blanchfield pairings. In Section 5, we outline additional constructions in terms of $\mathrm{SU}(2)$ representations and braids.

2 Definition and properties

In this section, we review the definition of the Levine-Tristram and nullity using Seifert matrices (Subsection 2.1) before listing several properties of these invariants (Subsections 2.2, 2.3 and 2.4). Knot theory textbooks which mention the Levine-Tristram signature include [61, 49, 46, 67].

2.1 The definition via Seifert surfaces

A *Seifert surface* for an oriented link L is a compact oriented surface F whose oriented boundary is L. While a Seifert surface may be disconnected, we require that it has no closed components. Since F is oriented, it admits a regular neighborhood homeomorphic to $F\times[-1,1]$ in which F is identified with $F\times\{0\}$. For $\varepsilon=\pm1$, the *push off maps* $i^\varepsilon\colon H_1(F;\mathbb{Z})\to H_1(S^3\setminus F;\mathbb{Z})$ are defined by sending a (homology class of a) curve x to $i^\varepsilon(x):=x\times\{\varepsilon\}$. The *Seifert pairing* of F is the bilinear form

$$\begin{aligned} H_1(F;\mathbb{Z})\times H_1(F;\mathbb{Z})&\to\mathbb{Z}\\ (a,b)&\mapsto \ell k(i^-(a),b). \end{aligned}$$

A *Seifert matrix* for an oriented link L is any matrix representing the Seifert pairing. Although Seifert matrices do not provide link invariants, their so-called *S-equivalence class* does [61, Chapter 8]. Given a Seifert matrix A, observe that the matrix $(1-\omega)A+(1-\overline{\omega})A^T$ is Hermitian for all ω lying in S^1.

Definition 1. Let L be an oriented link, let F be a Seifert surface for L with $\beta_0(F)$ components and let A be a matrix representing the Seifert pairing of F. Given $\omega\in S^1$, the *Levine-Tristram signature* and *nullity* of L at ω are defined as

$$\begin{aligned} \sigma_L(\omega)&:=\operatorname{sign}((1-\omega)A+(1-\overline{\omega})A^T),\\ \eta_L(\omega)&:=\operatorname{null}((1-\omega)A+(1-\overline{\omega})A^T)+\beta_0(F)-1. \end{aligned}$$

These signatures and nullities are well defined (i.e. they are independent of the choice of the Seifert surface) [61, Theorem 8.9] and, varying ω along S^1, produce functions $\sigma_L,\eta_L\colon S^1\to\mathbb{Z}$. The Levine-Tristram signature is sometimes called the *ω-signature* (or the *equivariant signature* or the *Tristram-Levine signature*), while $\sigma_L(-1)$ is referred to as *the* signature of L or as the *Murasugi signature* of L [78]. The definition of $\sigma_L(\omega)$ goes back to Tristram [94] and Levine [57].

Remark 1. We note that σ_L and η_L are piecewise constant: both observations follow from the fact that the Alexander polynomial $\Delta_L(t)$ can be computed (up to its indeterminacy) by the formula $\Delta_L(t)=\det(tA-A^T)$. This latter fact also shows that, given $\omega\in S^1\setminus\{1\}$, the nullity $\eta_L(\omega)$ vanishes if and only if $\Delta_L(\omega)\neq0$. Moreover, the discontinuities of σ_L only occur at zeros of $(t-1)\Delta_L^{\mathrm{tor}}(t)$ [38, Theorem 2.1].

Several authors assume Seifert surfaces to be connected, and the nullity is then simply defined as the nullity of the matrix $H(\omega) = (1-\omega)A + (1-\overline{\omega})A^T$. The extra flexibility afforded by disconnected Seifert surfaces can for instance be taken advantage of when studying the behavior of the signature and nullity of boundary links.

Remark 2. Since the matrix $(1-\omega)A + (1-\overline{\omega})A^T$ vanishes at $\omega = 1$, we shall frequently think of σ_L and η_L as functions on $S^1_* := S^1 \setminus \{1\}$. Note nevertheless that for a knot K, the function σ_K vanishes in a neighborhood of $1 \in S^1$ [59, page 255], while for a μ-component link, one can only conclude that the limits of $|\sigma_L(\omega)|$ are at most $\mu - 1$ as ω approaches 1.

2.2 Properties of the signature and nullity

This subsection discusses the behaviour of the signature and nullity under operations such as orientation reversal, mirror image, connected sums and satellite operations.

The following proposition collects several properties of the Levine-Tristram signature.

Proposition 1. *Let L be a μ-component oriented link and let $\omega \in S^1$.*

1. *The Levine-Tristram signature is symmetric: $\sigma_L(\overline{\omega}) = \sigma_L(\omega)$.*
2. *The integer $\sigma_L(\omega) + \eta_L(\omega) - \mu + 1$ is even.*
3. *If $\Delta_L(\omega) \neq 0$, then $\sigma_L(\omega) = \mu - \operatorname{sgn}(i^\mu \nabla_L(\sqrt{\omega}))$ mod 4.* [1]
4. *If mL denotes the mirror image of L, then $\sigma_{mL}(\omega) = -\sigma_L(\omega)$.*
5. *If rL is obtained by reversing the orientation of each component of L, then $\sigma_{rL}(\omega) = \sigma_L(\omega)$.*
6. *Let L' and L'' be two oriented links. If L is obtained by performing a single connected sum between a component of L' and a component of L'', then $\sigma_L(\omega) = \sigma_{L'}(\omega) + \sigma_{L''}(\omega)$.*
7. *The signature is additive under the disjoint sum operation: if L is the link obtained by taking the disjoint union of two oriented links L' and L'', then $\sigma_L(\omega) = \sigma_{L'}(\omega) + \sigma_{L''}(\omega)$.*
8. *If S is a satellite knot with companion knot C, pattern P and winding number n, then*
$$\sigma_S(\omega) = \sigma_P(\omega) + \sigma_C(\omega^n).$$

Proof. The first assertion is immediate from Definition 1. The proof of the second and third assertions can be found respectively in [89]; see also [15, Lemmas 5.6 and 5.7]. The proof of the third assertion can be found in [61, Theorem 8.10]; see

[1] Here $\nabla_L(t)$ denotes the *one variable potential function* of L. Given a Seifert matrix A for L, the *normalized Alexander polynomial* is $D_L(t) = \det(-tA + t^{-1}A^T)$ and $\nabla_L(t)$ can be defined as $\nabla_L(t) = D_L(t)/(t - t^{-1})$. In what follows, for a complex number $\omega = e^{i\theta}$ with $0 < \theta < 2\pi$, we write $\sqrt{\omega}$ for the complex number $e^{i\theta/2}$.

also [15, Proposition 2.10]. The proof of the fifth, sixth and seventh assertions can be respectively be found in [15, Corollary 2.9, Proposition 2.12, Proposition 2.13]. For the proof of the last assertion, we refer to [66, Theorem 2]; see also [89, Theorem 9] (and [72, Theorem 3]) for the case $\omega = -1$.

Note that the second and third assertions of Proposition 1 generalize the well known fact that the Murasugi signature of a knot K is even. The behavior of σ_L under splicing (a generalization of the satellite operation) is discussed in [23, 24]. For discussions on the (Murasugi) signature of covering links, we refer to [78, 41] and [42] (which also provides a signature obstruction to a knot being periodic).

The following proposition collects the corresponding properties of the nullity.

Proposition 2. *Let L be an oriented link and let $\omega \in S^1_* := S^1 \setminus \{1\}$.*

1. *The nullity is symmetric: $\eta_L(\overline{\omega}) = \eta_L(\omega)$.*
2. *The nullity $\eta_L(\omega)$ is nonzero if and only if $\Delta_L(\omega) = 0$.*
3. *If mL denotes the mirror image of L, then $\eta_{mL}(\omega) = \eta_L(\omega)$.*
4. *If rL is obtained by reversing the orientation of each component of L, then $\eta_{rL}(\omega) = \eta_L(\omega)$.*
5. *Let L' and L'' be two oriented links. If L is obtained by performing a single connected sum between a component of L' and a component of L'', then $\eta_L(\omega) = \eta_{L'}(\omega) + \eta_{L''}(\omega)$.*
6. *If L is the link obtained by taking the disjoint union of two oriented links L' and L'', then we have $\eta_L(\omega) = \eta_{L'}(\omega) + \eta_{L''}(\omega) + 1$.*
7. *The nullity $\eta_L(\omega)$ is equal to the dimension of the twisted homology $\mathbb{C}$-vector space $H_1(X_L; \mathbb{C}_\omega)$, where $\mathbb{C}_\omega$ is the right $\mathbb{Z}[\pi_1(X_L)]$-module arising from the map $\mathbb{Z}[\pi_1(X_L)] \to \mathbb{C}, \gamma \to \omega^{\ell k(\gamma, L)}$.*
8. *If S is a satellite knot with companion knot C, pattern P and winding number n, then*
$$\eta_S(\omega) = \eta_P(\omega) + \eta_C(\omega^n).$$

Proof. The first assertion is immediate from Definition 1, while the second assertion was already discussed in Remark 1. The proof of assertions $(3) - (6)$ can respectively be found in [15, Proposition 2.10, Corollary 2.9, Proposition 2.12, Proposition 2.13]. To prove the penultimate assertion, pick a connected Seifert surface F for L, let A be an associated Seifert matrix and set $H(\omega) = (1-\omega)A + (1-\overline{\omega})A^T$. Since $tA - A^T$ presents the Alexander module $H_1(X_L; \mathbb{Z}[t^{\pm 1}])$, some homological algebra (as for instance in [21, proof of Proposition 3.4]) shows that $H(\omega)$ presents $H_1(X_L; \mathbb{C}_\omega)$; the assertion follows. The satellite formula can be deduced from [24, Theorem 5.2], or by using the equality $\eta_S(\omega) = \dim_{\mathbb{C}} H_1(X_S; \mathbb{C}_\omega)$ and running a Mayer-Vietoris argument for $H_1(X_S; \mathbb{C}_\omega)$.

We conclude this subsection by mentioning some additional facts about the signature function. Livingston provided a complete characterization of the functions $\sigma\colon S^1 \to \mathbb{Z}$ that arise as the Levine-Tristram signature function of a knot [70]. The corresponding question for links appears to be open. If $\Delta_L(t)$ is not identically zero, then it has at least $\sigma(L)$ roots on the unit circle [63, Appendix]. Finally, we describe the Murasugi signature for some particular classes of links.

Remark 3. Rudolph showed that the Murasugi signature of the closure of a nontrivial positive braid is negative (or positive, according to conventions) [88]. This result was later independently extended to positive links [93, 85] (see also [17]) and to almost positive links [86]. Later, Stoimenow improved Rudolph's result by showing that the Murasugi signature is bounded by an increasing function of the first Betti number [91]. Subsequent improvements of this result include [27, 3]. Formulas for the Levine-Tristram signature of torus knots can be found in [66].

2.3 Lower bounds on the unlinking number

In this subsection, we review some applications of signatures to unlinking and splitting links.

The *unlinking number* $u(L)$ of a link L is the minimal number of crossing changes needed to turn L into an unlink. The *splitting number* $\mathrm{sp}(L)$ of L is the minimal number of crossing changes between different components needed to turn L into the split union of its components. The Levine-Tristram signature and nullity are known to provide lower bounds on both these quantities:

Theorem 1. *Set $S^1_* = S^1 \setminus \{1\}$. Let $L = L_1 \cup \ldots \cup L_\mu$ be an oriented link and let $\omega \in S^1_*$.*

1. *The signature provides lower bounds on the unlinking number:*

$$|\sigma_L(\omega)| + |\eta_L(\omega) + \mu - 1| \leq 2u(L).$$

2. *The signature and nullity provide lower bounds on the splitting number:*

$$\left|\sigma_L(\omega) + \sum_{i<j} \ell k(L_i, L_j) - \sum_{i=1}^{\mu} \sigma_{L_i}(\omega)\right| + \left|\mu - 1 - \eta_L(\omega) + \sum_{i=1}^{\mu} \eta_{L_i}(\omega)\right| \leq \mathrm{sp}(L).$$

At the time of writing, the second inequality can only be proved using the multivariable signature [14]. A key step in proving the first inequality is to understand the behavior of the signature and nullity under crossing changes. The next proposition collects several such results:

Proposition 3. *Given, $\omega \in S^1_*$, the following assertions hold.*

1. *If L_+ is obtained from L_- by changing a single negative crossing change, then*

$$(\sigma_{L_+}(\omega) \pm \eta_{L_+}(\omega)) - (\sigma_{L_-}(\omega) \pm \eta_{L_-}(\omega)) \in \{0, -2\}.$$

2. *If, additionally, we let μ denote the number of components of L_+ (and L_-) and assume that ω is neither a root of $\Delta_{L_-}(t)$ nor of $\Delta_{L_+}(t)$, then*

$$\sigma_{L_+}(\omega) - \sigma_{L_-}(\omega) = \begin{cases} 0 & \text{if } (-1)^\mu \nabla_{L_+}(\sqrt{\omega}) \nabla_{L_-}(\sqrt{\omega}) > 0, \\ -2 & \text{if } (-1)^\mu \nabla_{L_+}(\sqrt{\omega}) \nabla_{L_-}(\sqrt{\omega}) < 0. \end{cases}$$

3. If L and L' differ by a single crossing change, then

$$|\eta_L(\omega) - \eta_{L'}(\omega)| \leq 1.$$

Proof. The proof of the first and third assertions can be found in [79, Lemma 2.1] (the proof is written for $\omega = -1$, but also holds for general ω). The proof of the second assertion now follows from the second item of Proposition 1 which states that modulo 4, the signature $\sigma_L(\omega)$ is congruent to $\mu + 1$ or $\mu - 1$ according to the sign of $i^\mu \nabla_L(\sqrt{\omega})$.

Note that similar conclusions hold if L_- is obtained from L_+ by changing a single negative crossing change; we refer to [79, Lemma 2.1] for the precise statement. Although Proposition 3 is well known, it seems that the full statement is hard to find in the literature: subsets of the statement for knots appear for instance in [26, 35, 45] (away from the roots of $\Delta_K(t)$) and for links in [79, Lemma 2.1] (for $\omega = -1$, without the statement involving ∇_L), and [15, Section 5] (in which various local relations are described; see also [22, Section 7.10] and [81, Lemma 3.1]).

We conclude this subsection with two additional remarks in the knot case.

Remark 4. In the knot case, the second assertion of Proposition 3 is fairly well known (e.g. [35, Lemma 2.2] and [45, Equation (10)]). Indeed, under the same assumptions as in Proposition 3, and using the normalized Alexander polynomial $D_L(t)$ (which satisfies $D_L(t) = (t - t^{-1})\nabla_L(t)$), it can be rewritten as

$$\sigma_{K_+}(\omega) - \sigma_{K_-}(\omega) = \begin{cases} 0 & \text{if } D_{K_+}(\sqrt{\omega})D_{K_-}(\sqrt{\omega}) > 0, \\ -2 & \text{if } D_{K_+}(\sqrt{\omega})D_{K_-}(\sqrt{\omega}) < 0. \end{cases}$$

Finally, note that for knots, the lower bound on the unknotting number can be significantly improved upon by using the jumps of the signature function [71]. Other applications of the Levine-Tristram signature to unknotting numbers can be found in [90] (as well as a relation to finite type invariants).

2.4 Concordance invariance and the Murasugi-Tristram inequalities

In this subsection, we review properties of the Levine-Tristram signature related to 4-dimensional topology. Namely we discuss the conditions under which the signature is a concordance invariant, and gives lower bounds on the 4-genus.

Two oriented μ-component links L and J are smoothly (resp. topologically) *concordant* if there is a smooth (resp. locally flat) embedding into $S^3 \times I$ of a disjoint union of μ annuli $A \hookrightarrow S^3 \times I$, such that the oriented boundary of A satisfies

$$\partial A = -L \sqcup J \subset -S^3 \sqcup S^3 = \partial(S^3 \times I).$$

The integers $\sigma_L(\omega)$ and $\eta_L(\omega)$ are known to be concordance invariants for any root of unity ω of prime power order [78, 94]. However, it is only recently that Nagel and Powell gave a precise characterization of the $\omega \in S^1$ at which σ_L and η_L are concordance invariants [80] (see also [97]). To describe this characterization, we say that a complex number $\omega \in S^1_*$ is a *Knotennullstelle* if it is the root of a Laurent polynomial $p(t) \in \mathbb{Z}[t^{\pm 1}]$ satisfying $p(1) = \pm 1$. We write $S^1_!$ for the set of $\omega \in S^1$ which do *not* arise as a Knotennullstelle.

The main result of [80] can be stated as follows.

Theorem 2. *The Levine-Tristram signature σ_L and nullity η_L are concordance invariants at $\omega \in S^1_*$ if and only if $\omega \in S^1_!$.*

In the knot case, Cha and Livingston had previously shown that for any Knotennullstelle ω, there is a slice knot K with $\sigma_K(\omega) \neq 0$ and $\eta_K(\omega) \neq 0$ [13]. Here, recall that a knot $K \subset S^3$ is smoothly (resp. topologically) *slice* if it is smoothly (resp. topologically) concordant to the unknot or, equivalently, if it bounds a smoothly (resp. locally flat) properly embedded disk in the 4-ball. Still restricting to knots, the converse can be established as follows.

Remark 5. The Levine-Tristram signature of an oriented knot K vanishes at $\omega \in S^1_!$ whenever K is algebraically slice i.e. whenever it admits a metabolic Seifert matrix A. To see this, first note that since A is metabolic, the matrix $H(\omega) = (1-\omega)A + (1-\overline{\omega})A^T$ is congruent to one which admits a half size block of zeros in its upper left corner. Furthermore the definition of $S^1_!$ and the equality $H(t) = (t^{-1}-1)(tA - A^T)$ imply that $H(\omega)$ is nonsingular for $\omega \in S^1_!$: indeed, since K is a knot, $\Delta_K(1) = \pm 1$. Combining these facts, $\sigma_K(\omega) = \operatorname{sign}(H(\omega))$ vanishes for $\omega \in S^1_!$. As slice knots are algebraically slice (see e.g. [61, Proposition 8.17]), we have established that if K is slice, then σ_K vanishes on $S^1_!$.

Using Remark 5 and Theorem 1, one sees that the Levine-Tristram signature actually provides lower bounds on the *slicing number* of a knot K i.e. the minimum number of crossing changes required to convert K to a slice knot [68, 83]. In a somewhat different direction, the Levine-Tristram signature is also a lower bound on the algebraic unknotting number [30, 77, 6, 8, 7].

Several steps in Remark 5 fail to generalize from knots to links: there is no obvious notion of algebraic sliceness for links and, if L has two components or more, then $\Delta_L(1) = 0$. In fact, even the notion of a slice link deserves some comments.

Remark 6. An oriented link $L = L_1 \cup \ldots \cup L_\mu$ is smoothly (resp. topologically) *slice in the strong sense* if there are disjointly smoothly (resp. locally flat) properly embedded disks $D_1, \ldots, D_\mu$ with $\partial D_i = L_i$. As a corollary of Theorem 2, one sees that if L is topologically slice in the strong sense, then $\sigma_L(\omega) = 0$ and $\eta_L(\omega) = \mu - 1$ for all $\omega \in S^1_!$.

On the other hand, an oriented link is smoothly (resp. topologically) *slice in the ordinary sense* if it is the cross-section of a single smooth (resp. locally flat) 2-sphere in S^4. It is known that if L is slice in the ordinary sense, then $\sigma_L(\omega) = 0$ for all ω of prime power order [15, Corollary 7.5] (see also [47, Theorem 3.13]). There is little

doubt that this result should hold for a larger subset of S^1 and in the topological category.

In a similar spirit, the Levine-Tristram signatures can be used to provide restrictions on the surfaces a link can bound in the 4-ball. Such inequalities go back to Murasugi [78] and Tristram [94]. Since then, these inequalities have been generalized in several directions [37, Corollary 4.3], [28, Theorem 5.19], [15, Theorem 7.2], [69], [97, Section 4] and [21, Theorem 1.2 and Corollary 1.4]. Applications to the study of algebraic curves can be found in [28, 81, 82].

The following theorem describes such a *Murasugi-Tristram inequality* in the topological category which holds for a large subset of S^1.

Theorem 3. *If an oriented link L bounds an m-component properly embedded locally flat surface $F \subset D^4$ with first Betti number $b_1(F)$, then for any $\omega \in S^1_!$, the following inequality holds:*

$$|\sigma_L(\omega)| + |\eta_L(\omega) - m + 1| \leq b_1(F).$$

Observe that if L is a strongly slice link, then m is equal to the number of components of L and $b_1(F) = 0$ and thus $\sigma_L(\omega) = 0$ and $\eta_L(\omega) = m - 1$ for all $\omega \in S^1_!$, recovering the result mentioned in Remark 6. On the other hand, if K is a knot, then Theorem 3 can be expressed in terms of the topological 4-genus $g_4(K)$ of K: the minimal genus of a locally flat surface in D^4 cobounding K. An article studying the sharpness of this bound includes [62].

In order to obtain results which are valid on the whole of the unit circle S^1, it is possible to consider the average of the one-sided limits of the signature and nullity. Namely for $\omega = e^{i\theta} \in S^1$ and any Seifert matrix A, one sets $H(\omega) = (1-\omega)A + (1-\overline{\omega})A^T$ and considers

$$\sigma_L^{\mathrm{av}}(\omega) = \frac{1}{2}\big(\lim_{\eta \to \theta_+} \operatorname{sign}(H(e^{i\eta})) + \lim_{\eta \to \theta_-} \operatorname{sign}(H(e^{i\eta}))\big),$$
$$\eta^{\mathrm{av}}(\omega) = \frac{1}{2}\big(\lim_{\eta \to \theta_+} \operatorname{null}(H(e^{i\eta})) + \lim_{\eta \to \theta_-} \operatorname{null}(H(e^{i\eta}))\big).$$

The earliest explicit observation that these *averaged Levine-Tristram signatures* are smooth concordance invariants seems to go back to Gordon's survey [39]. Working with the averaged Levine-Tristram signature and in the topological locally flat category, Powell [84] recently proved a Murasugi-Tristram type inequality which holds for *each* $\omega \in S^1_*$.

We conclude this subsection with three remarks on knots.

Remark 7. A knot K is smoothly (resp. topologically) *doubly slice* if it is the cross section of an unknotted smoothly (resp. locally flat) embedded 2-sphere S^2 in S^4. It is known that if K is topologically doubly slice, then $\sigma_K(\omega)$ vanishes for *all* $\omega \in S^1$; no averaging is needed [92, 53, 59]. Is there a meaningful statement for links?

The Levine-Tristram signature also appears in knot concordance in relation to a particular von Neumann ρ-invariant (or L^2-signature). This invariant associates a

real number to any closed 3-manifold together with a map $\phi\colon \pi_1(M) \to \Gamma$, with Γ a PTFA group. When M is the 0-framed surgery along a knot K and ϕ is the abelianization map, then this invariant coincides with the (normalized) integral of $\sigma_K(\omega)$ along the circle [18, Proposition 5.1]. Computations of this invariant on (iterated) torus knots can be found in [55, 5, 20].

The Levine-Tristram signature is an invariant of *rational* knot concordance at prime order roots of unity; see [12, Theorem 1.1] and [16, Proposition 4.2] for further details.

3 4-dimensional definitions of the signature

In this section, we describe 4-dimensional definitions of the Levine-Tristram signature using embedded surfaces in the 4-ball (Subsection 3.1) and as a bordism invariant of the 0-framed surgery (Subsection 3.2).

3.1 Signatures via exteriors of surfaces in the 4-ball

We relate the Levine-Tristram signature to signature invariants of the exterior of embedded surfaces in the 4-ball. Historically, the first approach of this kind involved branched covers [96] (see also [11, 47]) while more recent results make use of twisted homology [18, 97, 84].

Given a smoothly properly embedded connected surface $F \subset D^4$, denote by W_F the complement of a tubular neighborhood of F. A short Mayer-Vietoris argument shows that $H_1(W_F;\mathbb{Z})$ is infinite cyclic and one may consider the covering space $W_k \to W_F$ obtained by composing the abelianization homomorphism with the quotient map $H_1(W_F;\mathbb{Z}) \cong \mathbb{Z} \to \mathbb{Z}_k$. The restriction of this cover to $F \times S^1$ is $\mathrm{id} \times p$, where $p\colon S^1 \to S^1$ is the k-fold cover of the circle. Extending p to a cover $D^2 \to D^2$ branched along 0, and setting

$$\overline{W}_F := W_k \cup_{F\times S^1} (F \times D^2)$$

produces a cover $\overline{W}_F \to D^4$ branched along $F = F \times \{0\}$. Denote by t a generator of the finite cyclic group $\mathbb{Z}_k$. The $\mathbb{C}[\mathbb{Z}_k]$-module structure of $H_2(\overline{W}_F, \mathbb{C})$ gives rise to a complex vector space

$$H_2(\overline{W}_F, \mathbb{C})_\omega = \{x \in H_2(\overline{W}_F, \mathbb{C}) \mid tx = \omega x\}$$

for each root of unity ω of order k. Restricting the intersection form on $H_2(\overline{W}_F, \mathbb{C})$ to $H_2(\overline{W}_F, \mathbb{C})_\omega$ produces a Hermitian pairing whose signature we denote by $\sigma_\omega(\overline{W}_F)$.

The next result, originally due to Viro [96], was historically the first 4-dimensional interpretation of the Levine-Tristram signature; see also [47].

Theorem 4. *Assume that an oriented link L bounds a smoothly properly embedded compact oriented surface $F \subset D^4$ and let $\overline{W}_F$ be the k-fold cover of D^4 branched along F. Then, for any root of unity $\omega \in S^1_*$ of order k, the following equality holds:*

$$\sigma_L(\omega) = \sigma_\omega(\overline{W}_F).$$

As for the results described in Subsection 2.4, Theorem 4 can be sharpened by working in the topological category and using arbitrary $\omega \in S^1_*$. As the next paragraphs detail, the idea is to rely on twisted homology instead of branched covers [84, 97, 21].

Let $\omega \in S^1_*$. From now on, we assume that $F \subset D^4$ is a locally flat properly embedded (possibly disconnected) compact oriented surface. Since $H_1(W_F;\mathbb{Z})$ is free abelian, there is a map $H_1(W_F;\mathbb{Z}) \to \mathbb{C}$ obtained by sending each meridian of F to ω. Precomposing with the abelianization homomorphism, gives rise to a right $\mathbb{Z}[\pi_1(W_F)]$-module structure on $\mathbb{C}$ which we denote by $\mathbb{C}_\omega$ for emphasis. We can therefore consider the twisted homology groups $H_*(W_F;\mathbb{C}_\omega)$ and the corresponding $\mathbb{C}$-valued intersection form $\lambda_{W_F,\mathbb{C}_\omega}$ on $H_2(W_F;\mathbb{C}_\omega)$.

The following result can be seen as a generalization of Theorem 4.

Theorem 5. *Assume that an oriented link L bounds a properly embedded locally flat compact oriented surface $F \subset D^4$. Then the following equality holds for any $\omega \in S^1_*$:*

$$\sigma_L(\omega) = \operatorname{sign}(\lambda_{W_F,\mathbb{C}_\omega}).$$

A key feature of Theorems 4 and 5 lies in the fact that the signature invariants associated to W_F do not depend on the choice of F. This plays a crucial role in the 4-dimensional proofs of Murasugi-Tristram type inequalities. This independance statement relies on the Novikov-Wall addivity as well as on the G-signature theorem (for Theorem 4) and on bordisms considerations over the classifying space $B\mathbb{Z}$ (for Theorem 5).

3.2 Signatures as invariants of the 0-framed surgery

In this subsection, we outline how the Levine-Tristram signature of a link L can be viewed as a bordism invariant of the 0-framed surgery along L. To achieve this, we describe bordism invariants of pairs consisting of a closed connected oriented 3-manifold together with a map from $\pi_1(M)$ to $\mathbb{Z}_m$ or $\mathbb{Z}$.

Let M be an oriented closed 3-manifold and let $\chi\colon \pi_1(M) \to \mathbb{Z}_m$ be a homomorphism. Since the bordism group $\Omega_3(\mathbb{Z}_m)$ is finite, there exists a non-negative integer r, a 4-manifold W and a map $\psi\colon \pi_1(W) \to \mathbb{Z}_m$ such that the boundary of W consists of the disjoint union of r copies of M and the restriction of ψ to ∂W coincides with χ on each copy of M. If these conditions are satisfied, we write $\partial(W,\psi) = r(M,\chi)$ for brevity. Mapping the generator of $\mathbb{Z}_m$ to $\omega := e^{\frac{2\pi i}{m}}$ gives rise to a map $\mathbb{Z}[\mathbb{Z}_m] \to \mathbb{Q}(\omega)$. Precomposing with ψ, we

obtain a $(\mathbb{Q}(\omega), \mathbb{Z}[\pi_1(W)])$-bimodule structure on $\mathbb{Q}(\omega)$ and twisted homology groups $H_*(W;\mathbb{Q}(\omega))$. The $\mathbb{Q}(\omega)$-vector space $H_2(W;\mathbb{Q}(\omega))$ is endowed with a $\mathbb{Q}(\omega)$-valued Hermitian form $\lambda_{W,\mathbb{Q}(\omega)}$ whose signature is denoted $\operatorname{sign}^{\psi}(W) := \operatorname{sign}(\lambda_{W,\mathbb{Q}(\omega)})$. In this setting, the *Casson-Gordon σ-invariant* of (M,χ) is

$$\sigma(M,\chi) := \frac{1}{r}\left(\operatorname{sign}^{\psi}(W) - \operatorname{sign}(W)\right) \in \mathbb{Q}.$$

We now focus on the case where $M = M_L$ is the closed 3-manifold obtained by performing 0-framed surgery on a link L. In this case, a short Mayer-Vietoris argument shows that $H_1(M_L;\mathbb{Z})$ is freely generated by the meridians of L.

Casson and Gordon proved the following theorem [11, Lemma 3.1].

Theorem 6. *Let $\chi\colon H_1(M_L;\mathbb{Z}) \to \mathbb{Z}_m \subset \mathbb{C}^*$ be the character mapping each meridian of L to ω^r, where $\omega = e^{\frac{2\pi i}{m}}$ and $0 < r < m$. Then the Casson-Gordon σ-invariant satisfies*

$$\sigma(M_L,\chi) = \sigma_L(\omega^r).$$

Note that Casson and Gordon proved a version of Theorem 6 for arbitrary surgeries on links; we also refer to [37, Theorem 3.6] and [15, Theorem 6.7] for generalizations to more general characters. The idea of defining link invariants using the Casson-Gordon invariants is pursued further in [28, 29].

Remark 8. The Casson-Gordon σ-invariant (and thus the Levine-Tristram signature) can be understood as a particular case of the Atiyah-Patodi-Singer ρ-invariant [2] which associates a real number to pairs (M,α), with M a closed connected oriented 3-manifold and $\alpha\colon \pi_1(M) \to U(k)$ a unitary representation. For further reading on this point of view, we refer to [60, 56, 31, 32, 33].

Next, we describe how to circumvent the restriction that ω be of finite order. Briefly, the idea is to work in the infinite cyclic cover as long as possible, delaying the appearance of ω [65, Section 2]; see also [18, Section 5]. Following [84], the next paragraphs describe the resulting construction.

Let M be a closed connected oriented 3-manifold, and let $\phi\colon \pi_1(M) \to \mathbb{Z}$ be a homomorphism. Since $\Omega_3^{STOP}(\mathbb{Z})$ is zero, M bounds a connected topological 4-manifold W and there is a map $\psi\colon \pi_1(W) \to \mathbb{Z}$ which extends ϕ. This map endows $\mathbb{Q}(t)$ with a $(\mathbb{Q}(t), \mathbb{Z}[\pi_1(W)])$-bimodule structure and therefore gives rise to a $\mathbb{Q}(t)$-valued intersection form $\lambda_{W,\mathbb{Q}(t)}$ on $H_2(W;\mathbb{Q}(t))$. It can be checked that $\lambda_{W,\mathbb{Q}(t)}$ induces a nonsingular Hermitian form $\lambda^{\text{nonsing}}_{W,\mathbb{Q}(t)}$ on the quotient of $H_2(W;\mathbb{Q}(t))$ by $\operatorname{im}(H_2(M;\mathbb{Q}(t)) \to H_2(W;\mathbb{Q}(t)))$ [84, Lemma 3.1]. As a consequence, $\lambda^{\text{nonsing}}_{W,\mathbb{Q}(t)}$ gives rise to an element $[\lambda^{\text{nonsing}}_{W,\mathbb{Q}(t)}]$ of the Witt group $W(\mathbb{Q}(t))$. Taking the averaged signature at $\omega \in S^1$ of a representative of an element in $W(\mathbb{Q}(t))$ produces a well defined homomorphism $\operatorname{sign}_\omega\colon W(\mathbb{Q}(t)) \to \mathbb{C}$. Thus, for $\omega \in S^1_*$ and $(M,\phi) = \partial(W,\psi)$ as above, one can set

$$\sigma^{\text{av}}_{M,\phi}(\omega) = \operatorname{sign}_\omega([\lambda^{\text{nonsing}}_{W,\mathbb{Q}(t)}]) - \operatorname{sign}(W).$$

It can be checked that $\sigma^{\mathrm{av}}_{M,\phi}$ does not depend on W and ψ [84, Section 3]. We now return to links: we let L be an oriented link, assume that M is the 0-framed surgery M_L and that ϕ is the map ϕ_L which sends each meridian of L to 1.

The following result is due to Powell [84, Lemma 4.1].

Theorem 7. *For any oriented link L and any $\omega \in S^1_*$, the following equality holds:*

$$\sigma^{\mathrm{av}}_{M_L,\phi_L}(\omega) = \sigma^{\mathrm{av}}_L(\omega).$$

Describing σ_L as a 3-manifold invariant as in Theorem 7 provides a useful tool to work in the topological category; see for instance Powell's proof a Murasugi-Tristram type inequality [84, Theorem 1.4].

4 Signatures via pairings on infinite cyclic covers

In this section, we review two additional intrinsic descriptions of the Levine-Tristram signature of a knot K. Both constructions make heavy use of the algebraic topology of the infinite cyclic cover of the exterior of K: the first uses the Milnor pairing (Subsection 4.1), while the second relies on the Blanchfield pairing (Subsection 4.2).

4.1 Milnor signatures

In this subsection, we recall the definition of a pairing which was first described by Milnor [76]. We then outline how the resulting "Milnor signatures" are related to (the jumps of) the Levine-Tristram signature.

Given an oriented knot K in S^3, use $X_K = S^3 \setminus \nu K$ to denote its exterior. The kernel of the abelianization homomorphism $\pi_1(X_K) \to H_1(X_K;\mathbb{Z}) \cong \mathbb{Z}$ gives rise to an infinite cyclic cover $X^\infty_K \to X_K$. Milnor showed that the cup product

$$H^1(X^\infty_K;\mathbb{R}) \times H^1(X^\infty_K, \partial X^\infty_K;\mathbb{R}) \to H^2(X^\infty_K, \partial X^\infty_K;\mathbb{R}) \cong \mathbb{R}$$

defines a nonsingular skew-symmetric $\mathbb{R}$-bilinear form [76, Assertion 9]. Since the canonical inclusion $(X_K,\emptyset) \to (X_K,\partial X_K)$ induces an isomorphism $H^1(X^\infty_K;\mathbb{R}) \to H^1(X^\infty_K,\partial X^\infty_K;\mathbb{R})$, the aforementioned cup product pairing gives to rise to a nonsingular skew-symmetric form

$$\cup\colon H^1(X^\infty_K;\mathbb{R}) \times H^1(X^\infty_K;\mathbb{R}) \to \mathbb{R}.$$

Use t^* to denote the automorphism induced on $H^1(X^\infty_K;\mathbb{R})$ by the generator of the deck transformation group of X^∞_K. Milnor defines the *quadratic form of K* as the pairing

$$b_K\colon H^1(X_K^\infty;\mathbb{R})\times H^1(X_K^\infty;\mathbb{R})\to\mathbb{R}$$
$$(x,y)\mapsto (t^*x)\cup y+(t^*y)\cup x.$$

This pairing is symmetric and nonsingular [76, Assertion 10] and Milnor defines the signature of K as the signature of b_K. Erle later related $\operatorname{sign}(b_K)$ to the Murasugi signature of K [25]:

Theorem 8. *Let K be an oriented knot. The signature of the symmetric form b_K is equal to the Murasugi signature of K:*

$$\operatorname{sign}(b_K)=\sigma(K).$$

Next, we describe the so-called Milnor signatures. As $\mathbb{R}$ is a field, the ring $\mathbb{R}[t^{\pm1}]$ is a PID and therefore the torsion $\mathbb{R}[t^{\pm1}]$-module $H:=H_1(X_K^\infty;\mathbb{R})$ decomposes as a direct sum over its $p(t)$-primary components, where $p(t)$ ranges over the irreducible polynomials of $\mathbb{R}[t^{\pm1}]$. [2] As explained in [76, proof of Assertion 11], the symmetric form b_K decomposes orthogonally once we distinguish symmetric polynomials (i.e. $p(t)=rt^{\pm i}p(t^{-1})$; written $p(t)\doteq p(t^{-1})$) from non-symmetric ones:[3]

$$(H^1(X_K^\infty;\mathbb{R}),b_K)=\bigoplus_{p(t)\doteq p(t^{-1})}(H_{p(t)},b_K|_{H_{p(t)}})$$
$$\oplus\bigoplus_{p(t)\neq p(t^{-1})}(H_{p(t)}\oplus H_{p(t^{-1})},b_K|_{H_{p(t)}\oplus H_{p(t^{-1})}}).$$

In a nutshell, for $p(t)$ irreducible and symmetric, the restrictions of $b_K|_{H_{p_\theta}(t)}$ produce additional signature invariants. If $p(t)$ and $q(t)$ differ by multiplication by a unit, then their corresponding primary summands are equal. From now on, a polynomial is thus understood to be symmetric if $p(t)=p(t^{-1})$. As we are working over $\mathbb{R}[t^{\pm1}]$, the irreducible symmetric polynomials are of the form $p_\theta(t)=t-2\cos(\theta)+t^{-1}$ with $0<\theta<\pi$ [76, page 129].

Definition 2. For $0<\theta<\pi$, the *Milnor signature* $\sigma_\theta(K)$ is the signature of the restriction of b_K to the $p_\theta(t)$-primary summand of $H:=H^1(X_K^\infty;\mathbb{R})$:

$$\sigma_\theta(K):=\operatorname{sign}(b_K|_{H_{p_\theta}(t)}).$$

Note that $\sigma_\theta(K)$ is zero if $p_\theta(t)$ does not divide the Alexander polynomial $\Delta_K(t)$ of K. In particular, by Erle's result, the Murasugi signature $\sigma(K)$ is equal to the sum of the $\sigma_\theta(K)$ over all θ such that $p_\theta(t)$ divides $\Delta_K(t)$. Thus, recalling that ±1 can not be a root of the Alexander polynomial of a knot, one can write

[2] By a *$p(t)$-primary component*, we mean $H_{p(t)}=\{x\in H\mid p(t)^nx=0 \text{ for some } n>0\}$. Observe that $H_{p(t)}\neq 0$ only if $p(t)$ is a factor of $\Delta_K(t)$.

[3] Here, since we use the same notation for direct sums and orthogonal sums, the underlying algebraic fact to keep in mind is "if $p(t)\neq q(t^{-1})$, then $H_{p(t)}$ and $H_{q(t)}$ are orthogonal".

$$\sigma(K) = \sum_{0<\theta<\pi} \sigma_\theta(K) = \sum_{\{\theta\colon p_\theta | \Delta_K\}} \sigma_\theta(K). \tag{1}$$

Next, following Matumoto, we relate the Milnor signatures to the Levine-Tristram signatures [73]. First, note that Erle proves a stronger result than the equality $\sigma(K) = \operatorname{sign}(b_K)$: indeed he shows that b_K is represented by $W + W^T$, where W is a nonsingular matrix over $\mathbb{Z}$ which is S-equivalent to a Seifert matrix of K; he calls such a matrix a *reduced Seifert matrix* [25, Section 3.4]. As a consequence, Matumoto considers an arbitrary nonsingular bilinear form on a $\mathbb{R}$-vector space V, represented by a matrix A and compares the signature of $(1-\omega)A + (1-\overline{\omega})A^T$ (for $\omega \in S^1$) with the signatures of $A + A^T$ restricted to the $p(t)$-primary summands of V (here t is thought alternatively as an indeterminate and as the $\mathbb{R}$-automorphism $(A^T)^{-1}A$) [73]).

A particular case of one of Matumoto's results can be now be stated as follows [73, Theorem 2].

Theorem 9. *Let K be an oriented knot and let $\omega = e^{i\varphi}$ with $0 < \varphi \leq \pi$. If ω is not a root of $\Delta_K(t)$, then the following equality holds:*

$$\sigma_K(\omega) = \sum_{0<\theta<\varphi} \sigma_\theta(K) + \frac{1}{2}\sigma_\varphi(K).$$

Observe that if $\omega = e^{i\varphi}$ is not a root of $\Delta_K(t)$, then the Milnor signature $\sigma_\varphi(K)$ vanishes. In particular, since -1 is never a root of the Alexander polynomial of a knot, Theorem 9 recovers (1). The Milnor pairing can also be considered over $\mathbb{C}$ in which case the statement is somewhat different [73, Theorem 1]. Informally, Theorem 9 states that the Milnor signatures measure the jumps of $\sigma_K\colon S^1 \to \mathbb{Z}$ at the roots of $\Delta_K(t)$ which lie on S^1. The situation for links is more complicated [50]; see also [48].

We conclude by mentioning some further properties of the Milnor signatures.

Remark 9. The Milnor signatures are concordance invariants [76, p.129]. Milnor establishes this result by showing that his signatures vanish on slice knots and are additive under connected sums. A satellite formula for the Milnor signatures is stated without proof in [51].

4.2 Signatures via the Blanchfield pairing

In this subsection, we review how the Levine-Tristram signature of a knot can be recovered from the Blanchfield pairing. Note that while the Blanchfield pairing is known to determine the S-equivalence type of K [95], the approaches we discuss here are arguably more concrete.

Given an oriented knot K, recall that X_K^∞ denotes the infinite cyclic cover of the exterior X_K. Since $\mathbb{Z} = \langle t \rangle$ acts on X_K^∞, the homology group $H_1(X_K^\infty; \mathbb{Z})$ is naturally

endowed with a $\mathbb{Z}[t^{\pm 1}]$-module structure. This $\mathbb{Z}[t^{\pm 1}]$-module is called the *Alexander module* and is known to be finitely generated and torsion [58]. Using $\mathbb{Q}(t)$ to denote the field of fractions of $\mathbb{Z}[t^{\pm 1}]$, the *Blanchfield form* of a knot is a Hermitian and nonsingular sesquilinear pairing

$$\mathrm{Bl}_K\colon H_1(X_K^\infty;\mathbb{Z})\times H_1(X_K^\infty;\mathbb{Z})\to \mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}].$$

In order to define Bl_K, we describe its adjoint

$$\mathrm{Bl}_K^\bullet\colon H_1(X_K^\infty;\mathbb{Z})\to \overline{\mathrm{Hom}_{\mathbb{Z}[t^{\pm 1}]}(H_1(X_K^\infty;\mathbb{Z}),\mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}])}$$

so that $\mathrm{Bl}_K(x,y)=\mathrm{Bl}_K^\bullet(y)(x)$. [4] Using local coefficients, the Alexander module can be written as $H_1(X_K;\mathbb{Z}[t^{\pm 1}])$. The short exact sequence of coefficients

$$0\to\mathbb{Z}[t^{\pm 1}]\to\mathbb{Q}(t)\to\mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}]\to 0$$

gives rise to a Bockstein homomorphism

$$\mathrm{BS}\colon H^1(X_K;\mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}])\to H^2(X_K;\mathbb{Z}[t^{\pm 1}]).$$

Since the Alexander module is torsion, BS is in fact an isomorphism. Composing the map induced by the inclusion $\iota\colon (X_K,\emptyset)\to(X_K,\partial X_K)$ with Poincaré duality, BS^{-1} and the Kronecker evaluation map yields the desired $\mathbb{Z}[t^{\pm 1}]$-linear map:

$$\begin{aligned}\mathrm{Bl}_K^\bullet\colon H_1(X_K;\mathbb{Z}[t^{\pm 1}]) &\xrightarrow{\iota_*} H_1(X_K,\partial X_K;\mathbb{Z}[t^{\pm 1}]) \xrightarrow{\mathrm{PD}} H^2(X_K;\mathbb{Z}[t^{\pm 1}])\\ &\xrightarrow{\mathrm{BS}^{-1}} H^1(X_K;\mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}])\\ &\xrightarrow{\mathrm{ev}} \overline{\mathrm{Hom}_{\mathbb{Z}[t^{\pm 1}]}(H_1(X_K;\mathbb{Z}[t^{\pm 1}]),\mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}])}.\end{aligned} \tag{2}$$

Following Kearton [52, 54], we outline how signatures can be extracted from the (real) Blanchfield pairing

$$\mathrm{Bl}_K\colon H_1(X_K;\mathbb{R}[t^{\pm 1}])\times H_1(X_K;\mathbb{R}[t^{\pm 1}])\to \mathbb{R}(t)/\mathbb{R}[t^{\pm 1}].$$

Let $p(t)$ be a real irreducible symmetric factor of $\Delta_K(t)$, and let $H_{p(t)}$ be the $p(t)$-primary summand of $H:=H_1(X_K;\mathbb{R}[t^{\pm 1}])$. There is a decomposition $H_{p(t)}=\bigoplus_{i=1}^m H_{p(t)}^r$, where each $H_{p(t)}^r$ is a free module over $\mathbb{R}[t^{\pm 1}]/p(t)^r\mathbb{R}[t^{\pm 1}]$. For $i=1,\dots,m$, consider the quotient

$$V_{p(t)}^r:=H_{p(t)}^r/p(t)H_{p(t)}^r$$

as a vector space over $\mathbb{C}\cong\mathbb{R}(\xi)\cong\mathbb{R}[t^{\pm 1}]/p(t)\mathbb{R}[t^{\pm 1}]$, where ξ is a root of $p(t)$. The Blanchfield pairing Bl_K now induces the following well defined Hermitian pairing:

[4] Given a ring R with involution, and given an R-module M, we denote by $\overline{M}$ the R-module that has the same underlying additive group as M, but for which the action by R on M is precomposed with the involution on R.

$$\begin{aligned}\mathrm{bl}_{r,p(t)}(K)\colon V^r_{p(t)}\times V^r_{p(t)} &\to \mathbb{C}\\ ([x],[y]) &\mapsto \mathrm{Bl}_K(p(t)^{r-1}x,y).\end{aligned}$$

As above, we write $p_\theta(t) = t - 2\cos(\theta) + t^{-1}$: this way for each $\theta \in (0,\pi)$ and every integer r, we obtain additional signature invariants.

Definition 3. For $0 < \theta < \pi$ and $r > 0$, the *Blanchfield signature* $\sigma_{r,\theta}(K)$ is the signature of the Hermitian pairing $\mathrm{bl}_{r,p_\theta(t)}(K)$:

$$\sigma_{r,\theta}(K) := \mathrm{sign}(\mathrm{bl}_{r,p_\theta(t)}(K)).$$

Kearton [52, Section 9] relates the signatures $\sigma_{r,\theta}(K)$ to the Milnor signatures, while Levine relates the $\sigma_{r,\theta}(K)$ to the Levine-Tristram signature [59, Theorem 2.3]:

Theorem 10. *Given an oriented knot K and $0 < \theta < \pi$, one has*

$$\sigma_\theta(K) = \sum_{r\ odd} \sigma_{r,\theta}(K).$$

If $\omega := e^{i\theta}$ is a root of $\Delta_K(t)$, and if $\omega_+, \omega_- \in S^1_* \setminus \{\omega \in S^1_* \mid \Delta_K(\omega) = 0\}$ are such that ω is the only root of $\Delta_K(t)$ lying on an arc of S^1 connecting them, then

$$\begin{aligned}\sigma_K(\omega^+) - \sigma_K(\omega^-) &= 2\sum_{r\ odd} \sigma_{r,\theta}(K),\\ \sigma_K(\omega) &= \frac{1}{2}(\sigma_K(\omega^+) - \sigma_K(\omega^-)) - \sum_{r\ even} \sigma_{r,\theta}(K).\end{aligned}$$

As we already mentioned in the previous subsection, Theorem 10 (and Theorem 9) shows that the Blanchfield and Milnor signatures measure the jumps of $\sigma_K\colon S^1 \to \mathbb{Z}$ at the roots of $\Delta_K(t)$.

Next, we mention some further properties of the Blanchfield signatures.

Remark 10. For each $0 < \theta < \pi$, the sum $\sum_{r\ \mathrm{odd}} \sigma_{r,\theta}(K)$ of Blanchfield signatures is a concordance invariant: this can either be seen directly [59] or by relating this sum to the Milnor signature $\sigma_\theta(K)$ (recall Theorem 10) and using its concordance invariance (recall Remark 9). Combining this fact with Theorem 10 yields a proof that the Levine-Tristram signature function σ_K vanishes away from the roots of Δ_K if K is (algebraically) slice (recall Subsection 2.4).

While the Blanchfield signatures $\sigma_{r,\theta}(K)$ are not concordance invariants, they do vanish if K is doubly slice [53, 59]. Combining this fact with Theorem 10 yields a proof that the Levine-Tristram signature function σ_K vanishes identically if K is doubly slice (recall Remark 7).

Next, following Borodzik-Friedl, we describe a second way of extracting signatures from the Blanchfield pairing [8]. The Blanchfield pairing is known to be *representable*: as shown in [8, Proposition 2.1] there exists a non-degenerate Hermitian matrix $A(t)$ over $\mathbb{Z}[t^{\pm 1}]$ such that Bl_K is isometric to the pairing

$$\lambda_{A(t)}\colon \operatorname{coker}(A(t)^T)\times\operatorname{coker}(A(t)^T)\to \mathbb{Q}(t)/\mathbb{Z}[t^{\pm 1}]$$
$$([x],[y])\mapsto x^T A(t)^{-1}\overline{y}.$$

In this case, we say that the Hermitian matrix $A(t)$ *represents* Bl_K. These representing matrices provide an alternative way of defining the Levine-Tristram signature [8, Lemma 3.2]:

Proposition 4. *Let K be an oriented knot and let $\omega\in S^1$. For any Hermitian matrix $A(t)$ which represents the Blanchfield pairing Bl_K, the following equalities hold:*

$$\sigma_K(\omega)=\operatorname{sign}(A(\omega))-\operatorname{sign}(A(1)),$$
$$\eta_K(\omega)=\operatorname{null}(A(\omega)).$$

In the case of links, even though the Blanchfield form can be defined in a way similar to (2), no generalization of Proposition 4 appears to be known at the time of writing. Similarly, the Blanchfield signatures described in Definition 3 do not appear to have been generalized to links.

5 Two additional constructions

We briefly discuss two additional constructions of the Levine-Tristram signature. In Subsection 5.1, we review a construction (due to Lin [64]) which expresses the Murasugi signature of a knot as a signed count of traceless $\mathrm{SU}(2)$-representations. In Subsection 5.2, we discuss Gambaudo and Ghys' work, a corollary of which expresses the Levine-Tristram signature in terms of the Burau representation of the braid group and the Meyer cocycle.

5.1 The Casson-Lin invariant

Let K be an oriented knot. Inspired by the construction of the Casson invariant, Lin defined a knot invariant $h(K)$ via a signed count of conjugacy classes of traceless irreducible representations of $\pi_1(X_K)$ into $\mathrm{SU}(2)$ [64]. Using the behavior of $h(K)$ under crossing changes, Lin additionally showed that $h(K)$ is equal to half the Murasugi signature $\sigma(K)$. The goal of this subsection is to briefly review Lin's construction and to mention some later generalizations.

Let X be a topological space. The *representation space* of X is the set $R(X):=\operatorname{Hom}(\pi_1(X),\mathrm{SU}(2))$ endowed with the compact open topology. A representation is *abelian* if its image is an abelian subgroup of $\mathrm{SU}(2)$ and we let $S(X)$ denote the set of abelian representations. Note that an $\mathrm{SU}(2)$-representation is abelian if and only if it is reducible. The group $\mathrm{SU}(2)$ acts on $R(X)$ by conjugation and its turns out that $\mathrm{SO}(3)=\mathrm{SU}(2)/\pm\mathrm{id}$ acts freely and properly on the set $R(X)\setminus S(X)$

of irreducible (i.e. non abelian) representations. The space of conjugacy classes of irreducible $\mathrm{SU}(2)$-representations of X is denoted by

$$\widehat{R}(X) = (R(X) \setminus S(X))/\mathrm{SO}(3).$$

Given an oriented knot K whose exterior is denoted X_K, the goal is now to make sense of a signed count of the elements $\widehat{R}(X_K)$. The next paragraphs outline the idea underlying Lin's construction.

The braid group B_n can be identified with the group of isotopy classes of orientation preserving homeomorphisms of the punctured disk D_n that fix the boundary pointwise. In particular, each braid β can be represented by a homeomorphism $h_\beta\colon D_n \to D_n$ which in turn induces an automorphism of the free group $F_n \cong \pi_1(D_n)$. In turn, since $R(D_n) \cong \mathrm{SU}(2)^n$, the braid β gives rise to a self-homeomorphism $\beta\colon \mathrm{SU}(2)^n \to \mathrm{SU}(2)^n$. We can therefore consider the spaces

$$\begin{aligned}\Lambda_n &= \{(A_1,\dots,A_n,A_1,\dots,A_n) \mid A_i \in \mathrm{SU}_2^n,\ \mathrm{tr}(A_i) = 0\},\\ \Gamma_n &= \{(A_1,\dots,A_n,\beta(A_1),\dots,\beta(A_n)) \mid A_i \in \mathrm{SU}_2^n,\ \mathrm{tr}(A_i) = 0\}.\end{aligned}$$

Use $\widehat{\beta}$ to denote the link obtained as the closure of a braid β. The representation space $R^0(X_{\widehat{\beta}})$ of traceless $\mathrm{SU}(2)$ representations of $\pi_1(X_{\widehat{\beta}})$ can be identified with $\Lambda_n \cap \Gamma_n$ i.e. the fixed point set of the homeomorphism $\beta\colon \mathrm{SU}(2)^n \to \mathrm{SU}(2)^n$ [64, Lemma 1.2]. Therefore, Lin's idea is to make sense of an algebaic intersection of Λ_n with Γ_n inside the ambient space

$$H_n = \{(A_1,\dots,A_n,B_1,\dots,B_n) \in \mathrm{SU}(2)^n \times \mathrm{SU}(2)^n,\ \mathrm{tr}(A_i) = \mathrm{tr}(B_i) = 0\}.$$

Next, we briefly explain how Lin manages to make sense of this algebraic intersection number. The space $\mathrm{SU}(2) \cong S^3$ is 3-dimensional and the subspace of traceless matrices is homeomorphic to a 2-dimensional sphere. As a consequence, Λ_n and Γ_n are both $2n$-dimensional smooth compact manifolds, and Lin shows that $\widehat{H}_n$ is $4n-3$ dimensional [64, Lemma 1.5]. The $\mathrm{SO}(3)$ action restricts to the spaces Λ_n, Γ_n, H_n and one sets

$$\widehat{H}_n = H_n/\mathrm{SO}(3),\ \ \widehat{\Lambda}_n = \Lambda_n/\mathrm{SO}(3),\ \ \widehat{\Gamma}_n = \Gamma_n/\mathrm{SO}(3).$$

After carefully assigning orientations to these spaces, it follows that $\widehat{\Lambda}_n, \widehat{\Gamma}_n$ are half dimensional smooth oriented submanifolds of the smooth oriented manifold $\widehat{H}_n$. The intersection $\widehat{\Lambda}_n \cap \widehat{\Gamma}_n$ is compact whenever $\widehat{\beta}$ is a knot [64, Lemma 1.6] and therefore, after arranging transversality, one can define the *Casson-Lin invariant* of the braid β as the algebraic intersection

$$h(\beta) := \langle \widehat{\Lambda}_n, \widehat{\Gamma}_n \rangle_{\widehat{H}_n}.$$

Lin proves the invariance of $h(\beta)$ under the Markov moves and shows that the resulting knot invariant is equal to half the Murasugi signature [64, Theorem 1.8 and Corollary 2.10]:

Theorem 11. *The Casson-Lin invariant $h(\beta)$ is unchanged under the Markov moves and thus, setting $h(K) = h(\widehat{\beta})$ for any braid β such that $K = \widehat{\beta}$ defines a knot invariant. Furthermore, $h(K)$ is equal to half the Murasugi signature of K:*

$$h(K) = \frac{1}{2}\sigma(K).$$

Lin's work was later generalized by Herald [43] and Heusener-Kroll [45] to show that the Levine-Tristram signature $\sigma_K(e^{2i\theta})$ can be obtained as a signed count of conjugacy classes of irreducible $\mathrm{SU}(2)$-representations with trace $2\cos(\theta)$. Herald obtained this result via a gauge theoretic interpretation of the Casson-Lin invariant (to do so, he used a 4-dimensional interpretation of the signature), while Heusener-Kroll generalized Lin's original proof (which studies the behavior of $h(K)$ under crossing changes and uses Remark 4).

We also refer to [44] for an interpretation of Lin's construction using the plat closure of a braid (the result is closer to Casson's original construction in terms of Heegaard splittings [1]), and to [19] for a construction of an instanton Floer homology theory whose Euler characteristic is the Levine-Tristram signature. Is there a formula for links? [5] Can Theorem 11 be understood using the constructions of Section 4?

5.2 The Gambaudo-Ghys formula

Since the Alexander polynomial can be expressed using the Burau representation of the braid group [10], one might wonder whether a similar result holds for the Levine-Tristram signature. This subsection describes work of Gambaudo and Ghys [34], a consequence of which answers this question in the positive.

Let B_n denote the n-stranded braid group. Given $\omega \in S^1$, Gambaudo and Ghys study the map $B_n \to \mathbb{Z}, \beta \mapsto \sigma_{\widehat{\beta}}(\omega)$ obtained by sending a braid to the Levine-Tristram signature of its closure. While this map is not a homomorphism, these authors express the homomorphism defect $\sigma_{\widehat{\alpha\beta}}(\omega) - \sigma_{\widehat{\alpha}}(\omega) - \sigma_{\widehat{\beta}}(\omega)$ in terms of the reduced Burau representation

$$\overline{\mathcal{B}}_t \colon B_n \to GL_{n-1}(\mathbb{Z}[t^{\pm 1}]).$$

We briefly recall the definition of $\overline{\mathcal{B}}_t$. Any braid $\beta \in B_n$ can be represented by (an isotopy class of) a homeomorphism $h_\beta \colon D_n \to D_n$ of the punctured disk D_n. This punctured disk has a canonical infinite cyclic cover D_n^∞ (corresponding to the kernel of the map $\pi_1(D_n) \to \mathbb{Z}$ sending the obvious generators of $\pi_1(D_n)$ to 1) and, after fixing basepoints, the homeomorphism h_β lifts to a homeomorphism $\widetilde{h}_\beta \colon D_n^\infty \to D_n^\infty$. It turns out that $H_1(D_n^\infty; \mathbb{Z})$ is a free $\mathbb{Z}[t^{\pm 1}]$-module of rank $n-1$ and the *reduced Bu-*

[5] see [4] for a formula in the case of 2-component links with linking number 1.

rau representation is the $\mathbb{Z}[t^{\pm 1}]$-linear automorphism of $H_1(D_n^\infty;\mathbb{Z})$ induced by $\widetilde{h}_\beta$. This representation is unitary with respect to the equivariant skew-Hermitian form on $H_1(D_n^\infty;\mathbb{Z})$ which is defined by mapping $x,y \in H_1(D_n^\infty;\mathbb{Z})$ to

$$\xi(x,y) = \sum_{n\in\mathbb{Z}} \langle x, t^n y\rangle t^{-n}.$$

In particular, evaluating any matrix for $\overline{\mathcal{B}}_t(\beta)$ at $t=\omega$, the matrix $\overline{\mathcal{B}}_\omega(\alpha)$ preserves the skew-Hermitian form obtained by evaluating a matrix for ξ at $t=\omega$. Therefore, given two braids $\alpha,\beta \in B_n$ and $\omega \in S^1$, one can consider the Meyer cocycle of the two unitary matrices $\overline{\mathcal{B}}_\omega(\alpha)$ and $\overline{\mathcal{B}}_\omega(\beta)$. Here, given a skew-Hermitian form ξ on a complex vector space $\mathbb{C}$ and two unitary automorphisms γ_1,γ_2 of (V,ξ), the *Meyer cocycle* $\mathrm{Meyer}(\gamma_1,\gamma_2)$ is computed by considering the space $E_{\gamma_1,\gamma_2} = \mathrm{im}(\gamma_1^{-1}-\mathrm{id}) \cap \mathrm{im}(\mathrm{id}-\gamma_2)$ and taking the signature of the Hermitian form obtained by setting $b(e,e') = \xi(x_1+x_2,e')$ for $e = \gamma_1^{-1}(x_1)-x_1 = x_2-\gamma_2(x_2) \in E_{\gamma_1,\gamma_2}$ [75, 74].

The following result is due to Gambaudo and Ghys [34, Theorem A].

Theorem 12. *For all $\alpha,\beta \in B_n$ and $\omega \in S^1$ of order coprime to n, the following equation holds:*

$$\sigma_{\widehat{\alpha\beta}}(\omega) - \sigma_{\widehat{\alpha}}(\omega) - \sigma_{\widehat{\beta}}(\omega) = -\mathrm{Meyer}(\overline{\mathcal{B}}_\omega(\alpha), \overline{\mathcal{B}}_\omega(\beta)). \tag{3}$$

In fact, since both sides of (3) define locally constant functions on S^1, Theorem 12 holds on a dense subset of S^1. The proof of Theorem 12 is 4-dimensional; can it also be understood using the constructions of Section 4? The answer ought to follow from [9], where a result analogous to Theorem 12 is established for Blanchfield pairings; see also [36].

We conclude this survey by applying Theorem 12 recursively in order to provide a formula for the Levine-Tristram signature purely in terms of braids. Indeed, using $\sigma_1,\dots,\sigma_{n-1}$ to denote the generators of the braid group B_n (and recalling that the signature vanishes on trivial links), the next result follows from Theorem 12:

Corollary 1. *If an oriented link L is the closure of a braid $\sigma_{i_1}\cdots\sigma_{i_l}$, then the following equality holds on a dense subset of S^1:*

$$\sigma_L(\omega) = -\sum_{j=1}^{l-1} \mathrm{Meyer}(\overline{\mathcal{B}}_\omega(\sigma_{i_1}\cdots\sigma_{i_j}), \overline{\mathcal{B}}_\omega(\sigma_{i_{j+1}})).$$

Acknowledgements I thank Durham University for its hospitality and was supported by an early Postdoc.Mobility fellowship funded by the Swiss National Science Foundation.

References

1. Akbulut, S., McCarthy, J.D.: Casson's invariant for oriented homology 3-spheres, *Mathematical Notes*, vol. 36. Princeton University Press, Princeton, NJ (1990). URL https://doi.org/10.1515/9781400860623. An exposition

2. Atiyah, M., Patodi, V., Singer, I.: Spectral asymmetry and Riemannian geometry. I. Math. Proc. Cambridge Philos. Soc. **77**, 43–69 (1975)
3. Baader, S., Dehornoy, P., Liechti, L.: Signature and concordance of positive knots. Bull. Lond. Math. Soc. **50**(1), 166–173 (2018). URL https://doi.org/10.1112/blms.12124
4. Bénard, L., Conway, A.: A multivariable Casson-Lin type invariant. Annales de l'institut Fourier To appear
5. Borodzik, M.: Abelian ρ-invariants of iterated torus knots. In: Low-dimensional and symplectic topology, *Proc. Sympos. Pure Math.*, vol. 82, pp. 29–38. Amer. Math. Soc., Providence, RI (2011). URL https://doi.org/10.1090/pspum/082/2768651
6. Borodzik, M., Friedl, S.: On the algebraic unknotting number. Trans. London Math. Soc. **1**(1), 57–84 (2014)
7. Borodzik, M., Friedl, S.: The unknotting number and classical invariants II. Glasg. Math. J. **56**(3), 657–680 (2014). URL http://dx.doi.org/10.1017/S0017089514000081
8. Borodzik, M., Friedl, S.: The unknotting number and classical invariants, I. Algebr. Geom. Topol. **15**(1), 85–135 (2015). URL http://dx.doi.org/10.2140/agt.2015.15.85
9. Bourrigan, M.: Quasimorphismes sur les groupes de tresses et forme de blanchfield. PhD thesis (2013)
10. Burau, W.: Über Zopfgruppen und gleichsinnig verdrillte Verkettungen. Abh. Math. Sem. Univ. Hamburg **11**(1), 179–186 (1935). URL http://dx.doi.org/10.1007/BF02940722
11. Casson, A., Gordon, C.: On slice knots in dimension three. In: Algebraic and geometric topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, Calif., 1976), Part 2, Proc. Sympos. Pure Math., XXXII, pp. 39–53. Amer. Math. Soc., Providence, R.I. (1978)
12. Cha, J.C.: The structure of the rational concordance group of knots. Mem. Amer. Math. Soc. **189**(885), x+95 (2007). URL https://doi.org/10.1090/memo/0885
13. Cha, J.C., Livingston, C.: Knot signature functions are independent. Proc. Amer. Math. Soc. **132**(9), 2809–2816 (2004). URL https://doi.org/10.1090/S0002-9939-04-07378-2
14. Cimasoni, D., Conway, A., Zacharova, K.: Splitting numbers and signatures. Proc. Amer. Math. Soc. **144**(12), 5443–5455 (2016). URL https://doi.org/10.1090/proc/13156
15. Cimasoni, D., Florens, V.: Generalized Seifert surfaces and signatures of colored links. Trans. Amer. Math. Soc. **360**(3), 1223–1264 (electronic) (2008)
16. Cochran, T.D., Franklin, B.D., Hedden, M., Horn, P.D.: Knot concordance and homology cobordism. Proc. Amer. Math. Soc. **141**(6), 2193–2208 (2013). URL https://doi.org/10.1090/S0002-9939-2013-11471-1
17. Cochran, T.D., Gompf, R.E.: Applications of Donaldson's theorems to classical knot concordance, homology 3-spheres and property *P*. Topology **27**(4), 495–512 (1988). URL https://doi.org/10.1016/0040-9383(88)90028-6
18. Cochran, T.D., Orr, K.E., Teichner, P.: Structure in the classical knot concordance group. Comment. Math. Helv. **79**(1), 105–123 (2004). URL https://doi.org/10.1007/s00014-001-0793-6
19. Collin, O., Steer, B.: Instanton Floer homology for knots via 3-orbifolds. J. Differential Geom. **51**(1), 149–202 (1999). URL http://projecteuclid.org/euclid.jdg/1214425027
20. Collins, J.: The L^2 signature of torus knots (2010). Https://arxiv.org/pdf/1001.1329.pdf
21. Conway, A., Nagel, M., Toffoli, E.: Multivariable signatures, genus bounds and 1-solvable cobordisms. ArXiv:1703.07540 (2017)
22. Conway, J.H.: An enumeration of knots and links, and some of their algebraic properties. In: Computational Problems in Abstract Algebra (Proc. Conf., Oxford, 1967), pp. 329–358. Pergamon, Oxford (1970)

23. Degtyarev, A., Florens, V., Lecuona, A.G.: The signature of a splice. Int. Math. Res. Not. IMRN (8), 2249–2283 (2017). URL https://doi.org/10.1093/imrn/rnw068
24. Degtyarev, A., Florens, V., Lecuona, A.G.: Slopes and signatures of links (2018). Https://arxiv.org/pdf/1802.01836.pdf
25. Erle, D.: Quadratische Formen als Invarianten von Einbettungen der Kodimension 2. Topology **8**, 99–114 (1969)
26. Feller, P.: Gordian adjacency for torus knots. Algebr. Geom. Topol. **14**(2), 769–793 (2014). URL https://doi.org/10.2140/agt.2014.14.769
27. Feller, P.: The signature of positive braids is linearly bounded by their first Betti number. Internat. J. Math. **26**(10), 1550,081, 14 (2015). URL https://doi.org/10.1142/S0129167X15500810
28. Florens, V.: Signatures of colored links with application to real algebraic curves. J. Knot Theory Ramifications **14**(7), 883–918 (2005). URL http://dx.doi.org/10.1142/S0218216505004093
29. Florens, V., Gilmer, P.M.: On the slice genus of links. Algebr. Geom. Topol. **3**, 905–920 (2003). URL https://doi.org/10.2140/agt.2003.3.905
30. Fogel, M.E.: The algebraic unknotting number. ProQuest LLC, Ann Arbor, MI (1993). URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:9407954. Thesis (Ph.D.)–University of California, Berkeley
31. Friedl, S.: Eta invariants as sliceness obstructions and their relation to Casson-Gordon invariants. Algebr. Geom. Topol. **4**, 893–934 (2004). URL http://dx.doi.org/10.2140/agt.2004.4.893
32. Friedl, S.: L^2-eta-invariants and their approximation by unitary eta-invariants. Math. Proc. Cambridge Philos. Soc. **138**(2), 327–338 (2005). URL http://dx.doi.org/10.1017/S0305004104008084
33. Friedl, S.: Link concordance, boundary link concordance and eta-invariants. Math. Proc. Cambridge Philos. Soc. **138**(3), 437–460 (2005). URL https://doi.org/10.1017/S0305004105008455
34. Gambaudo, J.M., Ghys, É.: Braids and signatures. Bull. Soc. Math. France **133**(4), 541–579 (2005)
35. Garoufalidis, S.: Does the Jones polynomial determine the signature of a knot? (2003). Https://arxiv.org/pdf/math/0310203.pdf
36. Ghys, E., Ranicki, A. (eds.): Six papers on signatures, braids and Seifert surfaces, *Ensaios Matemáticos [Mathematical Surveys]*, vol. 30. Sociedade Brasileira de Matemática, Rio de Janeiro (2016)
37. Gilmer, P.M.: Configurations of surfaces in 4-manifolds. Trans. Amer. Math. Soc. **264**(2), 353–380 (1981). URL https://doi.org/10.2307/1998544
38. Gilmer, P.M., Livingston, C.: Signature jumps and Alexander polynomials for links. Proc. Amer. Math. Soc. **144**(12), 5407–5417 (2016). URL http://dx.doi.org/10.1090/proc/13129
39. Gordon, C.: Some aspects of classical knot theory. In: Knot theory (Proc. Sem., Plans-sur-Bex, 1977), *Lecture Notes in Math.*, vol. 685, pp. 1–60. Springer, Berlin (1978)
40. Gordon, C., Litherland, R.: On the signature of a link. Invent. Math. **47**(1), 53–69 (1978). URL https://doi.org/10.1007/BF01609479
41. Gordon, C., Litherland, R.: On a theorem of Murasugi. Pacific J. Math. **82**(1), 69–74 (1979). URL http://projecteuclid.org/euclid.pjm/1102785061
42. Gordon, C., Litherland, R., Murasugi, K.: Signatures of covering links. Canad. J. Math. **33**(2), 381–394 (1981). URL https://doi.org/10.4153/CJM-1981-032-3
43. Herald, C.M.: Flat connections, the Alexander invariant, and Casson's invariant. Comm. Anal. Geom. **5**(1), 93–120 (1997). URL https://doi.org/10.4310/CAG.1997.v5.n1.a2

44. Heusener, M.: An orientation for the SU(2)-representation space of knot groups. In: Proceedings of the Pacific Institute for the Mathematical Sciences Workshop "Invariants of Three-Manifolds" (Calgary, AB, 1999), vol. 127, pp. 175–197 (2003). URL https://doi.org/10.1016/S0166-8641(02)00059-7
45. Heusener, M., Kroll, J.: Deforming abelian SU(2)-representations of knot groups. Comment. Math. Helv. **73**(3), 480–498 (1998). URL https://doi.org/10.1007/s000140050065
46. Kauffman, L.H.: On knots, *Annals of Mathematics Studies*, vol. 115. Princeton University Press, Princeton, NJ (1987)
47. Kauffman, L.H., Taylor, L.R.: Signature of links. Trans. Amer. Math. Soc. **216**, 351–365 (1976)
48. Kawauchi, A.: On quadratic forms of 3-manifolds. Invent. Math. **43**(2), 177–198 (1977). URL https://doi.org/10.1007/BF01390003
49. Kawauchi, A.: A survey of knot theory. Birkhäuser Verlag, Basel (1996). Translated and revised from the 1990 Japanese original by the author
50. Kawauchi, A.: The quadratic form of a link. In: Low-dimensional topology (Funchal, 1998), *Contemp. Math.*, vol. 233, pp. 97–116. Amer. Math. Soc., Providence, RI (1999). URL https://doi.org/10.1090/conm/233/03426
51. Kearton, C.: The Milnor signatures of compound knots. Proc. Amer. Math. Soc. **76**(1), 157–160 (1979). URL http://dx.doi.org/10.2307/2042936
52. Kearton, C.: Signatures of knots and the free differential calculus. Quart. J. Math. Oxford Ser. (2) **30**(118), 157–182 (1979). URL http://dx.doi.org/10.1093/qmath/30.2.157
53. Kearton, C.: Hermitian signatures and double-null-cobordism of knots. J. London Math. Soc. (2) **23**(3), 563–576 (1981). URL https://doi.org/10.1112/jlms/s2-23.3.563
54. Kearton, C.: Quadratic forms in knot theory. In: Quadratic forms and their applications (Dublin, 1999), *Contemp. Math.*, vol. 272, pp. 135–154. Amer. Math. Soc., Providence, RI (2000). URL https://doi.org/10.1090/conm/272/04401
55. Kirby, R., Melvin, P.: Dedekind sums, μ-invariants and the signature cocycle. Math. Ann. **299**(2), 231–267 (1994). URL https://doi.org/10.1007/BF01459782
56. Letsche, C.F.: An obstruction to slicing knots using the eta invariant. Math. Proc. Cambridge Philos. Soc. **128**(2), 301–319 (2000). URL http://dx.doi.org./10.1017/S0305004199004016
57. Levine, J.: Knot cobordism groups in codimension two. Comment. Math. Helv. **44**, 229–244 (1969)
58. Levine, J.: Knot modules. I. Trans. Amer. Math. Soc. **229**, 1–50 (1977)
59. Levine, J.: Metabolic and hyperbolic forms from knot theory. J. Pure Appl. Algebra **58**(3), 251–260 (1989). URL https://doi.org/10.1016/0022-4049(89)90040-6
60. Levine, J.: Link invariants via the eta invariant. Comment. Math. Helv. **69**(1), 82–119 (1994). URL http://dx.doi.org/10.1007/BF02564475
61. Lickorish, W.B.R.: An introduction to knot theory, *Graduate Texts in Mathematics*, vol. 175. Springer-Verlag, New York (1997). URL http://dx.doi.org/10.1007/978-1-4612-0691-0
62. Liechti, L.: Positive braid knots of maximal topological 4-genus. Math. Proc. Cambridge Philos. Soc. **161**(3), 559–568 (2016). URL https://doi.org/10.1017/S0305004116000670
63. Liechti, L.: Signature, positive Hopf plumbing and the Coxeter transformation. Osaka J. Math. **53**(1), 251–266 (2016). URL http://projecteuclid.org/euclid.ojm/1455892632. With an appendix by Peter Feller and Liechti
64. Lin, X.S.: Representations of knot groups and twisted Alexander polynomials. Acta Math. Sin. (Engl. Ser.) **17**(3), 361–380 (2001). URL http://dx.doi.org/10.1007/s101140100122
65. Litherland, R.: Cobordism of satellite knots. Four-manifold theory, Proc. AMS-IMS-SIAM Joint Summer Res. Conf., Durham/N.H. 1982, Contemp. Math. 35, 327-362 (1984). (1984)

66. Litherland, R.A.: Signatures of iterated torus knots. In: Topology of low-dimensional manifolds (Proc. Second Sussex Conf., Chelwood Gate, 1977), *Lecture Notes in Math.*, vol. 722, pp. 71–84. Springer, Berlin (1979)
67. Livingston, C.: Knot theory, *Carus Mathematical Monographs*, vol. 24. Mathematical Association of America, Washington, DC (1993)
68. Livingston, C.: The slicing number of a knot. Algebr. Geom. Topol. **2**, 1051–1060 (2002). URL https://doi.org/10.2140/agt.2002.2.1051
69. Livingston, C.: Knot 4-genus and the rank of classes in $W(\mathbb{Q}(t))$. Pacific J. Math. **252**(1), 113–126 (2011). URL https://doi.org/10.2140/pjm.2011.252.113
70. Livingston, C.: Signature functions of knots (2017). Https://arxiv.org/pdf/1709.00732.pdf
71. Livingston, C.: Signature invariants related to the unknotting number (2017). Https://arxiv.org/pdf/1710.10530.pdf
72. Livingston, C., Melvin, P.: Abelian invariants of satellite knots. In: Geometry and topology (College Park, Md., 1983/84), *Lecture Notes in Math.*, vol. 1167, pp. 217–227. Springer, Berlin (1985). URL https://doi.org/10.1007/BFb0075225
73. Matumoto, T.: On the signature invariants of a non-singular complex sesquilinear form. J. Math. Soc. Japan **29**(1), 67–71 (1977)
74. Meyer, W.: Die Signatur von lokalen Koeffizientensystemen und Faserbündeln. Bonn. Math. Schr. (53), viii+59 (1972)
75. Meyer, W.: Die Signatur von Flächenbündeln. Math. Ann. **201**, 239–264 (1973)
76. Milnor, J.W.: Infinite cyclic coverings. In: Conference on the Topology of Manifolds (Michigan State Univ., E. Lansing, Mich., 1967), pp. 115–133. Prindle, Weber & Schmidt, Boston, Mass. (1968)
77. Murakami, H.: Algebraic unknotting operation. In: Proceedings of the Second Soviet-Japan Joint Symposium of Topology (Khabarovsk, 1989), vol. 8, pp. 283–292 (1990)
78. Murasugi, K.: On a certain numerical invariant of link types. Trans. Amer. Math. Soc. **117**, 387–422 (1965)
79. Nagel, M., Owens, B.: Unlinking information from 4-manifolds. Bull. Lond. Math. Soc. **47**(6), 964–979 (2015). URL https://doi.org/10.1112/blms/bdv072
80. Nagel, M., Powell, M.: Concordance invariance of Levine-Tristram signatures of links. Doc. Math. **22**, 25–43 (2017)
81. Orevkov, S.Y.: Plane real algebraic curves of odd degree with a deep nest. J. Knot Theory Ramifications **14**(4), 497–522 (2005). URL https://doi.org/10.1142/S0218216505003920
82. Orevkov, S.Y.: Some examples of real algebraic and real pseudoholomorphic curves. In: Perspectives in analysis, geometry, and topology, *Progr. Math.*, vol. 296, pp. 355–387. Birkhäuser/Springer, New York (2012). URL https://doi.org/10.1007/978-0-8176-8277-4_15
83. Owens, B.: On slicing invariants of knots. Trans. Amer. Math. Soc. **362**(6), 3095–3106 (2010). URL https://doi.org/10.1090/s0002-9947-09-04904-6
84. Powell, M.: The four-genus of a link, Levine-Tristram signatures and satellites. J. Knot Theory Ramifications **26**(2), 1740,008, 28 (2017). URL http://dx.doi.org/10.1142/S0218216517400089
85. Przytycki, J.H.: Positive knots have negative signature. Bull. Polish Acad. Sci. Math. **37**(7-12), 559–562 (1990) (1989)
86. Przytycki, J.H., Taniyama, K.: Almost positive links have negative signature. J. Knot Theory Ramifications **19**(2), 187–289 (2010). URL https://doi.org/10.1142/S0218216510007838
87. Rinat, K.: On symmetric matrices associated with oriented link diagrams (2018). Https://arxiv.org/pdf/1801.04632.pdf
88. Rudolph, L.: Nontrivial positive braids have positive signature. Topology **21**(3), 325–327 (1982). URL https://doi.org/10.1016/0040-9383(82)90014-3

89. Shinohara, Y.: On the signature of knots and links. Trans. Amer. Math. Soc. **156**, 273–285 (1971)
90. Stoimenow, A.: Some applications of Tristram-Levine signatures and relation to Vassiliev invariants. Adv. Math. **194**(2), 463–484 (2005). URL https://doi.org/10.1016/j.aim.2004.07.003
91. Stoimenow, A.: Bennequin's inequality and the positivity of the signature. Trans. Amer. Math. Soc. **360**(10), 5173–5199 (2008). URL https://doi.org/10.1090/S0002-9947-08-04410-3
92. Sumners, D.W.: Invertible knot cobordisms. Comment. Math. Helv. **46**, 240–256 (1971). URL https://doi.org/10.1007/BF02566842
93. Traczyk, P.h.: Nontrivial negative links have positive signature. Manuscripta Math. **61**(3), 279–284 (1988). URL https://doi.org/10.1007/BF01258439
94. Tristram, A.: Some cobordism invariants for links. Proc. Cambridge Philos. Soc. **66**, 251–264 (1969)
95. Trotter, H.: On *S*-equivalence of Seifert matrices. Invent. Math. **20**, 173–207 (1973)
96. Viro, O.: Branched coverings of manifolds with boundary, and invariants of links. I. Izv. Akad. Nauk SSSR Ser. Mat. **37**, 1241–1258 (1973)
97. Viro, O.: Twisted acyclicity of a circle and signatures of a link. J. Knot Theory Ramifications **18**(6), 729–755 (2009). URL http://dx.doi.org/10.1142/S0218216509007142

PD_4-complexes and 2-dimensional duality groups

Jonathan A. Hillman

Abstract This paper is a synthesis and extension of three earlier papers on PD_4-complexes X such that $\pi = \pi_1(X)$ has one end and $c.d.\pi = 2$. The basic notion is that of strongly minimal PD_4-complex, one for which the equivariant intersection pairing λ_X on $\pi_2(X)$ is null. The first main result is that two PD_4-complexes with the same strongly minimal model are homotopy equivalent if and only if their intersection pairings are isometric. If $c.d.\pi \leq 2$ every such complex has a strongly minimal model, and the second half of the paper focuses largely on determining the minimal models. In particular, if π is a surface group or is a semidirect product $F(r) \rtimes \mathbb{Z}$ then the homotopy type of X is determined by π, the Stiefel-Whitney classes and λ_X. Although we expect that the strategy in the surface group case should extend to all π such that $c.d.\pi = 2$ and π has one end, we do not yet have a unified proof that covers the known cases. We conclude with an application to 2-knots and a short list of questions for further research.

1 Introduction

It remains an open problem to give a homotopy classification of closed 4-manifolds or PD_4-complexes, in terms of standard invariants such as the fundamental group, characteristic classes and intersection pairings. Hambleton and Kreck showed that if X is orientable and $H_2(X;\mathbb{Q}) \neq 0$ the homotopy type of X is determined by its Postnikov 2-stage $P_2(X)$ and the image of the fundamental class $[X]$ in $H_4(P_2(X);\mathbb{Z})$, and if $\pi_1(X)$ is finite and of cohomological period dividing 4 this image is in turn determined by the equivariant intersection pairing on $\pi_2(X)$ [27]. Baues and Bleile have extended the first part of this result to all PD_4-complexes: two PD_4-complexes X and Y are homotopy equivalent if and only if there is a homotopy equivalence $h : P_2(X) \to P_2(Y)$ such that $h^*w_1(Y) = w_1(X)$, and which carries the image of $[X]$

School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
e-mail: jonathan.hillman@sydney.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_3

in $H_4(P_2(X);\mathbb{Z}^{w_1(X)})$ to the image of $\pm[Y]$ in $H_4(P_2(Y);\mathbb{Z}^{w_1(Y)})$. (Here $w_1(X)$ and $w_1(Y)$ are the orientation characters and $\mathbb{Z}^{w_1(X)}$ and $\mathbb{Z}^{w_1(X)}$ the associated twisted coefficient modules.) They also give a homotopy classification of PD_4-complexes (up to 2-torsion) in terms of homotopy classes of chain complexes with a homotopy commutative diagonal and an additional quadratic structure [5]. However, there is still the question of how to characterize the classes in $H_4(P_2(X);\mathbb{Z}^{w_1(X)})$ which correspond to PD_4-complexes.

We shall extend the work of [27] to relate such classes to intersection pairings, for certain cases with $\pi = \pi_1(X)$ infinite. The central idea is that of "strongly minimal PD_4-complex", one for which the equivariant intersection pairing is identically 0. (We shall in fact use the equivalent cohomological pairing.) If there is a 2-connected degree-1 map $f : X \to Z$, with Z strongly minimal, and if the orientation character $w = w_1(X) : \pi \to \mathbb{Z}^\times$ does not split then the homotopy type of X is determined by the homotopy type of Z and the equivariant intersection pairing. Every PD_4-complex X with fundamental group π has such a "strongly minimal model" Z if and only if $c.d.\pi \leq 2$. (See Theorem 21 below.) This class of groups is both tractable and of direct interest to low-dimensional geometric topology, as it includes all surface groups, knot groups and the groups of many other bounded 3-manifolds. We expect that if $c.d.\pi \leq 2$ the homotopy type of Z is determined by π, w and the Wu class $v_2(Z)$, and that if $v_2(X)$ is induced from π then the minimal model is unique. (In the latter case, the homotopy type of X is determined by π, w, $v_2(X)$ and the equivariant intersection pairing.) However, this is only known for π a free group, a surface group, a semidirect product $F(r) \rtimes \mathbb{Z}$ or a solvable Baumslag-Solitar group $\mathbb{Z}*_m$.

We shall now outline the paper in more detail. The first two sections are algebraic. In particular, Theorem 1 (in §2) establishes a connection between hermitian pairings and the Whitehead quadratic functor Γ_W. Sections 3–8 consider the homotopy classification of PD_4-complexes, and introduce several notions of minimality. The first main result is Theorem 7 in §7, where it is shown that two PD_4-complexes with the same strongly minimal model and $\pm$isometric intersection pairings are homotopy equivalent, provided $w : \pi \to \mathbb{Z}^\times$ does not split. Sections 9 and 10 determine the strongly minimal PD_4-complexes with $\pi_2 = 0$ and for which π has finitely many ends. Strongly minimal PD_4-complexes with π a semidirect product $\nu \rtimes \mathbb{Z}$ (with ν finitely presentable) are shown to be mapping tori in §11. When ν is a free group the homotopy type of such a mapping torus is determined by π and the Stiefel-Whitney classes, by Theorem 22. The next five sections lead to the second main result, Theorem 27 (in §16), which extends the result of Theorem 22 to the case when π has one end and $c.d.\pi = 2$ *provided* that the image of the symmetric square $\Pi \odot \Pi$ in $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$ is 2-torsion free, where $\Pi = \pi_2(X) \cong \overline{H^2(\pi;\mathbb{Z}[\pi])}$. This theorem is modelled on the much simpler case analyzed in §14, in which π is a PD_2-group. Apart from the notion of minimality, the main technical points are the connection between hermitian pairings and Γ_W, the fact that a certain "cup product" defines an isomorphism, and the 2-torsion condition. In [40], we showed that the cup-product condition held for surface groups, torus knot groups and solvable Baumslag-Solitar groups. Here we show that it holds for all finitely presentable groups π with one end and $c.d.\pi = 2$ (Theorem 26). The 2-torsion condition is only known for π a PD_2-

group or π a solvable Baumslag-Solitar group (Theorem 30), and does not hold for all the cases covered by Theorem 22. The penultimate section considers the classification up to TOP s-cobordism or homeomorphism of closed 4-manifolds with groups as in Theorem 27. In particular, it is shown that a remarkable 2-knot discovered by Fox is determined up to TOP isotopy and reflection by its knot group. In the body of the text we raise a number of questions, some on points of detail, that we have not been able to settle. The most significant of these have been collected in the final section.

The interactions of cohomology of groups, Poincaré duality and the lower stages of Postnikov towers are central to the arguments. We refer particularly to [9, 51, 54, 62] for more on these topics. Some of the other techniques invoked, such as L^2-homology (used in §5 to compare various notions of minimality) or Farrell cohomology (used in §10 in connection with PD_4-complexes whiose universal covers have two ends) may seem more recondite, but these are mostly used in excursions aside the main theme, and familiarity with such notions is not essential.

The theme of Hambleton, Kreck and Teichner [29] is close to ours, although their methods are very different. They use Kreck's modified surgery theory to classify up to s-cobordism closed orientable 4-manifolds with fundamental groups of geometric dimension 2 (subject to some K- and L-theoretic hypotheses), and they show also that every automorphism of the algebraic 2-type is realized by an s-cobordism, in many cases. (They do not require that π have one end, which is a restriction imposed by our arguments. However, when π is a free group there is a simpler, more homological approach, which also uses the ideas of §2 below [37].)

This paper is a synthesis and extension of three papers [38, 39, 40] which explored the role of minimality in the classification of PD_4-complexes, in particular, those with fundamental group π such that $c.d.\pi = 2$ and π has one end. (Some aspects were considered much earlier [35, 36].) Apart from the benefits of revision, the main novelties are in showing that strongly minimal finite PD_4-complexes have minimal Euler characteristic (Corollary 3), strong minimality is equivalent to order minimality if and only if $c.d.\pi \leq 2$ (Theorem 21), verification that cup product defines an isomorphism for all 2-dimensional duality groups (Theorem 26), clarification of the role of the refined v_2-type, and relaxation of some of the hypotheses.

I would like to thank the referee for his close reading of the original submission, and for his suggestions for the improvement of the exposition.

2 Modules and group rings

Let π be a finitely presentable group and $w : \pi \to \mathbb{Z}^\times = \{\pm 1\}$ be a homomorphism. (This shall represent the orientation character for a PD_n-complex with fundamental group π.) We shall at times view w as a class in $H^1(\pi;\mathbb{F}_2)$, since this cohomology group may be identified with $Hom(\pi,\mathbb{Z}^\times)$. Define an involution on $\mathbb{Z}[\pi]$ by $\bar{g} = w(g)g^{-1}$, for all $g \in \pi$. Let $\mathbb{Z}$ and $\mathbb{Z}^w$ be the augmentation and w-twisted augmentation rings, and $\varepsilon : \mathbb{Z}[\pi] \to \mathbb{Z}$ and $\varepsilon_w : \mathbb{Z}[\pi] \to \mathbb{Z}^w$ be the augmentation and

the w-twisted augmentation, defined by $\varepsilon(g) = 1$ and $\varepsilon_w(g) = w(g)$, for all $g \in \pi$, respectively. Let $I_w = \mathrm{Ker}(\varepsilon_w)$.

All modules considered here shall be left modules, unless otherwise noted. However, if L is a left $\mathbb{Z}[\pi]$-module the dual $Hom_{\mathbb{Z}[\pi]}(L,\mathbb{Z}[\pi])$ and the higher extension groups $Ext^i_{\mathbb{Z}[\pi]}(L,\mathbb{Z}[\pi])$ are naturally right modules. If R is a right $\mathbb{Z}[\pi]$-module let $\overline{R}$ be the corresponding left $\mathbb{Z}[\pi]$-module with the conjugate structure given by $g.r = r.\bar{g}$, for all $g \in \mathbb{Z}[\pi]$ and $r \in R$. Let $L^\dagger = \overline{Hom_{\mathbb{Z}[\pi]}(L,\mathbb{Z}[\pi])}$ and $E^iL = \overline{Ext^i_{\mathbb{Z}[\pi]}(L,\mathbb{Z}[\pi])}$, for $i \geq 0$ be the conjugate dual left modules. If L is free, stably free or projective then so is $E^0L = L^\dagger$. We shall consider $\mathbb{Z}$ and $\mathbb{Z}^w$ to be bimodules, with the same left and right π-structures. (Note that $\overline{\mathbb{Z}} = \mathbb{Z}^w$.)

The modules $E^q\mathbb{Z} = \overline{H^q(\pi;\mathbb{Z}[\pi])}$ with $q \leq 3$ shall recur throughout this paper. In particular, $E^0\mathbb{Z} \cong \mathbb{Z}^w$ if π is finite and is 0 otherwise, while $E^1\mathbb{Z}$ reflects the number of ends of π. It is 0 if π is finite or has one end, infinite cyclic if π has two ends (i.e., is virtually infinite cyclic) and is free abelian of infinite rank otherwise.

Lemma 1. *Let M be a $\mathbb{Z}[\pi]$-module with a finite resolution of length n and such that $E^iM = 0$ for $i < n$. Then $Aut(M) \cong Aut(E^nM)$.*

Proof. Since $E^iM = 0$ for $i < n$ the dual of a resolution of length n for M is a finite resolution for E^nM. Taking duals again recovers the original resolution, and so $E^nE^nM \cong M$. If $f \in Aut(M)$ it extends to an endomorphism of the resolution inducing an automorphism E^nf of E^nM. Taking duals again gives $E^nE^nf = f$. Thus $f \mapsto E^nf$ determines an isomorphism $Aut(M) \cong Aut(E^nM)$.

A group π is an *n-dimensional duality group* over $\mathbb{Z}$ if the augmentation $\mathbb{Z}[\pi]$-module $\mathbb{Z}$ has a finite projection resolution of length n, $H^i(\pi;\mathbb{Z}[\pi]) = 0$ for $i < n$ and the *dualizing module* $\mathscr{D} = H^n(\pi;\mathbb{Z}[\pi])$ is torsion free as an abelian group. (See [9, Theorem VIII.10.1].) We then have $Aut(E^n\mathbb{Z}) = \mathbb{Z}^\times$, by Lemma 1. Finitely generated free groups are duality groups of dimension 1. If π is finitely presentable and $c.d.\pi = 2$ then $H^2(\pi;\mathbb{Z}[\pi]) \neq 0$, and it is torsion free, by [25, Proposition 13.7.1]. Hence π is a 2-dimensional duality group if and only if it has one end.

In general, $H^2(\pi;\mathbb{Z}[\pi])$ is 0, $\mathbb{Z}$ or not finitely generated ([21] – see [25, Proposition 13.7.12]). In the latter case, $H^2(\pi;\mathbb{Z}[\pi])$ must have infinite rank, by the main result of [8]. It remains open whether $H^2(\pi;\mathbb{Z}[\pi])$ must be free as an abelian group.

We shall use the "free differential calculus" of Fox and Lyndon to provide partial resolutions of augmentation modules. (See [23] and [46].) Let $F(n)$ be the free group with basis $\{x_1,\dots,x_n\}$. The augmentation ideal of $\mathbb{Z}[F(n)]$ is freely generated by $\{x_1-1,\dots,x_n-1\}$ as a left $\mathbb{Z}[F(n)]$-module and so we may write

$$r - 1 = \Sigma_{1\leq i\leq n} \frac{\partial r}{\partial x_i}(x_i - 1),$$

for $r \in F(n)$. Since $rs - 1 = r - 1 + r(s-1)$, for all $r,s \in F(\mu)$, the Leibniz conditions

$$\frac{\partial rs}{\partial x_i} = \frac{\partial r}{\partial x_i} + r\frac{\partial s}{\partial x_i}$$

hold for all $r,s \in F(\mu)$ and $1 \le i \le n$. In particular, $\frac{\partial 1}{\partial x_i} = 0$ and $\frac{\partial r^{-1}}{\partial x_i} = -r^{-1}\frac{\partial r}{\partial x_i}$, for $1 \le i \le n$. We may extend these functions linearly to "derivations" of $\mathbb{Z}[F(n)]$.

Now let π be a group with a finite presentation

$$\mathscr{P} = \langle x_1, \dots, x_g | r_1, \dots, r_h \rangle^{\varphi},$$

where $\varphi : F(g) \to \pi$ is an epimorphism with kernel the normal closure of $\{r_1, \dots, r_h\}$. Let $def(\mathscr{P}) = g - h$ be the deficiency and $C(\mathscr{P})$ be the 2-complex corresponding to this presentation. Then $\chi(C(\mathscr{P})) = 1 - def(\mathscr{P})$. A choice of lifts of the q-cells of $C(\mathscr{P})$ to the universal cover $\widetilde{C(\mathscr{P})}$ determines a basis for $C_q(\widetilde{C(\mathscr{P})})$ as a free left $\mathbb{Z}[\pi]$-module. We view these as modules of column vectors. The differentials are given by $\partial_1(c_1^{(i)}) = (\varphi(x_i) - 1)c_0$ and $\partial_2(c_2^{(j)}) = \Sigma_{1 \le i \le g}\varphi(\frac{\partial r_j}{\partial x_i})c_1^{(i)}$. (We extend φ linearly to the group rings.) The module of 0-cycles $Z_0(\widetilde{C(\mathscr{P})})$ is isomorphic to $I(\pi)$, and so $I(\pi)$ has a $g \times h$ presentation matrix with (i,j)th entry $\varphi(\frac{\partial R_j}{\partial x_i})$. (We shall refer to $C_*(\widetilde{C(\mathscr{P})})$ as the *Fox-Lyndon resolution* of $\mathbb{Z}$ associated to $\mathscr{P}$.)

Lemma 2. *Let $\pi = G * F(n)$, where $G = *_{i=1}^{m} G_i$ is the free product of $m \ge 1$ finitely generated, one-ended groups G_i and $n \ge 0$. Then $E^1\mathbb{Z} \cong \mathbb{Z}[\pi]^{m+n-1}$.*

Proof. If $n = 0$ the result follows from the Mayer-Vietoris sequence for the free product, with coefficients $\mathbb{Z}[\pi]$.

In general, let $C_*(G)$ be a resolution of the augmentation module by free $\mathbb{Z}[G]$-modules with $C_0(G) = \mathbb{Z}[G]$. Then there is a corresponding resolution $C_*(\pi)$ with $C_q(\pi) \cong \mathbb{Z}[\pi] \otimes_{\mathbb{Z}[G]} C_q(G)$ if $q \ne 1$ and $C_1(\pi) \cong \mathbb{Z}[\pi] \otimes_{\mathbb{Z}[G]} C_q(G) \oplus \mathbb{Z}[\pi]^n$. Hence there is a short exact sequence of chain complexes

$$0 \to \mathbb{Z}[\pi] \otimes_{\mathbb{Z}[G]} C_*(G) \to C_*(\pi) \to \mathbb{Z}[\pi]^n \to 0,$$

where the third term is concentrated in degree 1. The exact sequence of cohomology with coefficients $\mathbb{Z}[\pi]$ and conjugation give a short exact sequence

$$0 \to \mathbb{Z}[\pi]^s \to \overline{H^1(\pi; \mathbb{Z}[\pi])} \to \overline{H^1(Hom_{\mathbb{Z}[\pi]}(\mathbb{Z}[\pi] \otimes_{\mathbb{Z}[G]} C_*(G), \mathbb{Z}[\pi])} \to 0.$$

We may identify the right-hand term with $\mathbb{Z}[\pi] \otimes_{\mathbb{Z}[G]} \overline{H^1(G; \mathbb{Z}[G])} \cong \mathbb{Z}[\pi]^{m-1}$, since G is finitely generated. The middle term is $E^1\mathbb{Z}$, and so the lemma follows easily.

The hypothesis of this lemma holds if π is torsion free but not free. On the other hand, if π is a nontrivial free group then $E^1\mathbb{Z}$ has projective dimension 1 as a $\mathbb{Z}[\pi]$-module, and so the conclusion fails.

If M is a $\mathbb{Z}[\pi]$-module and ν is a subgroup of π then $M|_\nu$ shall denote the $\mathbb{Z}[\nu]$-module obtained by restriction of scalars.

3 The Whitehead functor and hermitian pairings

Let A and B be abelian groups. A function $f : A \to B$ is *quadratic* if $f(-a) = f(a)$ for all $a \in A$ and if $f(a+b) - f(a) - f(b)$ defines a bilinear function from $A \times A$ to B. The *Whitehead quadratic functor* Γ_W assigns to each abelian group A an abelian group $\Gamma_W(A)$ and a quadratic function $\gamma_A : A \to \Gamma_W(A)$ which is universal for quadratic functions with domain A. The natural epimorphism from A onto $A/2A = \mathbb{F}_2 \otimes A$ is quadratic, and so induces a canonical epimorphism q_A from $\Gamma_W(A)$ to $A/2A$. Let $A \odot A = A \otimes A/\langle a \otimes b - b \otimes a | \forall a, b \in A \rangle$ be the *symmetric square* of A. Then the kernel of q_A is the image of $A \odot A$ under the homomorphism s from $A \odot A$ to $\Gamma_W(A)$ given by $s(a \odot b) = \gamma_A(a+b) - \gamma_A(a) - \gamma_A(b)$. Thus there is an exact sequence

$$A \odot A \xrightarrow{s} \Gamma_W(A) \xrightarrow{q_A} /2A \to 0.$$

Moreover, $2\gamma_A(a) = s(a \odot a)$, for all $a \in A$. (Topologically, if $\eta : S^3 \to S^2$ is the Hopf map and $x \in \pi_2(X)$ then $2x \circ \eta = [x,x]$, the Whitehead product in $\pi_3(X)$.) This sequence is short exact if A is torsion free [4, §1.2].

If A and B are abelian groups the inclusions into $A \oplus B$ induce a canonical splitting $\Gamma_W(A \oplus B) \cong \Gamma_W(A) \oplus \Gamma_W(B) \oplus (A \otimes B)$. Since $\Gamma(\mathbb{Z}) \cong \mathbb{Z}$ it follows by a finite induction that if $A \cong \mathbb{Z}^r$ then $\Gamma_W(\mathbb{Z}^r)$ is finitely generated and free, and that s is injective. If A is any free abelian group, every finitely generated subgroup of such a group lies in a finitely generated direct summand, and so $\Gamma_W(A)$ is again free, and s is injective.

A w-hermitian pairing on a finitely generated $\mathbb{Z}[\pi]$-module M is a function $b : M \times M \to \mathbb{Z}[\pi]$ which is linear in the first variable and such that $b(n,m) = \overline{b(m,n)}$, for all $m,n \in M$. The *adjoint homomorphism* $\widetilde{b} : M \to M^\dagger$ is given by $\widetilde{b}(n)(m) = b(m,n)$, for all $m,n \in M$. The pairing b is *nonsingular* if $\widetilde{b}$ is an isomorphism.

Let $Her_w(M)$ be the group of w-hermitian pairings on M. Let $ev_M(m)(n,n') = \overline{n(m)}n'(m)$ for all $m \in M$ and $n,n' \in M^\dagger$. Then $ev_M(m)(n,n')$ is quadratic in m and w-hermitian in n and n' and $ev_M(gm) = w(g)ev_M(m)$ for all $g \in \pi$ and $m \in M$. Hence ev_M determines a homomorphism

$$B_M : \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M) \to Her_w(M^\dagger).$$

Let $M \odot M$ have the diagonal π-action, given by $g(m \odot n) = gm \odot gn$, for all $g \in \pi$ and $m,n \in M$, and let $M \odot_\pi M = \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} (M \odot M)$.

Theorem 1. *Let π be a group, $w : \pi \to \mathbb{Z}^\times$ a homomorphism and M a finitely generated projective $\mathbb{Z}[\pi]$-module. If* $\mathrm{Ker}(w)$ *has no element of order* 2 *then B_M is surjective, while if there is no element $g \in \pi$ of order* 2 *such that $w(g) = -1$ then B_M is injective.*

Proof. Since M is a free abelian group there is a short exact sequence

$$0 \to M \odot M \to \Gamma_W(M) \to M/2M \to 0,$$

and $\Gamma_W(M)$ is free as an abelian group. This is a sequence of $\mathbb{Z}[\pi]$-modules and homomorphisms. Since M is projective, $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} M$ is also free as an abelian group. Hence the sequence

$$0 \to M \odot_\pi M \to \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M) \to \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} M/2M = \mathbb{F}_2 \otimes_{\mathbb{Z}[\pi]} M \to 0$$

is also exact, since $Tor_1^{\mathbb{Z}[\pi]}(\mathbb{Z}^w, M/2M) = \mathrm{Ker}(2.id_{\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} M}) = 0$.

Let $\eta_M : M \to \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M)$ be the composite of γ_M with the reduction from $\Gamma_W(M)$ to $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M)$. Then the composite of η_M with the projection to $\mathbb{F}_2 \otimes_{\mathbb{Z}[\pi]} M$ is the canonical epimorphism. Let $[m \odot n]$ be the image of $m \odot n$ in $M \odot_\pi M$.

Suppose first that M is a free $\mathbb{Z}[\pi]$-module, with basis $e_1, \ldots, e_r$, and let $e_1^*, \ldots, e_r^*$ be the dual basis for $M^\dagger$, defined by $e_i^*(e_i) = 1$ and $e_i^*(e_j) = 0$ if $i \neq j$. Since

$$[m \odot gn] = [g(g^{-1}m \odot n)] = [\bar{g}m \circ n] \quad \text{in } M \odot_\pi M,$$

the typical element of $M \odot_\pi M$ may be expressed in the form $\mu = \Sigma_{i \leq j}(r_{ij}e_i) \odot e_j$. For such an element

$$B_M(\mu)(e_k^*, e_l^*) = r_{kl}, \quad \text{for } k < l,$$

and

$$B_M(\mu)(e_k^*, e_l^*) = r_{kk} + \bar{r}_{kk}, \quad \text{for } k = l.$$

In particular, $B_M(\mu)$ is *even*: if $\varepsilon_2 : \mathbb{Z}[\pi] \to \mathbb{F}_2$ is the composite of the augmentation with reduction *mod* (2) then $\varepsilon_2(B_M(\mu)(n,n)) = 0$ for all $n \in M^\dagger$. If $m \in M$ has non-trivial image in $\mathbb{F}_2 \otimes_{\mathbb{Z}[\pi]} M$ then $\varepsilon_2(e_i^*(m)) \neq 0$ for some $i \leq r$. Hence $B_M(\eta_M(m))$ is not even, and it follows easily that $\mathrm{Ker}(B_M) \leq M \odot_\pi M$. If $B_M(\mu) = 0$, for some $\mu = \Sigma_{i \leq j}(r_{ij}e_i) \odot e_j$, then $r_{kl} = 0$, if $k < l$, and $r_{ii} + \bar{r}_{ii} = 0$, for all i.

If π has no orientation reversing element of order 2 and $B_M(\mu) = 0$, where $\mu = \Sigma_{i \leq j}(r_{ij}e_i) \odot e_j$, then $r_{ii} = \Sigma_{g \in F(i)} a_{ig}(g - \bar{g})$, where $F(i)$ is a finite subset of π, for $1 \leq i \leq r$. Since $((g - \bar{g})e_i) \odot e_i = 0$ it follows easily that $\mu = \Sigma(r_{ii}e_i) \odot e_i = 0$. Hence B_M is injective.

To show that B_M is surjective when $\mathrm{Ker}(w)$ has no element of order 2 it shall suffice to assume that M has rank 1 or 2, since h is determined by the values $h_{ij} = h(e_i^*, e_j^*)$. Let $\varepsilon_w[m, m']$ be the image of $m \odot m'$ in $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M)$. Then

$$B_M(\varepsilon_w[m,m'])(n,n') = \overline{n(m)}n'(m') + \overline{n(m')}n'(m),$$

for all $m, m' \in M$ and $n, n' \in M^\dagger$. Suppose first that M has rank 1. Since $h_{11} = \bar{h}_{11}$ and $\mathrm{Ker}(w)$ has no element of order 2 we may write $h_{11} = 2b + \delta + \Sigma_{g \in F}(g + \bar{g})$, where $b = \bar{b}$, $\delta = 1$ or 0 and F is a finite subset of π. Let

$$\mu = \varepsilon_w[(b + \delta + \Sigma_{g \in F} g)e_1, e_1] + \delta\eta_M(e_1).$$

Then $B_M(\mu)(e_1^*, e_1^*) = h_{11}$. If M has rank 2 and $h_{11} = h_{22} = 0$ let $\mu = \varepsilon_w[h_{12}e_1, e_2]$. Then $B_M(\mu)(e_i^*, e_j^*) = h_{ij}$. In each case $B_M(\mu) = h$, since each side of the equation is a w-hermitian pairing on $M^\dagger$.

Now suppose that M is projective, and that P is a finitely generated projective complement to M, so that $M \oplus P \cong \mathbb{Z}[\pi]^r$ for some $r \geq 0$. The inclusion of M into the direct sum induces a split monomorphism from $\Gamma_W(M)$ to $\Gamma_W(\mathbb{Z}[\pi]^r)$ which is clearly compatible with B_M and $B_{\mathbb{Z}[\pi]^r}$. We may extend an hermitian pairing h on $M^\dagger$ to a pairing h_1 on $M^\dagger \oplus P^\dagger$ by setting $h_1(n,p) = h_1(p',p) = 0$ for all $n \in M^\dagger$ and $p, p' \in P^\dagger$. Clearly $h_1|_{M\times M} = h$ and so this extension determines a split monomorphism from $Her_w(M^\dagger)$ to $Her_w((\mathbb{Z}[\pi]^r)^\dagger)$. If $h_1 = B_{\mathbb{Z}[\pi]^r}(\theta)$ then $h = B_M(\theta_M)$, where θ_M is the image of θ under the homomorphism induced by the projection from $M \oplus P$ onto M. Thus if $B_{\mathbb{Z}[\pi]^r}$ is a monomorphism or an epimorphism so is B_M.

In particular, if π has no 2-torsion then B_M is an isomorphism, for any projective $\mathbb{Z}[\pi]$-module M. The restriction on 2-torsion is necessary, as can be seen by considering the group $G = \mathbb{Z}/2\mathbb{Z} = \langle g \mid g^2 \rangle$ with w trivial and h the pairing on $M = \mathbb{Z}[G]$ determined by $h(m,n) = mg\bar{n}$.

Let E be another left $\mathbb{Z}[\pi]$-module. Then the summand $M \otimes E$ of $\Gamma_W(M \oplus E)$ has the diagonal left $\mathbb{Z}[\pi]$-module structure. Let $d : M \to M^{\dagger\dagger}$ and $t : \mathbb{Z} \otimes_{\mathbb{Z}[\pi]} (M \otimes E) \to Hom_{\mathbb{Z}[\pi]}(M,E)$ be given by $d(m)(\mu) = \overline{\mu(m)}$ and $t(\mu \otimes e)(m) = \mu(m)e$, for all $m \in M$, $\mu \in M^\dagger$ and $e \in E$. If M is finitely generated and projective these functions are isomorphisms (of left $\mathbb{Z}[\pi]$-modules and abelian groups, respectively). Let $\widetilde{B_M}(\gamma)$ be the adjoint of $B_M(1 \otimes \gamma)$, for all $\gamma \in \Gamma_W(M)$.

Lemma 3. *Let M be a finitely generated projective $\mathbb{Z}[\pi]$-module and $\theta : M \to E$ be a $\mathbb{Z}[\pi]$-module homomorphism. Let $d : M \to M^{\dagger\dagger}$ and $t : \mathbb{Z} \otimes_{\mathbb{Z}[\pi]} (M \otimes E) \to Hom_{\mathbb{Z}[\pi]}(M,E)$ be the isomorphisms defined above, and let*

$$\alpha_\theta(m,e) = (m, e + \theta(m)),$$

for all $(m,e) \in M \oplus E$. Then α_θ is an automorphism of $M \oplus E$ and

$$\Gamma_W(\alpha_\theta)(\gamma) - \gamma \equiv (d \otimes 1)^{-1}[(\widetilde{B_M}(\gamma) \otimes 1)(t^{-1}(\theta))] \quad mod \quad \Gamma_W(E),$$

for all $\gamma \in \Gamma_W(M)$.

Proof. The homomorphism α_θ is clearly an automorphism of $M \oplus E$ which restricts to the identity on the summands E and M, and

$$\Gamma_W(\alpha_\theta)(\gamma_{M\oplus E}(m)) = \gamma_{M\oplus E}(m) + \gamma_{M\oplus E}(\theta(m)) + m \otimes \theta(m),$$

for all $m \in M$ [4, 1.2.7].

Let $\beta_m = B_M(1 \otimes \gamma_M(m))$, for $m \in M$. Now the adjoint homomorphism $\widetilde{\beta_m}$ is given by $\widetilde{\beta_m}(\mu) = \overline{\mu(m)}d(m)$. Since t is surjective we have $\theta = t(\Sigma \mu_i \otimes e_i)$, for some $\mu_i \in M^\dagger$ and $e_i \in E$. Then $(\widetilde{\beta_m} \otimes 1)(t^{-1}(\theta)) =$

$$\Sigma \widetilde{\beta_m}(\mu_i) \otimes e_i = \Sigma d(m) \otimes \mu_i(m)e_i = d(m) \otimes \theta(m) = (d \otimes 1)(m \otimes \theta(m)).$$

Since

$$\Gamma_W(\alpha_\theta)(\gamma_{M\oplus E}(m)) - \gamma_{M\oplus E}(m) \equiv (d\otimes 1)^{-1}[(\widetilde{\beta_m}\otimes 1)(t^{-1}(\theta))] \ \ mod \ \ \Gamma_W(E),$$

for all $m \in M$, and since each side is quadratic in m, we have

$$\Gamma_W(\alpha_\theta)(\gamma) - \gamma \equiv (d\otimes 1)^{-1}[(\widetilde{B_M}(\gamma)\otimes 1)(t^{-1}(\theta))] \quad mod \quad \Gamma_W(E),$$

for all $\gamma \in \Gamma_W(M)$.

4 Postnikov stages

Let X be a based, connected cell complex with fundamental group π, and let $p_X : \widetilde{X} \to X$ be its universal covering projection. Let $E_0(X)$ be the group of based homotopy classes of based self-homotopy equivalences of X, and $E_\pi(X)$ be the subgroup which induces the identity on π. If we fix a basepoint for $\widetilde{X}$ over the basepoint of X then there are well-defined Hurewicz homomorphisms

$$hwz_q : \pi_q(X) = \pi_q(\widetilde{X}) \to H_q(\widetilde{X};\mathbb{Z}), \quad \text{for all } q \geq 2.$$

Let $f_{X,k} : X \to P_k(X)$ be the k^{th} stage of the Postnikov tower for X. We may construct $P_k(X)$ by adjoining cells of dimension at least $k+2$ to kill the higher homotopy groups of X. The map $f_{X,k}$ is then given by the inclusion of X into $P_k(X)$, and is a $(k+1)$-connected map. In particular, $P_1(X) \simeq K = K(\pi,1)$ and $c_X = f_{X,1}$ is the classifying map for the fundamental group $\pi = \pi_1(X)$.

If M is a left $\mathbb{Z}[\pi]$-module let $L_\pi(M,n)$ be the *generalized Eilenberg-Mac Lane space* over $K = K(\pi,1)$ realizing the given action of π on M. Thus the classifying map for $L = L_\pi(M,n)$ is a principal $K(M,n)$-fibration with a section $\sigma : K \to L$. The pair (c_L, σ) is an object in the category *ex-K* of spaces over K with sections, and we may view $L_\pi(M,n)$ as the *ex-K* loop space $\overline{\Omega}L_\pi(M,n+1)$ [53], with section σ and projection c_L. Let $\mu : L \times_K L \to L$ be the (fibrewise) loop multiplication. Then $\mu(id_L, \sigma c_L) = \mu(\sigma c_L, id_L) = id_L$ in $[L;L]_K$. Let $\iota_{M,n} \in H^n(L;M)$ be the characteristic element.

Let $[X;Y]_K$ be the set of homotopy classes over $K = K(\pi,1)$ of maps $f : X \to Y$ such that $c_X = c_Y f$. (These may also be considered as π-equivariant homotopy classes of π-equivariant maps from $\widetilde{K}$ to $\widetilde{L}$.) The function $\theta : [X,L]_K \to H^n(X;M)$ given by $\theta(f) = f^*\iota_{M,n}$ is a isomorphism with respect to the addition on $[X,L]_K$ determined by μ. Thus $\theta(id_L) = \iota_{M,n}$, $\theta(\sigma c_X) = 0$ and $\theta(\mu(f,f')) = \theta(f) + \theta(f')$ [2, §V.2].

Let $k_1(X) \in H^3(\pi;\pi_2(X))$ be the first k-invariant which may be defined as the primary obstruction to constructing a left inverse to the classifying map c_X. (It may also be identified with the class in $Ext^3_{\mathbb{Z}[\pi]}(\mathbb{Z},\Pi)$ of the iterated extension

$$0 \to \pi_2(X) \to C_2/\partial C_3 \to C_1 \to C_0 \to \mathbb{Z} \to 0.$$

This was surely known to Eilenberg, Mac Lane and Whitehead, and appears closely related to the Homotopy Addition Theorem [62, Theorem IV.6.1] or [54, Proposition 7.5.3], but it is difficult to find an accessible published proof. See [10, Theorem 12.2.10] or [49].)

Let $f_X = f_{X,2}$ be the second stage of the Postnikov tower for X. The *algebraic* 2*-type* $[\pi, \pi_2(X), k_1(X)]$ and the Postnikov 2-stage determine each other. More precisely, $P_2(X) \simeq P_2(Y)$ if and only if there are isomorphisms $\alpha : \pi \cong \pi_1(Y)$ and $\beta : \pi_2(X) \cong \pi_2(Y)$ such that β is α-semilinear and $\alpha^* k_1(Y) = \beta_\# k_1(X)$ in $H^3(\pi; \pi_2(Y))$. Moreover,

$$k_1(X) = 0 \Leftrightarrow c_{P_2(X)} \text{ has a section } \Leftrightarrow P_2(X) \simeq L_\pi(\pi_2(X), 2).$$

Let $L = L_\pi(M, 2)$. Then $E_\pi(L)$ is the group of units of $[L, L]_K$ with respect to composition. We shall use the following special case of a result of Tsukiyama [56]; we give only the part that we need below.

Lemma 4. *There is an exact sequence*

$$1 \to H^2(\pi; M) \to E_\pi(L) \to Aut(M) \to 1.$$

Proof. Let $\theta : [K, L]_K \to H^2(\pi; M)$ be the isomorphism given by $\theta(s) = s^* \iota_{M,2}$, and let $\theta^{-1}(\phi) = s_\phi$ for $\phi \in H^2(\pi; M)$. Then s_ϕ is a homotopy class of sections of c_L, $s_0 = \sigma$ and $s_{\phi+\psi} = \mu(s_\phi, s_\psi)$, while $\phi = s_\phi^* \iota_{M,2}$. (Recall that $\mu : L \times_K L \to L$ is the fibrewise loop multiplication.)

Let $h_\phi = \mu(s_\phi c_L, id_L)$. Then $c_L h_\phi = c_L$ and so $h_\phi \in [L; L]_K$. Clearly $h_0 = \mu(\sigma c_L, id_L) = id_L$ and $h_\phi^* \iota_{M,2} = \iota_{M,2} + c_L^* \phi \in H^2(L; M)$. We also see that

$$\begin{aligned} h_{\phi+\psi} &= \mu(\mu(s_\phi, s_\psi) c_L, id_L) \\ &= \mu(\mu(s_\phi c_L, s_\psi c_L), id_L) \\ &= \mu(s_\phi c_L, \mu(s_\psi c_L, id_L)) \end{aligned}$$

(by homotopy associativity of μ) and so

$$h_{\phi+\psi} = \mu(s_\phi c_L, h_\psi) = \mu(s_\phi c_L h_\psi, h_\psi) = h_\phi h_\psi.$$

Therefore h_ϕ is a homotopy equivalence for all $\phi \in H^2(\pi; M)$, and $\phi \mapsto h_\phi$ defines a homomorphism from $H^2(\pi; M)$ to $E_\pi(L)$.

The lift of h_ϕ to the universal cover $\widetilde{L}$ is (non-equivariantly) homotopic to the identity, since the lift of c_L is (non-equivariantly) homotopic to a constant map. Therefore h_ϕ acts as the identity on $M = \pi_2(L)$.

The homomorphism $h : \phi \mapsto h_\phi$ is in fact an isomorphism onto the kernel of the action of $E_\pi(L)$ on M [56], and the extension splits: $E_\pi(L)$ is isomorphic to a semidirect product $H^2(\pi; M) \rtimes Aut(M)$ [3, Corollary 8.2.7]. More generally, if $P = P_2(X)$, $\Pi = \pi_2(X)$ and H is the subgroup of $Aut_\pi(\Pi) \rtimes Aut(\pi)$ which fixes $k_1(X) \in H^3(\pi; \Pi)$ then

$$E_0(P) \cong H^2(\pi; \Pi) \rtimes H$$

(see [53, Part II]). Thus if $P = L_\pi(\Pi)$ every automorphism of π lifts to a self-homotopy equivalence of L, and $E_0(L) \cong E_\pi(L) \rtimes Aut(\pi)$.

Let $X^{[k]}$ be the k-skeleton of X, for all $k \geq 0$, and let $\Pi = \pi_2(X)$. The image of $\pi_3(X^{[2]})$ in $\pi_3(X^{[3]})$ is isomorphic to $\Gamma_W(\Pi)$, and the inclusion of the 3-skeleton induces a homomorphism $\iota_X : \Gamma_W(\Pi) \to \pi_3(X)$. The composite of ι_X with the natural map from $\Pi \odot \Pi$ to $\Gamma_W(\Pi)$ is the Whitehead product $[-,-]$, and there is a natural *Whitehead exact sequence* of abelian groups

$$\pi_4(X) \xrightarrow{hwz_4} H_4(\widetilde{X};\mathbb{Z}) \xrightarrow{b_X} \Gamma_W(\Pi) \xrightarrow{\iota_X} \pi_3(X) \xrightarrow{hwz_3} H_3(\widetilde{X};\mathbb{Z}) \to 0,$$

where b_X is the *secondary boundary homomorphism* [61]. (See [4, 2.1.17].) This is an exact sequence of left $\mathbb{Z}[\pi]$-modules, by naturality. (Note also that the Whitehead sequence for $K(\Pi,2)$ gives $H_4(\Pi,2;\mathbb{Z}) \cong \Gamma_W(\Pi)$.)

The homology spectral sequence for $P_3(\widetilde{X})$ as a fibration over $K(\Pi,2)$ with fibre $K(\pi_3(X),3)$ gives an exact sequence

$$0 \to H_4(P_3(\widetilde{X});\mathbb{Z}) \to H_4(\Pi,2;\mathbb{Z}) \xrightarrow{d^2_{4,0}} H_3(\pi_3(X),3;\mathbb{Z}) \to H_3(P_3(\widetilde{X});\mathbb{Z}) \to 0,$$

in which $d^2_{4,0}$ is the homology transgression. Composing $d^2_{4,0}$ with the inverse of the Hurewicz isomorphism hwz_3 for $K(\pi_3(X),3)$ gives the image of the second k-invariant $k_2(\widetilde{X}) \in H^4(\Pi,2;\pi_3(X))$ in $Hom(H_4(\Pi,2;\mathbb{Z}),\pi_3(X))$ under the evaluation homomorphism, by the interpretation of k-invariants in terms of transgression [47]. In fact $d^2_{4,0} = hwz_3\iota_X$ [4, Theorem 2.5.10].

5 PD_4-complexes and intersection pairings

Let X be a based finitely dominated cell complex, with the natural left $\mathbb{Z}[\pi]$-module structure. The equivariant cellular chain complex $C_* = C_*(X;\mathbb{Z}[\pi])$ of $\widetilde{X}$ is a complex of left $\mathbb{Z}[\pi]$-modules, and is $\mathbb{Z}[\pi]$-chain homotopy equivalent to a finitely generated complex of projective modules. Let $B_q \leq Z_q \leq C_q$ be the submodules of q-boundaries and q-cycles, respectively. Let $C^q = Hom_{\mathbb{Z}[\pi]}(C_q,\mathbb{Z}[\pi])$, for all $q \geq 0$, and let $\Pi = H_2(\widetilde{X};\mathbb{Z}) = H_2(C_*)$. Recall that the choice of a basepoint for $\widetilde{X}$ determines an isomorphism $\pi_2(X) \cong \Pi$.

Let $ev : \overline{H^2(X;\mathbb{Z}[\pi])} \to \Pi^\dagger$ be the evaluation homomorphism, given by

$$ev([c])([z]) = [c] \cap [z] = c(z) \quad \forall\, c \in C^2 \text{ and } z \in C_2.$$

This homomorphism sits in the *evaluation* exact sequence

$$0 \to E^2\mathbb{Z} \to \overline{H^2(X;\mathbb{Z}[\pi])} \xrightarrow{ev} \Pi^\dagger \to E^3\mathbb{Z} \to \overline{H^3(X;\mathbb{Z}[\pi])}.$$

(See [34, Lemma 3.3].) If X is a PD_4-complex then $H^3(X;\mathbb{Z}[\pi]) = H_1(\widetilde{X};\mathbb{Z}) = 0$, and the evaluation sequence is a 4-term exact sequence.

We assume henceforth that X is a PD_4-complex, with orientation character $w = w_1(X)$. Let X^+ be the orientable covering space associated to $\pi^+ = \mathrm{Ker}(w)$. The complex X is finitely dominated and is homotopy equivalent to $X_o \cup_\phi e^4$, where X_o is a complex of dimension at most 3 and $\phi \in \pi_3(X_o)$ [60]. In particular, π is finitely presentable. In [37] and [38] cellular decompositions were used to study the homotopy types of PD_4-complexes. Here we shall rely more consistently on the dual Postnikov approach.

Lemma 5. *If π is infinite the homotopy type of X is determined by $P_3(X)$.*

Proof. If X and Y are two such PD_4-complexes and $h : P_3(X) \to P_3(Y)$ is a homotopy equivalence then $hf_{X,3}$ is homotopic to a map $g : X \to Y$. Since π is infinite $H_4(\widetilde{X};\mathbb{Z}) = H_4(\widetilde{Y};\mathbb{Z}) = 0$, by Poincaré duality. Since $\pi_i(g)$ is is an isomorphism for $i \leq 3$ any lift $\tilde{g} : \widetilde{X} \to \widetilde{Y}$ is a homotopy equivalence, by the Hurewicz and Whitehead theorems, and so g is a homotopy equivalence.

In particular, if π is torsion free but not free then $H_3(X;\mathbb{Z}[\pi]) \cong E^1\mathbb{Z}$ is a free $\mathbb{Z}[\pi]$-module, by Lemma 2, and so $\pi_3(X) \cong \Gamma_W(\Pi) \oplus E^1\mathbb{Z}$. Hence the homotopy type of X is determined by π, w, Π and the first two k-invariants.

Let $H = \overline{H^2(X;\mathbb{Z}[\pi])}$. A choice of generator $[X]$ for $H_4(X;\mathbb{Z}^w) \cong \mathbb{Z}$ determines a Poincaré duality isomorphism $D : H \to \Pi$ by $D(u) = u \cap [X]$, for all $u \in H$. Moreover $H^3(X;\mathbb{Z}[\pi]) = 0$. The *cohomology intersection pairing* $\lambda : H \times H \to \mathbb{Z}[\pi]$ is defined by

$$\lambda(u,v) = ev(v)(D(u)) \quad \text{for all } u, v \in H.$$

This pairing is w-hermitian: $\lambda(gu,hv) = g\lambda(u,v)\bar{h}$ and $\lambda(v,u) = \overline{\lambda(u,v)}$ for all $u, v \in H$ and $g, h \in \pi$. If X is a closed 4-manifold this pairing is equivalent under Poincaré duality to the equivariant intersection pairing on Π. (See [51, page 82].) Replacing $[X]$ by $-[X]$ changes the sign of the pairing. Since $\lambda(u,e) = 0$ for all $u \in H$ and $e \in E = E^2\mathbb{Z}$ the pairing λ induces a pairing

$$\lambda_X : H/E \times H/E \to \mathbb{Z}[\pi].$$

The adjoint $\widetilde{\lambda_X}$ is a monomorphism, since $\mathrm{Ker}(ev) = E$. The PD_4-complex X is *strongly minimal* if $\lambda_X = 0$.

The next lemma relates nonsingularity of λ_X, projectivity of Π and H/E and conditions on $E^2\mathbb{Z}$ and $E^3\mathbb{Z}$.

Lemma 6. *Let X be a PD_4-complex with fundamental group π, and let $E = E^2\mathbb{Z}$, $H = \overline{H^2(X;\mathbb{Z}[\pi])}$ and $\Pi = \pi_2(X)$. Then*

1. *$\lambda_X = 0$ if and only if $H = E$, and then $E^3\mathbb{Z} \cong E^\dagger$;*
2. *if λ_X is nonsingular and H/E is a projective $\mathbb{Z}[\pi]$-module then $E^3\mathbb{Z} \cong E^\dagger$;*
3. *if λ_X is nonsingular and $E^\dagger = 0$ then $E^3\mathbb{Z} = 0$;*
4. *if $E^3\mathbb{Z} = 0$ then λ_X is nonsingular;*
5. *if $E^3\mathbb{Z} = 0$ and Π is a projective $\mathbb{Z}[\pi]$-module then $E = 0$;*
6. *if $\pi = G * F(n)$, where $G = *_{i=1}^m G_i$ is the free product of $m \geq 1$ one-ended groups and Π is a projective $\mathbb{Z}[\pi]$-module then $c.d.\pi \leq 4$, with equality if π has one end.*

Proof. Let $p : \Pi \to \Pi/D(E)$ and $q : H \to H/E$ be the canonical epimorphisms. Poincaré duality induces an isomorphism $\gamma : H/E \cong \Pi/D(E)$. It is straightforward to verify that $p^\dagger(\gamma^\dagger)^{-1}\widetilde{\lambda_X q} = ev$, and (1) is clear.

If λ_X is nonsingular then $\widetilde{\lambda_X}$ is an isomorphism, and so $\mathrm{Coker}(p^\dagger) = \mathrm{Coker}(ev)$. If moreover $\Pi/D(E) \cong H/E$ is projective then Π is a direct sum: $\Pi \cong (\Pi/D(E)) \oplus D(E)$. Hence $\Pi^\dagger \cong (\Pi/D(E))^\dagger \oplus E^\dagger$, and so $E^\dagger \cong \mathrm{Coker}(p^\dagger) = E^3\mathbb{Z}$.

If λ_X is nonsingular and $E^\dagger = 0$ then $\widetilde{\lambda_X}$ and $p^\dagger$ are isomorphisms, and so $ev = p^\dagger(\gamma^\dagger)^{-1}\widetilde{\lambda_X q}$ is an epimorphism. Hence $E^3\mathbb{Z} = 0$.

If $E^3\mathbb{Z} = 0$ then $H/E = \Pi^\dagger$ and $ev = q$. Since q is an epimorphism it follows that $p^\dagger(\gamma^\dagger)^{-1}\widetilde{\lambda_X} = id_{\Pi^\dagger}$, and so $p^\dagger$ is an epimorphism. Since $p^\dagger$ is also a monomorphism it is an isomorphism. Therefore $\widetilde{\lambda_X} = \gamma^\dagger(p^\dagger)^{-1}$ is also an isomorphism.

If Π is projective then so is $\Pi^\dagger$. If, moreover, $E^3\mathbb{Z} = 0$ then $H \cong E \oplus \Pi^\dagger$. Hence E is projective, since it is a direct summand of $H \cong \Pi$, and so $E \cong E^{\dagger\dagger} = 0$.

If π is a free product of $m \geq 1$ one-ended groups and n copies of $\mathbb{Z}$ then $E^1\mathbb{Z} \cong \mathbb{Z}[\pi]^{m+n-1}$, by Lemma 2. If, moreover, Π is projective then so are $C_3' = C_3 \oplus \Pi$ and $C_4' = C_4 \oplus E^1\mathbb{Z}$. We may easily extend the differentials of C_* to obtain a projective resolution C_*' of length 4 for $\mathbb{Z}$. Hence $c.d.\pi \leq 4$. If π has one end and Π is projective then $H^4(\pi;\mathbb{Z}[\pi]) = E^4\mathbb{Z} \cong H^4(X;\mathbb{Z}[\pi]) \cong \mathbb{Z}$, by the Universal Coefficient spectral sequence and Poincaré duality, and so $c.d.\pi = 4$.

Parts (3) and (4) together imply that if $E^2\mathbb{Z} = 0$ then λ_X is nonsingular if and only if $E^3\mathbb{Z} = 0$ also. Does this remain the case without any conditions on $E^2\mathbb{Z}$? If Π is projective and λ_X is nonsingular then $\pi \cong \pi_1(Z)$ for some PD_4-complex Z with $\pi_2(Z) = 0$, by Theorem 5 below, and so $E^2\mathbb{Z} = E^3\mathbb{Z} = 0$.

We shall say that a based map $f : X \to Y$ between PD_4-complexes is a *degree-1 map* and write $f_*[X] = \pm[Y]$ if $f^*w_1(Y) = w_1(X) = w$ and the lift of f to a based map of universal covers induces an isomorphism $H_4(X;\mathbb{Z}^w) \cong H_4(Y;\mathbb{Z}^w)$. (Note that if we do not work with based maps the homomorphisms induced by different lifts may differ by sign – see [55] for an investigation of the subtleties involved.) The homomorphism $\pi_1(f)$ is then surjective, and Poincaré duality in X and Y determine *umkehr* homomorphisms $f_! : H_*(Y;\mathbb{Z}[\pi_1(Y)]) \to H_*(X;f^*\mathbb{Z}[\pi_1(Y)])$, which split the homomorphisms induced by f. The umkehr homomorphisms are well-defined up to sign [51, §10.3]. If $f : X \to Z$ is a 2-connected degree-1 map then cap product with $[X]$ induces an isomorphism from the *surgery cokernel* $K^2(f) = \overline{\mathrm{Cok}(H^2(f;\mathbb{Z}[\pi]))}$ to $K_2(f)$, and the induced pairing λ_f on $K^2(f) \times K^2(f)$ is nonsingular [60, Theorem 5.2].

We shall not usually specify a fundamental class $[X]$, and so we shall allow orientation-reversing homotopy equivalences of oriented PD_4-complexes, and isomorphisms of modules with pairings which are isometries after a change of sign. In particular, if Y is a second PD_4-complex we write $\lambda_X \cong \lambda_Y$ if there is an isomorphism $\theta : \pi_1(X) \cong \pi_1(Y)$ such that $w_1(X) = w_1(Y) \circ \theta$ and a $\mathbb{Z}[\pi]$-module isomorphism $\Theta : \pi_2(X) \cong \theta^*\pi_2(Y)$ inducing an isometry of cohomology intersection pairings (after changing the sign of $[Y]$, if necessary).

In [5] it is shown that a PD_4-complex X is determined by its algebraic 2-type (i.e., by $P_2(X)$) together with $w_1(X)$ and $f_{X*}[X]$. (The main step involves showing that if $h : P_2(X) \to P_2(Y)$ is a homotopy equivalence such that $h^*w_1(Y) = w_1(X)$ and $h_*f_{X*}[X] = f_{Y*}[Y]$ (up to sign) then $h = P_2(g)$ for some map $g : X \to Y$ such that $H_4(g;\mathbb{Z}^w)$ is an isomorphism.) Our goal is to show that under suitable conditions X is determined by the more accessible invariants encapsulated in the sextuple $[\pi, w, v_2(X), \Pi, k_1(X), \lambda_X]$. (This is the *quadratic 2-type* of X, as in [27], enhanced by the Wu classes; equivalently, by the Stiefel-Whitney classes.) If $\lambda_X \neq 0$ then λ_X determines w, since $\lambda_X(gu, gv) = w(g)g\lambda_X(u,v)g^{-1}$ for all u, v and g.

The *Wu classes* of a PD_n-complex P are the classes $v_i(P) \in H^i(P;\mathbb{F}_2)$ determined by Poincaré duality from the condition

$$u \smile v_i(P) = Sq^i u, \quad for\ all\ \ u \in H^{n-i}(P;\mathbb{F}_2).$$

If P is a manifold these are equivalent to the tangential Stiefel-Whitney classes, by the "Wu Formula" [54, Theorem 6.10.7]. (Spanier writes V_i for our v_i, and does not use the term Wu class.) In dimension 4 we can be quite explicit; if X is a PD_4-complex then $v_1 = w_1$ is the orientation character, and $v_2 = w_2 + w_1^2$. We choose to use v_2 rather than w_2 since it is the characteristic element for the intersection pairing on $H^2(X;\mathbb{F}_2)$.

It shall be useful to distinguish three "v_2-types" of PD_4-complexes:

I. $v_2(\widetilde{X}) \neq 0$ (i.e., $v_2(X)$ is not in the image of $H^2(\pi;\mathbb{F}_2)$ under c_X^*);
II. $v_2(X) = 0$;
III. $v_2(X) \neq 0$ but $v_2(\widetilde{X}) = 0$ (i.e., $v_2(X)$ is in $c_X^*(H^2(\pi;\mathbb{F}_2)) \setminus \{0\}$).

This trichotomy is due to Kreck, who formulated it in terms of Stiefel-Whitney classes of the stable normal bundle of a closed 4-manifold. The *refined* v_2-type (II and III) is given by the orbit of v_2 in $H^2(\pi;\mathbb{F}_2)$ under the action of automorphisms of π which fix the orientation character.

6 Minimal models

A *model* for a PD_4-complex X is a 2-connected degree-1 map $f : X \to Z$ to a PD_4-complex Z. (We shall also say that Z is a model for X.) The *surgery kernel* $K_2(f) = \mathrm{Ker}(\pi_2(f))$ is a finitely generated projective $\mathbb{Z}[\pi]$-module, and is an orthogonal direct summand of $\pi_2(X)$ with respect to the intersection pairing [60, Theorem 5.2]. If both complexes are finite then $K_2(f)$ is stably free. The PD_4-complex X is *order-minimal* if every such map is a homotopy equivalence, i.e., if X is minimal with respect to the order determined by such maps. It is *strongly minimal* if $\lambda_X = 0$, and is χ*-minimal* if $\chi(X) \leq \chi(Y)$, for Y any PD_4-complex with $(\pi_1(Y), w_1(Y)) \cong (\pi, w)$. We then let $q(\pi, w) = \chi(X)$ be this minimal value. (The definition of "strongly minimal" used here may be broader than the one used in [38], where we said that Z was strongly minimal if $\pi_2(Z)^\dagger = 0$. The two definitions are equivalent if $(E^2\mathbb{Z})^\dagger = 0$.)

Order minimality is the most natural property, and χ-minimality perhaps the one most easily established. It is clear that strongly minimal PD_4-complexes are order-minimal. We shall show that χ-minimality interpolates between these notions, when the L^2-Euler characteristic formula $\chi(X) = \Sigma(-1)^i \beta_i^{(2)}(X)$ applies. (Here $\beta_i^{(2)}(X)$ and $\beta_i^{(2)}(\pi)$ are the ith L^2-Betti numbers of the space X and group π. The book [45] is the definitive reference for L^2 homology; a brief outline is given in Sections 1.9 and 2.2 of [34].)

Theorem 2. *A PD_4-complex X with fundamental group π is strongly minimal if and only if $\beta_2^{(2)}(X) = \beta_2^{(2)}(\pi)$.*

Proof. The module $C^2(X;\mathbb{C}[\pi])$ may be identified with the group of cellular 2-cochains with compact support on $\widetilde{X}$, while the corresponding module $C^2_{(2)}(\widetilde{X})$ of L^2-cochains is the group of square-summable cellular 2-cochains on $\widetilde{X}$. The compactly supported cochains are dense in the square-summable cochains. For each $z \in \pi_2(X)$ the evaluation $ev_z : f \to f(z)$ is continuous as a linear map from $C^2_{(2)}(\widetilde{X})$ to $\mathbb{C}$. (See the proof of [34, Theorem 3.4]. If X is strongly minimal then $ev_z(f) = 0$ for all $f \in C^2(X;\mathbb{C}[\pi])$. Hence $ev_z = 0$ for all $z \in \pi_2(M)$. The L^2 analogue of the evaluation sequence (as in [19, §1.4]) then shows that c_X induces an isomorphism on the unreduced L^2-cohomology modules, and so $\beta_2^{(2)}(X) = \beta_2^{(2)}(\pi)$. The converse is part (3) of [34, Theorem 3.4].

The next two corollaries need a further hypothesis at present.

Corollary 3 *Suppose that either X is finite or π satisfies the Strong Bass Conjecture. Then if X is strongly minimal it is χ-minimal, and if it is χ-minimal it is order minimal.*

Proof. If X is finite or π satisfies the Strong Bass Conjecture we may use the L^2-Euler characteristic formula then $\chi(X) = \beta_2^{(2)}(X) - 2\beta_1^{(2)}(X)$ [20]. Since we may construct a $K(\pi,1)$ complex by adjoining cells of dimension > 2 to X, we have $\beta_2^{(2)}(X) \geq \beta_2^{(2)}(\pi)$, in general. Hence X strongly minimal implies that X is χ-minimal, by the Theorem.

Suppose that $f : X \to Y$ is a 2-connected degree-1 map and $\chi(X) = \chi(Y)$. Then $K_2(f)$ is a finitely generated projective $\mathbb{Z}[\pi]$-module and $\mathbb{Z} \otimes_{\mathbb{Z}[\pi]} K_2(f) = 0$. If X is finite then X is a stably free $\mathbb{Z}[\pi]$-module, so $K_2(f) = 0$, by a result of Kaplansky [52]. This also holds if π satisfies the Weak Bass Conjecture [18]. In either case, f is a homotopy equivalence, and so χ-minimality implies order minimality.

In particular, every sequence of 2-connected degree-1 maps

$$X_1 \to X_2 \to X_3 \to \dots$$

eventually becomes a sequence of homotopy equivalences. If $f : X \to Z$ is a 2-connected degree-1 map and Z is strongly minimal then $\lambda_f = \lambda_X$.

Corollary 4 *Suppose that either X is finite or π satisfies the Strong Bass Conjecture. If $\beta_1^{(2)}(X) = \chi(X) = 0$ then X is strongly minimal.*

Proof. In this case the L^2 Euler characteristic formula gives $\beta_2^{(2)}(X) = 0$. Hence $\beta_2^{(2)}(X) = \beta_2^{(2)}(\pi)$.

Strong minimality has the disadvantage of limited applicability. However, the case of greatest interest to us is when $c.d.\pi \leq 2$. The three notions of minimality are then equivalent, and order minimality is equivalent to strong minimality if and only if $c.d.\pi \leq 2$. (See Theorems 18 and 21 below, and [37] for π a free group.)

If $\pi \cong \mathbb{Z}^r$ and X is χ-minimal then X is order minimal. However, X can only be strongly minimal if $r = 1$, 2 or 4. The 4-torus $\mathbb{R}^4/\mathbb{Z}^4$ is the unique strongly minimal PD_4-complex with fundamental group $\mathbb{Z}^4$, since $E^s\mathbb{Z} = 0$ if $s \leq 3$ for this group. Hence $q(\mathbb{Z}^4) = 0$. Let K be the 2-complex corresponding to the standard presentation of $\mathbb{Z}^4$ with four generators and six relators, and let N be a regular neighbourhood of an embedding of K in $\mathbb{R}^5$. Then $M = \partial N$ is an orientable 4-manifold with $\pi_1(M) \cong \mathbb{Z}^4$ and $\chi(M) = 6$. If a 2-connected degree-1 map $f : M \to Y$ is not a homotopy equivalence then $\chi(Y) < \chi(M)$ and so $\beta_2(Y) < 12$. Since $c_Y^* H^2(\mathbb{Z}^4;\mathbb{Z})$ has rank 6 it follows easily from Poincaré duality in Y that $c_Y^* H^2(\mathbb{Z}^4;\mathbb{Z})$ cannot be self-annihilating with respect to cup product, and so c_Y has nonzero degree. However $c_{M*}[M] = 0$, since c_M factors through N, and so there can be no such map f. Thus M is order-minimal, but not χ-minimal, and not strongly minimal.

If Z is strongly minimal and $\pi \cong G_1 * G_2$ does Z decompose accordingly as a connected sum? If so, the hypothesis that π have one end would not be needed in our consideration later of groups of cohomologicial dimension 2. If M is a closed 4-manifold and $\pi_1(M) \cong G_1 * G_2$ then there is a simply-connected 4-manifold N such that $M\#N \cong P_1\#P_2$, where $\pi_1(P_i) \cong G_i$ for $i = 1, 2$ [34, Theorem 14.10]. If $p_i : P_i \to Z_i$ are strongly minimal models then $p = p_1\#p_2 : M\#N \to Z_1\#Z_2$ is a strongly minimal model for $M\#N$. The image of $\pi_2(N)$ generates a projective direct summand of $\pi_2(M\#N)$ on which the intersection pairing is nonsingular, and so p factors through M, by the construction of Theorem 5 below. Thus M has a strongly minimal model which is a connected sum.

A strongly minimal 4-manifold M must be of type II or III, since $\alpha^* v_2(\widetilde{M})$ is the normal Stiefel-Whitney class $w_2(\nu_\alpha)$, for α an immersion of S^2 in $\widetilde{M}$ with normal bundle ν_α, and so $v_2(\widetilde{M})([\alpha])$ is the *mod*-2 self-intersection number of $[\alpha] \in \pi_2(M)$. Is there a purely homotopy-theoretic argument showing that all strongly minimal PD_4-complexes are of type II or III? (This is so if $c.d.\pi = 2$, by Theorem 20 below.)

Lemma 7. *Let $f : X \to Z$ be a 2-connected degree-1 map of PD_4-complexes with fundamental group π. If X is of type* II *or* III *then so is Z.*

Proof. Since f is 2-connected, $c_X = gc_Z f$, for some self homotopy equivalence g of $K(\pi,1)$. If $v_2(X) = c_X^* V$ for some $V \in H^2(\pi;\mathbb{F}_2)$ then

$$f^*(v_2(Z) \smile \alpha) = f^*(\alpha^2) = v_2(X) \smile f^*\alpha = f^*(c_Z^* g^* V \smile \alpha),$$

for all $\alpha \in H^2(Z;\mathbb{F}_2)$. Hence $v_2(Z) = c_Z^* g^* V$, since $H^4(f;\mathbb{F}_2)$ is an isomorphism.

The converse is false. For instance, the blowup of a ruled surface is of type I, but its minimal models are of type II or III. (See §14 below.)

If X has v_2-type I and $c.d.\pi = 2$ is there a model $f : X \to Z$ with $v_2(Z) = 0$?

Lemma 8. *Let Z be a PD_4-complex with fundamental group π, and let Z_ρ be the covering space associated to a subgroup ρ of finite index in π. Then Z is strongly minimal if and only if Z_ρ is strongly minimal.*

Proof. Let $\Pi = \pi_2(Z)$. Then $\pi_2(Z_\rho) \cong \Pi|_\rho$. Moreover, $H^2(\pi;\mathbb{Z}[\pi])|_\rho \cong H^2(\rho;\mathbb{Z}[\rho])$ and $Hom_{\mathbb{Z}[\pi]}(\Pi,\mathbb{Z}[\pi])|_\rho \cong Hom_{\mathbb{Z}[\rho]}(\Pi|_\rho,\mathbb{Z}[\rho])$, as right $\mathbb{Z}[\rho]$-modules, since $[\pi:\rho]$ is finite. The lemma follows from these observations.

7 Existence of strongly minimal models

In this section we shall obtain a criterion for the existence of a strongly minimal model, as a consequence of the following theorem, which may be thought of as a converse to the 4-dimensional case of Wall's Lemma 2.2 and Theorem 5.2.

Theorem 5. *Let X be a PD_4-complex with fundamental group π. If K is a finitely generated projective direct summand of $H^2(X;\mathbb{Z}[\pi])$ such that λ_X induces a nonsingular pairing on $K \times K$ then there is a PD_4-complex Z and a 2-connected degree-1 map $f : X \to Z$ with $K_2(f) = D(K)$.*

Proof. Suppose first that K is stably free and choose maps $m_i : S^2 \to X$ for $1 \le i \le s$ representing generators of $D(K)$, and such that the kernel of the corresponding epimorphism $m : \mathbb{Z}[\pi]^s \to D(K)$ is free of rank t. Attach s 3-cells to X along the m_i to obtain a cell complex Y with $\pi_1(Y) \cong \pi$, $\pi_2(Y) \cong \Pi/D(K)$ and $H_3(Y;\mathbb{Z}[\pi]) \cong H_3(X;\mathbb{Z}[\pi]) \oplus \mathbb{Z}[\pi]^t$. Since the Hurewicz map is onto in degree 3 for 1-connected spaces (such as $\tilde{Y}$) we may then attach t 4-cells to Y along maps whose Hurewicz images form a basis for $H_3(Y,X;\mathbb{Z}[\pi])$ to obtain a cell complex Z with $\pi_1(Z) \cong \pi$ and $\pi_2(Z) \cong \Pi/D(K)$.

If K is not stably free then $K \oplus F \cong F$, where F is free of countable rank, and we first construct Y by attaching countably many 2- and 3-cells to X, and then attach countably many 4-cells to Y to obtain Z as before.

The inclusion $f : X \to Z$ is 2-connected and $\mathrm{Ker}(H_2(f;\mathbb{Z}[\pi])) = D(K)$. Comparison of the equivariant chain complexes for X and Z shows that $H_i(f;\mathbb{Z}[\pi])$ is an isomorphism for all $i \ne 2$, while $H^j(f;\mathbb{Z}[\pi])$ is an isomorphism for all $j \ne 2$ or 3, and $H^2(f;\mathbb{Z}[\pi])$ is a monomorphism. The connecting homomorphism in the long exact sequence for the cohomology of (Z,X) with coefficients $\mathbb{Z}[\pi]$ induces an isomorphism from the summand $K \le H^2(X;\mathbb{Z}[\pi])$ to $H^3(Z,X;\mathbb{Z}[\pi]) = Hom_{\mathbb{Z}[\pi]}(D(K),\mathbb{Z}[\pi])$. Therefore $H^3(Z;\mathbb{Z}[\pi]) = 0$. Let $[Z] = f_*[X] \in H_4(Z;\mathbb{Z}^w)$. Cap product with $[Z]$ gives isomorphisms $\overline{H^j(Z;\mathbb{Z}[\pi])} \cong H_{4-j}(Z;\mathbb{Z}[\pi])$ for $j \ne 2$, by the projection formula $f_*(f^*\alpha \frown [X]) = \alpha \frown [Z]$. This is also true when $j = 2$, for then $H^2(f;\mathbb{Z}[\pi])$ identifies $H^2(Z;\mathbb{Z}[\pi])$ with the orthogonal complement of K in

$H^2(X;\mathbb{Z}[\pi])$, and $f_*(- \frown [X])$ carries this isomorphically to $H_2(Z;\mathbb{Z}[\pi])$. Therefore Z is a PD_4-complex with fundamental class $[Z]$, f has degree 1 and $K_2(f) = D(K)$.

This construction derives from [38], via [39]. The main theorem of [32] includes a similar result, for X a closed orientable 4-manifold and K a free module.

Corollary 6 *The PD_4-complex X has a strongly minimal model if and only if H/E is a finitely generated projective $\mathbb{Z}[\pi]$-module and λ_X is nonsingular.*

Proof. If $f : X \to Z$ is a 2-connected degree-1 map then $K^2(f) = \mathrm{Cok}(H^2(f;\mathbb{Z}[\pi]))$ is a finitely generated projective direct summand of $H^2(X;\mathbb{Z}[\pi])$ [60, Lemma 2.2]. If Z is strongly minimal the inclusion $E \to \overline{H^2(Z;\mathbb{Z}[\pi])}$ is an isomorphism, and so $H/E \cong \overline{K_2(f)}$. Hence the conditions are necessary. If they hold the construction of Theorem 5 gives a strongly minimal model for X.

The above conditions hold if $\Pi^\dagger$ is a finitely generated projective $\mathbb{Z}[\pi]$-module and $E^3\mathbb{Z} = 0$. In particular, they hold if $c.d.\pi \leq 2$, by an elementary argument using Schanuel's Lemma and duality. (See Theorem 18 below). On the other hand, if $c.d.\pi = 3$ and $\mathbb{Z}$ has a finite projective resolution then no PD_4-complex with fundamental group π is strongly minimal. For if $\lambda_X = 0$ then $E^3\mathbb{Z} \cong (E^2\mathbb{Z})^\dagger$, by Lemma 6, and this condition cannot hold, by the next lemma.

Lemma 9. *Let π be a group such that the augmentation module $\mathbb{Z}$ has a finite projective resolution of length ≤ 3, and let $E = E^2\mathbb{Z}$. If $E^3\mathbb{Z} \cong E^\dagger$ then $c.d.\pi \leq 2$.*

Proof. Let P_* be a projective resolution of $\mathbb{Z}$, of length 3. Then $\partial_3^\dagger : P_2^\dagger \to P_3^\dagger$ is a presentation for $E^3\mathbb{Z}$. Hence $(E^3\mathbb{Z})^\dagger = \mathrm{Ker}(\partial_3^{\dagger\dagger}) = \mathrm{Ker}(\partial_3) = 0$. But then $E^3\mathbb{Z} \cong E^\dagger \cong E^{\dagger\dagger\dagger} = 0$. Hence ∂_3 is a split injection, and so $c.d.\pi \leq 2$.

Surgery on a factor of the 4-torus $\mathbb{R}^4/\mathbb{Z}^4$ gives a closed 4-manifold M with $\pi \cong \mathbb{Z}^3$ and $\chi(M) = 2$. This 4-manifold is χ-minimal [34, Lemma 3.11], and is order minimal, by Corollary 4, but cannot be strongly minimal, since $c.d.\pi = 3$.

The condition $E^3\mathbb{Z} \cong (E^2\mathbb{Z})^\dagger$ is far from characterizing the fundamental groups of strongly minimal PD_4-complexes. In §9–§14 we shall determine such groups within certain subclasses. In all cases considered, π has finitely many ends (i.e., π is virtually cyclic or $E^1\mathbb{Z} = 0$) and $E^3\mathbb{Z} = 0$.

Lemma 10. *Let $f : X \to Z$ be a 2-connected degree-1 map of PD_4-complexes with fundamental group π. Then $k_1(Z) = f_\#(k_1(X))$ and $k_1(X) = f_{!\#}k_1(Z)$, where $f_\#$ and $f_{!\#}$ are the change-of-coefficients homomorphisms induced by $\pi_2(f)$ and the umkehr homomorphism. If $E^3\mathbb{Z} = 0$ then these are mutually inverse isomorphisms.*

Proof. Since $K_2(f)$ is projective, $\pi_2(X) \cong \pi_2(Z) \oplus K_2(f)$, where the projection onto the first factor is given by $\pi_2(f)$ and is split by the umkehr map $f_!$.

Let $q : Q \to Z$ be the pullback of $P_3(f) : P_3(X) \to P_3(Z)$ over the inclusion of Z into $P_3(Z)$. Then q is a fibration with homotopy fibre $K(K_2(f),2)$ and $f = qg$, where $g : X \to Q$ and $P_3(g)$ is a homotopy equivalence. Hence $\pi_2(g)$ is an isomorphism and $k_1(Q) = g_\#k_1(X)$. The fibration q is determined by Z and a k-invariant in $H^3(Z;K_2(f)) \cong H_1(Z;K_2(f))$, which is 0 since $K_2(f)$ is projective.

Hence $k_1(Q) = g_\# f_{!\#} k_1(Z)$. Therefore $k_1(X) = f_{!\#} k_1(Z)$, since $g_\#$ is an isomorphism, and so $f_\# k_1(X) = f_\# f_{!\#} k_1(Z) = k_1(Z)$.

The second assertion follows easily from the fact that $\pi_2(f)$ is an epimorphism with kernel $K_2(f)$ a finitely generated projective direct summand of $\Pi = \pi_2(X)$ and the hypothesis $E^3\mathbb{Z} = 0$, which implies that $H^3(\pi; K_2(f)) = 0$.

In particular, if Z is strongly minimal then $k_1(X)$ derives from $H^3(\pi; E^2\mathbb{Z})$. Are there such examples with $k_1(X) \neq 0$? The simplest examples for testing that we have found are the groups $\pi = A_3^2 *_C A_2^3$, where $A_n = \mathbb{Z}^n * \mathbb{Z}^n$ and C is either trivial or $\mathbb{Z}^4$. These groups have $c.d.\pi = 6$. Mayer-Vietoris arguments show that if $C = 1$ then $E^1\mathbb{Z} \cong E^2\mathbb{Z} \cong E^3\mathbb{Z} \cong \mathbb{Z}[\pi]$, while if $C = \mathbb{Z}^4$ then $E^1\mathbb{Z} = 0$ (i.e., π has one end) and $E^2\mathbb{Z} \cong E^3\mathbb{Z} \cong \mathbb{Z}[\pi]$. In each case it follows that $H^3(\pi; E^2\mathbb{Z}) \cong \mathbb{Z}[\pi]$. These groups are right angled Artin groups. Perhaps the "smallest" such group with similar cohomological properties is the one given by the 1-skeleton of a minimal triangulation of $S^2 \times S^1$, which has 10 generators and 40 relators but is less easily described explicitly. (This group has one end and $c.d. = 4$.)

8 Reduction

The main result of this section implies that when a PD_4-complex X has a strongly minimal model Z its homotopy type is determined by Z and λ_X. Recall the notation $B_M(-)$ from §2.

Lemma 11. *Let $\beta_\xi = B_{\mathbb{Z}^n}(b_{(\mathbb{CP}^\infty)^n}(\xi))$, for $\xi \in H_4((\mathbb{CP}^\infty)^n; \mathbb{Z})$, and let G be a group. Let $u = \Sigma u_g g$ and $v = \Sigma v_h h \in H^2((\mathbb{CP}^\infty)^n; \mathbb{Z}[G]) \cong H^2((\mathbb{CP}^\infty)^n; \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{Z}[G]$. Then*

$$v(u \frown \xi) = \Sigma_{g,h\in G} \beta_\xi(u_g, v_h) g\bar{h}.$$

Proof. As each side of the equation is linear in ξ and $H_4((\mathbb{CP}^\infty)^n; \mathbb{Z})$ is generated by the images of homomorphisms induced by maps from $\mathbb{CP}^\infty$ or $(\mathbb{CP}^\infty)^2$, it suffices to assume $n = 1$ or 2. Since moreover each side of the equation is bilinear in u and v we may reduce to the case $G = 1$. As these functions have integral values and $2(x \otimes y) = (x+y) \otimes (x+y) - x \otimes x - y \otimes y$ in $H_4((\mathbb{CP}^\infty)^2; \mathbb{Z})$, for all $x, y \in \Pi \cong \mathbb{Z}^2$, we may reduce further to the case $n = 1$, which is easy.

Lemma 12. *Let M be a finitely generated projective $\mathbb{Z}[\pi]$-module and $L = L_\pi(M, 2)$. The secondary boundary homomorphism b_L determines an epimorphism b' from $H_4(L; \mathbb{Z}^w)$ to $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M)$ such that*

$$B_M(b'(x))(u, v) = v(u \frown x) \quad \text{for all } u, v \in M^\dagger \text{ and } x \in H_4(L; \mathbb{Z}^w).$$

Proof. The homomorphism from $H_4(L; \mathbb{Z}^w)$ to $H_4(\pi; \mathbb{Z}^w)$ induced by c_L is an epimorphism, since c_L has a section σ. Since $\widetilde{L} \simeq K(M, 2)$ the homomorphism $b_{\widetilde{L}}$ is an isomorphism and $H_3(\widetilde{L}; \mathbb{Z}) = 0$, while since M is projective $H_p(\pi; M) = 0$ for all

$p > 0$. Therefore it follows from the spectral sequence for the universal covering $\widetilde{L} \to L$ that the kernel of the epimorphism induced by c_L is $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} H_4(\widetilde{L};\mathbb{Z})$. Let $b'(x) = (1 \otimes b_{\widetilde{L}})(x - \sigma_* c_{L*}(x))$ for all $x \in H_4(L;\mathbb{Z}^w)$. Then b' is an epimorphism onto $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M)$.

Let $x \in H_4(L;\mathbb{Z}^w)$ and $u, v \in M^\dagger \cong H^2(L;\mathbb{Z}[\pi])$. Since M is the union of its finitely generated free abelian subgroups and homology commutes with direct limits there is an $n > 0$ and a map $k : (\mathbb{CP}^\infty)^n \to \widetilde{L}$ such that $b'(x)$ is the image of $k_*(\xi)$ for some $\xi \in H_4((\mathbb{CP}^\infty)^n;\mathbb{Z})$. Then $B_M(b'(x))(u,v) = ev_M(k_*\xi)(u,v)$.

Suppose that $k^*u = \Sigma u_g g$ and $k^*v = \Sigma v_h h$ in $H^2((\mathbb{CP}^\infty)^n;\mathbb{Z}[\pi])$. Then we have $ev_M(k_*\xi)(u,v) = \Sigma_{g,h\in G} \beta_\xi(u_g, v_h) g\bar{h}$, which is equal to $v(u \frown k_*\xi) = k^*v(k^*u \frown \xi)$, by Lemma 11. Now $x = k_*\xi + \sigma^* u \frown c_{L*}x$ and $u \frown \sigma_* c_{L*} x = \sigma_*(\sigma^* u \frown c_{L*}x) = 0$, since $H_2(\pi;\mathbb{Z}[\pi]) = 0$. Hence $B_M(b'(x))(u,v) = v(u \frown x)$, for all $u, v \in M^\dagger$ and $x \in H_4(L;\mathbb{Z}^w)$.

Theorem 7. *Let $g_X : X \to Z$ and $g_Y : Y \to Z$ be 2-connected degree-1 maps of PD_4-complexes with fundamental group π. If $w = w_1(Z)$ is trivial on elements of order 2 in π then there is a homotopy equivalence $h : X \to Y$ such that $g_Y h = g_X$ if and only if $\lambda_{g_X} \cong \lambda_{g_Y}$ (after changing the sign of $[Y]$, if necessary).*

Proof. The condition $\lambda_{g_X} \cong \lambda_{g_Y}$ is clearly necessary. Suppose that it holds.

Since g_X and g_Y induce isomorphisms on π_1, we may assume that $c_X = c_Z g_X$ and $c_Y = c_Z g_Y$. Since g_X and g_Y are 2-connected degree-1 maps, there are canonical splittings $\pi_2(X) = K_2(g_X) \oplus N$ and $\pi_2(Y) = K_2(g_Y) \oplus N$, where $N = \pi_2(Z)$, and $K_2(g_X)$ and $K_2(g_Y)$ are projective. The projections $\pi_2(g_X)$ and $\pi_2(g_Y)$ onto the second factors are split by the umkehr homomorphisms. We may identify $K_2(g_X)^\dagger$ and $K_2(g_Y)^\dagger$ with direct summands of $H^2(X;\mathbb{Z}[\pi])$ and $H^2(X;\mathbb{Z}[\pi])$, respectively [60, Lemma 2.2]. The homomorphism θ induces an isomorphism $K_2(Y) \cong M = K_2(X)$ such that $\lambda_{g_Y} = \lambda_{g_X}$ as pairings on $M^\dagger \times M^\dagger$. Hence $\pi_2(X) \cong \pi_2(Y) \cong \Pi = M \oplus N$. We may also assume that $M \neq 0$, for otherwise g_X and g_Y are homotopy equivalences.

Let $g : P = P_2(X) \to P_2(Z)$ be the 2-connected map induced by g_X. Then g is a fibration with fibre $K(M,2)$, and the inclusion of N as a direct summand of Π determines a section s for g. Since $\pi_2(X) \cong \pi_2(Y)$, and $k_1(X) = (g_{X!})_\#(k_1(Z))$ and $k_1(Y) = (g_{Y!})_\#(k_1(Z))$, by Lemma 10, we see that $P_2(Y) \simeq P$. We may choose the homotopy equivalence so that composition with g is homotopic to the map induced by g_Y. (This uses our knowledge of $E_\pi(P)$, as recorded in §3 above.)

The splitting $\Pi = M \oplus N$ also determines a projection $q : P \to L = L_\pi(M,2)$. We may construct L by adjoining 3-cells to X to kill the kernel of projection from Π onto M and then adjoining higher dimensional cells to kill the higher homotopy. Let $j : X \to L$ be the inclusion. Then $B_M(b'(j_*[X]))(u,v) = v(u \frown j_*[X])$ for all $u, v \in M^\dagger$, by Lemma 12. Using the projection formula and identifying $M^\dagger = H^2(L;\mathbb{Z}[\pi])$ with $K^2(X)$ we may equate this with $\lambda_{g_X}(u,v)$. Hence $f_{X*}[X]$ and $f_{Y*}[Y]$ have the same image $\lambda_{g_X} = \lambda_{g_Y}$ in $Her_w(M^\dagger)$.

Since $P_2(Z)$ is a retract of P comparison of the Cartan-Leray spectral sequences for the classifying maps c_P and $c_{P_2(Z)}$ shows that

$$\mathrm{Cok}(H_4(s;\mathbb{Z}^w)) \cong \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} (\Gamma_W(\Pi)/\Gamma_W(N)).$$

Since π has no orientation reversing element of order 2 the homomorphism B_M is injective, by Theorem 1, and therefore since $\lambda_{gX} = \lambda_{gY}$ the images of $f_{X*}[X]$ and $f_{Y*}[Y]$ in $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} (\Gamma_W(\Pi)/\Gamma_W(N))$ differ by an element of the subgroup $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} (M \otimes N)$. Let $c \in M \otimes N$ represent this difference, and let $\gamma \in \Gamma_W(M)$ represent $b'(f_{X*}[X])$. Since $B_M(1 \otimes \gamma) = \lambda_{gX}$ is nonsingular $\widetilde{B_M}(\gamma)$ is surjective, and so we may choose a homomorphism $\theta : M \to N$ such that $(\widetilde{B_M}(\gamma) \otimes 1)(t^{-1}(\theta)) = (d \otimes 1)(c)$. Hence $\Gamma_W(\alpha_\theta)(\gamma) - \gamma \equiv c \bmod \Gamma_W(N)$, by Lemma 3. Let $P(\theta)$ be the corresponding self homotopy equivalence of P. Then $gP(\theta) = g$ and $P(\theta)_* f_{Y*}[Y] = f_{X*}[X]$ *mod* $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(N)$. It follows that $P(\theta)_* f_{Y*}[Y] = f_{X*}[X]$ in $H_4(P;\mathbb{Z}^w)$, since $g_{X*}[X] = g_{Y*}[Y]$ in $H_4(Z;\mathbb{Z}^w)$ and so $(gf_X)_*[X] = (gf_Y)_*[Y]$ in $H_4(P_2(Z);\mathbb{Z}^w)$.

There is then a map $h : X \to Y$ with $f_Y h = f_X$, by the argument of [27, Lemma 1.3]. Since the orientation characters of X and Y are compatible, h lifts to a map $h^+ : X^+ \to Y^+$. Since f_X and f_Y are 3-connected $\pi_1(h^+)$, $\pi_2(h^+)$ and $H_2(h^+;\mathbb{Z})$ are isomorphisms. Since M is projective and nonzero, $\mathbb{Z} \otimes_{\mathrm{Ker}(w)} M$ is a nontrivial torsion free direct summand of $H_2(X^+;\mathbb{Z})$, and so h^+ has degree 1, by Poincaré duality. Hence h^+ is a homotopy equivalence, and therefore so is h.

The original version of this result [39, Theorem 11] assumed that $k_1(X) = k_1(Y) = 0$. This was relaxed to the condition that "$k_1(X) = (g_{X!})_\# k_1(Z)$ and $k_1(Y) = (g_{Y!})_\# k_1(Z)$" in an earlier version of the present paper [arXiv: 1303.5486v2]. The final step is due to Hegenbarth, Pamuk and Repovš, who noted that Poincaré duality in Z may be used to establish an equivalent condition [31]. (This observation has been used in the current version of Lemma 10 above.)

The argument for Theorem 7 breaks down when $\pi = \mathbb{Z}/2\mathbb{Z}$ and w is nontrivial, for then $B_M : \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(M) \to Her_w(M^\dagger)$ is no longer injective, and the intersection pairing is no longer a complete invariant [28]. Thus the condition on 2-torsion is in general necessary.

Corollary 8 *If X has a strongly minimal model Z and π has no 2-torsion then the homotopy type of X is determined by Z and λ_X.* □

Corollary 9 *[32] If $g : X \to Z$ is a 2-connected degree-1 map of PD_4-complexes such that $w_1(Z)$ is trivial on elements of order 2 in $\pi_1(Z)$ then X is homotopy equivalent to $M\#Z$ with M 1-connected if and only if λ_g is extended from a nonsingular pairing over $\mathbb{Z}$.* □

The result of [32] assumes that X is orientable, π is infinite and either $E^2\mathbb{Z} = 0$ or π acts trivially on $\pi_2(Z)$. (In the latter case $Hom_{\mathbb{Z}[\pi]}(\pi_2(Z),\mathbb{Z}[\pi]) = 0$, and so Z is strongly minimal.)

9 Realization of pairings

In this short section we shall show that if Z is a strongly minimal PD_4-complex and $\mathrm{Ker}(w)$ has no element of order 2 every nonsingular w-hermitian pairing on a finitely generated projective $\mathbb{Z}[\pi]$-module is realized as λ_X for some PD_4-complex X with minimal model Z. This is an immediate consequence of the following stronger result.

Theorem 10. *Let Z be a PD_4-complex with fundamental group π and let $w = w_1(Z)$. Assume that* $\mathrm{Ker}(w)$ *has no element of order* 2*. Let N be a finitely generated projective $\mathbb{Z}[\pi]$-module and Λ be a nonsingular w-hermitian pairing on $N^\dagger$. Then there is a PD_4-complex X and a* 2*-connected degree-*1 *map $f : X \to Z$ such that $\lambda_f \cong \Lambda$.*

Proof. Suppose $N \oplus F_1 \cong F_2$, where F_1 and F_2 are free $\mathbb{Z}[\pi]$-modules with countable bases I and J, respectively. (These may be assumed finite if N is stably free.) We may assume $Z = Z_o \cup_\theta e^4$ is obtained by attaching a single 4-cell to a 3-complex Z_o [60, Lemma 2.9]. Construct a 3-complex X_o with $\pi_2(X_o) \cong \pi_2(Z_o) \oplus N$ by attaching J 3-cells to $Z_o \vee (\vee^I S^2)$, along sums of translates under π of the 2-spheres in $\vee^I S^2$, as in Theorem 5. Let $i : Z_o \to X_o$ be the natural inclusion. Collapsing $\vee^I S^2$ gives $X_o / \vee^I S^2 \simeq Z_o \vee (\vee^J S^3)$, and so there is a retraction $q : X_o \to Z_o$. Let $p : \Pi = \pi_2(X_o) \to N$ be the projection with kernel $\mathrm{Im}(\pi_2(i))$, and let $j : X_o \to L = L_\pi(N,2)$ be the corresponding map. Then $\pi_2(ji) = 0$ and so ji factors through $K(\pi,1)$. The map $B_N : \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(N) \to Her_w(N^\dagger)$ is an epimorphism, by Theorem 1. Therefore we may choose $\psi \in \pi_3(X_o)$ so that $B_N([j(\psi)]) = \Lambda$.

Let $\phi = \psi - iq\psi + i\theta$. Then $q\phi = \theta$ and $j(\phi) = j(\psi)$, so $B_N([j(\phi)]) = \Lambda$. Let $X = X_o \cup_\phi D^4$. The retraction q extends to a map $f : X \to Z$. Comparison of the exact sequences for these pairs shows that f induces isomorphisms on homology and cohomology in degrees $\neq 2$. In particular, $H_4(X;\mathbb{Z}^w) \cong H_4(Z;\mathbb{Z}^w)$. Let $[X] = f_*^{-1}[Z]$. Then $f_*(f^*(\alpha) \frown [X]) = \alpha \frown [Z]$ for all cohomology classes α on Z, by the projection formula. Therefore cap product with $[X]$ induces the Poincaré duality isomorphisms for Z in degrees other than 2. As it induces an isomorphism $\overline{H^2(X;\mathbb{Z}[\pi])} \cong H_2(X;\mathbb{Z}[\pi])$, by the assumption on Λ, X_ϕ is a PD_4-complex with $\lambda_X \cong \Lambda$.

10 Strongly minimal models with $\pi_2 = 0$

A PD_4-complex Z with $\pi_2(Z) = 0$ is clearly strongly minimal.

Lemma 13. *Let X be a PD_4-complex with fundamental group π. Then*

1. *$\Pi = 0$ if and only if X is strongly minimal and $E^2\mathbb{Z} = 0$, and then $E^3\mathbb{Z} = 0$;*
2. *if $\Pi = 0$ and π is infinite then the homotopy type of X is determined by π, w and $k_2(X) \in H^4(\pi;E^1\mathbb{Z})$.*

Proof. Part (1) follows from part (1) of Lemma 6. If $\Pi = 0$ then $P_2(X) \simeq K(\pi,1)$ and $\pi_3(Z) \cong E^1\mathbb{Z}$, by Poincaré duality. Hence (2) follows from Lemma 5.

Theorem 11. *Let π be a finitely presentable group with no 2-torsion and such that $E^2\mathbb{Z} = E^3\mathbb{Z} = 0$, and let $w : \pi \to \mathbb{Z}^\times$ be a homomorphism. Then two PD_4-complexes X and Y with fundamental group π, $w_1(X) = c_X^* w$, $w_1(Y) = c_Y^* w$ and $\pi_2(X)$ and $\pi_2(Y)$ projective $\mathbb{Z}[\pi]$-modules are homotopy equivalent if and only if*

1. *$c_{X*}[X] = \pm g^* c_{Y*}[Y]$ in $H_4(\pi;\mathbb{Z}^w)$, for some $g \in Aut(\pi)$ with $wg = w$; and*
2. *$\lambda_X \cong \lambda_Y$.*

Proof. The hypotheses imply that X and Y have strongly minimal models Z_X and Z_Y with $\pi_2(Z_X) = \pi_2(Z_Y) = 0$, and hence $P_2(Z_X) \simeq P_2(Z_Y) \simeq K(\pi,1)$. Moreover $H^3(\pi;\pi_2(X)) = H^3(\pi;\pi_2(Y)) = 0$, since $E^3\mathbb{Z} = 0$, and so the result follows by the argument of Theorem 7.

In particular, $Z_X \simeq Z_Y$. If π also has one end then the minimal model is aspherical. See Theorem 15 below.

Connected sums of complexes with $\pi_2 = 0$ again have $\pi_2 = 0$, and the fundamental groups of such connected sums usually have infinitely many ends. (The sole nontrivial exception is $\mathbb{RP}^4\#\mathbb{RP}^4$.) The arguments of [57] can be extended to this situation, to show that if π splits as a free product then Z has a corresponding connected sum decomposition [7]. (In particular, if π is torsion free then its free factors are one-ended or infinite cyclic, and so the summands are either aspherical or copies of $S^3 \times S^1$ or $S^3 \tilde{\times} S^1$, the non-orientable S^3-bundle space over S^1.)

In the next two sections we shall determine the groups π with finitely many ends which are fundamental groups of strongly minimal PD_4-complexes Z with $\pi_2(Z) = 0$. (Little is known about such complexes with π indecomposable and having infinitely many ends. It follows from the results of [15] that the centralizer of any element of finite order is finite or has two ends.)

11 Strongly minimal models with π virtually free

If π is virtually free (in particular, if it is finite or two-ended) then $E^s\mathbb{Z} = 0$ for all $s > 1$, and so a strongly minimal PD_4-complex Z with fundamental group π must have $\pi_2(Z) = 0$, by Lemma 13. Thus if π is finite $\widetilde{Z} \simeq S^4$, and so $Z \simeq S^4$ or $\mathbb{RP}^4$ [34, Lemma 12.1]. Every orientable PD_n-complex admits a degree-1 map to S^n. It is well known that the (oriented) homotopy type of a 1-connected PD_4-complex is determined by its intersection pairing and that every such pairing is realized by some 1-connected topological 4-manifold [24, page 161]. Thus the only finite group we need to consider is $\pi = \mathbb{Z}/2\mathbb{Z}$.

Theorem 12. *Let X be a PD_4-complex with $\pi_1(X) = \mathbb{Z}/2\mathbb{Z}$ and let $w = w_1(X)$. Then $\mathbb{RP}^4$ is a model for X if and only if $w^4 \neq 0$.*

Proof. The condition is clearly necessary. Conversely, we may assume that $X = X_o \cup e^4$ is obtained by attaching a single 4-cell to a 3-complex X_o [60, Lemma 2.9]. The map $c_X : X \to \mathbb{RP}^\infty = K(\mathbb{Z}/2\mathbb{Z},1)$ factors through a map $f : X \to \mathbb{RP}^4$, and $w = f^*w_1(\mathbb{RP}^4)$, since $w \neq 0$. The degree of f is well-defined up to sign, and is odd since $w^4 \neq 0$. We may arrange that f is a degree-1 map, after modifying f on a disc, if necessary. (See [48].)

In particular, $\pi_2(X)$ is projective if and only if $w^4 \neq 0$. Can this be seen directly? The two $\mathbb{RP}^2$-bundles over S^2 provide contrasting examples. If $X = S^2 \times \mathbb{RP}^2$ then $w^3 = 0$ and $\Pi \cong \mathbb{Z} \oplus \mathbb{Z}^w$, which has no nontrivial projective $\mathbb{Z}[\mathbb{Z}/2\mathbb{Z}]$-module summand. Thus $S^2 \times \mathbb{RP}^2$ is order minimal but not strongly minimal. On the other hand, if X is the nontrivial bundle space then $w^4 \neq 0$ and $\Pi \cong \mathbb{Z}[\mathbb{Z}/2\mathbb{Z}]$.

Non-orientable topological 4-manifolds with fundamental group $\mathbb{Z}/2\mathbb{Z}$ are classified up to homeomorphism in [28], and it is shown there that the homotopy types are determined by the Euler characteristic, w^4, the v_2-type and an Arf invariant (for v_2-type III). The authors remark that their methods show that λ_X together with a quadratic enhancement $q : \Pi \to \mathbb{Z}/4\mathbb{Z}$ due to [42] is also a complete invariant for the homotopy type of such a manifold.

If $\pi = \pi_1(Z)$ has two ends and $\pi_2(Z) = 0$ then $\widetilde{Z} \simeq S^3$. Since π has two ends it is an extension of $\mathbb{Z}$ or the infinite dihedral group $D_\infty = \mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}$ by a finite normal subgroup F. Since F acts freely on $\widetilde{Z}$ it has cohomological period dividing 4 and acts trivially on $\pi_3(Z) \cong H_3(Z;\mathbb{Z}[\pi])$, while the action $u : \pi \to \{\pm 1\} = Aut(\pi_3(Z))$ induces the usual action of π/F on $H^4(F;\mathbb{Z})$. The action u and the orientation character $w_1(Z)$ determine each other, and every such group π and action u is realized by some PD_4-complex Z with $\pi_2(Z) = 0$. The homotopy type of Z is determined by π, u and the first nontrivial k-invariant in $H^4(\pi;\mathbb{Z}^u)$. (See [34, Chapter 11].)

We shall use Farrell cohomology to show that any PD_4-complex X with $\pi_1(X) \cong \pi$ satisfying corresponding conditions has a strongly minimal model. We refer to the final chaper of [9] for more information on Farrell cohomology.

It is convenient to use the following notation. If R is a noetherian ring and M is a finitely generated R-module let $\Omega^1 M = \mathrm{Ker}(\phi)$, where $\phi : R^n \to M$ is any epimorphism, and define $\Omega^k M$ for $k > 1$ by iteration, so that $\Omega^{n+1}M = \Omega^1\Omega^n M$. We shall say that two finitely generated R-modules M_1 and M_2 are projectively equivalent ($M_1 \simeq M_2$) if they are isomorphic up to direct sums with a finitely generated projective module. Then these "syzygy modules" $\Omega^k M$ are finitely generated, and are well-defined up to projective equivalence, by Schanuel's Lemma.

Theorem 13. *Let X be a PD_4-complex such that $\pi = \pi_1(X)$ has two ends. Then X has a strongly minimal model if and only if π and the action u of π on $H_3(X;\mathbb{Z}[\pi]) \cong \mathbb{Z}$ are realized by some PD_4-complex Z with $\pi_2(Z) = 0$.*

Proof. If $\pi_2(Z) = 0$ then $\widetilde{Z} \simeq S^3$, by Poincaré duality and the Hurewicz and Whitehead Theorems, and the conditions on π are necessary, by Theorem 11.1 and Lemma 11.3 of [34].

Conversely, since π is virtually infinite cyclic the conditions imply that the Farrell cohomology of π has period dividing 4 [22]. We may assume that the chain complex

C_* for $\widetilde{X}$ is a complex of finitely generated $\mathbb{Z}[\pi]$-modules. Then the modules B_2, Z_2 Z_3 and Π are finitely generated, since $\mathbb{Z}[\pi]$ is noetherian. The chain complex C_* gives rise to four exact sequences:

$$0 \to Z_2 \to C_2 \to C_1 \to C_0 \to \mathbb{Z} \to 0,$$

$$0 \to Z_3 \to C_3 \to B_2 \to 0,$$

$$0 \to B_2 \to Z_2 \to \Pi \to 0$$

and

$$0 \to C_4 \to Z_3 \to \mathbb{Z}^u \to 0.$$

It is clear that $Z_2 \simeq \Omega^3\mathbb{Z}$ and $Z_3 \simeq \Omega^1 B_2$, while $\Omega^1 Z_3 \simeq \Omega^1(\mathbb{Z}^u)$. The standard construction of a resolution of the middle term of a short exact sequence from resolutions of its extremes, applied to the third sequence, gives a projective equivalence $\Omega^1 Z_2 \simeq \Omega^1 B_2 \oplus \Omega^1 \Pi$. The corresponding sequences for a strongly minimal complex with the same group π and action u give an equivalence $\Omega^1(\mathbb{Z}^u) \simeq \Omega^1(\Omega^4\mathbb{Z})$. (This is in turn equivalent to $\Omega^1\mathbb{Z}$, by periodicity.) Together these equivalences give

$$\Omega^5\mathbb{Z} \simeq \Omega^2 Z_2 \simeq \Omega^2 B_2 \oplus \Omega^2\Pi \simeq \Omega^1 Z_3 \oplus \Omega^2\Pi \simeq \Omega^5\mathbb{Z} \oplus \Omega^2\Pi.$$

Hence $Ext^q_{\mathbb{Z}[\pi]}(\Omega^5\mathbb{Z}, N) \cong Ext^q_{\mathbb{Z}[\pi]}(\Omega^5\mathbb{Z}, N) \oplus Ext^q_{\mathbb{Z}[\pi]}(\Omega^2\Pi, N)$, for all $q > v.c.d.\pi = 1$, and any $\mathbb{Z}[\pi]$-module N. If N is finitely generated so is $Ext^q_{\mathbb{Z}[\pi]}(\Omega^1\mathbb{Z}, N)$, and so $Ext^{q+2}_{\mathbb{Z}[\pi]}(\Pi, N) = Ext^q_{\mathbb{Z}[\pi]}(\Omega^2\Pi, N) = 0$, for all $q > 1$. Since Π is finitely generated $Ext^r_{\mathbb{Z}[\pi]}(\Pi, -)$ commutes with direct limits and so is 0, for all $r > 3$. Therefore Π has finite projective dimension [9, Theorem X.5.3]. There is a Universal Coefficient spectral sequence

$$E_2^{pq} = Ext^q_{\mathbb{Z}[\pi]}(H_p(X; \mathbb{Z}[\pi]), \mathbb{Z}[\pi]) \Rightarrow H^{p+q}(X; \mathbb{Z}[\pi]).$$

Here $E_2^{pq} = 0$ unless $p = 0$, 2 or 3, and $E_2^{0q} = E_2^{3q} = 0$ if $q > 1$, since π is virtually infinite cyclic and $\Omega^1(\mathbb{Z}^u) \simeq \Omega^1\mathbb{Z}$. It follows easily from this spectral sequence and Poincaré duality that $Ext^s_{\mathbb{Z}[\pi]}(\Pi, \mathbb{Z}[\pi]) = 0$ for all $s \geq 1$. Since Π also has finite projective dimension it is projective. Hence X has a strongly minimal model, by Theorem 5.

Thus, for instance, an orientable PD_4-complex with fundamental group D_∞ does not have a strongly minimal model.

We shall summarize here the results of [37] on the case when $\pi \cong F(n)$, for some $n \geq 1$. All epimorphisms $w : F(n) \to \mathbb{Z}^\times$ are equivalent up to composition with an automorphism of $F(n)$. The ring $\mathbb{Z}[F(n)]$ is a coherent domain of global dimension 2, for which all projectives are free. There are just two homotopy types of χ-minimal PD_4-complexes Z with $\pi_1(Z) \cong F(n)$, namely $\#^n(S^3 \times S^1)$ (if $w = 0$) and $(S^3 \tilde{\times} S^1)\#(\#^{n-1}(S^3 \times S^1))$ (if $w \neq 0$). (These are strongly minimal, and so the notions of minimality coincide in this case.) If X is any PD_4-complex with $\pi_1(X) \cong F(n)$ then $\pi_2(X)$ is a finitely generated free $\mathbb{Z}[F(n)]$-module, and there is a degree-1 map

from X to the minimal model with compatible w. Every w-hermitian pairing on a finitely generated free $\mathbb{Z}[F(n)]$-module is realizable by some such PD_4-complex, and two such complexes X and Y realizing $(F(n),w)$ are homotopy equivalent if and only if λ_X and λ_Y are isometric.

The key observation is that if X is a PD_4-complex with $\pi_1(X) \cong F(n)$ then its 3-skeleton is standard: if $\beta_2(X) = \beta$ then $X \simeq X_\psi = X_o \cup_\psi e^4$, where $X_o = \vee^n(S^1 \vee S^3) \vee (\vee^\beta S^2)$ and $\psi \in \pi_3(X_o)$. (This is an easy homological argument, relying on Schanuel's Lemma and the theorems of Hurewicz and Whitehead.) The main results then follow on exploring how the group $E(X_o)$ acts on the attaching map ψ. This group is "large" and its action is easily analyzed. Most of these results (excepting for the determination of the minimal models) can also be proven by adapting the arguments of this paper.

Finitely generated virtually free groups provide a potentially broader class of examples. These groups are fundamental groups of finite graphs of finite groups. The arguments of [15] may be adapted to show that if Z is a strongly minimal PD_4-complex such that $\pi = \pi_1(Z)$ is virtually free (so $\pi_2(Z) = 0$) and if π has no dihedral subgroup of order > 2 then it is a free product of groups with two ends [7]. However, not much is known about criteria for 2-connected degree-1 maps to a specific minimal model.

12 Strongly minimal models with π one-ended

We begin this section with a general result on the case when π has one end.

Lemma 14. *Let G be a group. If T is a locally-finite normal subgroup of G then T acts trivially on $H^j(G;\mathbb{Z}[G])$, for all $j \geq 0$.*

Proof. If T is finite then $H^j(G;\mathbb{Z}[G]) \cong H^j(G/T;\mathbb{Z}[G/T]))$, for all j, and the result is clear. Thus we may assume that T and G are infinite. Hence $H^0(G;\mathbb{Z}[G]) = 0$, and T acts trivially. We may write $T = \cup_{n\geq 1} T_n$ as a strictly increasing union of finite subgroups. Then there are short exact sequences [41]

$$0 \to \varprojlim{}^1 H^{s-1}(T_n;\mathbb{Z}[\pi]) \to H^s(T;\mathbb{Z}[\pi]) \to \varprojlim H^s(T_n;\mathbb{Z}[\pi]) \to 0.$$

Hence $H^s(T;\mathbb{Z}[\pi]) = 0$ if $s \neq 1$ and $H^1(T;\mathbb{Z}[\pi]) = \varprojlim{}^1 H^0(T_n;\mathbb{Z}[\pi])$, and so the Lyndon-Hochschild-Serre spectral sequence collapses to give

$$H^j(G;\mathbb{Z}[G]) \cong H^{j-1}(G/T;H^1(T;\mathbb{Z}[G])), \quad \text{for all } j \geq 1.$$

Let $g \in T$. We may assume that $g \in T_n$ for all n, and so g acts trivially on $H^0(T_n;\mathbb{Z}G)$, for all j and n. But then g acts trivially on $\varprojlim{}^1 H^0(T_n;\mathbb{Z}[\pi])$, by the functoriality of the construction. Hence every element of T acts trivially on $H^{j-1}(G/T;H^1(T;\mathbb{Z}[G]))$, for all $j \geq 1$.

Theorem 14. *Let X be an orientable, strongly minimal PD_4-complex. If $\pi = \pi_1(X)$ has one end then π has no non-trivial locally-finite normal subgroup.*

Proof. Suppose that π has a nontrivial locally-finite normal subgroup T. Since π has one end, $H_s(X;\mathbb{Z}[\pi]) = 0$ for $s \neq 0$ or 2. Since X is strongly minimal, $\Pi = H_2(X;\mathbb{Z}[\pi]) \cong \overline{H^2(\pi;\mathbb{Z}[\pi])}$. Hence T acts trivially on Π, since it acts trivially on $H^2(\pi;\mathbb{Z}[\pi])$, by Lemma 14, and X is orientable.

Let $g \in T$ have prime order p, and let $C = \langle g \rangle \cong \mathbb{Z}/p\mathbb{Z}$. Then C acts freely on $\widetilde{X}$, which has homology only in degrees 0 and 2. On considering the homology spectral sequence for the classifying map $c_{\widetilde{X}/C} : \widetilde{X}/C \to K(C,1)$, we see that $H_{i+3}(C;\mathbb{Z}) \cong H_i(C;\Pi)$, for all $i \geq 2$. (See [34, Lemma 2.10].) Since C has cohomological period 2 and acts trivially on Π, there is an exact sequence

$$0 \to \mathbb{Z}/p\mathbb{Z} \to \Pi \to \Pi \to 0.$$

On the other hand, since π is finitely presentable, $\Pi \cong H^2(\pi;\mathbb{Z}[\pi])$ is torsion-free [25, Proposition 13.7.1]. Hence T has no such element g and so π has no such finite normal subgroup.

As an immediate consequence, if X is strongly minimal, but not orientable, and π has one end, then either π has no nontrivial locally-finite normal subgroup or $\pi \cong \pi^+ \times \mathbb{Z}/2\mathbb{Z}^-$, and π^+ has no nontrivial locally-finite normal subgroup.

A finitely presentable group G is a *PD_4-group* if $K(G,1)$ is a PD_4-complex. Such a group has one end and $E^2\mathbb{Z} = 0$, and $K(G,1)$ is clearly strongly minimal. Conversely, if X is a strongly minimal complex, $\pi = \pi_1(X)$ has one end and $E^2\mathbb{Z} = 0$ then X is aspherical. Hence π is a PD_4-group and $K(\pi,1)$ is the unique strongly minimal model. The next theorem gives several equivalent conditions for a PD_4-complex with such a group to have a strongly minimal model.

Theorem 15. *Let X be a PD_4-complex with fundamental group π such that π has one end and $E^2\mathbb{Z} = 0$. Then the following are equivalent:*

1. *X has a strongly minimal model;*
2. *π is a PD_4-group and $\Pi = \pi_2(X)$ is projective:*
3. *π is a PD_4-group, $w_1(X) = c_X^* w_1(\pi)$ and c_X is a degree-1 map;*
4. *π is a PD_4-group, $w_1(X) = c_X^* w_1(\pi)$ and $k_1(X) = 0$.*

Proof. The equivalence $(1) \Leftrightarrow (2)$ follows from Corollary 6.

If Z is strongly minimal and $E^1\mathbb{Z} = E^2\mathbb{Z} = 0$ then $\pi_2(Z) = 0$ and $\pi_3(Z) = E^1\mathbb{Z} = 0$. Hence Z is aspherical, so π is a PD_4-group and $Z \simeq K = K(\pi,1)$. Any 2-connected map $f : X \to K$ is homotopic to c_X (up to composition with a self homotopy equivalence of K). Thus $w_1(X) = c_X^* w_1(\pi)$ and c_X is a degree-1 map. Conversely, if (3) holds then $K = K(\pi,1)$ is the unique strongly minimal PD_4-complex with fundamental group π, and c_X is a 2-connected degree-1 map. Thus $(3) \Leftrightarrow (1)$.

If (2) or (3) holds then $\Pi = \mathrm{Ker}(\pi_2(c_X))$ is projective, Since π is a PD_4-group, $H^3(\pi;M) = 0$ for any projective module M, and so $k_1(X) = 0$. Conversely, if (4) holds the map $c_P : P = P_2(X) \to K$ has a section s, since $k_1(X) = 0$. We may assume

that $K = K_o \cup e^4$ and $X = X_o \cup e^4$, where K_o and X_o are 3-complexes. The restriction $s|_{K_o}$ factors through X_o, by cellular approximation, since $P = X_o \cup \{cells\ of\ dim \geq 4\}$. Thus K_o is a retract of X_o. The map c_X induces a commuting diagram of homomorphisms between the long exact sequences of the pairs (X, X_o) and (K, K_o), with coefficients $\mathbb{Z}[\pi]$. Hence the induced map from $H_4(X, X_0; \mathbb{Z}[\pi])$ to $H_4(K, K_0; \mathbb{Z}[\pi])$ is an isomorphism. The change of coefficients homomorphisms $\varepsilon_{w\#}$ induced by the w-twisted augmentation are epimorphisms. Since the natural maps from $H_4(X; \mathbb{Z}^w)$ to $H_4(X, X_o; \mathbb{Z}^w)$ and from $H_4(K; \mathbb{Z}^w)$ to $H_4 K, K_o; \mathbb{Z}^w)$ are isomorphisms, it follows that c_X has degree 1. Thus $(3) \Leftrightarrow (4)$.

If π has one end and Π is projective then $c.d.\pi = 4$ and $H^4(\pi; \mathbb{Z}[\pi]) \cong \mathbb{Z}$, by part (6) of Lemma 6. Must π be a PD_4-group? This is so if also $E^3\mathbb{Z} = 0$, for then X has a strongly minimal model, by Lemma 6 and Theorem 5, which must be aspherical. If X is strongly minimal and π is virtually an r-dimensional duality group then $r = 1, 2$ or 4, and in the latter case π is a PD_4-group.

The next result now follows from Corollary 20 and Theorem 15.

Corollary 16 *Let X and Y be PD_4-complexes with fundamental group π a PD_4-group, and such that $\pi_2(X)$ and $\pi_2(Y)$ are projective $\mathbb{Z}[\pi]$-modules, $w_1(X) = c_X^* w$ and $w_1(Y) = c_Y^* w$, where $w = w_1(\pi)$. Then X and Y are homotopy equivalent if and only if $\lambda_X \cong \lambda_Y$.* $\square$

This corollary and the equivalence of (3) and (4) in the Theorem are from [12]. (It is assumed there that X and π are orientable.) Theorems 15 and 7 give an alternative proof of the main result of [12], namely that a PD_4-complex X with fundamental group π a PD_4-group and $w_1(X) = w_1(\pi)$ is homotopy equivalent to $M\#K(\pi, 1)$, for some 1-connected PD_4-complex M if and only if $k_1(X) = 0$ and λ_X is extended from a nonsingular pairing over $\mathbb{Z}$.

13 Semidirect products and mapping tori

In this section we shall determine which semidirect products $\nu \rtimes_\alpha \mathbb{Z}$ with ν finitely presentable are fundamental groups of strongly minimal PD_4-complexes.

Theorem 17. *Let ν be a finitely presentable group and let X be a PD_4-complex with fundamental group $\pi \cong \nu \rtimes_\alpha \mathbb{Z}$, for some automorphism α of ν. Then the following are equivalent:*

1. *X is the mapping torus of a self homotopy equivalence of a PD_3-complex N with fundamental group ν;*
2. *X is strongly minimal;*
3. *$\chi(X) = 0$.*

In general, X has a strongly minimal model if and only if $\Pi^\dagger$ is projective.

Proof. Let X_ν be the covering space of X corresponding to ν. Then X_ν is the homotopy fibre of a map from X to S^1 which corresponds to the projection of π onto $\mathbb{Z}$, and $H_q(X_\nu;k) = 0$ for $q > 3$ and all coefficients k. The Lyndon-Hochschild-Serre spectral sequence gives an isomorphism $H^2(\pi;\mathbb{Z}[\pi])|_\nu \cong H^1(\nu;\mathbb{Z}[\nu])$ of right $\mathbb{Z}[\nu]$-modules. Since ν is finitely presentable it is accessible, and hence $H^1(\nu;\mathbb{Z}[\nu])$ is finitely generated as a right $\mathbb{Z}[\nu]$-module. (See Theorems VI.6.3 and IV.7.5 of [16].)

Suppose first that X is the mapping torus of a self homotopy equivalence of a PD_3-complex N. Since $\pi_2(X)|_\nu = \pi_2(N) \cong \overline{H^1(\nu;\mathbb{Z}[\nu])}$ is finitely generated as a left $\mathbb{Z}[\nu]$-module, $Hom_{\mathbb{Z}[\pi]}(\pi_2(X),\mathbb{Z}[\pi]) = 0$, and so X is strongly minimal.

If X is strongly minimal then $\pi_2(X) \cong \overline{H^2(X;\mathbb{Z}[\pi])} = \overline{H^2(\pi;\mathbb{Z}[\pi])}$, and so $\pi_2(X_\nu) = \pi_2(X)|_\nu$ is finitely generated as a left $\mathbb{Z}[\nu]$-module. Since ν is finitely presentable, it follows that $\beta_q(X_\nu;\mathbb{F}_2)$ is finite for $q \leq 2$. Poincaré duality in X gives an isomorphism $H_3(X_\nu;\mathbb{F}_2) \cong H^1(X;\mathbb{F}_2[\pi/\nu]) = \mathbb{F}_2$. Hence $\beta_q(X_\nu;\mathbb{F}_2)$ is finite for all q, and so $\chi(X) = 0$, by a Wang sequence argument applied to the fibration $X_\nu \to X \to S^1$.

If $\chi(X) = 0$ then X is a mapping torus of a self homotopy equivalence of a PD_3-complex N with $\pi_1(N) = \nu$. (See [34, Chapter 4].)

The indecomposable factors G_i of $\nu = *G_i$ are either PD_3-groups or virtually free [15], and in either case $H^2(G_i;\mathbb{Z}[G_i]) = 0$. Therefore $H^2(\nu;\mathbb{Z}[\nu]) = 0$ and so $E^3\mathbb{Z} = 0$. The final assertion now follows from the evaluation sequence, Lemma 6 and Theorem 5.

The condition that ν be the fundamental group of a PD_3-complex is quite restrictive. Mapping tori of self homotopy equivalences of PD_3-complexes are always strongly minimal, but other PD_4-complexes with such groups may be order-minimal but not χ-minimal, and so have no strongly minimal model. (See §5 above for an example with $\pi = \mathbb{Z}^4$ and $\chi(M) = 6$.)

If ν is finite then π has two ends, and if ν has one end then π is a PD_4-group. If ν is torsion free and has two ends it is $\mathbb{Z}$, and so $\pi \cong \mathbb{Z}^2$ or $\mathbb{Z} \rtimes_{-1} \mathbb{Z}$. More generally, when ν is a finitely generated free group $F(n)$ (with $n > 0$) then π has one end and $c.d.\pi = 2$. This broader class of groups is the focus of the rest of this paper.

14 Groups of cohomological dimension 2

When $c.d.\pi = 2$, we may drop the qualification "strongly", by the following theorem. (This is also so if π is a free group. The arguments below may be adapted to the latter case, which is well understood [37].)

Theorem 18. *Let X be a PD_4-complex with $\pi_1(X) \cong \pi$ such that $c.d.\pi = 2$, and let $w = w_1(X)$. Then*

1. *$C_*(X;\mathbb{Z}[\pi])$ is $\mathbb{Z}[\pi]$-chain homotopy equivalent to $D_* \oplus P[2] \oplus D^\dagger_{4-*}$, where D_* is a projective resolution of $\mathbb{Z}$, $P[2]$ is a finitely generated projective module P con-*

centrated in degree 2 *and* $D^{\dagger}_{4-*}$ *is the conjugate dual of* D_**, shifted to terminate in degree* 2*;*

2. $\Pi = \pi_2(X) \cong P \oplus E^2\mathbb{Z}$*;*
3. $\chi(X) \geq 2\chi(\pi)$*, with equality if and only if* $P = 0$*;*
4. $(E^2\mathbb{Z})^{\dagger} = 0$*;*
5. $\pi_3(X) \cong \Gamma_W(\Pi) \oplus E^1\mathbb{Z}$*.*

Moreover, $P_2(X) \simeq L = L_\pi(\Pi, 2)$*, and so the homotopy type of* X *is determined by* π*,* w*,* Π*, and the orbit of* $k_2(X) \in H^4(L; \pi_3(X))$ *under the actions of* $Aut_\pi(\pi_3(X))$ *and* $E_0(L)$*.*

Every nonsingular w*-hermitian pairing on a finitely generated projective* $\mathbb{Z}[\pi]$*-module is realized by some such* PD_4*-complex.*

Proof. Let $C_* = C_*(X; \mathbb{Z}[\pi])$, and let D_* be the chain complex with $D_0 = C_0$, $D_1 = C_1$, $D_2 = \mathrm{Im}(\partial_2^C)$ and $D_q = 0$ for $q > 2$. Then

$$0 \to D_2 \to D_1 \to D_0 \to \mathbb{Z} \to 0$$

is a resolution of the augmentation module. Since $c.d.\pi \leq 2$ and D_0 and D_1 are free modules D_2 is projective, by Schanuel's Lemma. Therefore the epimorphism from C_2 to D_2 splits, and so C_* is a direct sum $C_* \cong D_* \oplus (C/D)_*$. Since X is a PD_4-complex C_* is chain homotopy equivalent to the conjugate dual $C^{\dagger}_{4-*}$. Assertions (1) and (2) follow easily.

On taking homology with simple coefficients $\mathbb{Q}$, we see that $\chi(X) = 2\chi(\pi) + \dim_{\mathbb{Q}} \mathbb{Q} \otimes_\pi P$. Hence $\chi(X) \geq 2\chi(\pi)$. Since π satisfies the Weak Bass conjecture [18] and P is projective $P = 0$ if and only if $\dim_{\mathbb{Q}} \mathbb{Q} \otimes_\pi P = 0$.

Let $\delta : D_2 \to D_1$ be the inclusion. Then $E^2\mathbb{Z} = \mathrm{Cok}(\delta^{\dagger})$ and so $(E^2\mathbb{Z})^{\dagger} = \mathrm{Ker}(\delta^{\dagger\dagger})$. But $\delta^{\dagger\dagger} = \delta$ is injective, and so $(E^2\mathbb{Z})^{\dagger} = 0$.

The indecomposable free factors of π are either one-ended or infinite cyclic, and at least one factor has one end, since $c.d.\pi > 1$. Thus $H_3(\widetilde{X}; \mathbb{Z}) \cong E^1\mathbb{Z}$ is a free $\mathbb{Z}[\pi]$-module, by Lemma 2. Hence $\pi_3(X) \cong \Gamma_W(\Pi) \oplus E^1\mathbb{Z}$.

Since $c.d.\pi = 2$ the first k-invariant of X is trivial, and so $P_2(X) \simeq L = L_\pi(\Pi, 2)$. Hence the next assertion follows from Lemma 5.

The realization result follows from Theorem 10.

It follows immediately from (2), (3) and Theorem 5 that "χ-minimal", "order-minimal" and "strongly minimal" are equivalent, when $c.d.\pi = 2$. We shall henceforth use just "minimal" for such complexes.

It remains unknown whether every finitely presentable group π with $c.d.\pi = 2$ has a finite 2-dimensional $K(\pi, 1)$-complex. We shall write $g.d.\pi = 2$ if this is so.

Corollary 19 *Let* X *and* Y *be* PD_4*-complexes with fundamental group* π *such that* $c.d.\pi = 2$*, and* $w_1(X) = c_X^* w$ *and* $w_1(Y) = c_Y^* w$ *for some homomorphism* $w : \pi \to \mathbb{Z}^\times$*. Then* X *and* Y *are homotopy equivalent if and only if they have the same minimal model* Z *and* $\lambda_X \cong \lambda_Y$*.* □

The minimal model may not be uniquely determined! See §14 below.

Theorem 20. *Let Z be a minimal PD_4-complex with fundamental group π such that $c.d.\pi = 2$, and let $w = w_1(Z)$, $L = L_\pi(E^2\mathbb{Z}, 2)$ and $\pi_3 = \Gamma_W(E^2\mathbb{Z}) \oplus E^1\mathbb{Z}$. Then*

1. *the homotopy type of Z is determined by π, w and the orbit of $k_2(Z) \in H^4(L;\pi_3)$ under the actions of $Aut_\pi(\Gamma_W(E^2\mathbb{Z}) \oplus E^1\mathbb{Z})$ and $E_0(L)$;*
2. *if $\widehat{Z}$ is another such complex then $P_2(\widehat{Z}) \simeq P_2(Z)$ if and only if there is an isomorphism $f : \pi_1(\widehat{Z}) \cong \pi$ such that $w_1(\widehat{Z}) = f^*w$;*
3. *the v_2-type of Z is* II *or* III*, i.e., $v_2(Z) = c_Z^*V$ for some $V \in H^2(\pi;\mathbb{F}_2)$;*
4. *if Z is orientable then it has signature $\sigma(Z) = 0$;*
5. *for every $v \in H^2(\pi;\mathbb{F}_2)$ there is a minimal PD_4-complex Z with $\pi_1(Z) \cong \pi$, $w_1(Z) = c_Z^*w$ and $v_2(Z) = c_Z^*v$.*

Proof. The first assertion follows from Theorem 18, since $P_2(Z) \simeq L$.

If $f : \pi_1(\widehat{Z}) \cong \pi$ is an isomorphism such that $w_1(\widehat{Z}) = f^*w$ then $\pi_2(\widehat{Z}) \cong \Pi$ and so $P_2(\widehat{Z}) \simeq P_2(Z)$. Conversely, $Ext^2_{\mathbb{Z}[\pi]}(\Pi, \mathbb{Z}[\pi]) = \mathbb{Z}^w$, so π and Π determine w.

Let $H = c_Z^*H^2(\pi;\mathbb{F}_2)$. Then $\dim H^2(Z;\mathbb{F}_2) = 2\dim H$, since $\chi(Z) = 2\chi(\pi)$, and $H \cup H = 0$, since $c.d.\pi = 2$. In particular, $v_2(Z) \cup h = h \cup h = 0$ for all $h \in H$. Therefore $v_2(Z) \in H$, by the nonsingularity of Poincaré duality. If Z is orientable a similar argument with coefficients $\mathbb{Q}$ shows that $H^2(Z;\mathbb{Q})$ has a self-orthogonal summand of rank $\beta_2(\pi) = \frac{1}{2}\beta_2(Z)$, and so $\sigma(Z) = 0$.

We may use a finite presentation $\mathscr{P} = \langle X \mid R\rangle$ for π as a pattern for constructing a 5-dimensional handlebody $D^5 \cup_{x\in X} h^1_x \cup_{r\in R} h^2_r \simeq C(\mathscr{P})$, where the 1- and 2-handles are indexed by X and R, respectively, but we refine the construction by taking non-orientable 1-handles for generators x with $w(x) \neq 0$ and using $w_2 = v + w^2$ to twist the framings of the 2-handles corresponding to the relators. Let M be the boundary of the resulting 5-manifold. Then $\pi_1(M) \cong \pi$, $w_1(M) = c_M^*w$ and $v_2(M) = c_M^*v$. Since $E^3\mathbb{Z} = 0$ the pairing λ_M is nonsingular, by part (4) of Lemma 6. Hence M has a strongly minimal model Z, by Corollary 6. Since c_M factors through c_Z via a 2-connected degree-1 map, Z has the required properties.

The argument for realizing v is taken from [29], where it is shown that if $C(\mathscr{P})$ is aspherical then the manifold M is itself minimal.

How does $k_2(X)$ determine $v_2(X)$ (and conversely)? This seems to be a crucial question. We expect that the orbit of the k-invariant is detected by the refined v_2-type, but have only proven this in some cases. (See Theorems 25 and 27 below.)

Since the Postnikov third stage $f_{X,3}$ (defined in §3) is 4-connected, $H^4(f_{X,3};\mathbb{F}_2)$ is injective, and so it is an isomorphism if also $\beta_2(X;\mathbb{F}_2) > 0$, by the nondegeneracy of Poincaré duality. Thus the ring $H^*(X;\mathbb{F}_2)$ and hence $v_2(X)$ should be directly computable from $H^*(P_3(X);\mathbb{F}_2)$.

If X is of v_2-type II or III then any minimal model for X must have compatible v_2-type, by Lemma 7. What happens if $v_2(\widetilde{X}) \neq 0$? Does X have a minimal model Z with $v_2(Z) = 0$? (If π is a PD_2-group then X has minimal models of each type, by Theorem 24 below.)

We show next that the class of groups considered here is the largest for which every PD_4-complex with such a fundamental group has a strongly minimal model.

Theorem 21. *Let π be a finitely presentable group and $w : \pi \to \mathbb{Z}^\times$ be a homomorphism. Then the following are equivalent:*

1. *every PD_4-complex with fundamental group π and orientation character w has a strongly minimal model;*
2. *every order minimal PD_4-complex with fundamental group π and orientation character w is strongly minimal;*
3. *$c.d.\pi \leq 2$.*

Proof. The equivalence $(1) \Leftrightarrow (2)$ is clear.

Suppose that (1) holds, and let K be a finite 2-complex with $\pi_1(K) = \pi$. Then K has a 4-dimensional thickening N which is a handlebody with only 0-, 1- and 2-handles, and with $w_1(N) = c_N^* w$. (Cf. the final paragraph of Theorem 30.) Let $M = D(N)$ be the closed 4-manifold obtained by doubling N, and let $j : N \to M$ be one of the canonical inclusions. Then $(\pi_1(M), w_1(M)) \cong (\pi, w)$, and collapsing the double gives a retraction $r : M \to N$. We may assume that $c_M = c_N r$.

Since N is a retract of $M = D(N)$, we have

$$H^2(M;\mathbb{Z}[\pi]) \cong H^2(N;\mathbb{Z}[\pi]) \oplus H^2(M,N;\mathbb{Z}[\pi]).$$

Let $E = E^2\mathbb{Z}$, and $H = \overline{H^2(M;\mathbb{Z}[\pi])}$. Since $c_M \sim c_N r$, we have

$$H/E \cong (\overline{H^2(N;\mathbb{Z}[\pi])}/E) \oplus \overline{H^2(M,N;\mathbb{Z}[\pi])}.$$

Since M has a strongly minimal model H/E is projective, by Corollary 6. Hence so is the direct summand $\overline{H^2(M,N;\mathbb{Z}[\pi])}$. This summand is $\overline{H^2(M,N;\mathbb{Z}[\pi])} \cong H_2(N;\mathbb{Z}[\pi])$, by Poincaré-Lefshetz duality.

Now $H_2(N;\mathbb{Z}[\pi]) \cong P = H_2(K;\mathbb{Z}[\pi])$, since $K \simeq N$. Hence the augmentation $\mathbb{Z}[\pi]$-module $\mathbb{Z}$ has a projective resolution of length 3, given by $C_*(K;\mathbb{Z}[\pi])$ in degrees ≤ 2 and by the module P in degree 3, with differential ∂_3 given by the natural inclusion of P as the submodule of 2-cycles. Thus $c.d.\pi \leq 3$. We also have $E^3\mathbb{Z} \cong E^\dagger$, since there is a strongly minimal PD_4-complex realizing the pair (π, w). Therefore $c.d.\pi \leq 2$, by Lemma 9.

The converse implication $(3) \Rightarrow (1)$ follows from Theorem 20.

The group π is a PD_2-group if and only if $E^2\mathbb{Z}$ is infinite cyclic [8]. The minimal PD_4-complexes are then the total spaces of S^2-bundles over aspherical closed surfaces [34, Theorem 5.10]. We shall review this case in §14 below.

Otherwise $E^2\mathbb{Z}$ is not finitely generated. If $\pi \cong \nu \rtimes \mathbb{Z}$, with ν finitely presentable, then $\nu \cong F(n)$ for some $n > 0$ and π has one end. Let $S^2 \tilde{\times} S^1$ be the mapping torus of the antipodal map of S^2.

Theorem 22. *Let $\pi = F(n) \rtimes_\alpha \mathbb{Z}$, where $n > 0$, and let $w : \pi \to \mathbb{Z}^\times$ be a homomorphism. Then the minimal PD_4-complexes X with fundamental group π and $w_1(X) = c_X^* w$ are homotopy equivalent to mapping tori, and their homotopy types may be distinguished by their refined v_2-types.*

Proof. A PD_3-complex N with fundamental group $F(n)$ is homotopy equivalent to $\#^n(S^2 \times S^1)$ (if it is orientable) or $\#^n(S^2 \tilde{\times} S^1)$ (otherwise). There is a natural representation of $Aut(F(n))$ by isotopy classes of based homeomorphisms of N, and the group of based self homotopy equivalences $E_0(N)$ is a semidirect product $D \rtimes Aut(F(n))$, where D is generated by Dehn twists about nonseparating 2-spheres. If we identify D with $(\mathbb{Z}/2\mathbb{Z})^n = H^1(F(n);\mathbb{F}_2)$, we then see that $E_0(N) = (\mathbb{Z}/2\mathbb{Z})^n \rtimes Aut(F(n))$, with the natural action of $Aut(F(n))$ [33].

Thus a minimal PD_4-complex X with $\pi_1(X) \cong \pi$ is homotopy equivalent to the mapping torus $M(f)$ of a based self-homeomorphism f of such an N, with $w_1(N) = w|_{F(n)}$, and f has image (d,α) in $E_0(N)$. Let $\delta(f)$ be the image of d in $H^2(\pi;\mathbb{F}_2) = H^1(F(n);\mathbb{F}_2)/(\alpha-1)H^1(F(n);\mathbb{F}_2)$. If g is another based self-homeomorphism of N with image (d',α) and $\delta(g) = \delta(f)$ then $d - d' = (\alpha-1)(e)$ for some $e \in D$. Hence (d,α) and (d',α) are conjugate, and so $M(g) \simeq M(f)$.

All minimal PD_4-complexes X with $\pi_1(X) = \pi$ and $w_1(X) = w$ have the same Postnikov 2-stage $L = P_2(X)$, all have v_2-type II or III, and there is such a PD_4-complex X with $v_2(X) = V$, for every $V \in H^2(\pi;\mathbb{F}_2)$, by Theorem 18 and its corollary. Hence the refined v_2-type is a complete invariant.

If $\beta_1(\pi) > 1$ then N may not be determined by $M(f)$. For instance if $N = S^2 \tilde{\times} S^1$ then $M(id_N) = N \times S^1$ is also the mapping torus of an orientation reversing self homeomorphism of $S^2 \times S^1$. It is a remarkable fact that if $\pi = F(n) \rtimes_\alpha \mathbb{Z}$, $n > 1$ and $\beta_1(\pi) \geq 2$ then π is such a semidirect product for infinitely many distinct values of n [11]. However this does not affect our present considerations.

The refined v_2-type is also a complete invariant of the homotopy type of a minimal PD_4-complex when π is a PD_2-group. This case is treated in §15 below. The argument given there is generalized in Theorem 27 to other 2-dimensional duality groups, subject to a technical algebraic condition. This condition holds if $w = 1$ and π is an ascending HNN extension $\mathbb{Z}*_m$, by Theorem 30, while if m is even there is an unique minimal model, by Corollary 28.

15 Realizing k-invariants

For the rest of this paper we shall assume that π is a finitely presentable, 2-dimensional duality group (i.e., π has one end and $c.d.\pi = 2$). The homotopy type of a minimal PD_4-complex X with $\pi_1(X) = \pi$ is determined by π, w and the orbit of $k_2(X)$ under the actions of $E_0(L)$ and $Aut(\Gamma_W(\Pi))$, by Corollary 26. We would like to find more explicit and accessible invariants that characterize such orbits. We would also like to know which k-invariants give rise to PD_4-complexes. Note first that $H_3(\widetilde{X};\mathbb{Z}) = H_4(\widetilde{X};\mathbb{Z}) = 0$, since π has one end.

Theorem 23. *Let π be a finitely presentable, 2-dimensional duality group, and let $w : \pi \to \mathbb{Z}^\times$ be a homomorphism. Let $\Pi = E^2\mathbb{Z}$ and let $k \in H^4(L;\Gamma_W(\Pi))$. Then*

1. *there is a* 4*-complex* Y *with* $\pi_1(Y) \cong \pi$, $\pi_2(Y) \cong \Pi$, $\pi_3(Y) \cong \Gamma_W(\Pi)$, $k_2(Y) = k$ *and* $H_3(\widetilde{Y};\mathbb{Z}) = H_4(\widetilde{Y};\mathbb{Z}) = 0$ *if and only if the homomorphism determined by* p_L^*k *from* $H_4(K(\Pi,2);\mathbb{Z})$ *to* $\Gamma_W(\Pi)$ *is an isomorphism ;*
2. *any such complex* Y *is finitely dominated, and we may assume that* Y *is a finite complex if* π *is of type* FF*;*
3. $H^2(Y;\mathbb{Z}[\pi]) \cong \Pi$*;*
4. $H_4(Y;\mathbb{Z}^w) \cong \mathbb{Z}$ *and cap product with a generator induces isomorphisms* $H^p(Y;\mathbb{Z}[\pi]) \cong H_{4-p}(Y;\mathbb{Z}[\pi])$*, for* $p \neq 2$.

Proof. If Y is such a 4-complex then p_L^*k is an isomorphism, by the exactness of the Whitehead sequence.

Suppose, conversely, that p_L^*k is an isomorphism. Let $P(k)$ denote the Postnikov 3-stage determined by $k \in H^4(L;\Gamma_W(\Pi))$, and let $P = P(k)^{[4]}$. Let $C_* = C_*(\widetilde{P})$ be the equivariant cellular chain complex for $\widetilde{P}$, and let $B_q \leq Z_q \leq C_q$ be the submodules of q-boundaries and q-cycles, respectively. Clearly $H_1(C_*) = 0$ and $H_2(C_*) \cong \Pi$, while $H_3(C_*) = 0$, since p_L^*k is an isomorphism. Hence there are exact sequences

$$0 \to B_1 \to C_1 \to C_0 \to \mathbb{Z} \to 0,$$

$$0 \to B_3 \to C_3 \to Z_2 \to \Pi \to 0$$

and

$$0 \to H_4(C_*) = Z_4 \to C_4 \to B_3 \to 0.$$

Schanuel's Lemma implies that B_1 is projective, since $c.d.\pi = 2$. Hence $C_2 \cong B_1 \oplus Z_2$ and so Z_2 is also projective. It then follows that B_3 is also projective, and so $C_4 \cong B_3 \oplus Z_4$. Thus $H_4(C_*) = Z_4$ is a projective direct summand of C_4.

After replacing P by $P \vee W$, where W is a wedge of copies of S^4, if necessary, we may assume that $Z_4 = H_4(P;\mathbb{Z}[\pi])$ is free. Since $\Gamma_W(\Pi) \cong \pi_3(P)$ the Hurewicz homomorphism from $\pi_4(P)$ to $H_4(P;\mathbb{Z}[\pi])$ is onto, by the exactness of the Whitehead sequence. We may then attach 5-cells along maps representing a basis for Z_4 to obtain a countable 5-complex Q with 3-skeleton $Q^{[3]} = P(k)^{[3]}$ and with $H_q(\widetilde{Q};\mathbb{Z}) = 0$ for $q \geq 3$. The inclusion of P into $P(k)$ extends to a 4-connected map from Q to $P(k)$.

Let D_* be the finite projective resolution of $\mathbb{Z}$ determined by a finite presentation for π. Dualizing gives a finite projective resolution $E_* = D_{2-*}^{\dagger}$ for $\Pi = E^2\mathbb{Z}$. Then $C_*(\widetilde{Q})$ is chain homotopy equivalent to $D_* \oplus E_*[2]$, which is a finite projective chain complex. It follows from the finiteness conditions of Wall that Q is homotopy equivalent to a finitely dominated complex Y of dimension ≤ 4 [59]. (The splitting reflects the fact that c_Y is a retraction, since $k_1(Y) = 0$.) The homotopy type of Y is uniquely determined by the data, as in Lemma 5.

If π is of type FF then B_1 is stably free, by Schanuel's Lemma. Hence Z_2 is also stably free. Since dualizing a finite free resolution of $\mathbb{Z}$ gives a finite free resolution of $\Pi = E^2\mathbb{Z}$ we see in turn that B_3 must be stably free, and so $C_*(\widetilde{Y})$ is chain homotopy equivalent to a finite free complex. Hence Y is homotopy equivalent to a finite 4-complex [59].

Condition (3) follows immediately from the 4-term evaluation sequence, since $\Pi^\dagger = E^2\mathbb{Z}^\dagger = 0$, by part (4) of Theorem 18.

We see easily that $\overline{H^4(Y;\mathbb{Z}[\pi])} = E^2\Pi \cong \mathbb{Z}$ and $H^4(Y;\mathbb{Z}^w) \cong Ext^2(\Pi;\mathbb{Z}^w) \cong \mathbb{Z}$. The homomorphism $\varepsilon_{w\#} : H^4(Y;\mathbb{Z}[\pi]) \to H^4(Y;\mathbb{Z}^w)$ induced by ε_w is surjective, since Y is 4-dimensional, and therefore is an isomorphism. We also have $H_4(Y;\mathbb{Z}^w) \cong Tor_2(\mathbb{Z}^w;\Pi) \cong \mathbb{Z}^w \otimes_\pi \mathbb{Z}[\pi] \cong \mathbb{Z}$. Let $[Y]$ be a generator of $H_4(Y;\mathbb{Z}^w)$. Then evaluation on $[Y]$ induces an isomorphism from $\overline{H^4(Y;\mathbb{Z}[\pi])}$ to $H_0(Y;\mathbb{Z}[\pi])$. Hence $- \frown [Y]$ induces isomorphisms from $\overline{H^p(Y;\mathbb{Z}[\pi])}$ to $H_{4-p}(Y;\mathbb{Z}[\pi])$ for all $p \neq 2$, since $H^p(Y;\mathbb{Z}[\pi]) = H_{4-p}(Y;\mathbb{Z}[\pi]) = 0$ if $p \neq 2$ or 4.

Since $Hom_{\mathbb{Z}[\pi]}(\overline{H^2(Y;\mathbb{Z}[\pi])}, H_2(Y;\mathbb{Z}[\pi])) \cong Hom_{\mathbb{Z}[\pi]}(E^2\mathbb{Z}, E^2\mathbb{Z})$ and $End(E^2\mathbb{Z}) = \mathbb{Z}$, by Lemma 1, cap product with $[Y]$ in degree 2 is determined by an integer. The 4-complex Y is a PD_4-complex if and only if this integer is ± 1. The obvious question is: what is this integer? Is it always ± 1? The complex C_* is chain homotopy equivalent to its dual, but is the chain homotopy equivalence given by slant product with $[Y]$?

If π is either a semidirect product $F(n) \rtimes \mathbb{Z}$ or the fundamental group of a Haken 3-manifold M then $\widetilde{K}_0(\mathbb{Z}[\pi]) = 0$, i.e., projective $\mathbb{Z}[\pi]$-modules are stably free [58]. (This is not yet known for all torsion free one relator groups.) In such cases finitely dominated complexes are homotopy finite.

16 PD_2-groups

The case of most natural interest is when π is a PD_2-group, i.e., is the fundamental group of an aspherical closed surface F. If Z is the minimal model for such a PD_4-complex X then $\Pi = \pi_2(Z)$ and $\Gamma_W(\Pi)$ are infinite cyclic, and Z is homotopy equivalent to the total space of a S^2-bundle over a closed aspherical surface. (The action $u : \pi \to Aut(\Pi)$ is given by $u(g) = w_1(\pi)(g)w(g)$ for all $g \in \pi$ [34, Lemma 10.3], while the induced action on $\Gamma_W(\Pi)$ is trivial.) There are two minimal models for each pair (π, w), distinguished by their v_2-type. This follows easily from the fact that the inclusion of $O(3)$ into the monoid of self-homotopy equivalences $E(S^2)$ induces a bijection on components and an isomorphism on fundamental groups [34, Lemma 5.9]. It is instructive to consider this case from the point of view of k-invariants also, as we shall extend the argument of this section to other groups in Theorem 27 below. In this case we may take F as an exemplar of $K = K(\pi, 1)$.

Suppose first that π acts trivially on Π. Then $L \simeq K \times \mathbb{CP}^\infty$. Fix generators t, x, η and z for $H^2(\pi;\mathbb{Z})$, Π, $\Gamma_W(\Pi)$ and $H^2(\mathbb{CP}^\infty;\mathbb{Z}) = Hom(\Pi,\mathbb{Z})$, respectively, such that $z(x) = 1$ and $2\eta = [x,x]$. (These groups are all infinite cyclic, but we should be careful to distinguish the generators, as the Whitehead product pairing of Π with itself into $\Gamma_W(\Pi)$ is not the pairing given by multiplication.) Let t, z denote also the generators of $H^2(L;\mathbb{Z})$ induced by the projections to K and $\mathbb{CP}^\infty$, respectively. Then $H^2(\pi;\Pi)$ is generated by $t \otimes x$, while $H^4(L;\Gamma_W(\Pi))$ is generated by $tz \otimes \eta$ and $z^2 \otimes \eta$. (Note that t has order 2 if $w_1(\pi) \neq 0$.)

Lemma 15. *The k-invariant $k_2(S^2)$ generates $H^4(\mathbb{CP}^\infty;\mathbb{Z})$.*

Proof. Let $h:\mathbb{CP}^\infty \to K(\mathbb{Z},4)$ be the fibration with homotopy fibre $P_3(S^2)$ corresponding to $k_2(S^2)$. Since $P_3(S^2)$ may be obtained by adjoining cells of dimension ≥ 5 to S^2 we see that $H^4(P_3(S^2);\mathbb{Z}) = 0$. It follows from the spectral sequence of the fibration that h^* maps $H^4(K(\mathbb{Z},4);\mathbb{Z})$ onto $H^4(\mathbb{CP}^\infty;\mathbb{Z})$, and so $k_2(S^2) = h^*\iota_{\mathbb{Z},4}$ generates $H^4(\mathbb{CP}^\infty;\mathbb{Z})$.

Since $\widetilde{Z} \simeq S^2$, the image of $k_2(Z)$ in $H^4(\widetilde{L};\mathbb{Z}) \cong \mathbb{Z}$ generates this group. Hence $k_2(Z) = \pm(z^2\otimes\eta + mtz\otimes\eta)$ for some $m\in\mathbb{Z}$. The action of $[K,L]_K = [K,\mathbb{CP}^\infty] \cong H^2(\pi;\mathbb{Z})$ on $H^2(L;\mathbb{Z})$ is determined by $t\mapsto t$ and $z\mapsto z+t$, and so its action on $H^4(L;\Gamma_W(\Pi))$ is given by $tz\otimes\eta \mapsto tz\otimes\eta$ and $z^2\otimes\eta \mapsto z^2\otimes\eta + 2tz\otimes\eta$. There are thus two possible $E_0(L)$-orbits of k-invariants, and each is in fact realized by the total space of an S^2-bundle over the surface K.

If the action $u:\pi\to Aut(\Pi)$ is nontrivial these calculations go through essentially unchanged with coefficients $\mathbb{F}_2$ instead of $\mathbb{Z}$. There are again two possible $E_\pi(L)$-orbits of k-invariants, and each is realized by an S^2-bundle space.

In all cases the orbits of k-invariants correspond to the elements of $H^2(\pi;\mathbb{F}_2) = \mathbb{Z}/2\mathbb{Z}$. In fact the k-invariant may be detected by the Wu class. Let $[c]_2$ denote the image of a cohomology class under reduction *mod* (2). Since $k_2(Z)$ has image 0 in $H^4(Z;\Pi)$ it follows that $[z]_2^2 \equiv m[tz]_2$ in $H^4(Z;\mathbb{F}_2)$. This holds also if the PD_2-group π is non-orientable (i.e., the surface F is non-orientable) or the action u is nontrivial, and so $v_2(Z) = m[z]_2$ and the orbit of $k_2(Z)$ determine each other.

If X is not minimal and $v_2(\widetilde{X}) \neq 0$ then the minimal model Z is not uniquely determined by X. Nevertheless we have the following results.

Theorem 24. *Let E be the total space of an S^2-bundle over an aspherical closed surface F, and let X be a PD_4-complex with fundamental group $\pi\cong\pi_1(F)$. Let τ be the image of the generator of $H^2(\pi;\mathbb{F}_2)$ in $H^2(X;\mathbb{F}_2)$. Then there is a 2-connected degree-1 map $h:X\to E$ such that $c_E = c_X h$ if and only if*

1. *$(c_X^*)^{-1}w_1(X) = (c_E^*)^{-1}w_1(E)$; and*
2. *$\xi\smile\tau\neq 0$ for some $\xi\in H^2(X;\mathbb{F}_2)$ such that $\xi^2 = 0$ if $v_2(E) = 0$ and $\xi^2\neq 0$ if $v_2(E)\neq 0$.*

Proof. See Theorem 10.17 of the current version of [34].

This is consistent with Lemma 7, for if $v_2(X) = 0$ then $\xi^2 = 0$ and $v_2(E) = 0$, while if $v_2(X) = \tau$ then $\xi^2\neq 0$, and thus $v_2(E)\neq 0$ also.

If $w_1(X) = c_X^*w$, where $w = w_1(\pi)$, and $v_2(X) = 0$ then E must be $F\times S^2$, and we may construct a degree-1 map as follows. Let Ω generate $H^2(\pi;\mathbb{Z}^w)$. We may choose $y\in H^2(X;\mathbb{Z})$ so that $(y\smile c_X^*\Omega)\cap[X] = 1$, by Poincaré duality for X. Then $[y]_2^2 = 0$, since $v_2(X) = 0$. Therefore if F is non-orientable $y^2 = 0$ in $H^4(X;\mathbb{Z}) = \mathbb{Z}/2\mathbb{Z}$; if F is orientable then $y^2 = 2k(y\smile c_X^*\Omega)$ for some k, and we may replace y by $y' = y - kc_X^*\Omega$ to obtain a class with square 0. Such a class may be realized by a map $d:X\to S^2$ [54, Theorem 8.4.11], and we may set $h = (c_X,d):X\to F\times S^2$.

If $v_2(X) \neq 0$ or τ then there is a $\xi \in H^2(X;\mathbb{F}_2)$ such that $\xi \smile \tau \neq 0$ but $\xi^2 = 0$. There is also a class ζ such that $\zeta \smile (\tau - v_2(X)) = 0$ but $\zeta \smile \tau \neq 0$. Hence $\zeta^2 = \zeta \smile \tau \neq 0$. Thus X has minimal models of each v_2-type.

In particular, if C is a smooth projective complex curve of genus ≥ 1 and $X = (C \times \mathbb{CP}^1)\#\overline{\mathbb{CP}^2}$ is a blowup of the ruled surface $C \times \mathbb{CP}^1 = C \times S^2$ then each of the two orientable S^2-bundles over C is a minimal model for X. In this case they are also minimal models in the sense of complex surface theory. (See [1, Chapter VI].) Many of the other minimal complex surfaces in the Enriques-Kodaira classification are aspherical, and hence strongly minimal in our sense. However 1-connected complex surfaces are never minimal in our sense, since S^4 is the unique minimal 1-connected PD_4-complex and S^4 has no complex structure, by a classical result of Wu [1, Proposition IV.7.3].

Theorem 25. *The homotopy type of a PD_4-complex X with fundamental group π a PD_2-group is determined by π, $w_1(X)$, λ_X and the v_2-type.*

Proof. Let $v = w_1(\pi)$, $u = w_1(X) + c_X^* v$, and let Ω generate $H^2(\pi;\mathbb{Z}^v)$. Then $[\Omega]_2$ generates $H^2(\pi;\mathbb{F}_2)$, and $\tau = c_X^*[\Omega]_2 \neq 0$. If $v_2(X) = m\tau$ and $p : X \to Z$ is a 2-connected degree-1 map then $v_2(Z) = mc_Z^*[\Omega]_2$, and so there is an unique minimal model for X. Otherwise $\tau \neq v_2(X)$, and so there are elements $y, z \in H^2(X;\mathbb{F}_2)$ such that $y \smile \tau \neq y^2$ and $z \smile \tau \neq 0$. If $y \smile \tau = 0$ and $z^2 \neq 0$ then $(y+z) \smile \tau \neq 0$ and $(y+z)^2 = 0$. Taking $\xi = y, z$ or $y+z$ appropriately, we have $\xi \smile \tau \neq 0$ and $\xi^2 = 0$. Hence X has a minimal model Z with $v_2(Z) = 0$, by Theorem 24. In all cases the theorem now follows from Theorem 7.

If Z is strongly minimal and $E^2\mathbb{Z}$ is finitely generated but not 0 then $E^2\mathbb{Z}$ is infinite cyclic [8] and the kernel κ of the natural action of π on $\pi_2(Z) \cong \mathbb{Z}$ is a PD_2-group [34, Theorem 10.1]. Thus π is either a PD_2-group or a semidirect product $\kappa \rtimes (\mathbb{Z}/2\mathbb{Z})$. (In particular, π has one end).

17 Cup products

In Theorem 27 below we shall use a "cup-product" argument to relate cohomology in degrees 2 and 4. Let G be a group and let $\Gamma = \mathbb{Z}[G]$. Let C_* and D_* be chain complexes of left Γ-modules and $\mathscr{A}$ and $\mathscr{B}$ left Γ-modules. Using the diagonal homomorphism from G to $G \times G$ we may define *internal products*

$$H^p(Hom_\Gamma(C_*, \mathscr{A})) \otimes H^q(Hom_\Gamma(D_*, \mathscr{B})) \to H^{p+q}(Hom_\Gamma(C_* \otimes D_*, \mathscr{A} \otimes \mathscr{B}))$$

where the tensor products of Γ-modules taken over $\mathbb{Z}$ have the diagonal G-action. (See [14, Chapter XI.§4].) If C_* and D_* are resolutions of $\mathscr{C}$ and $\mathscr{D}$, respectively, we get pairings

$$Ext_\Gamma^p(\mathscr{C}, \mathscr{A}) \otimes Ext_\Gamma^q(\mathscr{D}, \mathscr{B}) \to Ext_\Gamma^{p+q}(\mathscr{C} \otimes \mathscr{D}, \mathscr{A} \otimes \mathscr{B}).$$

When $\mathscr{A}=\mathscr{B}=\mathscr{D}$, $\mathscr{C}=\mathbb{Z}$ and $q=0$ we get pairings

$$H^p(G;\mathscr{A})\otimes End_G(\mathscr{A})\to Ext^p_{\mathbb{Z}[G]}(\mathscr{A},\mathscr{A}\otimes\mathscr{A}).$$

If instead $C_*=D_*=C_*(\widetilde{S})$ for some space S with $\pi_1(S)\cong G$ composing with an equivariant diagonal approximation gives pairings

$$H^p(S;\mathscr{A})\otimes H^q(S;\mathscr{B})\to H^{p+q}(S;\mathscr{A}\otimes\mathscr{B}).$$

These pairings are compatible with the Universal Coefficient spectral sequences $Ext^q_\Gamma(H_p(C_*),\mathscr{A})\Rightarrow H^{p+q}(C^*;\mathscr{A})=H^{p+q}(Hom_\Gamma(C_*,\mathscr{A}))$, etc. We shall call these pairings "cup products", and use the symbol $\smile$ to express their values.

We wish to show that if π is a finitely presentable, 2-dimensional duality group then cup product with id_Π gives an isomorphism

$$c^2_{\pi,w}:H^2(\pi;\Pi)\to Ext^2_{\mathbb{Z}[\pi]}(\Pi,\Pi\otimes\Pi).$$

The next lemma shows that these groups are isomorphic; we state it in greater generality than we need, in order to clarify the hypotheses on the group.

Lemma 16. *Let G be a group for which the augmentation (left) module $\mathbb{Z}$ has a finite projective resolution P_* of length n, and such that $H^j(G;\Gamma)=0$ for $j<n$. Let $\mathscr{D}=H^n(G;\Gamma)$, $w:G\to\mathbb{Z}^\times$ be a homomorphism and $\mathscr{A}$ be a left Γ-module. Then there are natural isomorphisms*

1. $\alpha_\mathscr{A}:\mathscr{D}\otimes_\Gamma\mathscr{A}\to H^n(G;\mathscr{A})$; *and*
2. $e_\mathscr{A}:Ext^n_\Gamma(\overline{\mathscr{D}},\mathscr{A})\to\mathbb{Z}^w\otimes_\Gamma\mathscr{A}=\mathscr{A}/I_w\mathscr{A}$.

Hence $\theta_\mathscr{A}=\alpha_\mathscr{A}e_{\overline{\mathscr{D}}\otimes\mathscr{A}}:Ext^n_\Gamma(\overline{\mathscr{D}},\overline{\mathscr{D}}\otimes\mathscr{A})\to H^n(G;\mathscr{A})$ is an isomorphism.

Proof. If P is a finitely generated projective left Γ-module then $Q=Hom_\Gamma(P,\Gamma)$ is a finitely generated *right* module. There is a natural isomorphism $P\cong Hom_\Gamma(\overline{Q},\Gamma)$, given by $p\mapsto(:f\mapsto\overline{f(p)})$, for all $p\in P$ and $f\in\overline{Q}$. There are also bifunctorial natural isomorphisms of abelian groups $A_{P\mathscr{A}}:Hom_\Gamma(P,\Gamma)\otimes_\Gamma\mathscr{A}\to Hom_\Gamma(P,\mathscr{A})$ given by $A_{P\mathscr{A}}(q\otimes_\Gamma a)(p)=q(p)a$ for all $a\in\mathscr{A}$, $p\in P$ and $q\in Hom_\Gamma(P,\Gamma)$.

We may assume that $P_0=\Gamma$. Let $Q_j=Hom_\Gamma(P_{n-j},\Gamma)$ and $\partial^Q_i=Hom_\Gamma(\partial^P_{n-j},\Gamma)$. This gives a resolution Q_* for $\mathscr{D}$ with $Q_n=\Gamma$. The isomorphisms $A_{P_*\mathscr{A}}$ and $A_{\overline{Q_*}\mathscr{A}}$ induce isomorphisms of chain complexes $Q_*\otimes_\Gamma\mathscr{A}\to Hom_\Gamma(P_{n-*},\mathscr{A})$, and $\overline{P_*}\otimes_\Gamma\mathscr{A}\to Hom_\Gamma(\overline{Q_{n-*}},\mathscr{A})$, respectively, from which the first two isomorphisms follow. The final assertion follows since $\mathbb{Z}^w\otimes_\Gamma(\overline{\mathscr{D}}\otimes\mathscr{A})\cong\mathscr{D}\otimes_\Gamma\mathscr{A}$.

If G is finitely presentable, has one end and $n=2$ then G is a 2-dimensional duality group. It is not known whether all the groups considered in the lemma are duality groups.

Lemma 17. *If G satisfies the hypotheses of Lemma 16 and H is a subgroup of finite index in G then cup product with $id_{\overline{\mathscr{D}}}$ is an isomorphism for (G,w) if and only if it is so for $(H,w|_H)$.*

Proof. If $\mathscr{A}$ is a left $\mathbb{Z}[G]$-module then $H^n(G;\mathscr{A}) \cong H^n(H;\mathscr{A}|_H)$, by Shapiro's Lemma. Thus if G satisfies the hypotheses of Lemma 16 the corresponding module for H is $\overline{\mathscr{D}}|_H$. Further applications of Shapiro's Lemma then give the result.

In particular, it shall suffice to consider the orientable cases.

Let $\eta : Q_0 \to \mathscr{D}$ be the canonical epimorphism, and let $[\xi] \in H^n(G;\overline{\mathscr{D}})$ be the image of $\xi \in Hom_\Gamma(P_n, \overline{\mathscr{D}})$. Then $\xi \otimes \eta : P_n \otimes \overline{Q_0} \to \overline{\mathscr{D}} \otimes \overline{\mathscr{D}}$ represents $[\xi] \smile id_{\mathscr{D}}$ in $Ext^n_\Gamma(\overline{\mathscr{D}}, \overline{\mathscr{D}} \otimes \overline{\mathscr{D}})$. If $\xi = A_{P_n \overline{\mathscr{D}}}(q \otimes_\Gamma \delta)$ then $\alpha_{\overline{\mathscr{D}}}(\eta(q) \otimes_\Gamma \delta) = [\xi]$. There is a chain homotopy equivalence $j_* : \overline{Q}_* \to P_* \otimes \overline{Q}_*$, since P_* is a resolution of $\mathbb{Z}$. Given such a chain homotopy equivalence, $e_{\overline{\mathscr{D}} \otimes \overline{\mathscr{D}}}([\xi] \smile id_{\overline{\mathscr{D}}})$ is the image of $(\xi \otimes \eta)(j_n(1^*))$, where 1^* is the canonical generator of $\overline{Q}_n$, defined by $1^*(1) = 1$.

Theorem 26. *Let G be a finitely presentable, 2-dimensional duality group, and let $w : G \to \mathbb{Z}^\times$ be a homomorphism. Then $c^2_{G,w}$ is an isomorphism.*

Proof. Note first that G satisfies the hypothesis of Lemma 16, with $n = 2$. Let $\mathscr{P} = \langle X \mid R \rangle^\varphi$ be a finite presentation for G. (We shall suppress the defining epimorphism $\varphi : F(X) \to G$ where possible.) After introducing new generators x' and relators $x'x$, if necessary, we may assume that each relator is a product of distinct generators, with all the exponents positive. The new presentation $\mathscr{P}'$ has the same deficiency as $\mathscr{P}$. We may also assume that $w = 1$, after replacing G by $H = \mathrm{Ker}(w)$ if necessary, by Lemma 17.

The Fox-Lyndon resolution associated to $\mathscr{P}$ gives an exact sequence

$$0 \to P_3 = \pi_2(C(\mathscr{P})) \to P_2 \to P_1 \to P_0 = \Gamma \to \mathbb{Z} \to 0$$

in which P_1 and P_2 are free left Γ-modules with bases $\langle p^1_x; x \in X \rangle$ and $\langle p^2_r; r \in R \rangle$, respectively. The differentials are given by $\partial p^1_x = x - 1$ and $\partial p^2_r = \Sigma_{x \in X} r_x p^1_x$, where $r_x = \frac{\partial r}{\partial x}$, for $r \in R$ and $x \in X$. Moreover, P_3 is projective and ∂_3 is a split monomorphism, since $c.d.G = 2$.

Suppose first that the 2-complex $C(\mathscr{P})$ associated to the presentation is aspherical. (This assumption is not affected by our normalization of the presentations, for if $C(\mathscr{P})$ is aspherical then G is efficient, and $\chi(C(\mathscr{P}')) = def(\mathscr{P}') = \chi(C(\mathscr{P}))$. Hence $C(\mathscr{P}')$ is also aspherical [34, Theorem 2.8]. Then $P_3 = 0$ and the above sequence is a free resolution of $\mathbb{Z}$. Let $Q_j = Hom_\Gamma(P_{2-j}, \Gamma)$ and $\partial^Q_i = Hom_\Gamma(\partial^P_{2-j}, \Gamma)$. Then $\overline{Q}_0 = P^\dagger_2$ and $\overline{Q}_1 = P^\dagger_1$ have dual bases $\{q^0_x\}$ and $\{q^1_r\}$, respectively. (Thus $q^1_x(p^1_y) = 1$ if $x = y$ and 0 otherwise, and $q^0_r(p^2_s) = 1$ if $r = s$ and 0 otherwise.) Then $\partial 1^* = \Sigma_{x \in X}(x^{-1} - 1)q^1_x$ and $\partial q^1_x = \Sigma_{r \in R} \overline{r_x} q^0_r$. After our normalization of the presentation, each r_x is either 0 or in $F(X)$, for all $r \in R$ and $x \in X$, and so $r_x - 1 = \partial(\Sigma_{y \in X} \frac{\partial r_x}{\partial y} p^1_y)$.

Define homomorphisms $j_i : \overline{Q}_i \to (P_* \otimes \overline{Q}_*)_i$, for $i = 0, 1, 2$, by setting

$$j_0(q^0_r) = 1 \otimes q^0_r \quad \text{for} \quad r \in R,$$

$$j_1(q^1_x) = 1 \otimes q^1_x - \Sigma_{r,y} \overline{r_x}(\frac{\partial r_x}{\partial y} p^1_y \otimes q^0_r) \quad \text{for} \quad x \in X, \quad \text{and}$$

$$j_2(1^*) = 1 \otimes 1^* - \Sigma_{x\in X} x^{-1}(p^1_x \otimes q^1_x) - \Sigma_{r\in R}(p^2_r \otimes q^0_r).$$

Then

$$\partial j_1(q^1_x) - j_0(\partial q^1_x) = \Sigma_{r\in R}(1 \otimes \overline{r_x} q^0_r) - \Sigma_{r,y}\overline{r_x}(\frac{\partial r_x}{\partial y}(y-1) \otimes q^0_r) - \Sigma_{r\in R}\overline{r_x}(1 \otimes q^0_r)$$

$$= \Sigma_{r\in R}[(1 \otimes \overline{r_x} q^0_r) - \overline{r_x}((r_x - 1) \otimes q^0_r) - \overline{r_x}(1 \otimes q^0_r)] = 0,$$

and so $\partial j_1 = j_0 \partial$. Similarly,

$$\partial j_2(1^*) - j_1(\partial 1^*) = \Sigma_x[1 \otimes (x^{-1} - 1)q^1_x - x^{-1}((x-1) \otimes q^1_x)] +$$

$$\Sigma_x \Sigma_r[(x^{-1}(p^1_x \otimes \overline{r_x} q^0_r) - r_x p^1_x \otimes q^0_r)] - \Sigma_x(x^{-1} - 1)[1 \otimes q^1_x - \Sigma_{r,y}\overline{r_x}(\frac{\partial r_x}{\partial y} p^1_y \otimes q^0_r)]$$

$$= \Sigma_{r,x}[x^{-1}(p^1_x \otimes \overline{r_x} q^0_r) - r_x p^1_x \otimes q^0_r + \Sigma_y(x^{-1} - 1)\overline{r_x}(\frac{\partial r_x}{\partial y} p^1_y \otimes q^0_r)].$$

It shall clearly suffice to show that the summand corresponding to each relator r is 0. After our normalization of the presentation, we may assume that $r = x_1 \dots x_m$ for some distinct $x_1, \dots, x_m \in X$. Let $r_i = r_{x_i}$, for $1 \le i \le m$. Then $r_i = x_1 \dots x_{i-1}$, for $1 \le i \le m$, so $r_i x_i = r_{i+1}$ if $i < m$ and $r_m x_m = r = 1$ in G. Moreover, $\frac{\partial r_i}{\partial y} = r_j$ if $y = x_j$, for some $1 \le j < i$, and is 0 otherwise. Let $S_{i,j} = r_i^{-1}(r_j p^1_{x_j} \otimes q^0_r)$, for $1 \le j \le i \le m$. Then $x_m^{-1} S_{m,j} = S_{1,j}$, for all $j \le m$, and so the summand corresponding to the relator r in $\partial j_2(1^*) - j_1(\partial 1^*)$ is

$$\Sigma_{i\le m}(x_i^{-1} S_{i,i} - S_{1,i} + \Sigma_{j<i}(x_i^{-1} S_{i,j} - S_{i,j}))$$

$$= \Sigma_{i<m}(S_{i+1,i} - S_{1,i}) + \Sigma_{i\le m}\Sigma_{j<i}(S_{i+1,j} - S_{i,j}).$$

This sum collapses to 0, and so $\partial j_2 = j_1 \partial$. Thus j_* is a chain homomorphism. Since $\overline{Q}_*$ and $P_* \otimes \overline{Q}_*$ are resolutions of $\mathbb{Z}$ and j_* induces the identity on $\mathbb{Z}$, it is a chain homotopy equivalence.

We then have

$$(A_{P_2 \overline{\mathscr{D}}}(q^0_s \otimes_\Gamma \delta) \otimes \eta)(j_*(1^*)) = -\Sigma_{r\in R}(q^0_s(p^2_r)\delta \otimes_\Gamma \eta(q^0_r)),$$

which has image $-\delta \otimes_\Gamma \eta(q^0_s)$ in $\mathscr{D} \otimes_\Gamma \overline{\mathscr{D}}$. Let τ be the ($\mathbb{Z}$-linear) involution of $H^2(G; \overline{\mathscr{D}})$ given by $\tau(\alpha_{\overline{\mathscr{D}}}(\rho \otimes_\Gamma \alpha)) = \alpha_{\overline{\mathscr{D}}}(\alpha \otimes_\Gamma \rho)$. Then

$$[\xi] \smile id_{\overline{\mathscr{D}}} = -\theta_{\overline{\mathscr{D}}}(\tau([\xi])) \quad for \quad \xi \in H^2(G; \overline{\mathscr{D}}),$$

and so $c^2_{G,w}$ is an isomorphism.

If $C(\widetilde{\mathscr{P}})$ is not aspherical we modify the definition of the dual complex Q_* by setting $Q_1 = Hom_\Gamma(P_1, \Gamma) \oplus Hom_\Gamma(P_3, \Gamma)$ and extending the differential by $s^\dagger$, where $s\partial_3 = id_{P_3}$. Let $f : P_3^\dagger \to \Gamma^s$ be a split monomorphism, with left inverse $g : \Gamma^s \to P_3^\dagger$.

Fix a basis $\{e_1,\dots,e_s\}$ for Γ^s, and define a homomorphism $h:\Gamma\to\Gamma\otimes\Gamma^s$ by $h(e_i)=1\otimes e_i$. Then we may extend j_1 by setting $j_1=(1\otimes g)hf$ on $P_3^\dagger$.

In [40] we gave closed formulae for $j_2(1^*)$ for some simple (un-normalized) presentations of groups of particular interest. We should have also given the appropriate form of j_1 explicitly, for there we used the relators to simplify the derivatives r_x, which in general are sums of monomials $\Sigma_k \pm r_{xk}$, and such simplifications affect the second derivatives $\frac{\partial r_{xk}}{\partial y}$. It is safer to calculate such derivatives in $\mathbb{Z}[F(X)]$ *before* using the relators to simplify their images in Γ.

Similar formulae show that $c^1_{F,w}$ is an isomorphism for F free of finite rank $r\geq 1$.

18 Orbits of the k-invariant

In this section we shall attempt to extend the argument sketched in §15 above for the case of PD_2-groups to other finitely presentable, 2-dimensional duality groups. The hypothesis on 2-torsion in Theorem 27 below seems necessary for our argument, but does not hold in some cases where the result is known by other means.

Lemma 18. *Let π be a finitely presentable group such that $c.d.\pi=2$, and let $w:\pi\to\mathbb{Z}^\times$ be a homomorphism. Let $\Pi=E^2\mathbb{Z}$. Then there is an exact sequence*

$$\Pi\odot_\pi\Pi\to\mathbb{Z}^w\otimes_{\mathbb{Z}[\pi]}\Gamma_W(\Pi)\to H^2(\pi;\mathbb{F}_2)\to 0.$$

If $\Pi\odot_\pi\Pi$ is 2-torsion free this sequence is short exact. If, moreover, for every $x\in\Pi$ either $x\in(2,I_w)\Pi$ or $x\odot x\notin(2,I_w)(\Pi\odot\Pi)$ then $\mathbb{Z}^w\otimes_{\mathbb{Z}[\Pi]}\Gamma_W(\pi)$ is 2-torsion free.

Proof. Since Π is torsion free as an abelian group. it is a direct limit of free abelian groups, and so the natural map from $\Pi\odot\Pi$ to $\Gamma_W(\Pi)$ is injective. Applying $\mathbb{Z}^w\otimes_{\mathbb{Z}[\pi]}-$ to the exact sequence

$$0\to\Pi\odot\Pi\xrightarrow{s}\Gamma_W(\Pi)\xrightarrow{q_\Pi}\Pi/2\Pi\to 0.$$

gives the above sequence, since $\mathbb{Z}^w\otimes_{\mathbb{Z}[\pi]}\Pi/2\Pi\cong\Pi/(2,I_w)\Pi\cong H^2(\pi;\mathbb{F}_2)$. The kernel on the left in this sequence is the image of $Tor_1^{\mathbb{Z}[\pi]}(\mathbb{Z}^w,\Pi/2\Pi)$, which is a 2-torsion group.

If $\Pi\odot_\pi\Pi$ is 2-torsion free this sequence is short exact, and nontrivial 2-torsion in $\mathbb{Z}^w\otimes_{\mathbb{Z}[\pi]}\Gamma_W(\Pi)$ has nontrivial image in $\Pi/(2,I_w)\Pi$. If there is such torsion there are $x,y_i,z_i\in\Pi$ such that $x\notin(2,I_w)\Pi$ but $2[\gamma_\Pi(x)+s(\Sigma y_i\odot z_i)]=0$ in $\Pi\odot_\pi\Pi$. Since $2\gamma_\Pi(x)=s(x\odot x)$ in $\Gamma_W(\Pi)$, we then have $s(x\odot x)\equiv 2(-s(\Sigma y_i\odot z_i))$ *mod* $I_w(\Pi\odot\Pi)$, and so $x\odot x\in(2,I_w)(\Pi\odot\Pi)$.

The final condition in the lemma depends only on the image of x in $\Pi/(2,I_w)\Pi$.

Let X be a PD_4-complex with $\pi_1(X)=\pi$ and $\pi_2(X)=\Pi$, and let $L=L_\pi(\Pi,2)$. Then $\widetilde{L}\simeq K(\Pi,2)$, and so it follows from the Whitehead sequence that $H_3(\widetilde{L};\mathbb{Z})=0$

and $H_4(\widetilde{L};\mathbb{Z}) \cong \Gamma_W(\Pi)$. Let $\mathscr{A}$ be a left $\mathbb{Z}[\pi]$-module. Since π is a 2-dimensional duality group with dualizing module $\overline{\Pi}$, Lemma 16 gives canonical isomorphisms

$$H^2(\pi;\mathscr{A}) = Ext^2_{\mathbb{Z}[\pi]}(\mathbb{Z},\mathscr{A}) \cong \overline{\Pi} \otimes_{\mathbb{Z}[\pi]} \mathscr{A}$$

and

$$Ext^2_{\mathbb{Z}[\pi]}(\Pi,\mathscr{A}) = \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \mathscr{A}.$$

Moreover, $H^2(\widetilde{L};\mathscr{A}) = Hom_{\mathbb{Z}}(\Pi,\mathscr{A})$ and $H^4(\widetilde{L};\mathscr{A}) = Hom_{\mathbb{Z}}(\Gamma_W(\Pi),\mathscr{A})$. Hence the spectral sequence for the universal covering $p_L : \widetilde{L} \to L$ gives exact sequences

$$0 \to H^2(\pi;\mathscr{A}) \to H^2(L;\mathscr{A}) \to Hom_{\mathbb{Z}[\pi]}(\Pi,\mathscr{A}) \to 0$$

(split by the homomorphism $H^2(\sigma;\mathscr{A})$ induced by a section σ for c_L), and

$$0 \to \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \mathscr{A} \to H^4(L;\mathscr{A}) \xrightarrow{p_L^*} Hom_{\mathbb{Z}[\pi]}(\Gamma_W(\Pi),\mathscr{A}) \to 0.$$

The right hand homomorphisms are induced by p_L, in each case. Since $H_q(\widetilde{X};\mathbb{Z}) = 0$ for $q > 2$, the spectral sequence for $p_X : \widetilde{X} \to X$ gives an isomorphism $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \mathscr{A} = Ext^2_{\mathbb{Z}[\pi]}(\Pi,\mathscr{A})) \cong H^4(X;\mathscr{A})$, and so $f_{X,2}$ induces a (non-canonical?) splitting of the second of these sequences.

In the next theorem and subsequent comments p_L^* is used variously for the homomorphisms determined by $H^4(p_L;\Gamma_W(\Pi))$, $H^2(p_L;\Pi)$ and $H^4(p_L;\Pi/2\Pi)$.

Theorem 27. *Let π be a finitely presentable, 2-dimensional duality group, and let $w : \pi \to \mathbb{Z}^\times$ be a homomorphism. Let $\Pi = E^2\mathbb{Z}$. Assume that the image of $\Pi \odot_\pi \Pi$ in $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$ is 2-torsion free. Then the homotopy type of a minimal PD_4-complex Z with $(\pi_1(Z), w_1(Z)) \cong (\pi, w)$ is determined by its refined v_2-type.*

Proof. Let Z be a minimal PD_4-complex with $\pi_1(Z) \cong \pi$ and $w_1(Z) = c_Z^* w$. Then $\pi_2(Z) \cong \Pi$ and $\pi_3(Z) \cong \Gamma_W(\Pi)$, since π has one end, and the homotopy type of Z is determined by $k = k_2(Z) \in H^4(L;\Gamma_W(\Pi))$, where $\Pi = E^2\mathbb{Z}$ and $L = P_2(Z) = L_\pi(\Pi)$. This class is only well defined up to the actions of $Aut(\Gamma_W(\Pi))$ and $E_0(L)$. Since $p_L^* k = k_2(\widetilde{Z})$ is an automorphism (considered as an endomorphism of $\Gamma_W(\Pi)$), by part (1) of Theorem 23, we may assume that $p_L^* k = id_{\Gamma_W(\Pi)}$, after applying an automorphism of $\Gamma_W(\Pi)$. Now $E_0(L) \cong E_\pi(L) \rtimes Aut(\pi)$ and $E_\pi(L) \cong H^2(\pi;\Pi) \rtimes Aut(\Pi)$. (See §3 above). We shall consider the action of $Aut(\pi)$ in the final paragraph of the proof. Since $Aut(\Pi) = \{\pm 1\}$ acts trivially on $\Gamma_W(\Pi)$, the main task is to consider the action of $H^2(\pi;\Pi)$ on k. We shall show that this action is closely related to the cup product homomorphism $c^2_{\pi,w}$. Note also that since Z is minimal, $v_2(Z) = c_Z^* v$ for some $v \in H^2(\pi;\mathbb{F}_2)$, by Theorem 20, and $E_\pi(L)$ fixes classes induced from $K = K(\pi,1)$, such as $c_L^* v$.

Let $\phi \in H^2(\pi;\Pi)$ and let $s_\phi \in [K,L]_K$ and $h_\phi \in [L,L]_K$ be as defined in Lemma 4. Let $M = L_\pi(\Pi,3)$. Then $[M,M]_K = H^3(M;\Pi) \cong End(\Pi)$, since $c.d.\pi = 2$. Let $\overline{\Omega} : [M,M]_K \to [L,L]_K$ be the loop map. Let $g \in [M,M]_K$ have image $[g] = \pi_3(g) \in End(\Pi)$ and let $f = \overline{\Omega} g$. Then $\omega([g]) = f^* \iota_{\Pi,2}$ defines a homomorphism

$\omega : End(\Pi) \to H^2(L;\Pi)$ such that $p_L^*\omega([g]) = [g]$ for all $[g] \in End(\Pi)$. Moreover $f\mu = \mu(f,f)$, since $f = \overline{\Omega}g$, and so $fh_\phi = \mu(fs_\phi c_L, f)$. Hence

$$h_\phi^*\xi = \xi + c_L^* s_\phi^* \xi$$

for $\xi = \omega([g]) = f^*\iota_{\Pi,2}$. Naturality of the isomorphisms $H^2(X;\mathscr{A}) \cong [X, L_\pi(\mathscr{A},2)]_K$ for X a space over K and $\mathscr{A}$ a left $\mathbb{Z}[\pi]$-module implies that

$$s_\phi^*\omega([g]) = [g]_\# s_\phi^* \iota_{\Pi,2} = [g]_\#\phi$$

for all $\phi \in H^2(\pi;\Pi)$ and $g \in [M,M]_K$. (See [2, Chapter 5.§4].)

Using our present hypotheses, the exact sequences above give sequences

$$0 \to H^2(\pi;\Pi) \xrightarrow{c_L^*} H^2(L;\Pi) \xrightarrow{p_L^*} End(\Pi) \to 0$$

(split by ω and the homomorphism $H^2(\sigma;\Pi)$ induced by a section σ for c_L), and

$$0 \to \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi) \to H^4(L;\Gamma_W(\Pi)) \xrightarrow{p_L^*} End(\Gamma_W(\Pi)) \to 0.$$

We shall identify the modules on the left with their images, to simplify the notation.

If $u \in H^2(\pi;\Pi)$ then $h_\phi^*(u) = u$, since $c_L h_\phi = c_L$. The induced automorphism of the quotient $End(\Pi) = H^0(\pi;(H^2(\widetilde{L};\Pi))$ is also the identity, since the lifts of h_ϕ are (non-equivariantly) homotopic to the identity in $\widetilde{L}$. Hence there is a homomorphism

$$\delta_\phi : End(\Pi) \to H^2(\pi;\Pi)$$

such that $h_\phi^*(\xi) = \xi + c_L^*\delta_\phi(p_L^*\xi)$ for all $\xi \in H^2(L;\Pi)$. Since $p_L^*c_L^* = 0$ and $h_{\phi+\psi} = h_\phi h_\psi$ it follows that δ_ϕ is additive as a function of ϕ. Since π is a 2-dimensional duality group, $H^2(\pi;\Pi) \cong \overline{\Pi} \otimes_{\mathbb{Z}[\pi]} \Pi$, and so $\phi = \rho \otimes_\pi \alpha$ for some $\rho \in \overline{\Pi}$ and $\alpha \in \Pi$. If $g \in [M,M]_K$ then

$$\delta_\phi([g]) = \delta_\phi(p_L^*\omega([g])) = s_\phi^*\omega([g]) = \rho \otimes_\pi [g](\alpha). \tag{1}$$

In particular, $\delta_\phi(id_\Pi) = \phi$.

Similarly, the automorphism of $H^4(L;\Gamma_W(\Pi))$ induced by h_ϕ fixes the subgroup $G = \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$, and induces the identity on the quotient $End(\Gamma_W(\Pi)) = H^0(\pi;H^4(\widetilde{L};\Gamma_W(\Pi)))$. Then there is a homomorphism

$$f_\phi : H^4(L;\Gamma_W(\Pi)) \to G$$

such that $h_\phi^*(u) = u + f_\phi(u)$ for all $u \in H^4(L;\Gamma_W(\Pi))$, and such that $f_\phi|_G = 0$. Moreover, f_ϕ is additive as a function of ϕ, so we may define $\widehat{f} : H^2(\pi;\Pi) \to G$ by

$$\widehat{f}(\phi) = f_\phi(k), \quad \text{for all } \phi \in H^2(\pi;\Pi).$$

When $S=L$, $\mathscr{A}=\mathscr{B}=\Pi$, and $p=q=2$ the construction of §15 gives a cup product pairing of $H^2(L;\Pi)$ with itself with values in $H^4(L;\Pi\otimes\Pi)$. Since $c.d.\pi=2$ this pairing is trivial on the image of $H^2(\pi;\Pi)\otimes H^2(\pi;\Pi)$. The maps c_L and σ induce a splitting $H^2(L;\Pi)\cong H^2(\pi;\Pi)\oplus End(\Pi)$, and this pairing restricts to the cup product pairing of $H^2(\pi;\Pi)$ with $End(\Pi)$ with values in $Ext^2_{\mathbb{Z}[\pi]}(\Pi,\Pi\otimes\Pi)$. We may also compose with the natural homomorphisms from $\Pi\otimes\Pi$ to $\Pi\odot\Pi$ and $\Gamma_W(\Pi)$ to get pairings with values in $H^4(L;\Pi\odot\Pi)$ and $H^4(L;\Gamma_W(\Pi))$.

Since $h^*_\phi(\xi\smile\xi')=h^*_\phi\xi\smile h^*_\phi\xi'$ we have also

$$f_\phi(\xi\smile\xi')=\delta_\phi(p_L^*\xi')\smile\xi+\delta_\phi(p_L^*\xi)\smile\xi', \tag{2}$$

for all $\xi,\xi'\in H^2(L;\Pi)$. On passing to $\widetilde{L}\simeq K(\Pi,2)$ we find that

$$p_L^*(\xi\smile\xi')(\gamma_\Pi(x))=p_L^*\xi(x)\odot p_L^*\xi'(x), \tag{3}$$

for all $\xi,\xi'\in H^2(L;\Pi)$ and $x\in\Pi$. (To see this, note that the inclusion of x determines a map from $\mathbb{CP}^\infty$ to $K(\Pi,2)$, since $[\mathbb{CP}^\infty,K(\Pi,2)]=Hom(\mathbb{Z},\Pi)$. Hence we may use naturality of cup products to reduce to the case when $K(\Pi,2)=\mathbb{CP}^\infty$ and x is a generator of $\Pi=\mathbb{Z}$.)

Let P be the image of $\Pi\odot_\pi\Pi$ in G. Since $c^2_{\pi,w}$ is an isomorphism, by Theorem 26, the induced map $\widehat{c}\colon H^2(\pi;\Pi)\to P$ is an epimorphism. Let $e=\widehat{f}-\widehat{c}$.

If $\Xi=\lambda\smile\lambda$ with $p_L^*\lambda=id_\Pi$ then $p_L^*(\Xi)(\gamma_\Pi(x))=x\odot x=2\gamma_\Pi(x)$, for all $x\in\Pi$, by Equation (5), while $f_\phi(\Xi)=2(\phi\smile\lambda)=2\phi\smile id_\Pi$, by Equation (4) and by the triviality of the cup product on the image of $H^2(\pi;\Pi)\otimes H^2(\pi;\Pi)$. Hence

$$p_L^*(\Xi)=2id_{\Gamma_W(\Pi)}\quad\text{and}\quad f_\phi(\Xi)=2\widehat{c}(\phi).$$

Since $p_L^*k=id_{\Gamma_W(\Pi)}$, we have $p_L^*(2k-\Xi)=0$, and so $2k-\Xi\in G$, by the exactness of sequence (2) above. Then

$$2e(\phi)=f_\phi(2k-\Xi)=0,$$

since $f_\phi|_G=0$. Hence e has image in the 2-torsion subgroup ${}_2G$.

We invoke the hypothesis on 2-torsion at this point. Since ${}_2G\cap P=0$, it follows easily that $|\mathrm{Cok}(\widehat{f})|\le|G/P|=|H^2(\pi;\mathbb{F}_2)|$. As ϕ varies in $H^2(\pi;\Pi)$ the values of $h_\phi(k)$ sweep out a coset of $\mathrm{Im}(\widehat{f})$ in $k+G=(p_L^*)^{-1}(id_{\Gamma_W(\Pi)})$, and there are at most 2^β cosets, where $\beta=\beta_2(\pi;\mathbb{F}_2)$.

For each $v\in H^2(\pi;\mathbb{F}_2)$ there is a minimal PD_4-complex Z such that $v_2(Z)=c_Z^*v$, by Theorem 18. The group $Aut(\pi)$ acts on K and L through based self-homotopy equivalences, and hence acts on the classifying maps c_Z and $f_{Z,2}$ by composition. These actions induce actions on $H^2(\pi;\mathbb{F}_2)$ and Π, and hence on $H^4(L;\Gamma_W(\Pi))$. The association $k\mapsto v_2(Z)$ defines a $Aut(\pi)$-equivariant surjection from $(p_L^*)^{-1}(id_{\Gamma_W(\Pi)})=k+G$ to $H^2(\pi;\mathbb{F}_2)$, which is constant on cosets of $\mathrm{Im}(\widehat{f})$, since $E_\pi(L)$ acts trivially on $H^2(\pi;\mathbb{F}_2)$. It follows that the refined v_2-type is a complete invariant for the homotopy types of such complexes.

If $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$ is 2-torsion free then $\widehat{f} = \widehat{c}$ (since $e = 0$), and the argument can be simplified slightly.

The hypothesis on 2-torsion holds if π is a PD_2-group, for then $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi) \cong \mathbb{Z}$ if $w = 1$ and has order 2 otherwise. (Note that in this case $\Pi \cong \mathbb{Z}^u$, where $u = w + w_1(\pi)$. We do not assume here that $w = w_1(\pi)$!) It holds also if $\pi = Z *_m$ with $|m| > 1$, by Theorem 30 below. On the other hand, if $\pi = F(r) \times \mathbb{Z}$ and $w(t) = -1$, where $t \in \pi$ generates the central $\mathbb{Z}$ factor, then $\Pi \odot_\pi \Pi$ and $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$ have exponent 2, since t acts through ± 1 on Π. If $r > 1$ these groups are not finitely generated, and so the hypothesis of Theorem 27 does not hold.

Corollary 28 *If $H^2(\pi;\mathbb{F}_2) = 0$ and $\Pi \odot_\pi \Pi$ is 2-torsion free there is an unique minimal PD_4-complex realizing (π, w).* □

Hence two PD_4-complexes X and Y with fundamental group π are homotopy equivalent if and only if $\lambda_X \cong \lambda_Y$ (i.e., there is an isomorphism $\theta : \pi_1(X) \cong \pi_1(Y)$ such that $w_1(X) = w_1(Y) \circ \theta$ and an isometry of the pairings, up to sign.)

The hypothesis $H^2(\pi;\mathbb{F}_2) = 0$ holds if π is the group of a link of 2-spheres in an homology 4-sphere, in particular, if it is a 2-knot group or is the fundamental group of an homology 4-sphere.

Corollary 29 *If $H^2(\pi;\mathbb{F}_2) = \mathbb{F}_2$ and the image of $\Pi \odot_\pi \Pi$ in $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$ is 2-torsion free there are two minimal PD_4-complexes realizing (π, w), distinguished by whether $v_2(X) = 0$ or not.* □

The work of [29] suggests that the refined v_2-type should be a complete homotopy invariant, without the technical hypothesis on 2-torsion or the restriction that π have one end. If, moreover, $g.d.\pi = 2$ then every such minimal PD_4-complex should be homotopy equivalent to a closed 4-manifold, by Theorem 18. This is so if π is a semidirect product $F(r) \rtimes \mathbb{Z}$ or a PD_2-group, by Theorems 17 and 25. Can the connection between k_2 and v_2 be made more explicit? The canonical epimorphism $q_\Pi : \Gamma_W(\Pi) \to \Pi/2\Pi$ determines a change of coefficients homomorphism $q_{\Pi\#}$ from sequence (2) above to the parallel sequence

$$0 \to H^2(\pi;\mathbb{F}_2) \to H^4(L;\Pi/2\Pi) \xrightarrow{p_L^*} Hom_{\mathbb{Z}[\pi]}(\Gamma_W(\Pi), \Pi/2\Pi) \to 0.$$

Thus $q_{\Pi\#}(k_2(Z))$ lies in the $H^2(\pi;\mathbb{F}_2)$-coset $(p_L^*)^{-1}(q_\Pi)$.

Does Theorem 24 have an analogue for other 2-dimensional duality groups? Let X and Z be PD_4-complexes with such a fundamental group π, with Z minimal, and such that $(c_X^*)^{-1} w_1(X) = (c_Z^*)^{-1} w_1(Z)$. Then $[X,Z]_K$ maps onto $[X,P_3(Z)]_K$, by cellular approximation, and hence onto $\{f \in [X,L]_K \mid f^* k_2(Z) = 0\}$. Can the condition $f^* k_2(Z) = 0$ be made more explicit? The map f corresponds to a class in $H^2(X;\Pi)$ and $H^4(X;\Gamma_W(\Pi)) \cong \mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi))$, by Poincaré duality for X. Theorem 24 suggests that we should consider the image of $f^* k_2(Z)$ in $H^2(\pi;\mathbb{F}_2)$, under the epimorphism of Lemma 18. Apart from this, we must determine when such a map f has a degree-1 representative $g : X \to Z$.

19 Verifying the torsion condition for $Z*_m$

If π is a 2-dimensional duality group but not a PD_2-group then $\Pi = E^2\mathbb{Z}$ is finitely generated as a left $\mathbb{Z}[\pi]$-module, but is not finitely generated as an abelian group. The associated groups $\Pi \odot_\pi \Pi$ and $\mathbb{Z}^w \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi))$ are infinitely generated abelian groups with no natural module structure. In this section we shall investigate the 2-torsion condition.

We consider first groups which have a one-relator presentation $\mathscr{P} = \langle X \mid r\rangle$. It is well-known that if the relator r is not conjugate to a proper power then the associated 2-complex $C(\mathscr{P})$ is aspherical, and so $g.d.\pi \leq 2$. (See §§9-11 of Chapter III of [46], or [17].)

Lemma 19. *Let π be a group with a finite one-relator presentation $\langle X \mid r\rangle$ and $c.d.\pi = 2$, and let $w = 1$. Let $\Pi = E^2\mathbb{Z}$. Then*

$$\Pi \odot_\pi \Pi \cong \mathbb{Z}[\pi]/(U + \overline{\Delta}),$$

where Δ is the right ideal generated by the free derivatives $\frac{\partial r}{\partial x}$, for all $x \in X$, and U is the subgroup of $\mathbb{Z}[\pi]$ generated by $g - g^{-1}$, for all $g \in \pi$.

Proof. On dualizing the Fox-Lyndon resolution of $\mathbb{Z}$ associated to the presentation $\langle X \mid r\rangle$, we see that $H^2(\pi;\mathbb{Z}[\pi]) \cong \mathbb{Z}[\pi]/\Delta$, and so $\Pi \cong \mathbb{Z}[\pi]/\overline{\Delta}$.

Define a function $T : \mathbb{Z}[\pi] \otimes \mathbb{Z}[\pi] \to \mathbb{Z}[\pi] \otimes \mathbb{Z}[\pi]$ by $T(s \otimes t) = \bar{s} \otimes t$, for all $s,t \in \mathbb{Z}[\pi]$. Then T is an additive bijection and $T(gs \otimes gt) = \bar{s}\bar{g} \otimes gt$, for all $g \in \pi$. Hence T induces an additive isomorphism from the quotient of $\mathbb{Z}[\pi] \otimes \mathbb{Z}[\pi]$ by the diagonal action of π to $\mathbb{Z}[\pi] \otimes_{\mathbb{Z}[\pi]} \mathbb{Z}[\pi] \cong \mathbb{Z}[\pi]$, which maps $s \otimes t$ to $\bar{s}t$. The images of $\mathbb{Z}[\pi] \otimes \overline{\Delta}$ and $\overline{\Delta} \otimes \mathbb{Z}[\pi]$ under T are $\overline{\Delta}$ and Δ, respectively. We obtain the symmetric product $\mathbb{Z}[\pi] \odot \mathbb{Z}[\pi]$ by factoring out the tensor square $\mathbb{Z}[\pi] \otimes \mathbb{Z}[\pi]$ by all sums of terms of the form $s \otimes t - t \otimes s$. The image of all such sums in $\mathbb{Z}[\pi]$ is the subgroup U. (Note that U is not usually an ideal!) Since $\mathbb{Z}[\pi] \odot_{\mathbb{Z}[\pi]} \mathbb{Z}[\pi] \cong \mathbb{Z}[\pi]/U$ and $U + \overline{\Delta} = U + \Delta$, we see that $\Pi \odot_\pi \Pi \cong \mathbb{Z}[\pi]/(U + \overline{\Delta})$.

This may be extended to other 2-dimensional duality groups as follows. Suppose that P is an $a \times b$ presentation matrix for Π. View $\mathbb{Z}[\pi]^b$ as a module of row vectors, with standard basis $\{e_1, \ldots, e_b\}$. Define a function $T : \mathbb{Z}[\pi]^b \otimes \mathbb{Z}[\pi]^b \to M_b(\mathbb{Z}[\pi])$ by $T(se_i \otimes te_j) = \bar{s}te_{ij}$, the matrix with (i,j) entry $\bar{s}t$ and all other entries 0. Then $T(\mathbb{Z}[\pi]^b \otimes \mathrm{Im}(P))$ is $Row(P)$, the left ideal in $M_b(\mathbb{Z}[\pi])$ consisting of matrices with all rows in $\mathrm{Im}(P)$, while $T(\mathrm{Im}(P) \otimes \mathbb{Z}[\pi]^b)$ is the right ideal $Row(P)^\dagger$, the conjugate transpose of $Row(P)$. Let V be the subgroup generated by $M - M^\dagger$, for all M in $M_b(\mathbb{Z}[\pi])$. Then $\Pi \otimes_\pi \Pi \cong M_b(\mathbb{Z}[\pi])/(V + Row(P) + Row(P)^\dagger)$.

Suppose now that π is solvable. Then it is a Baumslag-Solitar group $\mathbb{Z}*_m$, with a one-relator presentation $\langle a,t \mid tat^{-1}a^{-m}\rangle$, for some $m \neq 0$ [26]. In this case we have a more explicit model for $\Pi \odot_\pi \Pi$.

Theorem 30. *Let $\pi = \mathbb{Z}*_m$ and let $w : \pi \to \mathbb{Z}^\times$ be a homomorphism. Let $\Pi = E^2\mathbb{Z}$. If $|m| > 1$ then $\Pi \odot_\pi \Pi$ is torsion free.*

Proof. We may assume that π has the presentation $\langle a,t \mid tat^{-1}a^{-m}\rangle$. Let $A = \langle\langle a\rangle\rangle$. Then $\pi \cong A \rtimes \mathbb{Z}$. Let $a_n = t^n a t^{-n}$ in A, for all $n \in \mathbb{Z}$, and let $a^x = a^k_{-n}$, for all $x = \frac{k}{m^n} \in \mathbb{Z}[\frac{1}{m}]$. Then $a^0 = 1$, $a^1 = a$ and $a^x a^y = a^{x+y}$ for all $x,y \in \mathbb{Z}[\frac{1}{m}]$, and $x \mapsto a^x$ determines an isomorphism from $\mathbb{Z}[\frac{1}{m}]$ to A. Every element of π is uniquely of the form $t^p a^x$, for some $p \in \mathbb{Z}$ and $x \in \mathbb{Z}[\frac{1}{m}]$, and $(t^p a^x)^{-1} = t^{-p} a^{-m^p x}$. If m is even then $w(a^x) = 1$ for all x; if m is odd then $w(a^x) = w(a^{mx})$ for all x.

The function which sends a_n to a_{n+1} determines an automorphism α of the commutative domain $D = \mathbb{Z}[A] \cong \mathbb{Z}[a_n | n \in \mathbb{Z}]/(a_{n+1} - a_n^m)$, and $\mathbb{Z}[\pi]$ is isomorphic to the twisted Laurent extension $D_\alpha[t,t^{-1}]$. (An explicit isomorphism is given by the function which sends $t^p a_n \in \oplus_{p\in\mathbb{Z}} t^p D$ to $t^{n+p} a t^{-n} \in \mathbb{Z}[\pi]$ for all $n,p \in \mathbb{Z}$.)

We shall assume henceforth that m is positive, for simplicity of notation. Let $J_0 = \{1,\dots,m-1\}$, let $J_s = \{\frac{d}{m^s} \mid 0 < d < m^{s+1},\ (d,m) = 1\}$, for all $s \geq 1$, and let $J = \cup_{s\geq 0} J_s$. Then $E = D/D(a^m - w(a)^m)$ is freely generated as an abelian group by the image of $\{a^x \mid x \in J\}$.

The images of the free derivatives of the relator $r = tat^{-1}a^{-m}$ in $\mathbb{Z}[\pi]$ are $\frac{\partial r}{\partial a} = t - \mu_m$, where $\mu_m = \Sigma_{i=0}^{i=m-1} a^i$, and $\frac{\partial r}{\partial t} = 1 - a^m$. Hence

$$\Pi \cong \mathbb{Z}[\pi]/\mathbb{Z}[\pi](a^m - w(a)^m, t\overline{\mu_m} - w(t)) \cong (\oplus_{k\in\mathbb{Z}} t^k E)/\sim,$$

where

$$t^k a^x \sim w(t) t^k a^x t\overline{\mu_m} = w(t) t^{k+1} a^{\frac{x}{m}} \overline{\mu_m}, \quad \text{for all } k \in \mathbb{Z} \text{ and } x \in J.$$

As an abelian group, $\Pi \cong \varinjlim t^p E$, the direct limit as $p \to +\infty$ of the family of D-linear monomorphisms $\sigma : t^p E \to t^{p+1} E$ given by $\sigma(t^p a^x) = w(t) t^{p+1} a^{\frac{x}{m}} \overline{\mu_m}$, for all $p \in \mathbb{Z}$ and $x \in J$. It follows easily that

$$\Pi \odot \Pi \cong \varinjlim (t^k E \odot t^k E) = (\oplus_{p\in\mathbb{Z}} t^p E \odot t^p E)/\sim,$$

where $t^k a^x \odot t^k a^y \sim t^{k+1} a^{\frac{x}{m}} \overline{\mu_m} \odot t^{k+1} a^{\frac{y}{m}} \overline{\mu_m}$, for all $k \in \mathbb{Z}$ and $x,y \in J$.

Setting $z = y - x$ gives

$$t^k a^x (1 \odot a^z) \sim t^{k+1} a^{\frac{x}{m}} (\overline{\mu_m} \odot \overline{\mu_m} a^{\frac{z}{m}}).$$

(Here π acts diagonally on $\Pi \odot \Pi$.) We may expand the term in parentheses as

$$\overline{\mu_m} \odot \overline{\mu_m} a^{\frac{z}{m}} = \Sigma_{i,j=0}^{i,j=m-1} w(a)^i a^{-i} (1 \odot w(a)^{i-j} a^{i-j} a^{\frac{z}{m}}).$$

Define a function $f : E \to \Pi \odot \Pi$ by $f(e) = 1 \odot e = e \odot 1$ for $e \in E$. Then f is additive and $f(a^x) = w(a)^m a^x f(a^{m-x})$ for all x, since $a^x \odot 1 = a^x (1 \odot w(a)^m a^{m-x})$. The induced map from E to $\Pi \odot_\pi \Pi$ is onto, and

$$\Pi \odot_\pi \Pi \cong E/N,$$

where N is the subgroup generated by

$$\{a^z - w(a^{m-z})a^{m-z},\ a^z - w(t)m\Sigma_{k=0}^{m-1}w(a)^k a^{k+\frac{z}{m}},\ \forall z \in J\}.$$

Since $a^z - w(a^{m-z})a^{m-z} \in N$, the images $[a^z]$ of the elements a^z with $0 \leq z \leq \frac{m}{2}$ generate the quotient E/N. Given that $[a^z] = w(a)^{m-z}[a^{m-z}]$, the conditions

$$[a^z] = w(t)m\Sigma_{k=0}^{m-1}w(a)^k[a^{k+\frac{z}{m}}]$$

and

$$[a^{m-z}] = w(t)m\Sigma_{k=0}^{m-1}w(a)^k[a^{k+\frac{m-z}{m}}]$$

are equivalent.

Let F_s be the subgroup of $\Pi \odot_\pi \Pi$ generated by $\{[a^z] \mid m^{s-1}z \in \mathbb{Z}\}$, for $s \geq 1$. If $|m| > 1$ then the conditions $[a^z] = [w(t)m\Sigma_{k=0}^{m-1}w(a)^k[a^{k+\frac{z}{m}}]$ in E/N, for $z \in J$, imply that F_s is generated by $\{[a^0]\} \cup \{[a^z] \mid 0 < 2z \leq m,\ m^{s-1}z \in \mathbb{Z},\ m^{s-2}z \notin \mathbb{Z}\}$, for all $s \geq 1$, with a single relation of the form $(1 - w(t)m)[a^0] = m^s\sigma$, where σ is a sum of the generators $[a^z]$ with $z \in J_s$ such that $0 < 2z < m$, and coefficients not divisible by $(1-m)$. Hence F_s is torsion free, for all $s \geq 1$. Since $\Pi \odot_\pi \Pi$ is the increasing union $\cup_{s\geq 0}F_s$, it is also torsion free.

If $m = \pm 1$ and $w = 1$ then $\Pi \odot_\pi \Pi \cong \mathbb{Z}$. However, if $m = \pm 1$ and $w \neq 1$ then $\Pi \odot_\pi \Pi = \mathbb{Z}/2\mathbb{Z}$, and so the theorem does not extend to this case.

Note that the argument of the final paragraph implies that every generator of $\Pi \odot_\pi \Pi$ is m-divisible, and that $\Pi \odot_\pi \Pi$ is a free $\mathbb{Z}[\frac{1}{m}]$-module of infinite rank.

Corollary 31 *If $\pi = \mathbb{Z}*_m$ with $|m| > 1$ then $\mathbb{Z} \otimes_{\mathbb{Z}[\pi]} \Gamma_W(\Pi)$ is torsion free.*

Proof. If m is even this follows immediately from the theorem and the short exact sequence of Lemma 18, since $H^2(\pi;\mathbb{F}_2) = 0$ then. If m is odd we may apply the final part of Lemma 18. Letting x be the image of $1 \in \mathbb{Z}[\pi]$, we see that $\gamma_\Pi(x)$ generates $\Pi/(2, I_w)\Pi = H^2(\pi;\mathbb{F}_2)$, while the image of $f(1) = x \odot x$ in $\Pi \otimes_\pi \Pi$ is not 2-divisible.

It is not immediately obvious that the models for $\Pi \odot_\pi \Pi$ in Lemma 19 and Theorem 30 agree when $\pi \cong \mathbb{Z}*_m$. However (assuming for simplicity that $m \geq 1$ and $w = 1$), the relations

$$t^k a^x \sim_1 t^k a^x t\mu_m = t^{k+1}a^{\frac{x}{m}}\mu_m \quad \text{and} \quad t^k a^x \sim_2 (t^k a^x)^{-1} = t^{-k}a^{-m^k x}$$

together imply that $\Pi \odot_\pi \Pi$ is generated by the image of E and that

$$a^z \ \sim_1\ ta^{\frac{z}{m}}\mu_m = \Sigma_{i=0}^{i=m-1}ta^i a^{\frac{z}{m}} \ \sim_2\ \Sigma_{i=0}^{i=m-1}t^{-1}a^{-mi-z} = mt^{-1}a^{-z}$$

$$\sim_1\ ma^{-\frac{z}{m}}\mu_m \ \sim_2\ m\Sigma_{i=0}^{i=m-1}a^{-i}a^{\frac{z}{m}} = ma^{\frac{z}{m}}\mu_m,$$

for all $z \in J$. This is enough to see that $\mathbb{Z}[\pi]/(U + \overline{\Delta})$ is a quotient of E/N, as an abelian group, when $\Delta = (a^m - 1, t - \mu_m)\mathbb{Z}[\pi]$.

Can we extend the argument of Theorem 30 in any way? In particular, does the hypothesis of Theorem 27 hold for ascending HNN extensions $F*_\varphi$ with base F

a finitely generated free group and φ an endomorphism such that $p \prec \varphi(p)$ for all $1 \prec p$ with respect to some left ordering $\prec$ on F? When φ is an automorphism π is a semidirect product $F(r) \rtimes_\varphi \mathbb{Z}$, and the result of Theorem 27 holds by Theorem 22. If φ has odd order and $w = 1$ then it can be shown that $\Pi \odot_\pi \Pi$ is 2-torsion free. However, as we have seen, the argument of Theorem 27 itself must be changed in order to accommodate other semidirect products $F(r) \rtimes_\varphi \mathbb{Z}$ and orientation characters w.

20 4-manifolds and 2-knots

In this section we shall invoke surgery arguments, and so "4-manifold" and "s-cobordism" shall mean TOP 4-manifold and (5-dimensional) TOP s-cobordism, respectively. We continue to assume that π is a 2-dimensional duality group.

Suppose that π is either the fundamental group of a finite graph of groups, with all vertex groups $\mathbb{Z}$, or is square root closed accessible, or is a classical knot group. (This includes all PD_2-groups, semidirect products $F(n) \rtimes \mathbb{Z}$ and the solvable groups $\mathbb{Z}*_m$.) Then $Wh(\pi) = 0$, $L_5(\pi, w)$ acts trivially on the s-cobordism structure set $S^s_{TOP}(M)$ and the surgery obstruction map $\sigma_4(M) : [M, G/TOP] \to L_4(\pi, w)$ is onto, for any closed 4-manifold M realizing (π, w). (See Lemma 6.9 and Theorem 17.8 of [34].)

If, moreover, $w_2(\widetilde{M}) = 0$ then every 4-manifold homotopy equivalent to M is s-cobordant to M, by Theorem 6.7, Lemma 6.5 and Lemma 6.9 of [Hi]. If $w_2(\widetilde{M}) \neq 0$ there are at most two s-cobordism classes of homotopy equivalences. After stabilization by connected sum with copies of $S^2 \times S^2$ there are two s-cobordism classes, distinguished by their KS smoothing invariants (see [43]).

If π is solvable then 5-dimensional s-cobordisms are products and stabilization is unnecessary, so homotopy equivalent 4-manifolds with fundamental group π are homeomorphic if the universal cover is Spin, and there are two homeomorphism types otherwise, distinguished by their KS invariants.

The Baumslag-Solitar group $\mathbb{Z}*_m$ has such a graph-of-groups structure and is solvable, so the 5-dimensional TOP s-cobordism theorem holds. Thus if m is even the closed orientable 4-manifold M with $\pi_1(M) \cong \mathbb{Z}*_m$ and $\chi(M) = 0$ is unique up to homeomorphism. If m is odd there are two such homeomorphism types, distinguished by whether $v_2(M) = 0$ or $v_2(M) \neq 0$.

Let π be a finitely presentable group with $c.d.\pi = 2$. If $H_1(\pi; \mathbb{Z}) = \pi/\pi' \cong \mathbb{Z}$ and $H_2(\pi; \mathbb{Z}) = 0$ then $\mathrm{def}(\pi) = 1$ [34, Theorem 2.8]. If moreover π is the normal closure of a single element then $\pi \cong \pi K = \pi_1(S^4 \setminus K)$, for some 2-knot $K : S^2 \to S^4$. (If the Whitehead Conjecture is true every knot group of deficiency 1 has cohomological dimension at most 2.) Since π is torsion free it is indecomposable, by a theorem of Klyachko [44]. Hence π has one end.

Let $M = M(K)$ be the closed 4-manifold obtained by surgery on the 2-knot K. Then $\pi_1(M) \cong \pi = \pi K$ and $\chi(M) = \chi(\pi) = 0$, and so M is a minimal model for π. If K is reflexive it is determined by M and the orbit of its meridian under the automorphisms of π induced by self-homeomorphisms of M. If $\pi = F(n) \rtimes \mathbb{Z}$ the homotopy

type of M is determined by π, as explained in §4 above. Since $H^2(M;\mathbb{F}_2) = 0$ it follows that M is s-cobordant to the fibred 4-manifold with $\#^n(S^2 \times S^1)$ and fundamental group π. Knots with Seifert surface a punctured sum $\#^n(S^2 \times S^1)_o$ are reflexive. Thus if K is fibred (and $c.d.\pi = 2$) it is determined (among all 2-knots) up to s-concordance and change of orientations by π together with the orbit of its meridian under the automorphisms of π induced by self-homeomorphisms of the corresponding fibred 4-manifold. (This class of 2-knots includes all Artin spins of fibred 1-knots. See §6 of [34, Chapter 17] for more on 2-knots with $c.d.\pi = 2$.)

A stronger result holds for the group $\pi = \mathbb{Z}*_2$. This is the group of Fox's Example 10, which is a ribbon 2-knot [23]. In this case π determines the homotopy type of $M(K)$, by Theorems 30 and 27. Since metabelian knot groups have an unique conjugacy class of normal generators (up to inversion) Fox's Example 10 is the unique 2-knot (up to TOP isotopy and reflection) with this group. (If K is any other nontrivial 2-knot such that πK is torsion free and elementary amenable then $M(K)$ is homeomorphic to an infrasolvmanifold. See [34, Chapters 16-18].)

Let $\Lambda = \mathbb{Z}[\mathbb{Z}]$. There is a hermitian pairing B on a finitely generated free Λ-module which is not extended from the integers, and a closed orientable 4-manifold M_B with $\pi_1(M) \cong \mathbb{Z}$ and such that the intersection pairing on $\pi_2(M_B)$ is equivalent to B. In particular, M_B is not the connected sum of $S^1 \times S^3$ with a 1-connected 4-manifold [30]. Let $N_B \subset M_B$ be an open regular neighbourhood of a loop representing a generator of $\pi_1(M_B)$. Suppose that X is a closed 4-manifold with fundamental group π and that there is an orientation preserving loop $\gamma \subset X$ whose image in π/π' generates a free direct summand. (For instance, there is such a loop if X is the total space of an S^2-bundle over an aspherical closed surface F with $\beta_1(F) > 1$). Then γ has a regular neighbourhood homeomorphic to N_B, and we may identify these regular neighbourhoods to obtain $N = M_B \cup_{S^1 \times D^3} X$. The inclusion of $\langle g \rangle$ into π and the projection of π onto Z mapping g to 1 determines a monomorphism $\gamma : \Lambda \to \mathbb{Z}[\pi]$ and a retraction $\rho : \mathbb{Z}[\pi] \to \Lambda$. In particular, $\Lambda \otimes_{\mathbb{Z}[\rho]} (\mathbb{Z}[\pi] \otimes_{\mathbb{Z}[\gamma]} B) \cong B$. It follows that as B is not extended from $\mathbb{Z}$ neither is $\mathbb{Z}[\pi] \otimes_{\mathbb{Z}[\gamma]} B$. Therefore N is not the connected sum of E with a 1-connected 4-manifold.

21 Some questions

We shall collect here some of the questions that have arisen en route.

1. Are strongly minimal PD_4-complexes always of v_2-type II or III?
2. If X has v_2-type I and $c.d.\pi = 2$ is there a minimal model $f : X \to Z$ with $v_2(Z) = 0$?
3. Must a strongly minimal PD_4-complex with π a nontrivial free product be a connected sum?
4. Can we say more about PD_4-complexes with π infinitely ended and $\Pi = 0$?
5. Are there strongly minimal PD_4-complexes with $E^3\mathbb{Z} \neq 0$?
6. Do strongly minimal PD_4-complexes always have $k_1 = 0$?

7. If X is a PD_4-complex such that $\pi = \pi_1(X)$ has one end and $\Pi = \pi_2(X)$ is projective, must π be a PD_4-group?
8. To what extent do k_2 and v_2 determine each other?
9. In Theorem 23 must Y be a PD_4-complex?
10. Can we extend Theorems 27 and 30 to encompass the known results for π a semidirect product $F(r) \rtimes \mathbb{Z}$ (at least if $w = 1$)?
11. Can we relax the running hypothesis that π should have one end?

The final four questions are of most interest for the present work.

References

1. Barth, W., Peters, C. and Van de Ven, A. Compact Complex Surfaces. Ergebnisse der Mathematik und ihrer Grenzgebiete, 3 Folge, Bd 4, Springer-Verlag, Berlin – Heidelberg – New York (1984)
2. Baues, H.-J. Obstruction Theory. Lecture Notes in Mathematics 628, Springer-Verlag, Berlin – Heidelberg – New York (1977)
3. Baues, H.-J. Algebraic Homotopy. Cambridge University Press, Cambridge – New York (1989)
4. Baues, H.-J. Homotopy Type and Homology. Oxford Mathematical Monographs, Clarendon Press, Oxford (1996)
5. Baues, H.-J. and Bleile, B. Poincaré duality complexes in dimension four. Algebraic and Geometric Topology 8, 2355–2389 (2008)
6. Bieri, R. Normal subgroups in duality groups and in groups of cohomological dimension two. J. Pure Appl. Alg. 7, 35–56 (1976)
7. Bleile, B., Bokor, I. and Hillman, J.A. Poincaré duality complexes with highly connected universal cover. Alg. Geom. Top. 18, 3749–3788 (2018)
8. Bowditch, B.H. Planar groups and the Seifert conjecture. J. Reine u. Angew. Math. 576, 11–62 (2004)
9. Brown, K.S. Cohomology of Groups. Graduate Texts in Mathematics 87, Springer-Verlag, Berlin – Heidelberg – New York (1982)
10. Brown, R., Higgins, P. J. and Sivera, R. Nonabelian Algebraic Topology. Filtered spaces, crossed complexes, cubical homotopy groupoids, with contributions by C. D. Wensley and S. V. Soloviev. EMS Tracts in Mathematics 15, European Mathematical Society, Zürich (2011)
11. Button, J.O. Mapping tori with first Betti number at least two. J. Math. Soc. Japan 59, 351–370 (2007)
12. Cavicchioli, A. and Hegenbarth, F. On the homotopy classification of 4-manifolds having the fundamental group of an aspherical 4-manifold. Osaka J. Math. 37, 859–871 (2000)
13. Cavicchioli, A., Hegenbarth, F. and Repovš, D. Four-manifolds with surface fundamental groups. Trans. Amer. Math. Soc. 349, 4007–4019 (1997)
14. Cartan, H. and Eilenberg, S. Homological Algebra. Princeton University Press, Princeton (1956)
15. Crisp, J.S. The decomposition of 3-dimensional Poincaré complexes. Comment. Math. Helv. 75, 232–246 (2000)
16. Dicks, W. and Dunwoody, M.J. Groups acting on Graphs. Cambridge University Press (1989)
17. Dyer, E. and Vasquez, A.T. Some small aspherical spaces. J. Aust. Math. Soc. 16, 332–352 (1973)
18. Eckmann, B. Cyclic homology of groups and the Bass conjecture. Comment. Math. Helvetici 61, 193–202 (1986)
19. Eckmann, B. Manifolds of even dimension with amenable fundamental group. Comment. Math. Helvetici 69, 501–511 (1994)

20. Eckmann, B. Projective and Hilbert modules over group algebras. and finitely dominated spaces, Comment. Math. Helvetici 71, 453–462 (1996) Addendum, *ibid.* 72, 329 (1996)
21. Farrell, F.T. The second cohomology group of G with coefficients Z_2G. Topology 13, 313–326 (1974)
22. Farrell, F.T. An extension of Tate cohomology to a class of infinite groups, J. Pure Appl. Alg. 10, 153–161 (1977)
23. Fox, R.H. A quick trip through knot theory. In: Fort, M.K., Jr (ed.) Topology of 3-Manifolds and Related Topics Prentice-Hall, Englewood Cliffs, N.J., 120–167 (1962)
24. Freedman, M.H. and Quinn, F. Topology of 4-Manifolds. Princeton University Press, Princeton (1990)
25. Geoghegan, R. Topological Methods in Group Theory. Graduate Texts in Mathematics 243, Springer-Verlag, Berlin – Heidelberg – New York (2008)
26. Gildenhuys, D. Classification of soluble groups of cohomological dimension two. Math. Z. 166, 21–25 (1979)
27. Hambleton, I. and Kreck, M. On the classification of topological 4-manifolds with finite fundamental group. Math. Ann. 280, 85–104 (1988)
28. Hambleton, I., Kreck, M. and Teichner, P. Nonorientable four-manifolds with fundamental group of order 2. Trans. Amer. Math. Soc. 344, 649–665 (1994)
29. Hambleton, I., Kreck, M. and Teichner, P. Topological 4-manifolds with geometrically two-dimensional fundamental groups. J. Topol. Anal. 1, 123–151 (2009)
30. Hambleton, I. and Teichner, P. A non-extended hermitian form over $\mathbb{Z}[Z]$. Manus. Math. 93, 435–442 (1997)
31. Hegenbarth, F., Pamuk, M. and Repovš, D. Homotopy classification of PD_4-complexes relative to an order relation. Monatsh. Math. 177, 275–293 (2015)
32. Hegenbarth, F., Repovš, D. and Spaggiari, F. Connected sums of 4-manifolds, Top. Appl. 146/147, 209–225 (2005)
33. Hendriks, H. Applications de la théorie d'obstruction en dimension 3. Mem. Soc. Math. France 53, 1–86 (1977)
34. Hillman, J.A. Four-Manifolds, Geometries and Knots. GT Monographs, vol. 5, Geometry and Topology Publications, Warwick (2002 – revised 2007 and 2014).
35. Hillman, J.A. On 4-manifolds homotopy equivalent to surface bundles over surfaces, Top. Appl. 40, 275–286 (1991)
36. Hillman, J.A. Minimal 4-manifolds for groups of cohomological dimension 2. Proc. Edinburgh Math. Soc. 37, 455–462 (1994)
37. Hillman, J.A. PD_4-complexes with free fundamental group. Hiroshima Math. J. 34, 295–306 (2004)
38. Hillman, J.A. PD_4-complexes with fundamental group a PD_2-group. Top. Appl. 142, 49–60 (2004)
39. Hillman, J.A. PD_4-complexes with strongly minimal models. Top. Appl. 153, 2413–2424 (2006)
40. Hillman, J.A. Strongly minimal PD_4-complexes. Top. Appl. 156, 1565–1577 (2009)
41. Jensen, C.U. Les foncteurs dérivées de $\varprojlim$ et ses applications à la théorie des modules. Lecture Notes in Mathematics 254, Springer-Verlag, Berlin – Heidelberg – New York (1972)
42. Kim, M.H., Kojima, S. and Raymond, F. Homotopy invariants of nonorientable 4-manifolds. Trans. Amer. Math. Soc. 333, 71–83 (1992)
43. Kirby, R.C. and Taylor, L.R. A survey of 4-manifolds through the eyes of surgery. In: Cappell, S., Ranicki, A., Rosenberg, J. (eds.) Surveys on Surgery, vol. 2, Princeton University Press, Princeton, 387–421 (2001)
44. Klyachko, A. A funny property of sphere and equations over groups. Comm. Algebra 21, 2555–2575 (1993)
45. Lück, W. L^2-Invariants: Theory and Applications to Geometry and K-Theory. Ergebnisse der Mathematik und ihrer Grenzgebiete 3 Folge, Bd. 44, Springer-Verlag, Berlin - Heidelberg - New York (2002)
46. Lyndon, R.C. and Schupp. P.E. Combinatorial Group Theory. Ergenisse der Mathematik Bd 89, Springer-Verlag, Berlin – Heidelberg – New York (1977)

47. Nakaoka, M. Transgression and the invariant k_n^{q+1}. Proc. Japan Acad. 30, 363–368 (1954)
48. Olum, P. Mappings of manifolds and the notion of degree. Ann. Math. 58, 458–480 (1953)
49. Pirashvili, T. Category of eilenberg-maclane fibrations and cohomology of grothendieck constructions. Comm. Algebra 21, 309–341 (1993)
50. Plotnick, S.P. Equivariant intersection forms, knots in S^4 and rotations in 2-spheres. Trans. Amer. Math. Soc. 296, 543–575 (1986)
51. Ranicki, A. Algebraic and Geometric Surgery. Oxford Mathematical Monographs, Clarendon Press, Oxford (2002)
52. Rosset, S. A vanishing theorem for Euler characteristics. Math. Z. 185, 211–215 (1984)
53. Rutter, J.W. The group of homotopy self-equivalences of non-simply-connected spaces using Postnikov decompositions. Proc. Roy. Soc. Edinburgh Ser. A 120, 47–60; II, *ibid.* 122, 127–135 (1992)
54. Spanier, E. Algebraic Topology. McGraw-Hill, New York (1966).
55. Taylor, L.R. Unoriented geometric functors. Forum Math. 20, 457–469 (2008)
56. Tsukiyama, K. Self-homotopy-equivalences of a space with two nonvanishing homotopy groups. Proc. Amer. Math. Soc. 79, 134–138 (1980)
57. Turaev, V.G. 3-Dimensional Poincaré complexes: homotopy classification and splitting. Mat. Sbornik 180, 809–830 (1989). English translation: Math. USSR Sbornik 67, 261–282(1990)
58. Waldhausen, F. Algebraic K-theory of generalized free products. Ann. Math. 108, 135–256 (1978)
59. Wall, C.T.C. Finiteness conditions for CW complexes. II. Proc. Roy. Soc. London Series A 295, 129–139 (1966)
60. Wall, C.T.C. Surgery on Compact Manifolds. Academic Press, New York – London (1970)
61. Whitehead, J.H.C. A certain exact sequence. Ann. Math. 52, 51–110 (1950)
62. Whitehead, G.W. Elements of Homotopy Theory. Graduate Texts in Mathematics 61, Springer-Verlag, Berlin – Heidelberg – New York (1978)

Topologically flat embedded 2-spheres in specific simply connected 4-manifolds

Daniel Kasprowski, Peter Lambert-Cole, Markus Land, and Ana G. Lecuona

Abstract In this note we study whether specific elements in the second homology of specific simply connected closed 4-manifolds can be represented by smooth or topologically flat embedded spheres.

Introduction

Let X be a simply connected closed 4-manifold and consider an element in its second homology group. It is well known that any such element can be represented by an embedded closed oriented surface. Finding the minimal genus among all such representing surfaces is an interesting task; see [6] for a survey on this topic and [4, 9, 8] for some recent results. The aim of this note is to discuss specific examples in which it is possible (or not) to push the genus down to be zero, i.e. where it is possible (or not) to represent specific elements by embedded spheres. Notice that as X is simply connected, any element in the second homology can be represented by a sphere and since every topological 4-manifold is smoothable away from a point [10, Corollary 2.2.3], we can assume the sphere to be regularly immersed. Hence

Daniel Kasprowski
Rheinische Friedrich-Wilhelms-Universität Bonn, Mathematisches Institut, Endenicher Allee 60, 53115 Bonn, Germany, e-mail: kasprowski@uni-bonn.de

Peter Lambert-Cole
School of Mathematics, Georgia Institute of Technology, e-mail: plc@math.gatech.edu

Markus Land
Fakultät für Mathematik, Universität Regensburg, 93040 Regensburg, Germany
e-mail: markus.land@mathematik.uni-regensburg.de

Ana G. Lecuona
School of Mathematics & Statistics, University of Glasgow, University Place, Glasgow G12 8QQ
e-mail: Ana.Lecuona@glasgow.ac.uk

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_4

the above question is equivalent to asking whether this regularly immersed sphere is homotopic to an embedded one.

To be more precise, in this note we consider the manifolds $M = 8\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$, $M' = 8\mathbb{CP}^2\#_\star\overline{\mathbb{CP}^2}$ and $\mathbb{CP}^2\#M$. The groups $H_2(M;\mathbb{Z})$, $H_2(M';\mathbb{Z})$ and $H_2(\mathbb{CP}^2\#M;\mathbb{Z})$ will be considered with their 'evident' bases; for example, for $H_2(M;\mathbb{Z})$ the basis consists of 8 spheres of self-intersection 1, denoted $e_1,\dots,e_8$, and a last sphere e_9 of self-intersection -1. Within these groups and with respect to the evident bases, we will be interested in the elements $x = (1,\dots,1,3) \in H_2(M;\mathbb{Z})$, $x' = (1,\dots,1,3) \in H_2(M';\mathbb{Z})$ and $(0,x) \in H_2(\mathbb{CP}^2\#M;\mathbb{Z})$. The aim of this note is to show that x cannot be represented by a topologically flat embedded sphere while the elements x' and $(0,x)$ can be represented in such a way. At first we use the Kirby-Siebenmann invariant as an obstruction for x and Freedman's classification of simply connected manifolds to prove the statement for x' and $(0,x)$. In the second part of this paper, we reprove these statements using the Kervaire-Milnor invariant.

Existence of smooth or topologically flat embedded representatives

The intersection form of $M = 8\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$ is given by

$$\lambda_M = 8\langle 1\rangle \oplus \langle -1\rangle\ .$$

So we observe the following:

1. $x = (1,\dots,1,3)$ is a characteristic element for λ_M, i.e. $x\cdot y = y\cdot y \bmod 2$ for all $y \in H_2(M;\mathbb{Z})$, and $x\cdot x = -1$.
2. the orthogonal complement $\langle x\rangle^\perp$ of $\langle x\rangle$ is isomorphic to E_8: In fact the following elements give a basis of $\langle x\rangle^\perp$ whose representing matrix is the E_8-matrix: For $1 \le j \le 7$ take $f_j = e_{j+1} - e_j$, and let $f_8 = e_9 - e_6 - e_7 - e_8$.

Remark 1. The element x cannot be represented by a smoothly embedded sphere. Arguing by means of contradiction, we assume that x can be represented by a smoothly embedded sphere. Then the normal bundle of such an embedding is a complex line bundle with Euler class -1. Its associated disk bundle is thus diffeomorphic to $\overline{\mathbb{CP}}^2$ with an open disk removed and its sphere bundle is S^3. Removing the interior of a tubular neighbourhood of the embedded S^2 one obtains a manifold with boundary S^3 to which we can glue D^4. This construction is called a "blow-down" since it is the inverse to a "blow-up", i.e. taking connected sum with $\overline{\mathbb{CP}}^2$. The intersection form of the resulting smooth (and simply connected) manifold X is isometric to the orthogonal complement of $\langle x\rangle$, hence isometric to E_8. In particular, its intersection form is even, so that X is a smooth spin manifold. By Rokhlin's theorem the signature cannot be 8 and we obtain a contradiction.

In contrast to the smooth case considered in the previous remark, a topological spin 4-manifold with signature 8 exists by the work of Freedman. So one is led to ask whether x can be represented by a topologically flat embedding of S^2. We answer this question to the negative.

Theorem 1. *The element x cannot be represented by a topologically flat embedding $S^2 \to M$.*

Proof. As in the smooth case, a topologically flat embedding would give rise to a simply connected topological manifold with intersection form E_8. This manifold exists and is unique up to homeomorphism by Freedman's classification of simply connected topological 4-manifolds ([2, Theorem 1.5]). It is denoted by $E8$. More precisely, the topologically flat embedding would yield a homeomorphism $M \cong E8\#\overline{\mathbb{CP}^2}$. Recall that the Kirby–Siebenmann invariant KS is an obstruction for topological manifolds to admit a smooth (in fact PL) structure. It is a bordism invariant of 4-manifolds, hence we have

$$KS(M\#N) = KS(M) + KS(N)\,,$$

for all 4-manifolds M and N. As $E8$ is spin, we have $KS(E8) \equiv \sigma(E8)/8 \mod 2$ by [5, Theorem 13.1], where $\sigma(-)$ denotes the signature. As $\mathbb{CP}^2$ is smooth, we have $KS(\overline{\mathbb{CP}^2}) = 0$. Thus we find $KS(M) = 0$ but $KS(E8\#\mathbb{CP}^2) = 1$ so that these manifolds are not homeomorphic. It follows that x cannot be represented by a topologically flat embedding of S^2. □

Let us consider the following variant, namely the manifold $M' = 8\mathbb{CP}^2\#\overline{\star\mathbb{CP}^2}$, where $\star\mathbb{CP}^2$ is a fake $\mathbb{CP}^2$, i.e. a manifold homotopy equivalent to $\mathbb{CP}^2$, but with non-trivial Kirby–Siebenmann invariant. The existence of such a manifold again makes use of Freedman's theorem. One nice way to construct $\star\mathbb{CP}^2$, see [2, p. 370] where it is called the Chern manifold, is to consider the Poincaré 3-sphere, which by Freedman's theorem bounds a unique contractible 4-manifold. The trace of a $+1$-framed surgery on the trefoil knot produces a 4-manifold with boundary given by the Poincaré 3-sphere. Glueing together these two manifolds along their common boundary produces a simply connected 4-manifold with intersection form $\langle 1\rangle$, so this manifold is homotopy equivalent to $\mathbb{CP}^2$. However, one can check that its Kirby–Siebenmann invariant is non-trivial.

The intersection form of M' is thus also given by

$$\lambda_{M'} = 8\langle 1\rangle \oplus \langle -1\rangle\,,$$

and we consider again the element $x' = (1,\dots,1,3) \in H_2(M';\mathbb{Z})$.

Theorem 2. *The element x' can be represented by a topologically flat embedding $S^2 \to M'$.*

Proof. Since there is an isomorphism of forms $8\langle 1\rangle \oplus \langle -1\rangle \cong E8 \oplus \langle -1\rangle$ sending $x' = (1,\dots,1,3)$ to $(0,1)$ and $KS(8\mathbb{CP}^2\#\star\overline{\mathbb{CP}^2}) = KS(E8\#\overline{\mathbb{CP}^2}) = 1$, there

is a homeomorphism $8\mathbb{CP}^2\#\star\overline{\mathbb{CP}}^2 \cong E8\#\overline{\mathbb{CP}}^2$ which sends x' to a generator of $H_2(\overline{\mathbb{CP}}^2;\mathbb{Z})$ by Freedman's classification of simply connected topological manifolds ([2, Theorem 1.5 and its addendum]). The theorem now follows from the fact that the generator of $H_2(\overline{\mathbb{CP}}^2;\mathbb{Z})$ can be represented by a smoothly embedded sphere. □

Finally, we consider $\mathbb{CP}^2\#M = 9\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$ and the element $(0,x) = (0,1,\dots,1,3) \in H_2(\mathbb{CP}^2\#M;\mathbb{Z})$.

Theorem 3. *The element $(0,x)$ can be represented by a topologically flat embedding $S^2 \to \mathbb{CP}^2\#M$, but not by a smooth embedding.*

Proof. Again by Freedman's classification of simply connected topological manifolds, there is a homeomorphism $\mathbb{CP}^2\#M \cong \star\mathbb{CP}^2\#M'$ sending $(0,x)$ to $(0,x')$ and we know by Theorem 2 that $(0,x')$ can be represented by a topologically flat embedding.

As in Remark 1, if x were represented by a smooth embedding, there would be a smooth manifold with intersection form $E_8 \oplus \langle 1\rangle$. Since $E_8 \oplus \langle 1\rangle$ is definite but not diagonalizable, no compact, smooth, simply connected, orientable 4-manifold with this intersection form exists by Donaldson's theorem ([1, Theorem 1]). □

The Kervaire–Milnor invariant

The topological part of the above theorems can also be shown using the Kervaire–Milnor invariant km, introduced by Freedman and Quinn [3, Definition 10.8A]. For an immersed 2-sphere ι with an algebraically dual sphere g, i.e. $\lambda(\iota,g) = 0$, in a simply connected closed 4-manifold M the Kervaire–Milnor invariant takes values in $\mathbb{Z}/2$, if ι is s-characteristic, and it lives in the trivial group, if ι is not s-characteristic. Here, an immersed 2-sphere ι is called s-characteristic, if for every other immersed 2-sphere ι' one has $\lambda_M(\iota,\iota') \equiv \iota'\cdot\iota' \mod 2$.

One can describe $\mathrm{km}(\iota)$ as follows: Assume that the algebraic self-intersection number $\mu(\iota)$ vanishes (this can always be achieved by introducing local kinks, which only change the Euler number of the normal bundle by an even number). In this case, the geometric self-intersection points of ι can be paired up in couples with canceling signs. Therefore, one can choose a *framed* Whitney disc for each pair of self-intersections, and arrange that all the boundary arcs are disjoint. Then one counts the mod 2 number of intersection points of the interior of the Whitney disks with ι. The number obtained in this fashion then does not depend on the particular choice of Whitney disks, and the particular choices of changing ι to have algebraic self-intersection number 0.

Note that in general the Kervaire–Milnor invariant lives in the trivial group if ι is not r-characteristic, where ι is called r-characteristic if it is s-characteristic and for every immersion $\iota'\colon \mathbb{RP}^2 \to M$ we have for the $\mathbb{Z}/2$-intersection form that $\iota\cdot\iota' = \iota'\cdot\iota'$. However in a simply connected 4-manifold every immersion $\mathbb{RP}^2 \to M$

factors up to homotopy through $\mathbb{RP}^2/\mathbb{RP}^1 \simeq S^2$ and hence every s-characteristic sphere is r-characteristic.

Theorem 4 ([11, pp. 1310-1311]). *An immersed 2-sphere with an algebraically dual sphere is homotopic to a topologically flat embedded sphere if and only if it has trivial Kervaire–Milnor invariant.*

The problem of embedding spheres that represent homology classes of general odd divisibility was considered by Lee and Wilczyński [7].

Reproving Theorem 1

Recall that we are considering the manifold $M = 8\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$ and the element $x = (1,\dots,1,3) \in H_2(M;\mathbb{Z})$. The element x intersects the canonical 2-sphere representing $(1,0\dots,0) \in H_2(M;\mathbb{Z})$ in a single point and hence has an algebraically dual sphere. Furthermore, x is characteristic and hence s-characteristic. This implies that $\mathrm{km}(x)$ lives in $\mathbb{Z}/2$.

We can pick embedded spheres representing $1 \in H_2(\mathbb{CP}^2;\mathbb{Z})$ and an immersed sphere y in $\overline{\mathbb{CP}^2}$ representing $3 \in H_2(\overline{\mathbb{CP}^2};\mathbb{Z})$. We can add local kinks to y and pick Whitney disks inside $\overline{\mathbb{CP}^2}$ for y to compute $\mathrm{km}(y)$. The element x can be represented by taking a connected sum of the embedded spheres representing the generator in $H_2(\mathbb{CP}^2;\mathbb{Z})$ and y. We can take the connected sum in M avoiding the Whitney disks chosen for y. It follows that $\mathrm{km}(x) = \mathrm{km}(y)$. By [11, p. 1313], we have the formula

$$\mathrm{km}(\iota) \equiv (\iota\cdot\iota - \sigma(M))/8 + KS(M) \mod 2 \tag{1}$$

for an s-characteristic sphere ι in a simply connected 4-manifold. Thus

$$\mathrm{km}(y) \equiv (y\cdot y - \sigma(\overline{\mathbb{CP}^2}))/8 + KS(\overline{\mathbb{CP}^2}) = (-9-(-1))/8+0 = -1\,.$$

By Theorem 4, $\mathrm{km}(x) = \mathrm{km}(y) = 1$ implies that x cannot be represented by a topologically flat embedding.

Reproving Theorem 2

Recall that we are considering the manifold $M' = 8\mathbb{CP}^2\#\star\overline{\mathbb{CP}^2}$ and the element $x' = (1,\dots,1,3) \in H_2(M';\mathbb{Z})$. It has the same algebraically dual sphere as x. For x' we can consider the homeomorphism $M' \cong \star\mathbb{CP}^2\#7\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$. As for x, we see that $\mathrm{km}(x') = \mathrm{km}(y') + \mathrm{km}(y)$ where y' represents the generator of $H_2(\star\mathbb{CP}^2;\mathbb{Z})$. Using (1), we have

$$\mathrm{km}(y') \equiv (y'\cdot y' - \sigma(\star\mathbb{CP}^2))/8 + KS(\star\mathbb{CP}^2) = (1-1)/8+1 = 1\,.$$

Hence $\mathrm{km}(x') = 1 + 1 = 0$. By Theorem 4, x' can be represented by a topologically flat embedding.

Reproving Theorem 3

Recall that we are considering the manifold $\mathbb{CP}^2\#M = 9\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$ and the element $(0,x) = (0,1,\dots,1,3) \in H_2(\mathbb{CP}^2\#M;\mathbb{Z})$. Consider the element $z := (1,(1,0,\dots,0)) \in \mathbb{CP}^2\#M$. Then $\lambda(z,z) = 2$ and $\lambda(z,(0,x)) = 1$. Hence $(0,x)$ is not s-characteristic. Therefore $\mathrm{km}((0,x))$ lives in the trivial group and hence is itself trivial. By Theorem 4, $(0,x)$ can be represented by a topologically flat embedding.

Acknowledgements The authors thank the MATRIX Institute for hospitality during the workshop "Topology of Manifolds: Interactions Between High and Low Dimensions" where the question how to prove Theorem 1 came up during a discussion session. Most of the work on this article was done during this workshop. The authors also thank Peter Teichner for helpful discussions in particular about the Kervaire–Milnor invariant.

The first author was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - GZ 2047/1, Projekt-ID 390685813. The third author was supported by the SFB 1085 "Higher Invariants" in Regensburg.

References

1. Donaldson, S.K.: An application of gauge theory to four-dimensional topology. J. Differential Geom. **18**(2), 279–315 (1983). URL http://projecteuclid.org/euclid.jdg/1214437665
2. Freedman, M.H.: The topology of four-dimensional manifolds. J. Differential Geometry **17**(3), 357–453 (1982). URL http://projecteuclid.org/euclid.jdg/1214437136
3. Freedman, M.H., Quinn, F.: Topology of 4-manifolds, *Princeton Mathematical Series*, vol. 39. Princeton University Press, Princeton, NJ (1990)
4. Friedl, S., Vidussi, S.: Minimal genus in 4-manifolds with a free circle action. Adv. Math. **250**, 570–587 (2014). URL https://doi.org/10.1016/j.aim.2013.09.021
5. Kirby, R.C., Siebenmann, L.C.: Foundational essays on topological manifolds, smoothings, and triangulations. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo (1977). With notes by John Milnor and Michael Atiyah, Annals of Mathematics Studies, No. 88
6. Lawson, T.: The minimal genus problem. Exposition. Math. **15**(5), 385–431 (1997)
7. Lee, R., Wilczyński, D.M.: Locally flat 2-spheres in simply connected 4-manifolds. Comment. Math. Helv. **65**(3), 388–412 (1990). URL https://doi.org/10.1007/BF02566615
8. Nagel, M.: Minimal genus in circle bundles over 3-manifolds. J. Topol. **9**(3), 747–773 (2016). URL https://doi.org/10.1112/jtopol/jtw007
9. Nouh, M.A.: The minimal genus problem in $\mathbb{CP}^2\#\mathbb{CP}^2$. Algebr. Geom. Topol. **14**(2), 671–686 (2014). URL https://doi.org/10.2140/agt.2014.14.671
10. Quinn, F.: Ends of maps. III. Dimensions 4 and 5. J. Differential Geom. **17**(3), 503–521 (1982). URL http://projecteuclid.org/euclid.jdg/1214437139
11. Stong, R.: Existence of π_1-negligible embeddings in 4-manifolds. A correction to Theorem 10.5 of Freedmann and Quinn. Proc. Amer. Math. Soc. **120**(4), 1309–1314 (1994). URL https://doi.org/10.2307/2160253

Trisections of 5-Manifolds

Peter Lambert-Cole and Maggie Miller

Abstract Gay and Kirby introduced the notion of a trisection of a smooth 4-manifold, which is a decomposition of the 4-manifold into three ekementary pieces. Rubinstein and Tillmann later extended this idea to construct multisections of piecewise-linear manifolds in all dimensions. Given a PL manifold Y of dimension n, this is a decomposition of Y into $\lfloor n/2 \rfloor + 1$ PL submanifolds. We show that every smooth, oriented, compact 5-manifold admits a smooth trisection. Furthermore, given a smooth cobordism W between trisected 4-manifolds, there is a smooth trisection of W extending the trisections on its boundary.

Key words: trisection, topology, cobordism, 5-manifold

1 Introduction

We recall trisections of 4-manifolds, as introduced by Gay and Kirby [4].

Definition 1 ([4]). A $(g;k_1,k_2,k_3)$ *trisection* of a smooth, oriented, closed 4-manifold X is a triple (X_1,X_2,X_3) such that

- $X = X_1 \cup X_2 \cup X_3$ and $X_i \cap X_j = \partial X_i \cap \partial X_j$ for each $i \neq j$,
- Each $X_i \cong \natural_{k_i} S^1 \times B^3$ is a 4-dimensional 1-handlebody,
- Each double intersection $X_i \cap X_j$ is a 3-dimensional 1-handlebody, and
- The triple intersection $\Sigma = X_1 \cap X_2 \cap X_3$ is a closed, oriented surface of genus-g.

P. Lambert-Cole
University of Georgia, Athens, GA 30602, USA
e-mail: plc@uga.edu

M. Miller
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: maggiehm@mit.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_5

Moreover, every inclusion $X_i \hookrightarrow X$ is smooth in the interior of $\mathring{X}_i$ and piecewise smooth on ∂X_i.

Gay and Kirby proved that every closed, oriented smooth 4-manifold admits a trisection (which is unique up to a stabilization operation). The genesis was their study of Morse 2-functions, although they also showed that it is possible to build a trisection from a handle decomposition (see the proof of Theorem 4). Subsequently, Rubinstein and Tillmann generalized these ideas and found nice decompositions of piecewise-linear manifolds in all dimensions [8]. Given a PL manifold Y of dimension n, this is a decomposition of Y into $k = \lfloor n2 \rfloor + 1$ PL submanifolds, each of which is an n-dimensional 1-handlebody. The intersection of any j pieces, for $j = 2, \dots, k$, must satisfy further restrictions on their topology. When $n = 5$, the number of pieces is $\lfloor 5/2 \rfloor + 1 = 3$ and so PL 5-manifolds admit trisections as well.

Definition 2. A *smooth trisection* $\mathscr{M} = (Y_1, Y_2, Y_3)$ of a 5-manifold Y is a decomposition of Y into three pieces Y_1, Y_2, Y_3 so that the following conditions hold:

- $Y = Y_1 \cup Y_2 \cup Y_3$, where $Y_i \cap Y_j = \partial Y_i \cap \partial Y_j$ for $i \neq j$.
- Each Y_i is embedded in Y smoothly in its interior $\mathring{Y}_i$ and piecewise smoothly in ∂Y_i.
- For integers $k_1, k_2, k_3 \geq 0$, we have $Y_i \cong \natural_{k_i} S^1 \times B^4$.
- Each $Y_i \cap Y_j$ is a 4-manifold which is a regular neighborhood of its 2-skeleton.
- The triple intersection $Y_i \cap Y_j \cap Y_k$ is a 3-manifold properly embedded in Y.
- If $\partial Y \neq \emptyset$, then the triple $(Y_1 \cap \partial Y, Y_2 \cap \partial Y, Y_3 \cap \partial Y)$ is a trisection of ∂Y.

We refer to $Y_1 \cap Y_2 \cap Y_3$ as the *central submanifold* of $\mathscr{M}$.

Definition 2 agrees completely with the definition of [8] (in the 5-dimensional case), except where we require all inclusions to be smooth rather than piecewise linear.

From now on, "trisection" will always mean "smooth trisection." Our main theorems are the following:

Theorem 1. *Every closed, smooth, oriented 5-manifold Y admits a trisection.*

Theorem 2. *Let Y be smooth, oriented 5-manifold with positive boundary A and negative boundary B. Fix trisections $\mathscr{T}_A$ and $\mathscr{T}_B$ of A and B, respectively. There exist a trisection $\mathscr{T}_Y$ of Y whose restriction to A (resp. B) is $\mathscr{T}_A$ (resp. $\mathscr{T}_B$).*

Theorem 2 together with the fact that every closed 4-manifold admits a trisection [4] implies the following theorem.

Theorem 3. *Every compact, smooth, oriented 5-manifold Y admits a trisection.*

Similarly, Theorem 1 could be taken to be a consequence of Theorem 2.

Note that we do not consider the question of uniqueness of smooth trisections of 5-manifolds up to stabilization. In dimension 4, the central submanifold of a trisection is an orientable surface, and the stabilization operation increases the genus of this surface. Since any two orientable surfaces are related by such stabilization, it

is initially plausible that two trisections of a 4-manifold are related by stabilization. In contrast, a trisection of a 5-manifold has a 3-manifold as its central submanifold. The natural stabilization operation on the trisection adds a connect-summand of $S^1 \times S^2$ to this 3-manifold. In general, two 3-manifolds are not related by such stabilization, so we do not expect two arbitrary trisections of a 5-manifold to be related by stabilization.

Example 1. Fix coordinates (r,θ,x,y,z) on $\mathbb{R}^5$, where (r,θ) are polar coordinates and (x,y,z) are Cartesian. View S^5 as $\mathbb{R}^5 \cup \{\infty\}$. For $i = 1,2,3$, let $Y_i = \{2\pi(i-1)/3 \leq \theta \leq 2\pi i/3\} \cup \{\infty\}$. Then $\mathscr{T} = (Y_1, Y_2, Y_3)$ is a trisection of S^5.

On the other hand, we may view S^5 as $\partial(D^2 \times D^2 \times D^2)$. In this coordinate system, let $W_1 = S^1 \times D^2 \times D^2$, $W_2 = D^2 \times S^1 \times D^2$, $W_3 = D^2 \times D^2 \times S^1$. Then $\mathscr{T}' = (W_1, W_2, W_3)$ is a trisection of S^5.

The central submanifold of $\mathscr{T}$ is a a 3-sphere, while the central submanifold of $\mathscr{T}'$ is a 3-torus. Since $S^3 \#_m (S^1 \times S^2) \ncong T^3 \#_n (S^1 \times S^2)$ for any m,n, we conclude that $\mathscr{T}$ and $\mathscr{T}'$ have no common stabilization.

Question 1. Is there a suitable class of trisections of closed 5-manifolds and a "natural" set of stabilization operations which relate any two of these trisections which are of the same 5-manifold?

2 Trisecting Closed 5-Manifolds

In this section, we trisect closed 5-manifolds, which we might view as cobordisms between two empty manifolds. We begin by recalling how one trisects a closed 4-dimensional manifold.

2.1 Trisections of Closed 4-Manifolds

As preparation for the proof of Theorem 1, we review the construction in [4] of a trisection from a handle decomposition of a closed 4-manifold X. Roughly speaking, we partition the handles of X into three subsets — (1) the 0- and 1-handles; (2) the 2-handles; and (3) the 3-and 4-handles — and each group becomes one sector of the trisection.

Theorem 4 ([4]). *Every closed, oriented, smooth 4-manifold admits a trisection.*

Proof. Take a self-indexing Morse function f on X and let k_i be the the number of index-i critical points. Without loss of generality, assume $k_0 = k_4 = 1$. Identify the attaching link L of the 2-handles in the level set $f^{-1}(3/2)$ and choose a tubular neighborhood $\nu(L)$ in $f^{-1}(3/2)$. Choose a relative handle decomposition on the link complement $f^{-1}(3/2) \setminus \nu(L)$, which we can assume consists of 1-,2- and 3-handles.

Let H_1 be the union of the 2- and 3-handles of this decomposition and let H_2 be the union of $\nu(L)$ with the 1-handles. Then clearly H_1 and H_2 are 3-dimensional 1-handlebodies, meeting along a closed surface S. Equivalently, this gives a Heegaard splitting $f^{-1}(3/2) = H_1 \cup_S H_2$ of the level set.

The attaching link lies completely in H_2, so by flowing along a gradient vector field we can find the cylinder $H_1 \times [3/2, 5/2]$ in X. Define $X_1 = f^{-1}([0,3/2]) \cup H_1 \times [3/2,2]$; it retracts onto the sublevel set $f^{-1}([0,3/2])$ and so is a 1-handlebody. Similarly, define $X_3 = f^{-1}([5/2,4]) \cup H_1 \times [2,5/3]$; it retracts on the superlevel set $f^{-1}([5/2,4])$ and is also a 1-handlebody. Finally, let X_2 be the complement of $X_1 \cup X_3$ in X. Abstractly, it is diffeomorphic to $H_2 \times I \cup \{2\text{-handles}\}$. The manifold $H_2 \times I$ is a 1-handlebody and H_2 was obtained from $\nu(L)$ by attaching 1-handles. Thus each 2-handle cancels a unique 1-handle in $H_2 \times I$. Thus, the result is a 1-handlebody. Moreover, the double intersections $X_1 \cap X_2 = H_2$, $X_3 \cap X_1 = H_1$, $X_2 \cap X_3 =$ (the result of performing Dehn surgery to H_2 along the link $L \subset H_2$) are all 3-dimensional 1-handlebodies (note that L is contained in a 1-skeleton of H_2). The central submanifold is the surface $X_1 \cap X_2 \cap X_3 = S$.

2.2 Trisections of Closed 5-Manifolds

We can now describe how to obtain a trisection of a closed 5-manifold from a handle decomposition. The essential idea, as in the prequel, is to partition the handles into three sets: (1) the 0- and 1-handles; (2) the 2- and 3-handles; and (3) the 4- and 5-handles.

Proof (Proof of Theorem 1). Take a self-indexing Morse function f and let k_i be the the number of index-i critical points. Without loss of generality, assume $k_0 = k_5 = 1$. In the level set $f^{-1}(5/2)$, let S denote the attaching 2-spheres of the 3-handles above and let R denote the belt 2-spheres of the 2-handles below. We can assume they intersect transversely and then choose a tubular neighborhood $\nu(R \cup S)$. Choose a relative handle decomposition of $f^{-1}(5/2) \setminus \nu(R \cup S)$ which we can assume has no 0-handles. Let H_1 be the union of the 2-,3- and 4-handles of this handle decomposition and let H_2 be the union of $\nu(R \cup S)$ with the 1-handles. Clearly, H_1 and H_2 can be built with only 0-,1- and 2-handles and meet along a closed 3-manifold.

By flowing along a gradient vector field, we can find the cylinder $H_1 \times [3/2, 7/2]$ in Y. Define $Y_1 = f^{-1}([0,3/2]) \cup H_1 \times [3/2, 5/2]$; it retracts onto the sublevel set $f^{-1}([0,3/2])$ and so is a 1-handlebody. Similarly, define $Y_3 = f^{-1}([7/2,5]) \cup H_1 \times [5/2, 7/2]$; it retracts on the superlevel set $f^{-1}([7/2,5])$ and is also a 1-handlebody. Finally, let Y_2 be the complement of $Y_1 \cup Y_3$ in Y. Although it contains the 2- and 3-handles of Y, it is abstractly diffeomorphic to the union of $H_2 \times [0,1]$ with two collections of 3-handles. The 3-handles of Y are attached along the link $S \subset H_2 \times \{1\}$. By turning the 2-handles of Y upside down, we can view them as 3-handles attaching along $R \subset H_2 \times \{0\}$. Each of these three handles cancel a unique 2-handle in $H_2 \times [0,1]$. Moreover, these are the only 2-handles and so the result is a 1-handlebody. Moreover, the double intersections $Y_1 \cap Y_2 =$ (the result of surgering

H_2 along the belt spheres of the 2-handles), $Y_3 \cap Y_1 = H_1$, $Y_2 \cap Y_3 =$ (the result of surgering H_2 along the attaching spheres of the 3-handles) are all 4-dimensional 0, 1, 2-handlebodies (note that the belt spheres of the 2-handles and the attaching spheres of the 3-handles are contained in a 2-skeleton of H_2). The central submanifold is the 3-manifold $Y_1 \cap Y_2 \cap Y_3 = \partial H_1$.

3 Trisecting 5-Manifolds with Boundary

Tillmann and Rubinstein [8] do not fix a definition of a multisection of a manifold with boundary. A relative trisection of a 4-manifold X with boundary is well understood, having been originally introduced in [4] and fleshed out in [2]. A diagrammatic theory for relative trisections then appeared in [3], and has continued to appear throughout trisection literature. Briefly, a relative trisection of a 4-manifold with boundary is required to induce an open book on ∂X, so that relative trisections inducing the same boundary data can be glued to find trisections of the union. We give an analogous condition in this section.

3.1 Gluing Cobordisms

In order to build a trisection of Y from trisections of elementary pieces, we need to check that the topological conditions on a trisection hold after gluing together a pair of trisected cobordisms.

Let Y be a compact, smooth 5-manifold with boundary. Let $\mathscr{M} = (Y_1, Y_2, Y_3)$ be a trisection of Y, and let X be one of the boundary components of Y. A trisection $\mathscr{M}$ of Y is *compatible* with a trisection $\mathscr{T} = (X_1, X_2, X_3)$ of X if its restriction $\mathscr{M}|_X$ is equal to $\mathscr{T}$. A trisection $\mathscr{M}$ of Y is *strongly compatible* with $\mathscr{T}$ if it is compatible with $\mathscr{T}$ and the inclusion $X_i \hookrightarrow Y_i$ induces an injection $\pi_1(X_i) \hookrightarrow \pi_1(Y_i)$, in which each generator of the free group $\pi_1(X_i)$ maps to a distinct generator of the free group $\pi_1(Y_i)$.

We could alternatively state this condition as the inclusion maps a *core* of X_i to a *core* of Y_i, after isotopy. A *core* of an $(n > 3)$-dimensional genus-g handlebody H is a collection of g curves in the interior of H which generate $\pi_1(H)$. We may abuse notation and refer to one curve C in H as a core of H when $[C]$ is a generator of $\pi_1(H)$ (and thus a subset of a core of g curves).

If $\mathscr{M}$ is strongly compatible with its restriction to X, we also say $\mathscr{M}$ is strongly compatible with X. If $\mathscr{M}$ is strongly compatible with every boundary component, we say that $\mathscr{M}$ is strongly compatible with Y.

Let Y be a 5-manifold with boundary. To view Y as a cobordism of 4-manifolds, choose a decomposition $\partial Y = \partial Y_+ \sqcup (-\partial Y_-)$ (where one of $\partial Y_\pm$ may be empty). Suppose that W is another cobordism of 4-manifolds and there is a diffeomorphism

$\phi : \partial Y_+ \to \partial W_-$. Then we can *glue* Y to W and obtain a new cobordism $Y \cup_\phi W$ with boundary $\partial Y \cup_\phi \partial W = \partial W_+ \sqcup (-\partial Y_-)$.

Lemma 1. *Let Y and W be cobordisms of 4-manifolds and let $\phi : \partial Y_+ \to \partial W_-$ be a diffeomorphism. Suppose that $\mathscr{M}_Y = (Y_1, Y_2, Y_3)$ and $\mathscr{M}_W = (W_1, W_2, W_3)$ are strongly compatible trisections of Y and W (respectively) and that ϕ identifies $\mathscr{M}|_{\partial Y_+}$ with $\mathscr{M}|_{\partial W_-}$. Then $\mathscr{M}_{Y \cup_\phi W} = (Y_1 \cup W_1, Y_2 \cup W_2, Y_3 \cup W_3)$ is a trisection that is strongly compatible with $Y \cup_\phi W$.*

Proof. By definition, each of Y_i and W_i has a handle decomposition with only 0- and 1-handles. Since they are glued along a 1-handlebody, $Y_i \cup W_i$ has a handle decomposition with only 0-, 1-, and 2-handles. The 2-handles may be chosen to each run geometrically along a 1-handle of $Y_i \cap \partial Y_+$ and an identified 1-handle of $W_i \cap \partial W_-$ (as well as other 1-handles) once, since $\mathscr{M}_Y$ and $\mathscr{M}_W$ are strongly compatible with Y and W respectively. By assumption, the 2-handles can then be cancelled geometrically, so we conclude that $Y_i \cup W_i \cong \natural S^1 \times B^4$.

For $i \neq j$, we have $(Y_i \cup W_i) \cap (Y_j \cup W_j) = (Y_i \cap Y_j) \cup_{Y_i \cap Y_j \cap (\partial Y_+ \cong \partial W_-)} (W_i \cap W_j)$. Therefore, $(Y_i \cup W_i) \cap (Y_j \cup W_j)$ is obtained by gluing two 4-dimensional 0-,1-,2-handlebodies along a 3-dimensional handlebody. To glue along a handlebody, we need need only add 1- and 2-handles, so $(Y_i \cup W_i) \cap (Y_j \cup W_j)$ is a 0-, 1-, 2-handlebody, as desired.

The rest of Definition 2 follows easily.

3.2 Standard Trisections

For this subsection, refer to Figure 1. Our local models of trisections are obtained by pulling back a trisection on the unit disk D in $\mathbb{R}^2$. In radial coordinates, the symmetric trisection $D = D_1^s \cup D_2^s \cup D_3^s$ is defined by choosing the following subsets:

$$D_1^s = \left\{ 0 \leq \theta \leq \frac{2\pi}{3} \right\}, \quad D_2^s = \left\{ \frac{2\pi}{3} \leq \theta \leq \frac{4\pi}{3} \right\}, \quad D_3^s = \left\{ \frac{4\pi}{3} \leq \theta \leq 2\pi \right\}.$$

The symmetric trisection is symmetric under rotation by $2\pi/3$ (up to permuting indices). We also define a rectangular trisection $D = Y_1 \cup Y_2 \cup Y_3$ by setting

$$D_1^r = \{x \geq 0\}, \qquad D_2^r = \{x \leq 0, y \geq 0\}, \qquad D_3^r = \{x \leq 0, y \leq 0\}.$$

Geometrically, the rectangular trisection is asymmetric. Up to diffeomorphism, this trisection is equivalent to the symmetric trisection.

Definition 3. The *standard trisection* of S^k for $k \geq 2$ is the decomposition $\mathscr{T}_{std} = \{\pi^{-1}(D_i^s) \cap S^k\}$ where $\pi : \mathbb{R}^{k+1} \to \mathbb{R}^2$ is a coordinate projection and $D_1^s \cup D_2^s \cup D_3^s$ is the standard trisection of the unit disk in $\mathbb{R}^2$.

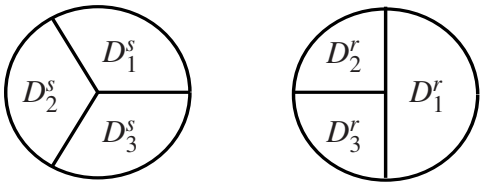


Fig. 1 Left: The symmetric trisection of the unit disk in $\mathbb{R}^2$. **Right:** The rectangular trisection of the unit disk in $\mathbb{R}^2$.

The *standard trisection* of B^k, for $k \geq 2$ is the decomposition $\mathscr{T}_{std} = \{\pi^{-1}(D_i^s) \cap B^k\}$ where $\pi : \mathbb{R}^k \to \mathbb{R}^2$ is a coordinate projection and $D_1^s \cup D_2^s \cup D_3^s$ is the standard trisection of the unit disk in $\mathbb{R}^2$.

When $k \notin \{4,5\}$, a multisection of S^k is not a trisection.

Note that we could have defined the standard trisection using the rectangular trisection of the disk rather than the symmetric trisection of the disk. Both definitions give equivalent (up to isotopy) trisections. Generally, we will use the coordinates of the rectangular trisection instead (for convenience). We will specify "standard symmetric trisection" or "standard rectangular trisection," but a reader comfortable with trisections may read this as "standard trisection."

Lemma 2. *Let $\mathscr{T}_{std} = (S_1, S_2, S_3)$ be the standard trisection of S^k. Then each S_i is diffeomorphic to B^k; each double intersection $S_i \cap S_j$ is diffeomorphic to B^{k-1} and the central surface is diffeomorphic to S^{k-2}.*

Let $\mathscr{T}_{std} = (D_1, D_2, D_3)$ be the standard trisection of B^k. Then each D_i is diffeomorphic to B^k; each double intersection $D_i \cap D_j$ is diffeomorphic to B^{k-1} and the central surface is diffeomorphic to B^{k-2}.

Proposition 1. *Let $\mathscr{T}_{std}$ be the standard trisection of S^4 and $\mathscr{M}_{std}$ be the standard trisection of B^5.*

1. *$\mathscr{T}_{std}$ is a trisection of S^4,*
2. *$\mathscr{M}_{std}$ is a trisection of B^5, and*
3. *$\mathscr{M}_{std}$ is strongly compatible with $\mathscr{T}_{std}$.*

The *standard trisection* of $S^1 \times S^3$ is obtained from the standard trisection of S^3 by taking the product of each sector with S^1. It is clear from Lemma 2 that this is a trisection.

3.3 Trisection Stabilization

Definition 4. Let $\mathscr{T} = (X_1, X_2, X_3)$ be a $(g; k_1, k_2, k_3)$-trisection of a 4-manifold X. We say that a $(g+1; k_1+1, k_2, k_3)$-trisection $\mathscr{T}'$ of X is *an elementary* 1-*stabilization of* $\mathscr{T}$ if there exists a boundary-parallel arc C properly embedded in $X_2 \cap X_3$ so that

$$X_1' = X_1 \cup \overline{\nu(C)}\,,$$
$$X_2' = X_2 \setminus \nu(C)\,,$$
$$X_3' = X_3 \setminus \nu(C)\,,$$

for some fixed open neighborhood $\nu(C)$ of C. We say that C is the *stabilization arc* of the stabilization $\mathscr{T} \mapsto \mathscr{T}'$.

We similarly define elementary 2- and 3-stabilization by permuting the indices $1,2$, and 3.

Lemma 3. *Let X be a closed 4-manifold. Fix a $(g;k_1,k_2,k_3)$-trisection $\mathscr{T} = (X_1,X_2,X_3)$ of X. Let $\mathscr{T}'$ be an elementary 1-stabilization of $\mathscr{T}$, so that $\mathscr{T}' = (X_1',X_2',X_3')$ is a $(g+1;k_1+1,k_2,k_3)$-trisection of X. There exists a smooth multisection $\mathscr{M} = (Y_1,Y_2,Y_3)$ of $X \times I$ whose restriction to $X \times \{0\}$ is $\mathscr{T}$ and whose restriction to $X \times \{1\}$ is $\mathscr{T}'$.*

Proof. See Figure 2 for a schematic.

Let C be the stabilization arc of $\mathscr{T} \mapsto \mathscr{T}'$. For $i = 1,2,3$, let $Y_i := (X_i \times [0,1/2]) \cup (X_i' \times [1/2,1])$, so that $Y = Y_1 \cup Y_2 \cup Y_3$. We immediately have $Y_i \cap (X \times 0) = X_i$ and $Y_i \cap (X \times 1) = X_i$. Obviously this implies that $\mathscr{M} = (Y_1,Y_2,Y_3)$ induces a trisection on ∂Y.

Note that for each i and j, Y_i and $Y_i \cap Y_j$ strongly deformation retract onto $Y_i \cap (X \times 1/2)$ and $(Y_i \cap Y_j) \cap (X \times 1/2)$, respectively. We will describe each of these intersections.

$$Y_1 \cap \left(X \times \frac{1}{2}\right) = X_1' \times \frac{1}{2}\,,$$
$$Y_2 \cap \left(X \times \frac{1}{2}\right) = X_2 \times \frac{1}{2}\,,$$
$$Y_3 \cap \left(X \times \frac{1}{2}\right) = X_3 \times \frac{1}{2}\,.$$

We conclude that $Y_1 \cong \natural_{k_1+1} S^1 \times B^4$, $Y_2 \cong \natural_{k_2} S^1 \times B^4$, and $Y_1 \cong \natural_{k_3} S^1 \times B^4$. Moreover,

$$Y_1 \cap Y_2 \cap \left(X \times \frac{1}{2}\right) = \left((X_1 \cap X_2) \cup \left(X_2 \cap \overline{\nu(C)}\right)\right) \times \frac{1}{2}\,,$$
$$Y_2 \cap Y_3 \cap \left(X \times \frac{1}{2}\right) = (X_2 \cap X_3) \times \frac{1}{2}\,,$$
$$Y_3 \cap Y_1 \cap \left(X \times \frac{1}{2}\right) = \left((X_1 \cap X_3) \cup \left(X_3 \cap \overline{\nu(C)}\right)\right) \times \frac{1}{2}\,.$$

Then immediately, $Y_2 \cap Y_3 \cong \natural_g S^1 \times B^3$. Moreover, we note that $Y_1 \cap Y_2 \cap (X \times 1/2)$ is obtained from the 3-dimensional handlebody $(X_1 \cap X_2) \times 1/2$ by attaching a 4-dimensional 1-handle, so strongly deformation retracts to a 1-skeleton. Therefore, $Y_1 \cap Y_2 \cong \natural_{g+1} S^1 \times B^3$. Similarly, $Y_3 \cap Y_1 \cong \natural_{g+1} S^1 \times B^3$.

Finally, we have that $Y_1 \cap Y_2 \cap Y_3$ is the 3-dimensional trace of the cobordism from $X_1 \cap X_2 \cap X_3$ to $X_1' \cap X_2' \cap X_3'$ obtained by attaching the 3-dimensional 1-handle $\overline{(\nu(C) \smallsetminus X_1)}$. That is, $Y_1 \cap Y_2 \cap Y_3$ is a compression body from a genus $(g+1)$-surface to a genus g-surface.

We will refer to the trisected 5-manifold Y of Lemma 3 as a stabilization cobordism. Figure 2 shows a schematic of a stabilization cobordism.

Proposition 2. *Let X be a closed, oriented, smooth 4-manifold and let $\mathscr{T}_0, \mathscr{T}_1$ be trisections of X. There exists a trisection $\mathscr{M} = (Y_1, Y_2, Y_3)$ of $X \times [0,1]$ whose restriction to $X \times \{i\}$ is $\mathscr{T}_i$.*

Proof. By [4, Theorem 11], there exists a common stabilization $\widetilde{\mathscr{T}}$ of $\mathscr{T}_0$ and $\mathscr{T}_1$. Therefore, the claim holds by induction on Lemmas 3 and 1.

3.4 Morse Theory for Manifolds with Boundary

For a comprehensive treatment, we refer the reader to [1].

Let f be a Morse function on X; for the sake of exposition we assume the critical values are all distinct. By the Morse lemma, we can choose coordinates around every nondegenerate critical point of f in which the function takes the form

$$f(x_1, \ldots, x_n) = x_1^2 + \cdots + x_{n-k}^2 - x_{n-k+1}^2 - \cdots - x_n^2 \tag{1}$$

for some k, which is called the *index* of the critical point. Let $X_s = f^{-1}((\infty, s])$. Up to diffeomorphism, the sublevel set X_ε can be obtained from the sublevel set $X_{-\varepsilon}$ by attaching a k-handle along some S^{k-1} in the level set $f^{-1}(-\varepsilon)$.

Now suppose that X has nonempty boundary and f is a Morse function on X that restricts to a Morse function on ∂X. If $z \in \partial X$ is a critical point of f, we can find Morse coordinates near z as in Equation 1 and such that ∂X is sent to the hyperplane $\{x_j = 0\}$ for some j. The critical point z is *boundary unstable* if $1 \le j \le n-k$ and is *boundary stable* if $n-k+1 \le j \le n$.

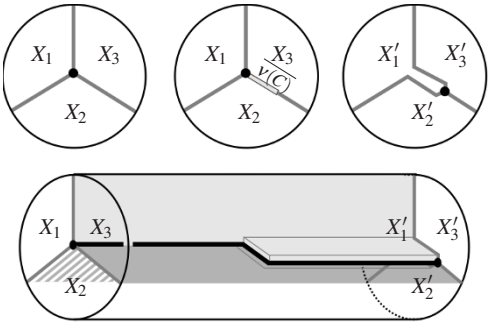


Fig. 2 A schematic of a stabilization cobordism corresponding to an elementary 1-stabilization about arc $C \subset X_2 \cap X_3$. This is a cobordism from X to X inducing trisection $\mathscr{T}$ on the left boundary and $\mathscr{T}'$ on the right boundary (where $\mathscr{T}'$ is obtained from $\mathscr{T}$ by 1-stabilization), as in Lemma 3.

The topological change to a sublevel set when f has a Morse critical point on the boundary depends on whether the critical point is boundary stable or boundary unstable.

Proposition 3. *Let $z \in \partial X$ be a Morse critical point of index-k.*

1. *If z is boundary stable, then X_ε is diffeomorphic to $X_{-\varepsilon}$. Furthermore, $(\partial X)_\varepsilon$ is obtained from $(\partial X)_{-\varepsilon}$ by attaching a handle of index-$k-1$.*
2. *If z is boundary unstable, then X_ε is obtained from $X_{-\varepsilon}$ by attaching a handle of index-k. Furthermore, $(\partial X)_\varepsilon$ is obtained from $(\partial X)_{-\varepsilon}$ by attaching a handle of index-k.*

Proof. The statements about the topology of ∂X are standard Morse theory. The statements about the topology of X are the combination of Lemmas 2.18 and 2.19 and Theorem 2.27 in [1].

3.5 Index-1

In this and the following subsections, let Y be a cobordism between closed 4-manifolds X and Z; let $\mathcal{T}_X = (X_1, X_2, X_3)$ and $\mathcal{T}_Z = (Z_1, Z_2, Z_3)$ be trisections of X and Z, respectively; and let $f : Y \to [0,1]$ be a relative Morse function so that $\partial Y = f^{-1}(\{0,1\})$, where $X = f^{-1}(0)$ and $Z = f^{-1}(1)$.

Suppose that S is an embedded S^0 in the central surface Σ of the trisection $\mathcal{T}_X$. Let $\nu_\Sigma(S)$ be a tubular neighborhood in the central surface. Then we can choose a tubular neighborhood $\nu_X(S) \cong \nu_\Sigma(S) \times D^2$ such that $\mathcal{T}_X$ restricts to a trisection obtained by pulling back the standard (which we view to be the rectangular) trisection of the disk under the projection $\nu_X(S) \to D^2$. Note that the trisection determines a framing of the bundle $\nu_\Sigma(S) \times D^2$, but this framing is unique up to isotopy since $\nu_\Sigma(S) \cong S^0 \times D^2$.

Proposition 4. *Suppose that there is a unique critical point of f of index-1 in the interior of Y. There exists a trisection $\mathcal{M} = (Y_1, Y_2, Y_3)$ of Y that is strongly compatible with the trisections $\mathcal{T}_X$ and $\mathcal{T}_Z$.*

Proof. First we describe the local model in Morse coordinates. Near a Morse critical point of index 1, we have Morse coordinates such that

$$f = x_1^2 + x_2^2 + x_3^2 + x_4^2 - x_5^2 \,.$$

We can view this as a function on $\mathbb{R}^2 \times \mathbb{R}^3$ and decompose f as $g + \widetilde{f}$, where

$$g(x_1, x_2) = x_1^2 + x_2^2 \,, \qquad \widetilde{f}(x_3, x_4, x_5) = x_3^2 + x_4^2 - x_5^2 \,.$$

Using the projection $\pi : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}^2$, we obtain a trisection near the Morse critical point (i.e. of a small 5-ball centered about the critical point) by pulling back the

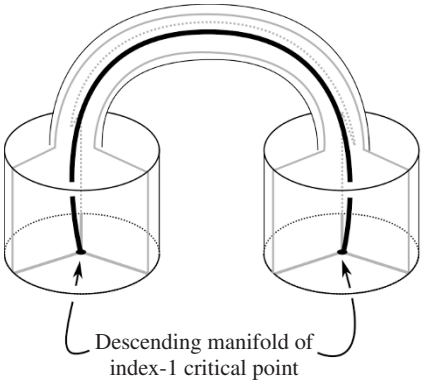


Fig. 3 The construction of Proposition 4. We trisect a cobordism Y from X to Z which includes an index-1 critical point. We isotope so that the descending manifold of the index-1 point intersects X in the central surface of a trisection $\mathcal{T}_X$ on X. We then trisect the 1-handle radially.

rectangular trisection of the disk, as described at the beginning of this subsection. In this model, the central submanifold is the hyperplane $\{x_1 = x_2 = 0\}$ and the restriction of $\widetilde{f}$ to the central submanifold is a Morse function with a critical point of index 1. We take this local model to agree with the restriction of $\mathcal{T}_X$ near the critical point.

Let g be the standard Euclidean metric on $\mathbb{R}^5$ and ∇f the gradient of f with respect to g. The desending manifold of the critical point, with respect to ∇f, is contained in the line $\{x_1 = x_2 = x_3 = x_4 = 0\}$ and intersects the level set $f^{-1}(-\varepsilon)$ in the 0-sphere $R = (0,0,0,0,\pm\sqrt{\varepsilon})$. Let $\widetilde{R} = (0,0,\pm\sqrt{\varepsilon}) \subset \mathbb{R}^3$ be its projection. Let $\nu(\widetilde{R}) \subset \mathbb{R}^3$ be a tubular neighborhood of $\widetilde{R}$ in $\widetilde{f}^{-1}(-\varepsilon)$. Flowing along $\nabla\widetilde{f}$, we obtain tubular neighborhoods of $(0,0,\pm(\sqrt{\varepsilon+\delta})$ in $\widetilde{f}^{-1}(-\varepsilon-\delta))$ for all $\delta > 0$. We can find a tubular neighborhood $\nu(R)$ of R in $f^{-1}(-\varepsilon)$ of the form

$$\nu(R) \cong \nu(\widetilde{R}) \times D^2 = \nu(\widetilde{R}) \times g^{-1}\left(\left[0, \frac{\varepsilon}{2}\right]\right) .$$

The trisection of the local model restricts a trisection of $\nu(R)$ obtained by pulling back the rectangular trisection of the disk under the projection $\nu(R) \to D^2$.

Via an identification

$$\nu_X(S) \cong \nu_\Sigma(S) \times D^2 \cong \nu(\widetilde{R}) \times D^2 \cong \nu(R) ,$$

we can use this model to extend a trisection from below the critical point to above the critical point. See Figure 3.

The topological result is as follows. In the local model, The central submanifold is the hyperplane $\{x_1 = x_2 = 0\}$ and the function f restricts to a Morse function with a critical point of index-1. Thus, moving from height $-\varepsilon$ to height ε results in surgery on $S \subset \Sigma$, increasing the genus by 1. The double intersection $H_1 = D_1^r \cap D_3^r$ is $\{x_1 \geq 0; x_2 = 0\}$ and the restriction of f restricts has a boundary unstable critical point of index 1. Thus, topologically moving from height $-\varepsilon$ to height ε adds a 1-handle to H_1 along S. By symmetry, this is also true of the remaining double intersections. Finally, the top dimensional sector D_1^r, in the rectangular model of the

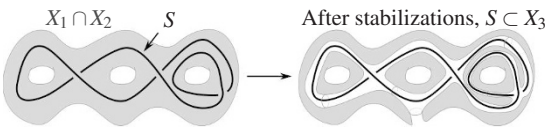


Fig. 4 In Lemma 4, we isotope S to lie in $X_1 \cap X_2$ and then 3-stabilize until S is contained in a core of X_3.

trisection, is the halfspace $\{x_1 \geq 0\}$. The restriction of f has a boundary unstable critical point of index 1 and so the topological effect when moving from height $-\varepsilon$ to ε is to add a 1-handle to D_1^r along S. Again by symmetry, this is true of the remaining sectors.

The local model gives a trisection $\mathscr{M}'$ on Y that restricts to $\mathscr{T}_X$ on X and some $\mathscr{T}_Z' = (Z_1', Z_2', Z_3')$ on Z. The sector X_i is a 4-dimensional 1-handlebody of genus k_i. The sector Y_i is obtained by thickening X_i and attaching a 1-handle. Thus clearly the inclusion $X_i \hookrightarrow Y_i$ includes the cores of X_i into the cores of Y_i. Moreover, the sector Y_i retracts onto Z_i'. Thus $\mathscr{M}'$ is strongly compatible. By Proposition 2, we can find a trisection on $Z \times I$ strongly compatible with $\mathscr{T}_Z'$ and $\mathscr{T}_Z$. Then by Lemma 1, we can glue these together to obtain the required trisection.

3.6 Index-2

Lemma 4. *Let S be some embedded S^1 in a 4-manifold X. After stabilizing the trisection $\mathscr{T}_X$ and isotopy of S, we can assume that S lies in the central surface Σ of $\mathscr{T}_X$ so that it includes in each 3D and 4D piece as a core, and that the framing induced on S by Σ may be chosen arbitrarily.*

Proof. Since $X_1 \cap X_2$ generates the unbased fundamental group of X, we may isotope S to lie in $X_1 \cap X_2$, disjoint from a 1-skeleton of $X_1 \cap X_2$. Let $\pi(S)$ denote the projection of S onto $\Sigma = \partial(X_1 \cap X_2)$. Take $\pi(S)$ to have only double points of self-intersection, and let $c(S)$ denote the number of such double points. Then we may perform $2 + c(S)$ 3-stabilizations (see Figure 4) so that S is contained in a core of X_3. Perform a 1- and a 2-stabilization, taking S to run through the core of each added genus to X_1 and X_2, and project S onto Σ to obtain an embedded curve C in Σ. By construction, this projection can be taken to be an isotopy.

Let A be the α curve in a triple of α, β, γ curves arising from the 1-stabilization, so A is parallel to a β curve, intersects a γ curve in point, and intersects C in one point. Since A bounds a disk whose interior is disjoint from Σ, $\mu(A) = 0$, where $\mu : H_1(\Sigma; \mathbb{Z}) \to \mathbb{Z}/2$ is the Rokhlin quadratic form. Let C' be a curve in Σ obtained by Dehn twisting C about A (in either direction). Then $\mu(C) \neq \mu(C')$. Note C' is isotopic to C in X, so C' is isotopic to S. Both C and C' include into each 3D and 4D piece of $\mathscr{T}$ as subsets of cores. Because there are only two possible framings on $S \subset X$, one of C or C' is the desired curve.

Proposition 5. *Suppose that there is a unique critical point of f of index-2 in the interior of Y. There exists a trisection $\mathscr{M} = (Y_1, Y_2, Y_3)$ of Y that is strongly compatible with the trisections $\mathscr{T}_X$ and $\mathscr{T}_Z$.*

Proof. The proof is analogous to the proof of Proposition 4, so we only sketch it. Near a Morse critical point of index 2, we have Morse coordinates such that

$$f = x_1^2 + x_2^2 + x_3^2 - x_4^2 - x_5^2 \, .$$

We can view this as a function on $\mathbb{R}^2 \times \mathbb{R}^3$ and decompose f as $g + \widetilde{f}$, where

$$g(x_1, x_2) = x_1^2 + x_2^2 \, , \qquad \widetilde{f}(x_3, x_4, x_5) = x_3^2 - x_4^2 - x_5^2 \, .$$

Using the projection $\pi : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}^2$, we obtain a trisection near the Morse critical point by pulling back the rectangular trisection of the disk. In this model, the central submanifold is the hyperplane $\{x_1 = x_2 = 0\}$ and the restriction of $\widetilde{f}$ to the central submanifold is a Morse function with a critical point of index 2. The desending manifold of the critical point, with respect to ∇f, is contained in the plane $\{x_1 = x_2 = x_3 = 0\}$ and intersects the level set $f^{-1}(-\varepsilon)$ in the 1-sphere R. Let $\widetilde{R} = \{(0, a, b) \mid a^2 + b^2 = \varepsilon \subset \mathbb{R}^3$ be its projection. Let $\nu(\widetilde{R}) \subset \mathbb{R}^3$ be a tubular neighborhood of $\widetilde{R}$ in $\widetilde{f}^{-1}(-\varepsilon)$. We can find a tubular neighborhood $\nu(R)$ of R in $f^{-1}(-\varepsilon)$ of the form

$$\nu(R) \cong \nu(\widetilde{R}) \times D^2 = \nu(\widetilde{R}) \times g^{-1}\left(\left[0, \frac{\varepsilon}{2}\right]\right) \, .$$

The trisection of the local model restricts a trisection of $\nu(R)$ obtained by pulling back the rectangular trisection of the disk under the projection $\nu(R) \to D^2$. By Lemma 4, we can take this trisection to agree with $\mathscr{T}_X$. Via an identification

$$\nu_X(S) \cong \nu_\Sigma(S) \times D^2 \cong \nu(\widetilde{R}) \times D^2 \cong \nu(R) \, ,$$

we can use this model to extend a trisection from below the critical point to above the critical point.

The topological result is as follows. In the local model, The central submanifold is the hyperplane $\{x_1 = x_2 = 0\}$ and the function f restricts to a Morse function with a critical point of index 2. Thus, moving from height $-\varepsilon$ to height ε results in surgery on $S \subset \Sigma$, decreasing the genus by 1. The double intersection $H_1 = D_1^r \cap D_3^r$ is $\{x_1 \geq 0; x_2 = 0\}$ and the restriction of f restricts has a boundary unstable critical point of index 2. Thus, topologically this adds a 2-handle to H_1 along S; however, by assumption, this handle is attached along a core and therefore cancels a 1-handle in H_1. By symmetry, this is also true of the remaining double intersections. Finally, the top dimensional sector D_1^r, in the rectangular model of the trisection, is the halfspace $\{x_1 \geq 0\}$. The restriction of f has a boundary unstable critical point of index 2 and so the topological effect of moving from height $-\varepsilon$ to ε is to add a 2-handle to D_1^r along S. Since S represents a core, this 2-handle cancels a 1-handle. Again by symmetry, this is true of the remaining sectors.

Strong compatibility follows as in the proof of Proposition 4. Then we can glue this local model to a model that changes the trisection on Z to obtain the required trisection of Y.

We remark that we now have all of the technology needed to prove Theorem 2. For now, if f is self-indexing with only index-1, -2, -3, -4 points, we may trisect $f^{-1}[0,5/2]$ and $f^{-1}[5/2,5]$ separately (by turning $f^{-1}[5/2,5]$ upside down) and then glue the two copies of $f^{-1}(5/2)$. However, we instead continue to build the trisection upward from level 0 for completion of the analogy between the handle structure and the trisection structure.

3.7 Index-3

Lemma 5. *Let S be an embedded $S^2 \subset X$. By an isotopy, we can assume that S is in 1-bridge position with respect to a stabilization of $\mathscr{T}_X$.*

Proof. This is a specialization of [7, Theorem 1.2].

Proposition 6. *Suppose that there is a unique critical point of f of index-3 in the interior of Y. There exists a trisection $\mathscr{M} = (Y_1, Y_2, Y_3)$ of Y that is strongly compatible with the trisections $\mathscr{T}_X$ and $\mathscr{T}_Z$.*

As hinted at the end of Subsection 3.6, this Proposition follows from Proposition 5 by replacing f with $-f$. However, here we give a direct proof without changing perspective.

Proof. This model can be obtained by turning the model in Proposition 5 upside-down. Near a Morse critical point of index 3, we have Morse coordinates such that

$$f = x_1^2 + x_2^2 - x_3^2 - x_4^2 - x_5^2 \, .$$

We can view this as a function on $\mathbb{R}^3 \times \mathbb{R}^2$ and decompose f as $\widetilde{f} - g$, where

$$\widetilde{f}(x_1,x_2,x_3) = x_1^2 + x_2^2 - x_3^2 \, , \qquad g(x_4,x_5) = x_4^2 + x_5^2 \, .$$

Using the projection $\pi : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^2$, we obtain a trisection near the Morse critical point by pulling back the rectangular trisection of the disk. In this model, the central submanifold is the hyperplane $\{x_4 = x_5 = 0\}$ and the restriction of $\widetilde{f}$ to the central submanifold is a Morse function with a critical point of index 1.

Now, however, the descending manifold intersects $f^{-1}(-\varepsilon)$ along the 2-sphere $R = \{-x_3^2 - x_4^2 - x_5^2 = -\varepsilon\}$. The trisection of the local model restricts to give the standard rectangular trisection of R. We can choose a neighborhood $\nu(R)$ and a splitting $\nu(R) = R \times D^2$ such that the trisection of local model, restricted to $\nu(R)$, is the same as the trisection obtained by pulling back the standard trisection of S^2 by the projection $\nu(R) \to R$.

Now suppose S is in 1-bridge position via Lemma 5. This means that $\mathscr{T}_X$ restricted to S is exactly the standard rectangular trisection of S^2 (up to isotopy). We can choose an identification $\nu(S) = S \times D^2$ such that $\mathscr{T}_X$ restricted to $\nu(S)$ is exactly the trisection obtained by pulling back the standard rectangular trisection of S^2 by the projection $\nu(S) \to S$. Via an identification

$$\nu_X(S) \cong \nu_\Sigma(S) \times D^2 \cong \nu(\widetilde{R}) \times D^2 \cong \nu(R) \,,$$

we can use this model to extend a trisection from below the critical point to above the critical point.

The topological result is as follows. In the local model, The central submanifold is the hyperplane $\{x_4 = x_5 = 0\}$ and the function f restricts to a Morse function with a critical point of index 1. Thus, moving from height $-\varepsilon$ to height ε results in surgery on $S \subset \Sigma$, increasing the genus by 1. The double intersection $H_1 = D_1^r \cap D_3^r$ is $\{x_1 \geq 0; x_2 = 0\}$ and the restriction of f restricts has a boundary stable critical point of index 2. Thus, moving from height $-\varepsilon$ to height ε does not change the topology. The top dimensional sector, in the rectangular model of the trisection, is the plane $\{x_5 \geq 0\}$. This is a boundary stable critical point of index 3, so moving from height $-\varepsilon$ to height ε does not change the topology.

Strong compatibility follows as in the proof of Proposition 4. Then we can glue this local model to a model that changes the trisection on Z to obtain the required trisection of Y.

3.8 Index-4

An $S^3 \subset X^4$ is in *standard position* with respect to some trisection $\mathscr{T} = (X_1, X_2, X_3)$ of X if $S \pitchfork \Sigma$ is a simple closed curve c that bounds disks in all three handlebodies $X_i \cap X_j$. When an embedded S^3 is in standard position with respect to $\mathscr{T}$, the restriction of $\mathscr{T}$ is exactly the standard rectangular trisection of S^3 (up to isotopy). Note that if the curve c is separating, this is a locally a model for connected sum; if the curve is nonseparating, this is a model for an $S^1 \times S^3$ factor.

Lemma 6. *Let S be an embedded $S^3 \subset X$. By an isotopy and a stabilization of the trisection $\mathscr{T}_X$, we can assume that S is in standard position with respect to $\mathscr{T}_X$.*

Proof. Suppose that S is separating. Then X decomposes as a connected sum $X = X_1 \#_S X_2$. Choose trisections $\mathscr{T}_1$ and $\mathscr{T}_2$ of X_1 and X_2, respectively. X therefore admits a trisection $\mathscr{T}_S = \mathscr{T}_1 \# \mathscr{T}_2$ and S is in standard position with respect to $\mathscr{T}_S$. The trisections $\mathscr{T}$ and $\mathscr{T}_S$ admit a common stabilization $\widetilde{T}$. Furthermore, 1-stabilization preserves the fact that S is in standard position.

Now suppose S is nonseparating. Let γ be a closed curve that intersects S transversely in a single point. Then X decomposes as a connected sum along the boundary of the tubular neighborhood $\nu(S \cup \gamma)$ into $X' \# S^1 \times S^3$. Let $\mathscr{T}'$ be a trisection of X' and let $\mathscr{T}_{sphere}$ be the standard trisection of $S^1 \times S^3$. The sphere S is isotopic to

$\{pt\} \times S^3$ and this sphere is in standard position with respect to $\mathscr{T}_{sphere}$. The connected sum $\mathscr{T}' \# \mathscr{T}_{sphere}$ is a trisection of X. Again, the trisections $\mathscr{T}$ and $\mathscr{T}' \# \mathscr{T}_{sphere}$ have a common stabilization $\widetilde{T}$ and S is in standard position with respect to this trisection.

Proposition 7. *Suppose that there is a unique critical point of f of index-4 in the interior of Y. There exists a trisection $\mathscr{M} = (Y_1, Y_2, Y_3)$ of Y that is strongly compatible with the trisections $\mathscr{T}_X$ and $\mathscr{T}_Z$.*

As hinted at the end of Subsection 3.6, this Proposition follows from Proposition 4 by replacing f with $-f$. However, here we give a direct proof without changing perspective.

Proof. This model can be obtained by turning the model in Proposition 4 upside-down. Near a Morse critical point of index 4, we have Morse coordinates such that

$$f = x_1^2 - x_2^2 - x_3^2 - x_4^2 - x_5^2 \, .$$

We can view this as a function on $\mathbb{R}^3 \times \mathbb{R}^2$ and decompose f as $\widetilde{f} - g$, where

$$\widetilde{f}(x_1, x_2, x_3) = x_1^2 - x_2^2 - x_3^2 \, , \qquad g(x_4, x_5) = x_4^2 + x_5^2 \, .$$

Using the projection $\pi : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^2$, we obtain a trisection near the Morse critical point by pulling back the standard rectangular trisection of the disk. In this model, the central submanifold is the hyperplane $\{x_4 = x_5 = 0\}$ and the restriction of $\widetilde{f}$ to the central submanifold is a Morse function with a critical point of index 2.

Now, however, the descending manifold intersects $f^{-1}(-\varepsilon)$ along the 3-sphere $R = \{-x_2^2 - x_3^2 - x_4^2 - x_5^2 = -\varepsilon\}$. The trisection of the local model restricts to give the standard rectangular trisection of R. We can choose a neighborhood $\nu(R)$ and a splitting $\nu(R) = R \times D^1$ such that the trisection of local model, restricted to $\nu(R)$, is the same as the trisection obtained by pulling back the standard trisection of S^3 by the projection $\nu(R) \to R$.

Now suppose S is in standard position (using Lemma 6). This means that $\mathscr{T}_X$ restricted to S is exactly the standard trisection of S^3. We can choose an identification $\nu(S) = S \times D^1$ such that $\mathscr{T}_X$ restricted to $\nu(S)$, is exactly the trisection obtained by pulling back the standard rectangular trisection of S^3 by the projection $\nu(S) \to S$. Via an identification

$$\nu_X(S) \cong \nu_\Sigma(S) \times D^1 \cong \nu(\widetilde{R}) \times D^1 \cong \nu(R) \, ,$$

we can use this model to extend a trisection from below the critical point to above the critical point.

The topological result is as follows. In the local model, The central submanifold is the hyperplane $\{x_4 = x_5 = 0\}$ and the function f restricts to a Morse function with a critical point of index 2. Thus, moving from height $-\varepsilon$ to height ε results in surgery on $S \subset \Sigma$, decreasing the genus by 1. The double intersection $H_1 = D_1^r \cap D_3^r$ is $\{x_1 \geq 0; x_2 = 0\}$ and the restriction of f restricts has a boundary stable critical

point of index 3. Thus, moving from height $-\varepsilon$ to height ε does not change the topology. The top dimensional sector, in the rectangular model of the trisection, is the plane $\{x_5 \geq 0\}$. This is a boundary stable critical point of index 4, so moving from height $-\varepsilon$ to height ε does not change the topology.

Strong compatibility follows as in the proof of Proposition 4. Then we can glue this local model to a model that changes the trisection on Z to obtain the required trisection of Y.

Theorem 5. *Let Y be a cobordism between smooth, closed, connected, nonempty 4-manifolds X and Z. Fix trisections $\mathcal{T}_X = (X_1, X_2, X_3)$ of X and $\mathcal{T}_Z = (Z_1, Z_2, Z_3)$ of Z.*

Then there exists a trisection $\mathcal{M} = (Y_1, Y_2, Y_3)$ of Y so that $Y_i \cap X = X_i$ and $Y_i \cap Z = Z_i$ for each $i = 1, 2, 3$. Moreover, $\mathcal{M}$ is strongly compatible with Y.

Proof. The theorem follows via induction on Propositions 4, 5, 6, and 7.

Note that Proposition 4 (for cobordisms including index-1 critical points) holds even when the two manifolds X and Z are disconnected, if we take a trisection of a disconnected manifold to consist of trisections on each connected component.

Theorem 6. *Let Y be a connected cobordism between smooth, closed, 4-manifolds X and Z, which are potentially disconnected or empty. Fix trisections $\mathcal{T}_X = (X_1, X_2, X_3)$ of X and $\mathcal{T}_Z = (Z_1, Z_2, Z_3)$ of Z.*

Then there exists a trisection $\mathcal{M} = (Y_1, Y_2, Y_3)$ of Y so that $Y_i \cap X = X_i$ and $Y_i \cap Z = Z_i$ for each $i = 1, 2, 3$. Moreover, $\mathcal{M}$ is strongly compatible with Y.

Proof. By the remark above the theorem statement, the claim follows as in Theorem 5 when X and Z are nonempty. If X is empty and Z is nonempty, take $X' = S^4$ and $\mathcal{T}_{X'}$ to be the standard trisection of $X' \cong S^4$. Puncture Y to obtain a cobordism Y' from X' to Z and apply the theorem to obtain a trisection of Y'. Glue a copy of B^5 with the standard trisection to Y' to obtain a trisection of Y extending $\mathcal{T}_Z$. Reversing the roles of X and Z, the claim similarly holds when $X \neq \emptyset$ and $Z = \emptyset$. If $X = Z = \emptyset$, then the claim follows by puncturing Y twice and proceeding in the same fashion.

The following corollary follows immediately from Theorem 5 and the definition of strongly compatible.

Corollary 1. *Let Y be a cobordism between smooth, closed, connected 4-manifolds A and B. Let W be a cobordism between closed, connected 4-manifolds B and C. Fix trisections $\mathcal{T}_A, \mathcal{T}_B, \mathcal{T}_C$ of A, B, and C respectively. Let $\mathcal{T}_Y, \mathcal{T}_W$ be trisections of Y and W as in Theorem 5 so that $\mathcal{T}_Y$ restricts to $\mathcal{T}_A$ on A, $\mathcal{T}_W$ restricts to $\mathcal{T}_C$ on C, and both $\mathcal{T}_Y$ and $\mathcal{T}_W$ restrict to $\mathcal{T}_B$ on B. Moreover, $\mathcal{T}_Y$ and $\mathcal{T}_W$ are strongly compatible with Y and W, respectively. Then we can glue $\mathcal{T}_Y$ and $\mathcal{T}_W$ to obtain a trisection $\mathcal{T}$ of the cobordism $Y \cup_B W$ from A to C.*

Note here that the gluing of $B \subset W$ to $B \subset Y$ is induced by $\mathcal{T}_B$.

Acknowledgements The work in this paper was initiated and mostly completed at *Topology of Manifolds: interactions between high and low dimensions* at MATRIX in January 2019. We thank Boris Lishak and Stephan Tillmann for interesting conversations at MATRIX and afterward about multisections. The second author's main takeaway was that trisections of closed 5-manifolds are significantly more complicated than those of closed 4-manifolds, and quadrisections of 6-manifolds are exponentially more so.

Thanks also to Mark Powell and an anonymous reviwer for helpful comments.

During the time of this project, the first author was a visiting assistant professor at the Georgia Institute of Technology and the second author was a graduate student at Princeton University. The second author was supported by NSF grant DGE-1656466.

References

1. Borodzik, M., Némethi, A., Ranicki, A.: Morse theory for manifolds with boundary. Algebr. Geom. Topol. **16**(2), 971–1023 (2016). URL https://doi-org.prx.library.gatech.edu/10.2140/agt.2016.16.971
2. Castro, N.A.: Relative trisections of smooth 4-manifolds with boundary. Ph.D. thesis, University of Georgia (2016)
3. Castro, N.A., Gay, D.T., Pinzón-Caicedo, J.: Diagrams for relative trisections. Pac. J. Math. **294**, 275–305 (2018)
4. Gay, D., Kirby, R.: Trisecting 4-manifolds. Geom. Topol. **20**(6), 3097–3132 (2016). URL https://doi-org.prx.library.gatech.edu/10.2140/gt.2016.20.3097
5. Gay, D., Meier, J.: Doubly pointed trisection diagrams and surgery on 2-knots. arXiv e-prints arXiv:1806.05351 (2018)
6. Mazur, M., Hirsch, M., Mazur, B.: Smoothings of Piecewise Linear Manifolds. Annals of mathematics studies. Princeton University Press (1974). URL https://books.google.com/books?id=31dWd26JIeQC
7. Meier, J., Zupan, A.: Bridge trisections of knotted surfaces in 4-manifolds. Proc. Natl. Acad. Sci. USA **115**(43), 10,880–10,886 (2018). URL https://doi-org.prx.library.gatech.edu/10.1073/pnas.1717171115
8. Rubinstein, J.H., Tillmann, S.: Generalized trisections in all dimensions. Proc. Natl. Acad. Sci. USA **115**(43), 10,908–10,913 (2018). URL https://doi-org.prx.library.gatech.edu/10.1073/pnas.1718961115
9. Whitehead, J.H.C.: On C^1-complexes. Ann. of Math. **41**, 809–824 (1940)

The octonionic projective plane

Malte Lackmann

1 Introduction

As mathematicians found out in the last century, there are only four normed division algebras[1] over $\mathbb{R}$: the real numbers themselves, the complex numbers, the quaternions and the octonions. Whereas the real and complex numbers are very well-known and most of their properties carry over to the quaternions (apart from the fact that these are not commutative), the octonions are very different and harder to handle since they are not even associative. However, they can be used for several interesting topological constructions, often paralleling constructions known for $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$.

In this article, we will construct $\mathbb{O}\mathrm{P}^2$, a space having very similar properties to the well-known two-dimensional projective spaces over $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$.

We will begin, in the second section, by recalling a construction of the octonions and discussing their basic algebraic properties. We will then move on to actually constructing the octonionic projective plane $\mathbb{O}\mathrm{P}^2$. In the fourth section, we will discuss properties of $\mathbb{O}\mathrm{P}^2$ and applications in algebraic topology. In particular, we will use it to construct a map $\mathrm{S}^{15} \longrightarrow \mathrm{S}^8$ of Hopf invariant 1. Moreover, it will be explained why there cannot be projective spaces over the octonions in dimensions higher than 2.

Malte Lackmann
Mathematisches Institut, Universität Bonn, Endenicher Allee 60, 53115 Bonn, e-mail: lackmann@math.uni-bonn.de

[1] A division algebra over $\mathbb{R}$ is a finite-dimensional unital $\mathbb{R}$-algebra without zero divisors, not necessarily commutative or associative. A normed algebra over $\mathbb{R}$ is an $\mathbb{R}$-algebra $\mathbb{A}$ together with a map $\|\cdot\| : \mathbb{A} \to \mathbb{R}$ coinciding with the usual absolute value on $\mathbb{R} \cdot 1 \cong \mathbb{R}$, satisfying the triangular inequality, positive definiteness and the rule $\|xy\| = \|x\| \, \|y\|$.

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_6

2 Construction of $\mathbb{O}$

In this section, we will explain how the octonions can be constructed. Of course, there are many ways to define them – the simplest way would be to choose a basis of $\mathbb{O}$ as a vector space and then specify the products of each pair of basis elements. We will try to explain a little better the "reason" for the existence of octonions by giving a construction which leads from the quaternions to the octonions, but which can also be used to construct $\mathbb{C}$ out of $\mathbb{R}$ and $\mathbb{H}$ out of $\mathbb{C}$, the so-called *Cayley-Dickson construction*.

As mentioned in the introduction, the octonions are neither commutative nor even associative. However, they have the following crucial property called *alternativity*:

For any two octonions x and y, the subalgebra of $\mathbb{O}$ generated by x and y is associative.

Here a subalgebra is always meant to contain the unit. By a nontrivial theorem of Emil Artin [Zor31, Sch95], the above condition is equivalent to requiring that the formulas

$$x(yy) = (xy)y \quad \text{and} \quad (xx)y = x(xy) \tag{1}$$

hold for any two elements $x, y \in \mathbb{O}$.

We will now begin to construct the octonions. To do this, note that the well-known division algebras $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ are not only normed $\mathbb{R}$-algebras, but they come together with a *conjugation*: an anti-involution * (i.e., a linear map * from the algebra to itself such that $(xy)^* = y^*x^*$ and $(x^*)^* = x$) with the property that $xx^* = x^*x = \|x\|^2$.

This extra structure goes into the following construction, called "doubling construction" or "Cayley-Dickson construction": Let $(\mathbb{A}, \|\cdot\|, ^*)$ be a normed real algebra with conjugation. Then we define the structure of a normed real algebra with conjugation on the real vector space $\mathbb{A}^2$ by

- $(a,b)\cdot(c,d) = (ac - d^*b, da + bc^*)$,
- $\|(a,b)\| = \sqrt{\|a\|^2 + \|b\|^2}$,
- $(a,b)^* = (a^*, -b)$.

It can be checked that this turns $\mathbb{A}^2$ indeed into a real algebra with conjugation. Furthermore, it also conserves most of the properties as an algebra that $\mathbb{A}$ has had, though not all:

- The real numbers are an associative and commutative division algebra and have the additional property that the conjugation is just the identity. Applying the Cayley-Dickson construction, we obtain the complex numbers which are still an associative and commutative division algebra, but the conjugation is not trivial any longer – it is the usual complex conjugation.
- Going from $\mathbb{C}$ to $\mathbb{H}$, we lose the property of commutativity: $\mathbb{H}$ is only an associative division algebra.

- Applying the construction to $\mathbb{H}$, we get the octonions $\mathbb{O}$, which are not even associative any more, but are still a normed division algebra and are alternative as explained above.
- Continuing to apply the Cayley-Dickson construction, one obtains a 16-dimensional real algebra called the *sedenions*. The sedenions have zero divisors, and therefore cannot have a multiplicative norm. They are by far less important than the other four division algebras. However, they still satisfy a property called *flexibility* which is a very weak form of associativity. Interestingly, if the Cayley-Dickson construction is applied again, this property is *not* lost, so there are flexible 2^n-dimensional normed real algebras for every n [Gui97].

The proofs of these statements are rather lengthy, but straightforward. We will carry them out for the passage from quaternions to octonions, since this is the case we are most interested in and also the most difficult one, and leave the remaining cases to the reader.

Proposition 1. $\mathbb{O}$ *is an alternative normed division algebra.*

Proof. For the alternativity, we prove (1) and then use Artin's theorem to deduce alternativity. We write $x = (a,b)$ and $y = (c,d)$ with $a,b,c,d \in \mathbb{H}$ and write out both sides of the first equation, using the definition of the multiplication displayed above. Doing the calculation and obvious cancellations, this leaves us to show the two identities

$$d^*bc + d^*da + d^*bc^* = add^* + c^*d^*b + cd^*b$$

and

$$dac - dd^*b + dac^* = dca + dc^*a - bdd^* .$$

Considering the first equation, note that $dd^* = d^*d = \|d\|^2$ is a real number and thus central, so that $d^*da = add^*$. The remaining terms can be regrouped as follows:

$$d^*b(c+c^*) = (c+c^*)d^*b .$$

However, $c+c^*$ is a real number as well, as can for instance be seen easily from the Cayley-Dickson construction, so that we have proved the first of the two identities. The second can be traced back similarly to the facts that $dd^*b = bdd^*$ and

$$da(c+c^*) = d(c+c^*)a .$$

We now prove that the octonion norm is multiplicative. This is also not completely formal, as can be seen from the fact that it is not true for the sedenions. With notation as above, the equation $\|(a,b)\|^2 \|(c,d)\|^2 = \|(a,b)(c,d)\|^2$ can be simplified to

$$acb^*d + d^*bc^*a^* = bacb^* + bc^*a^*d^* . \tag{2}$$

Following [KS89, p. 48], we consider two cases: If d is real, the equation holds true trivially. If d is purely imaginary in the sense that $d^* = -d$, then the equation is equivalent to

$$d(acb^* + bc^*a^*) = (acb^* + bc^*a^*)d\,,$$

which is true since

$$acb^* + bc^*a^* = acb^* + (acb^*)^*$$

is real. By linearity, (2) is true for all d.

The multiplicativity of the norm at hand directly implies the fact that $\mathbb{O}$ is a division algebra: If $xy = 0$, then

$$\|x\|\,\|y\| = \|xy\| = 0\,,$$

so $\|x\| = 0$ or $\|y\| = 0$ and thus $x = 0$ or $y = 0$ since $\|\cdot\|$ is a norm. □

Remark 1. (i) Note that all formulas of the above proof are written without parantheses, thus we used secretly that $\mathbb{H}$ is associative. This is essential: As mentioned above, the sedenions, constructed out of the non-associative octonions, are neither alternative, nor a division algebra (and, consequently, they cannot possess a multiplicative norm).

(ii) The book [CS03] gives a geometric argument that the octonions are alternative, which does not use Artin's theorem. See Section 6.8, in particular Theorem 2.

3 Construction of $\mathbb{O}\mathrm{P}^2$

This section discusses a construction of a projective space of dimension 2 over the octonions (which is very similar to the construction in [Bae02]). Our goal is to get spaces with similar properties as their analogues over the real and complex numbers and the quaternions. The naïve ansatz would be to define the n-dimensional octonionic projective space as a quotient of $\mathbb{O}^{n+1} \setminus \{0\}$, identifying every vector with its (octonionic) multiples. However, when one starts calculating, one sees that associativity is needed for this to be an equivalence relation. Since the octonions are not associative, we have to be more careful. In fact, the construction given in the following only works for $n \leq 2$ (thanks to the property of alternativity), and we will see that there are theoretical obstructions to the existence of higher $\mathbb{O}\mathrm{P}^n$.

One main difference to the other three projective planes is that we don't construct $\mathbb{O}\mathrm{P}^2$ as a quotient of $\mathbb{O}^3 \setminus \{0\}$, but we restrict ourselves to the subset

$$T = \Big\{(x,y,z) \in \mathbb{O}^3;\ \|x\|^2 + \|y\|^2 + \|z\|^2 = 1 \quad \text{and the subalgebra generated by } x, y \text{ and } z \text{ is associative}\Big\}\,.$$

We call two triples $(x,y,z), (\tilde{x},\tilde{y},\tilde{z}) \in T$ equivalent, $(x,y,z) \sim (\tilde{x},\tilde{y},\tilde{z})$, if and only if the six equations

$$xx^* = \tilde{x}\tilde{x}^*, \quad xy^* = \tilde{x}\tilde{y}^*, \quad xz^* = \tilde{x}\tilde{z}^*, \quad yy^* = \tilde{y}\tilde{y}^*, \quad yz^* = \tilde{y}\tilde{z}^*, \quad zz^* = \tilde{z}\tilde{z}^*$$

hold. (The three remaining relations of a similar form follow from these by the properties of the conjugation.) It is obvious that this is an equivalence relation, and thus we can set

$$\mathbb{O}\mathrm{P}^2 = T/\sim .$$

3.1 Manifold structure

A first observation is that our space $\mathbb{O}\mathrm{P}^2$ constructed like this is quasi-compact since it is the quotient of a quasi-compact space. We will now show that it's a 16-dimensional real manifold.

The proof uses the following construction. Let a,b,c be real numbers. Define an $\mathbb{R}$-linear map

$$\ell = \ell_{(a,b,c)} \colon \mathbb{O}^3 \longrightarrow \mathbb{O}, \quad \ell(x,y,z) = ax + by + cz.$$

Note that for two triples $(x,y,z) \sim (\tilde{x},\tilde{y},\tilde{z})$ in T, we have $\ell(x,y,z) = 0$ if and only if $\ell(\tilde{x},\tilde{y},\tilde{z}) = 0$, since $\ell(x,y,z)$ vanishes exactly if $\ell(x,y,z)\ell(x,y,z)^*$ vanishes, and $\ell(x,y,z)\ell(x,y,z)^* = \ell(\tilde{x},\tilde{y},\tilde{z})\ell(\tilde{x},\tilde{y},\tilde{z})^*$ by the definition of $\sim$ and since a,b,c are real.

Thus we get a well-defined open set

$$U_\ell = U_{(a,b,c)} = \{[x,y,z];\, \ell(x,y,z) \neq 0\} \subset \mathbb{O}\mathrm{P}^2.$$

Proposition 2. $\mathbb{O}\mathrm{P}^2$ *is a locally Euclidean topological space.*

Proof. Suppose that $c \neq 0$. Consider the maps

$$\varphi_\ell : U_\ell \to \mathbb{O}^2, \quad [x,y,z] \mapsto \left(\frac{x\ell^*}{\|\ell\|^2}, \frac{y\ell^*}{\|\ell\|^2} \right),$$

where $\ell = \ell(x,y,z)$. This map is well-defined (the argument given above shows that $\|\ell\|^2 = \ell\ell^*$ only depends on $[x,y,z]$ and not on (x,y,z); the same argument works for $x\ell^*$ and $y\ell^*$) and thus continuous by the universal property of the quotient topology. An inverse map is given by

$$\psi_\ell : \mathbb{O}^2 \to U_\ell,$$
$$(x,y) \mapsto \left[\frac{x}{r}, \frac{y}{r}, \frac{1-ax-by}{cr} \right], \quad r = \sqrt{\|x\|^2 + \|y\|^2 + \frac{1}{c^2}\,\|1-ax-by\|^2}\,.$$

Recall that in the definition of $\mathbb{O}\mathrm{P}^2$, we only consider triples whose entries lie in an associative subalgebra of $\mathbb{O}$. Thus this map is only well-defined since $\mathbb{O}$ is alternative, as defined in Section 2. Here we use that for every octonion y, the conjugate y^* lies in the subalgebra generated by y since it only differs from $-y$ by a real number. This can directly be seen from the Cayley-Dickson construction.

Note that equation (1) would not be sufficient at this point, but we use the nontrivial Artin theorem which states that (1) implies alternativity. Moreover, it is exactly at this point (and in the next paragraph) where our procedure breaks down if we want to construct higher $\mathbb{O}\mathrm{P}^n$ in the same way.

Checking that φ_ℓ and ψ_ℓ are inverse to each other is a simple calculation. The reader may amuse herself by reproducing them. Note that we evade associativity of $\mathbb{O}$ by carrying out calculations only with elements lying in an associative subalgebra of $\mathbb{O}$ – by definition of T for the one and by the alternativity of $\mathbb{O}$ for the other direction.

The same obviously works with a or b instead of c, thus for all triples $(a,b,c) \neq (0,0,0)$. Thus we have covered $\mathbb{O}\mathrm{P}^2$ by the three chart regions $U_{(1,0,0)}, U_{(0,1,0)}$ and $U_{(0,0,1)}$. □

Remark 2. (i) It is easy to check that the coordinate changes are smooth, such that $\mathbb{O}\mathrm{P}^2$ in fact becomes a smooth manifold.

(ii) The referee raised the question whether $\mathbb{O}\mathrm{P}^2$ admits the structure of a complex manifold. Since it is an $(n-1)$-connected $2n$-manifold for $n = 8$, we may apply the criterion of [Yan12, Thm. 1], which says that $\mathbb{O}\mathrm{P}^2$ doesn't even admit a stable almost complex structure. Note that $\mathbb{H}\mathrm{P}^2$ admits a stable almost complex structure, but no almost complex structure [Yan12, Thm. 1, 2].

Lemma 1. $\mathbb{O}\mathrm{P}^2$ *is Hausdorff.*

Proof. Let (x,y,z) and (x',y',z') two elements in T. We have to find $(a,b,c) \in \mathbb{R}^3$ such that $\ell_{(a,b,c)}(x,y,z) \neq 0$ and $\ell_{(a,b,c)}(x',y',z') \neq 0$, since then it follows that $[x,y,z]$ and $[x',y',z']$ both lie in the open subset $U_{(a,b,c)}$ which we already know to be Hausdorff.

To find (a,b,c) as above, note that given x, y and z, the set of all solutions to the equation $ax+by+cz = 0$ is a subspace of $\mathbb{R}^3$ of dimension at most 2, and the same is true for the equation $ax'+by'+cz' = 0$. Since the union of two planes is never the whole $\mathbb{R}^3$, we find a point that fails to satisfy both of these equations. □

Corollary 1. $\mathbb{O}\mathrm{P}^2$ *is a closed* 16*-dimensional real manifold.* □

3.2 The projective line $\mathbb{O}\mathbf{P}^1$

Consider the closed subset of $\mathbb{O}\mathrm{P}^2$ given by all equivalence classes of the form $[x,y,0]$ with $x,y \in \mathbb{O}$. It is immediate from the definitions that this is homeomorphic to the space

$$\{(x,y) \in \mathbb{O}^2; \|x\|^2 + \|y\|^2 = 1\}/\sim,$$

where $(x,y) \sim (\tilde{x},\tilde{y})$ if and and only if the three relations

$$xx^* = \tilde{x}\tilde{x}^*, xy^* = \tilde{x}\tilde{y}^*, yy^* = \tilde{y}\tilde{y}^*$$

hold. This is of course the analogue of our construction of $\mathbb{O}\mathrm{P}^2$ in one dimension lower, so we call the resulting space $\mathbb{O}\mathrm{P}^1$.

By the same argumentation as in the previous paragraph, $\mathbb{O}\mathrm{P}^1$ is a closed manifold. Moreover, there is a homeomorphism $\mathbb{O}\mathrm{P}^1 \cong \mathrm{S}^8$. This can be constructed in the very same way as in the familiar cases over $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$ (since the map in one direction involves only one element of $\mathbb{O}$ at a time).

3.3 CW structure

The octonionic projective plane has a very simple cell structure with one cell in each of the dimensions 0, 8 and 16. This cell structure is constructed in exactly the same way as the analogous structures for $\mathbb{R}\mathrm{P}^2$, $\mathbb{C}\mathrm{P}^2$ and $\mathbb{H}\mathrm{P}^2$.

A little lemma that we will need in the proof of the following statements is the observation that we can choose a vector space isomorphism $\mathbb{O} \cong \mathbb{R}^8$ such that the norm $\|\cdot\|$ becomes the usual Euclidean norm on $\mathbb{R}^8$. This follows directly from the definition of $\mathbb{O}$ via the Cayley-Dickson construction, but it can also be deduced from the formal properties of the conjugation map: $\langle x,y\rangle = \frac{1}{2}(x^*y + y^*x)$ defines a symmetric, positive definite bilinear form on the vector space $\mathbb{O}$, so we can find an orthonormal basis by Sylvester's law of inertia.

To begin with, consider the canonical inclusion

$$\mathbb{O}\mathrm{P}^1 \hookrightarrow \mathbb{O}\mathrm{P}^2, \quad [x,y] \mapsto [x,y,0]\,.$$

By our definition of $\mathbb{O}\mathrm{P}^1$, this map is a homeomorphism onto its image, which is closed in $\mathbb{O}\mathrm{P}^2$. Since $\mathbb{O}\mathrm{P}^1 \cong \mathrm{S}^8$, we can use $\mathbb{O}\mathrm{P}^1$ as the 8-skeleton in our cell decomposition. Now consider the map

$$f : \mathrm{S}^{15} \longrightarrow \mathbb{O}\mathrm{P}^1, \quad (x,y) \mapsto [x,y]\,,$$

where we think of S^{15} as a subset of $\mathbb{R}^{16} = \mathbb{R}^8 \times \mathbb{R}^8$.

Lemma 2. *We have* $\mathbb{O}\mathrm{P}^2 = \mathbb{O}\mathrm{P}^1 \cup_f D^{16}$.

Proof. A map from D^{16} to $\mathbb{O}\mathrm{P}^2$ coinciding with f on $\partial\mathrm{D}^{16}$ is given by

$$(x,y) \mapsto \left[x, y, \sqrt{1 - \|x\|^2 - \|y\|^2}\right].$$

It is easily checked that this map induces a bijective continuous map from $\mathbb{O}\mathrm{P}^1 \cup_f \mathrm{D}^{16}$ to $\mathbb{O}\mathrm{P}^2$ which is thus a homeomorphism since $\mathbb{O}\mathrm{P}^1 \cup_f \mathrm{D}^{16}$ is quasi-compact and $\mathbb{O}\mathrm{P}^2$ is Hausdorff. □

4 Cohomology of $\mathbb{O}\mathrm{P}^2$

After having constructed this very simple cell structure for $\mathbb{O}\mathrm{P}^2$, it is easy to compute the cohomology via the cellular cochain complex:

Corollary 2. *Let A be any abelian group. Then*

$$\mathrm{H}^k(\mathbb{O}\mathrm{P}^2, A) \cong \begin{cases} A, & k = 0, 8 \text{ or } 16, \\ 0, & \text{else.} \end{cases}$$

The homology groups are computed in exactly the same way and with the same result.

Similarly, we can compute the homotopy groups of $\mathbb{O}\mathrm{P}^2$: By cellular approximation, $\pi_n(\mathbb{O}\mathrm{P}^2, *) = 0$ for $n \leq 7$ and $\pi_n(\mathbb{O}\mathrm{P}^2, *) = \pi_n(\mathrm{S}^8)$ for $n \leq 14$.

The following question has been brought to the author's attention by Jens Reinhold.

Question 1. Is there a closed, 8-connected manifold of positive dimension with odd Euler characteristic?

The octonionic projective plane gives such a manifold which is 7-connected. An 8-connected example would have to have dimension divisible by 32, by a result of Hoekzema [Hoe18, Thm. 1.2, Cor. 4.2].

4.1 Connection with the Hopf invariant 1 *problem*

The Hopf invariant is a classical invariant for maps $f\colon \mathrm{S}^{2n-1} \to \mathrm{S}^n$, with $n > 1$. It goes back to work of Hopf in the 1930's. We quickly recall its definition from [MT08]. Let us consider the mapping cylinder of such a map f. By inspection of the cellular cochain complex, as in Corollary 2 above, it has cohomology groups in degree n and $2n$ which are cyclic with generators τ and σ. These are unique up to sign, depending on the orientation of the two spheres. The Hopf invariant $H(f)$ is defined by the formula

$$\tau^2 = H(f) \cdot \sigma\,.$$

It is unique up to sign, which only depends on the orientation of S^{2n-1} since τ appears squared.

The question in which dimensions there exists a map of Hopf invariant 1 was a famous open problem in the early days of algebraic topology, until Adams proved in 1960 that this is only the case for $d \in \{1, 2, 4, 8\}$. The sought maps for $d = 2, 4, 8$ can be constructed as the attaching maps of the top dimensional cell in the projective planes over the complex numbers, quaternionics and octonionics, and we will now prove this for the octonionics, by analysing the ring structure on the cohomology of $\mathbb{O}\mathrm{P}^2$.

Theorem 1. *The attaching map $f\colon S^{15} \longrightarrow S^8$ of the* 16*-cell in* $\mathbb{O}\mathrm{P}^2$ *has Hopf invariant* ± 1*, the sign depending on the orientation of* S^{15}.

Proof. Let τ and σ be generators of $\mathrm{H}^8(\mathbb{O}\mathrm{P}^2,\mathbb{Z})$ and $\mathrm{H}^{16}(\mathbb{O}\mathrm{P}^2,\mathbb{Z})$, respectively. Since $\mathbb{O}\mathrm{P}^2$ is the mapping cylinder of f, we just have to show that $\tau^2 = \sigma$ (up to sign) by the definition above. Since we have seen $\mathbb{O}\mathrm{P}^2$ to be a closed manifold which is orientable since it is simply-connected, we can profit of Poincaré duality to do so: Let $\mu \in \mathrm{H}_{16}(\mathbb{O}\mathrm{P}^2,\mathbb{Z})$ be a fundamental class. Using the universal coefficient theorem as well as Poincaré duality, we get isomorphisms

$$\mathrm{H}^n(\mathbb{O}\mathrm{P}^2,\mathbb{Z}) \cong \mathrm{Hom}(\mathrm{H}_n(\mathbb{O}\mathrm{P}^2,\mathbb{Z}),\mathbb{Z}) \cong \mathrm{Hom}(\mathrm{H}^{16-n}(\mathbb{O}\mathrm{P}^2,\mathbb{Z}),\mathbb{Z})\,,$$

where the map from the left to the right maps f to the linear map

$$g \mapsto \langle f, \mu \cap g\rangle = \langle g \cup f, \mu\rangle\,.$$

Now, $\mathrm{Hom}(\mathrm{H}^8(\mathbb{O}\mathrm{P}^2,\mathbb{Z}),\mathbb{Z})$ is isomorphic to $\mathbb{Z}$, generated by the two isomorphisms. Thus, τ has to be mapped to an isomorphism $\mathrm{H}^8(\mathbb{O}\mathrm{P}^2,\mathbb{Z}) \longrightarrow \mathbb{Z}$, which in turn has to map τ to a generator of $\mathbb{Z}$, so

$$\langle \tau^2, \mu\rangle = \pm 1\,.$$

Now, setting $\tau^2 = k\sigma$ with $k \in \mathbb{Z}$, we get

$$k \cdot \langle \sigma, \mu\rangle = \pm 1\,,$$

so k divides 1, giving $k = \pm 1$ and thus $\tau^2 = \sigma$. □

Note that the constructions we have carried out in the last two sections can be done in a much more general setting: Suppose that $\mathbb{A}$ is a normed real algebra $\mathbb{A}$ of dimension $d < \infty$ with conjugation, which is alternative and has no zero divisors. Then we can write down the same formulas as above to define a topological space $\mathbb{A}\mathrm{P}^2$, prove that it is a manifold and give it a cell structure with one cell in each of the dimensions 0, d and $2d$. The attaching map of the $2d$-cell will then be a map $\mathrm{S}^{2d-1} \longrightarrow \mathrm{S}^d$ of Hopf invariant 1.

In [EHH$^+$91, Sec. 8.1, 8.2, 9.1], it is shown that the existence of the two extra structures on $\mathbb{A}$ does not have to be claimed on its own: For any real alternative division algebra, there is a canonical norm and conjugation[2].

Summarising, we have argued that the following holds:

Theorem 2. *If there exists a real alternative division algebra of dimension d, then there is a map $S^{2d-1} \longrightarrow S^d$ of Hopf invariant* 1.

By Adams' result, this is only the case for $d \in \{1,2,4,8\}$. Thus the construction of the projective plane shows that a real alternative division algebra can only exist

[2] As the alert reader may have noticed, we have used exactly one more property of the conjugation, namely the fact that z^* always lies in the subalgebra generated by z. However, the canonical conjugation constructed in [EHH$^+$91] always has this property.

in dimensions $1, 2, 4$ and 8. Of course, this is still true if the alternativity claim is dropped, but one needs a different proof for this [Hat09, Sec. 2.3].

4.2 Non-existence of higher octonionic projective spaces

As pointed out above, our construction of the octonionic projective space $\mathbb{O}\mathrm{P}^2$ doesn't generalise to higher dimensional projective spaces since we have intensively used the fact that all calculations are done in associative subalgebras of $\mathbb{O}$. However, there is also a conceptual reason that there *can't* be a space which deserves to be called $\mathbb{O}\mathrm{P}^3$ (or $\mathbb{O}\mathrm{P}^n$ for some $n \geq 3$) which we will now explain.

We will only claim two properties of our wannabe projective octonionic 3-space: it should be a closed manifold, and it should have a cell structure with $\mathbb{O}\mathrm{P}^2$ as the 16-skeleton and only one more 24-cell. It then follows directly that the cohomology $\mathrm{H}^*(\mathbb{O}\mathrm{P}^3, \mathbb{Z})$ is $\mathbb{Z}$ in dimensions $0, 8, 16$ and 24 and trivial otherwise.

By a similar argument as for $\mathbb{O}\mathrm{P}^2$, we also get the ring structure on the cohomology:

$$\mathrm{H}^*(\mathbb{O}\mathrm{P}^3, \mathbb{Z}) \cong \mathbb{Z}[x]/(x^4), \quad |x| = 8.$$

To see this, note that the inclusion of the 16-skeleton induces an isomorphism on cohomology in degrees smaller than 23 which respects the multiplicative structure, thus it is sufficient to show that a fourth power of the generator of H^8 generates H^{24}. But this is done in a very similar way as for the octonionic projective plane, using Poincaré duality.

However, using Steenrod powers modulo 2 and 3, one can show that a space with cohomology $\mathbb{Z}[x]/(x^m)$, $m > 3$, can only exist if x has degree 2 or 4 [Hat02, Sec. 4.L].

4.3 The story continues

We have used the octonions to construct a map between spheres of Hopf invariant 1. There are other phenomena in the intersection of algebra, topology and geometry that show deep relations with the octonions. Examples include exotic spheres, Bott periodicity and exceptional Lie groups (these can be used to see that $\mathbb{O}\mathrm{P}^2$ is a homogeneous space, for instance). The article [Bae02] explains these and many more interesting examples.

Acknowledgements This note was originally written by the author as a term paper for the Algebraic Topology 1 course of Peter Teichner at the University of Bonn. I thank him for suggesting this interesting topic. At the time of publication, I am working as a PhD student at Bonn, supported by the ERC Advanced Grant "KL2MG-interactions" (no. 662400) of Wolfgang Lück. I thank Jasper Knyphausen, Stefan Friedl and the anonymous referee for helpful comments on this paper.

References

Bae02. John C. Baez. The octonions. *Bull. Amer. Math. Soc. (N.S.)*, 39(2):145–205, 2002.

CS03. John H. Conway and Derek A. Smith. *On quaternions and octonions: their geometry, arithmetic, and symmetry*. A K Peters, Ltd., Natick, MA, 2003.

EHH$^+$91. Heinz-Dieter Ebbinghaus, Hans Hermes, Friedrich Hirzebruch, Max Koecher, Klaus Lamotke, Jürgen Neukirch, Klaus Mainzer, Alexander Prestel, and Reinhold Remmert. *Numbers*, volume 123 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991.

Gui97. R. Guillermo Moreno. The zero divisors of the Cayley-Dickson algebras over the real numbers. *eprint arXiv:q-alg/971001*, pages q–alg/9710013, Oct 1997.

Hat02. Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

Hat09. Allen Hatcher. Vector bundles & k-theory, 2009. Available online at `http://www.math.cornell.edu/~hatcher/VBKT/VBpage.html`.

Hoe18. Renee S. Hoekzema. Manifolds with Odd Euler Characteristic and Higher Orientability. *International Mathematics Research Notices*, 2018.

KS89. Issai L. Kantor and Aleksandr S. Solodovnikov. *Hypercomplex numbers*. Springer-Verlag, New York, 1989. An elementary introduction to algebras, Translated from the Russian by A. Shenitzer.

MT08. Robert E. Mosher and Martin C. Tangora. *Cohomology operations and applications in homotopy theory*. Dover Pub., New York, 2008.

Sch95. Richard D. Schafer. *An introduction to nonassociative algebras*. Dover Publications, Inc., New York, 1995. Corrected reprint of the 1966 original.

Yan12. Huijun Yang. Almost complex structures on $(n-1)$-connected $2n$-manifolds. *Topology and Its Applications*, 159:1361–1368, 2012.

Zor31. Max Zorn. Theorie der alternativen ringe. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 8(1):123–147, Dec 1931.

Null-homologous twisting and the algebraic genus

Duncan McCoy

Abstract The algebraic genus of a knot is an invariant that arises when one considers upper bounds for the topological slice genus coming from Freedman's theorem that Alexander polynomial one knots are topologically slice. This paper develops null-homologous twisting operations as a tool for studying the algebraic genus and, consequently, for bounding the topological slice genus above. As applications we give new upper bounds on the algebraic genera of torus knots and satellite knots.

1 Introduction

In this paper we study the algebraic genus $g_{\mathrm{alg}}(L)$ of an oriented link L in S^3, as defined by Feller-Lewark [6]. It is a famous theorem of Freedman that a knot K in S^3 with Alexander polynomial $\Delta_K = 1$ is topologically slice [10]. It was first observed by Rudolph that this can be used to construct upper bounds on the topological slice genus of knots even when the Alexander polynomial is non-trivial [15]. If a knot K has a Seifert surface F containing a subsurface F' such that $\partial F'$ is a knot with Alexander polynomial one, then F' can be replaced by a locally flat disk in the 4-ball to show that K cobounds a locally flat surface of genus $g(F) - g(F')$. The algebraic genus can be defined as the optimal upper bound for $g_4^{\mathrm{top}}(L)$ that can be achieved by this method:

$$g_{\mathrm{alg}}(L) = \min\left\{ g(F) - g(F') \,\middle|\, \begin{array}{l} F \text{ is a Seifert surface for } L \text{ and } F' \subset F \text{ is a subsur-} \\ \text{face such that } \partial \Sigma' = K' \text{ is a knot with } \Delta_{K'}(t) = 1. \end{array} \right\}.$$

The main utility of the algebraic genus is that it has several equivalent formulations, including one that depends only on the S-equivalence class of the Seifert form of L

Duncan McCoy
Université du Québec à Montréal, Canada
e-mail: mc_coy.duncan@uqam.ca

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_7

[6]. These different formulations have made the algebraic genus a valuable tool for proving results about the topological slice genus [2, 5, 8, 13]. It turns out that, at least for knots, the algebraic genus has a pleasing topological interpretation as the minimal possible genus of a compact, locally flatly embedded surface $F \subseteq B^4$ such that $\partial F = K$ and $\pi_1(B^4 \setminus F) \cong \mathbb{Z}$ [7].

The purpose of this paper is to explore how the algebraic genus changes under certain twisting operations. Using these operations, we obtain new upper bounds for the algebraic genus of satellite knots and torus knots.

Null-homologous twisting

Given an oriented knot or link L in S^3 and an integer n, we perform a *null-homologous n-twist* by taking an unknotted curve C disjoint from L with $lk(C,L) = 0$ and performing $1/n$-surgery on C. Such a twist can always be performed locally by adding n full twists on some number of parallel strands with appropriate orientations. See Figure 1, for example.

It turns out that certain pairs of null-homologous twisting operations change the algebraic genus by at most one.

Theorem 1.1 *If L and L' are oriented links related by a null-homologous m-twist and a null-homologous n-twist for $m,n \in \mathbb{Z}$ such that $-mn$ is a square, then*

$$|g_{\text{alg}}(L) - g_{\text{alg}}(L')| \leq 1.$$

Most notably this shows that for any integer n, a single null-homologous n-twist changes the algebraic genus by at most one. It is also shows that a null-homologous $+1$-twist and a null-homologous -1-twist change the algebraic genus by at most one. This latter observation can be seen as an analogue of the well-known fact that changing a negative crossing and a positive crossing changes the smooth slice genus by at most one.

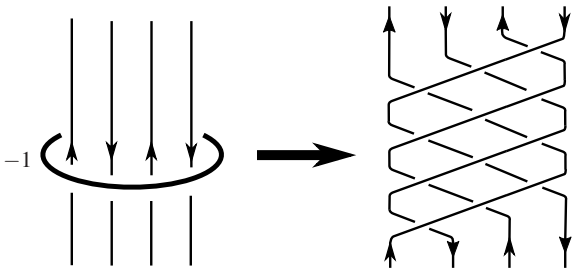


Fig. 1 A negative null-homologous -1-twist on 4 strands.

For any link one can always find pairs of null-homologous $+1$- and -1-twists which decrease the algebraic genus. This leads to the following description of the algebraic genus.

Theorem 1.2 *For any link L, we have*

$$g_{\mathrm{alg}}(L) = \min\left\{ \max\{n,p\} \,\middle|\, \begin{array}{l} L \text{ can be converted to a link } L' \text{ with } g_{\mathrm{alg}}(L') = \\ 0 \text{ by } p \text{ null-homologous } +1\text{-twists and } n \text{ null-} \\ \text{homologous } -1\text{-twists.} \end{array} \right\}.$$

For knots, a stronger formulation of Theorem 1.2 holds. Using the work Borodzik and Friedl on the Blanchfield form [3, 4], one can show that the null-homologous twists in Theorem 1.2 can be realized by crossing changes [7].

The condition that $-mn$ be a square turns out to be essential to the proof Theorem 1.1.

Proposition 1.3 *For any $m,n \in \mathbb{Z}$ such that $-mn$ is not a square, there is a knot K with $g_{\mathrm{alg}}(K) = g_4^{\mathrm{top}}(K) = 2$, which can be unknotted by performing a null-homologous m-twist and a null-homologous n-twist.*

Satellite knots

For satellite knots we prove the following upper bound on the algebraic genus. This bound was first obtained (using different ideas) by Feller, Miller and Pinzon-Caicedo [9].

Theorem 1.4 *For a satellite knot $P(K)$, we have*

$$g_{\mathrm{alg}}(P(K)) \leq g_{\mathrm{alg}}(P(U)) + g_{\mathrm{alg}}(K).$$

One striking feature of Theorem 1.4 is that the upper bound it establishes is independent of the winding number of the pattern P. This behaviour should be contrasted with that of both the classical Seifert genus and the smooth slice genus where dependence on the winding number of the pattern is unavoidable. For example, if one takes K_n to be the $(n,1)$-cable of the trefoil, then one can show that $g(K_n) = g_4(K_n) = n$. However, it follows from Theorem 1.4 that $g_{\mathrm{alg}}(K_n) = g_4^{\mathrm{top}}(K_n) = 1$.

It is natural to wonder whether there is an analogue of Theorem 1.4 for the topological slice genus. A detailed discussion of this question and related issues can be found in [9].

Torus knots

Whilst the smooth slice genera of torus knots have now been determined by a variety of methods, the topological slice genus remains far less well understood.

Rudolph showed that in general the topological slice genus of a torus knot is strictly smaller than the classical Seifert genus [15]. Later Baader-Feller-Lewark-Liechti constructed further upper bounds on the topological slice genus of torus knots, showing that with the exception of torus knots with $|\sigma(T_{p,q})| = 2g_4(T_{p,q})$ the topological slice genus satisfies $g_4^{\text{top}}(T_{p,q}) \leq \frac{6}{7} g_4(T_{p,q})$ [1].

Using null-homologous twisting operations we establish the following upper bound.

Theorem 1.5 *For any torus knot or link $T_{p,q}$ with $p, q > 1$ we have*

$$g_4^{\text{top}}(T_{p,q}) \leq g_{\text{alg}}(T_{p,q}) < \frac{pq}{3} + p\log_2 q + q\log_2 p.$$

This bound is particularly effective when p and q are both relatively large. One can measure the asymptotic difference between the smooth and topological slice genera of torus knots by considering the following limit:

$$\ell := \lim_{\min\{p,q\}\to\infty} \frac{g_4^{\text{top}}(T_{p,q})}{g_4(T_{p,q})}.$$

It is known that this limit exists and satisfies the bounds $\frac{1}{2} \leq \ell < \frac{3}{4}$ [1]. Theorem 1.5 provides an improved upper bound for ℓ by showing that $\ell \leq \frac{2}{3}$.

Structure

In Section 2 we set out the properties of the algebraic genus that will be used throughout the paper and prove Theorem 1.1. In Section 3, we show that there is always a null-homologous $+1$-twist and a -1-twist that can be used to decrease the algebraic genus. This gives the proof of Theorem 1.2. Then in Section 4 and Section 5 contain the results on the algebraic genera of satellite knots and torus knots respectively. Finally we conclude with Section 6 where we prove Proposition 1.3.

2 Properties of the algebraic genus

In this section we recap some of the necessary properties of the algebraic genus and prove Theorem 1.1. Throughout this paper, all knots and links will be oriented. A Seifert surface for a link L is a connected, oriented, embedded surface $F \subseteq S^3$ with $\partial F = L$. If L has r components, then a genus g Seifert surface has $H_1(F;\mathbb{Z}) \cong \mathbb{Z}^{2g+r-1}$. A Seifert surface comes equipped with its Seifert form $\theta : H_1(F;\mathbb{Z}) \times H_1(F;\mathbb{Z}) \to \mathbb{Z}$. A subgroup $H \leq H_1(F;\mathbb{Z})$ of rank $2n$ is said to be Alexander trivial, if for some (equivalently any) basis, the matrix M representing $\theta|_H$ has the property that $\det(tM - M^T) = t^n$. We record the following three equivalent definitions of

the algebraic genus. The equality of all three quantities were essentially proven by Feller-Lewark, where the third quantity is a variation of their characterization of the algebraic genus in terms of 3-dimensional cobordism distance [6].

Proposition 2.1 *Let $L \subset S^3$ be an oriented r-component link. The algebraic genus can be characterized in the following equivalent ways:*

1.

$$g_{\text{alg}}(L) = \min\left\{ g(F) - g(F') \,\middle|\, \begin{array}{l} F \text{ is a Seifert surface for } L \text{ and } F' \subseteq F \text{ is} \\ \text{a subsurface such that } \partial F' = K' \text{ is a knot} \\ \text{with } \Delta_{K'}(t) = 1. \end{array} \right\}$$

2.

$$g_{\text{alg}}(L) = \min\left\{ \frac{m-r+1}{2} - n \,\middle|\, \begin{array}{l} L \text{ has a Seifert form } \theta : H_1(F) \times H_1(F) \to \\ \mathbb{Z}, \text{ where } H_1(F) \cong \mathbb{Z}^m \text{ and } H_1(F) \text{ con-} \\ \text{tains an Alexander trivial subgroup of} \\ \text{rank } 2n. \end{array} \right\}$$

3.

$$g_{\text{alg}}(L) = \min\left\{ \frac{n-r+1}{2} \,\middle|\, \begin{array}{l} L \text{ can be obtained by } n \text{ oriented bands} \\ \text{moves on a knot } K' \text{ with } \Delta_{K'}(t) = 1. \end{array} \right\}$$

The following lemma shows the equivalence of the first two definitions. We refer the reader to [6, Proposition 9] for proof.

Lemma 2.2 *Given a link L with a Seifert surface F of genus g and corresponding Seifert form θ. There is an Alexander trivial subgroup of rank $2n$ in $H_1(F;\mathbb{Z})$ if and only if F contains a connected genus n subsurface F', where $\partial F' = K'$ is a knot with $\Delta_{K'} = 1$.* □

Although it is not known in general which Seifert surfaces for a link realize the algebraic genus, it turns out that any Seifert surface can be stabilized until it realizes the algebraic genus. The following is a consequence of the results of [6, Section 2]

Lemma 2.3 *Let L be an oriented link with r components and let F be a Seifert surface for L. Then F can be stabilized to yield a surface $\widetilde{F}$ containing a subsurface $\widetilde{F}'$ such that $\partial\widetilde{F}'$ is a knot with Alexander polynomial one and*

$$g_{\text{alg}}(L) = g(\widetilde{F}) - g(\widetilde{F}').$$

□

The following lemma shows how the algebraic genus changes under oriented band moves.

Lemma 2.4 *Let L be an r component link and L' an $r+1$ component link related by an oriented band move. Then*

$$g_{\text{alg}}(L') \leq g_{\text{alg}}(L) \leq g_{\text{alg}}(L') + 1.$$

Proof. Suppose that L_2 is obtained from L_1 by an oriented band move, where L_1 has m components and L_2 has $m \pm 1$ components. Choose a connected Seifert surface F for L_1 disjoint from the band B realizing the band move being performed. This is always possible, since we may choose a diagram for L_1 so that the band B appears as a short planar band between two strands. Applying Seifert's algorithm to such a diagram yields a Seifert surface which can be made disjoint from B. Furthermore, Lemma 2.3 shows that by stabilizing we may assume that F realizes the algebraic genus. Thus F contains a subsurface F' such that ∂F is a knot with Alexander polynomial one and $g_{\text{alg}}(L_1) = g(F) - g(F')$. Take F'' to be the Seifert surface for L_2 obtained by attaching the band B to F. Clearly F' is still a subsurface of F'' so

$$g_{\text{alg}}(L_2) \leq g(F'') - g(F').$$

If L_2 has $m+1$ components, then $g(F'') = g(F)$. Taking $L = L_1$ and $L' = L_2$ in this case shows that $g_{\text{alg}}(L') \leq g_{\text{alg}}(L)$. If L_2 has $m-1$ components, then $g(F'') = g(F)+1$. Taking $L = L_2$ and $L' = L_1$ in this case shows that $g_{\text{alg}}(L) \leq g_{\text{alg}}(L')+1$. This proves the two required inequalities. □

With these lemmas in hand we can prove Proposition 2.1

Proof (of Proposition 2.1). The equality of the first two definitions follows from Lemma 2.2. We prove equality between the first and third definitions. Suppose that a link L can be obtained from an Alexander polynomial one knot K' by n oriented band moves. Suppose that n_+ of these band moves increase the number of components and n_- of the moves decrease the number of components. Since $n_+ + n_- = n$ and $n_+ - n_- = r-1$, we see that $n_- = \frac{n-r+1}{2}$. Thus Lemma 2.4 shows $g_{\text{alg}}(L') \leq \frac{n-r+1}{2}$. Conversely, suppose that F is a Seifert surface for L and F' a subsurface cobounding an Alexander polynomial one knot K' realizing the algebraic genus. Consider the surface $\Sigma = F \setminus \operatorname{int} F'$. The surface Σ can be constructed by starting with K' and attaching bands. Since $g(\Sigma) = g_{\text{alg}}(L) = g(F) - g(F')$, the surface Σ can be constructed by attaching oriented $2g_{\text{alg}}(L) + r - 1$ bands to K'. Thus L can be constructed from K' by $2g_{\text{alg}}(L) + r - 1$ band moves, as required. □

We conclude the section by proving Theorem 1.1.

Theorem 1.1 *If L and L' are oriented links related by a null-homologous m-twist and a null-homologous n-twist for $m, n \in \mathbb{Z}$ such that $-mn$ is a square, then*

$$|g_{\text{alg}}(L) - g_{\text{alg}}(L')| \leq 1.$$

Proof. Suppose that L' is obtained from L by a null-homologous m-twist and a null-homologous n-twist. That is, L' is obtained from L by performing $1/m$-surgery and $1/n$-surgery on two unknotted curves, say C_1 and C_2. First we construct a nice Seifert surface for L. As shown in Figure 2, we can choose a diagram for L such that Seifert's algorithm yields a Seifert surface F for L which is disjoint from C_1 and C_2. Moreover we can choose a basis $H_1(F;\mathbb{Z})$ such that the classes linking non-trivially with C_1 and C_2 can be represented by a collection of disjoint curves forming an un-

link. Lemma 2.3 shows that by further stabilizing F we can assume that it realizes the algebraic genus of L.

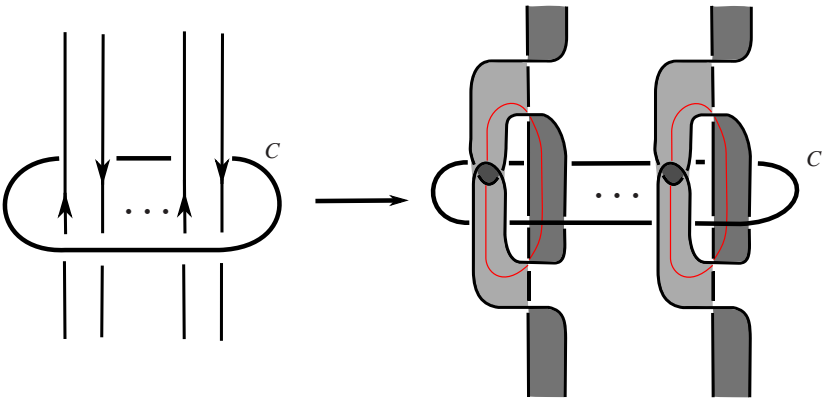


Fig. 2 Choosing a nice surface to twist. The red curves represent the only homology classes in the basis passing linking with the surgery curve.

Thus with respect to an appropriate ordering of the bases, we can assume that L and L' have Seifert matrices M and M' of the form

$$M = \begin{pmatrix} 0 \cdots 0 & & \\ \vdots \quad \vdots & B & F_1 \\ 0 \cdots 0 & & \\ & 0 \cdots 0 & \\ C & \vdots \quad \vdots & F_2 \\ & 0 \cdots 0 & \\ F_3 & F_4 & F_5 \end{pmatrix},$$

and

$$M' = \begin{pmatrix} -m \cdots -m & & \\ \vdots \quad \vdots & B & F_1 \\ -m \cdots -m & & \\ & -n \cdots -n & \\ C & \vdots \quad \vdots & F_2 \\ & -n \cdots -n & \\ F_3 & F_4 & F_5 \end{pmatrix}.$$

If $-mn$ is a square, then we can assume that m and n take the form $m = -ax^2$ and $n = ay^2$, for some integers x, y and a. By stabilizing M' we obtain a new Seifert matrix M'' for L':

$$M'' = \left(\begin{array}{ccccccc|cc} ax^2 & \cdots & ax^2 & & & & & -ax & 0 \\ \vdots & & \vdots & & B & & F_1 & \vdots & \vdots \\ ax^2 & \cdots & ax^2 & & & & & -ax & 0 \\ & & & -ay^2 & \cdots & -ay^2 & & 0 & 0 \\ & C & & \vdots & & \vdots & F_2 & \vdots & \vdots \\ & & & -ay^2 & \cdots & -ay^2 & & 0 & 0 \\ & & & & & & & 0 & 0 \\ & F_3 & & & F_4 & & F_5 & \vdots & \vdots \\ & & & & & & & 0 & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \cdots 0 & 0 & 1 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \cdots 0 & 0 & 0 \end{array}\right).$$

Consider the following matrix identity:

$$\begin{pmatrix} 1 & 0 & x & 0 \\ 0 & 1 & y & ay \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} A+ax^2 & B & -ax & 0 \\ C & D-ay^2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ x & y & 1 & 0 \\ 0 & ay & 0 & 1 \end{pmatrix} = \begin{pmatrix} A & B & -ax & x \\ C & D & 0 & y \\ 0 & ay & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1}$$

By replacing the entries of the matrices in (1) by identity matrices and block matrices of the appropriate size, we see that there is an invertible matrix P such that

$$P^T M'' P = \left(\begin{array}{ccccccc|cc} 0 & \cdots & 0 & & & & & -ax & x \\ \vdots & & \vdots & & B & & F_1 & \vdots & \vdots \\ 0 & \cdots & 0 & & & & & -ax & x \\ & & & 0 & \cdots & 0 & & 0 & y \\ & C & & \vdots & & \vdots & F_2 & \vdots & \vdots \\ & & & 0 & \cdots & 0 & & 0 & y \\ & & & & & & & 0 & 0 \\ & F_3 & & & F_4 & & F_5 & \vdots & \vdots \\ & & & & & & & 0 & 0 \\ \hline 0 & \cdots & 0 & ay & \cdots & ay & 0 \cdots 0 & 0 & 1 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \cdots 0 & 0 & 0 \end{array}\right).$$

Since the upper left submatrix of $P^T M'' P$ is precisely M, this shows that L' has a Seifert matrix obtained by adjoining two additional rows and columns to M. Since we started with a surface F realizing the algebraic genus, it follows that

$$g_{\text{alg}}(L') \leq g_{\text{alg}}(L) + 1. \tag{2}$$

Since L' can be obtained from L by a null-homologous $-m$-twist and a null-homologous $-n$-twist, we can reverse the roles of L and L' in (2). This gives the desired result:

$$|g_{\text{alg}}(L) - g_{\text{alg}}(L')| \leq 1. \qquad \square$$

3 Decreasing the algebraic genus

Theorem 1.1 accounts for half of Theorem 1.2. In order to complete the proof we need to show that there are always pairs of null-homologous twisting operations that decrease the algebraic genus. This can be done by adapting the argument used by Livingston to prove that any knot can be converted to the unknot using at most $2g$ null-homologous twists [14].

Proposition 3.1 *Given a link L with $g_{\mathrm{alg}}(L) > 0$, then L can be obtained from a link L' with $g_{\mathrm{alg}}(L') = g_{\mathrm{alg}}(L) - 1$ by a null-homologous $+1$-twist and a null-homologous -1-twist.*

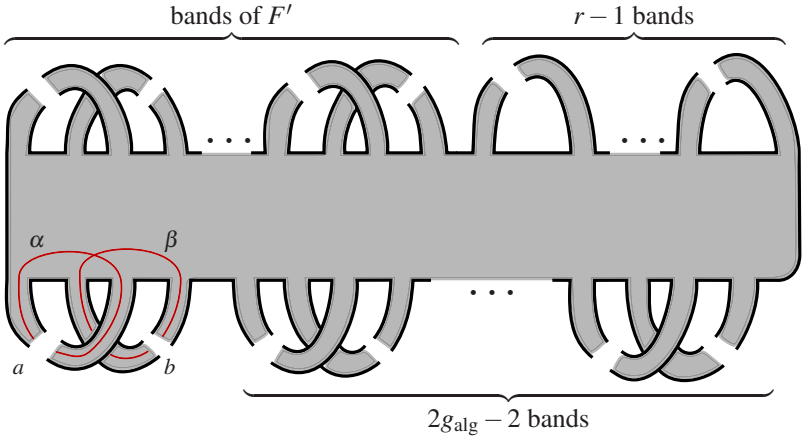


Fig. 3 Arranging the handles of the surface F. The gaps in the bands indicate that they may be knotted, linked together and twisted.

Proof. Suppose that L has r components. Consider a Seifert surface for F for L which realizes the algebraic genus. This contains a connected subsurface F' with $\partial F'$ a knot with Alexander polynomial one and $g_{\mathrm{alg}}(L) = g(F) - g(F')$. We may view F as obtained by attaching $2g_{\mathrm{alg}} + r - 1$ handles to F'. So if we present F' as a surface obtained by attaching $2g(F)$ bands to a disk, then we can present F as being obtained by attaching $2g_{\mathrm{alg}}(L) + r - 1$ further handles to this disk. Furthermore by performing handle slides, we can assume that the bands are grouped together into three groups: the $g(F')$ pairs of bands comprising F', the $r-1$ bands increasing the number of components and the g_{alg} pairs of bands contributing to the algebraic genus of L. This is illustrating in Figure 3. Let a and b be a pair of handles contributing non-trivially to $g_{\mathrm{alg}}(L)$. Let α and β be curves running over the cores of these handles as

shown in Figure 3. Let F'' be the surface obtained by deleting the handles a and b from F and take L' to be the boundary of F''. The existence of the surface F'' shows that $g_{\text{alg}}(L') \leq g_{\text{alg}}(L) - 1$. However L is obtained from L' by a pair of oriented band moves, so Lemma 2.4 shows that $g_{\text{alg}}(L') = g_{\text{alg}}(L) - 1$.

The aim is to find a pair of null-homologous twists which will transform L' into L. We will produce these twists by taking a surgery presentation for L and manipulating it until we find a surgery presentation for L which is a diagram for L' with the addition of two appropriately framed surgery curves.

By definition the framing of the curve α is given by $\theta(\alpha, \alpha)$, where θ is the Seifert form of F. We may assume that α has odd framing. If β has odd framing, then we can simply use b in place of a. If both α and β have even framing, then we can change our handle decomposition of F by sliding a over b. After such a slide the curve α' running over a has the homology class of $\alpha + \beta$. This curve has odd framing, since

$$\begin{aligned}\theta(\alpha+\beta, \alpha+\beta) &= \theta(\alpha,\alpha) + \theta(\beta,\alpha) + \theta(\alpha,\beta) + \theta(\beta,\beta) \\ &\equiv \theta(\beta,\alpha) - \theta(\alpha,\beta) \equiv 1 \bmod 2,\end{aligned}$$

where we have used that the anti-symmetrization of the Seifert form is the intersection form of $H_1(F;\mathbb{Z})$ in the second line.

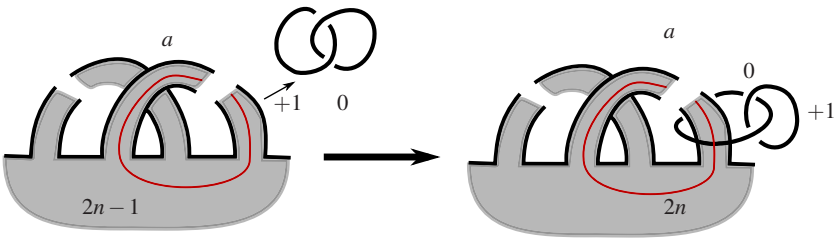


Fig. 4 Sliding the band a over the $+1$-framed component.

Now we produce our surgery diagram. Introduce a Hopf link to S^3 with one component 0-framed and the other $+1$-framed. This provides a surgery presentation for S^3. Slide the band a over the $+1$-framed curve. After this slide, the 0-framed curve forms a meridian of a and the framing of the core curve α becomes an even integer. Since the 0-framed curve forms a meridian of a, we can slide other bands over it to effect "crossing changes" between a and other bands in the handle decomposition of F and also to pass the band a through itself. Thus after some sequence of such moves we can assume that the curve α is unknotted and the band a lies entirely above the band b and is unlinked from all other bands. Moreover notice that sliding a over the 0-framed curve changes the framing of α by ± 2 and that sliding any other band over a does not change the framing of α. Thus the framing on α is still an even integer and moreover by performing further slides we can assume that α has framing 0. So by performing a sequence of isotopies and handle slides, we can

obtain a surface F' where a appears as in Figure 5 but F' is otherwise identical to F. Notice that the link $\partial F'$ is isotopic to L', the link bounding the surface F''.

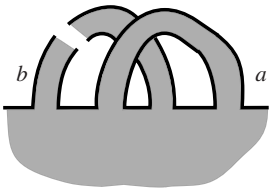


Fig. 5 The band a after simplification.

Now slide the 0-framed component of the Hopf link over the $+1$-framed component so that it becomes a two component unlink with a $+1$-framed component and a -1-framed component. Thus we have a surgery description showing that L can be obtained from L' by perfoming a null-homologous $+1$-twist and a null-homologous -1-twist as required. □

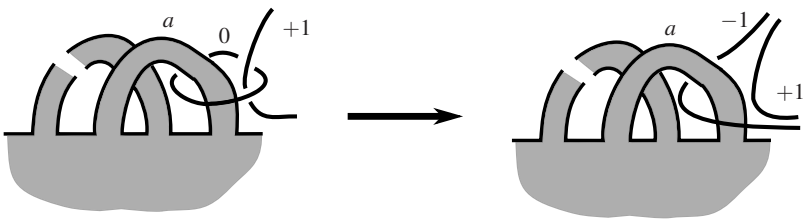


Fig. 6 Sliding the 0-framed component over the $+1$-framed component.

Thus we can prove Theorem 1.2.

Theorem 1.2 *For any link L, we have*

$$g_{\mathrm{alg}}(L) = \min\left\{ \max\{n,p\} \,\middle|\, \begin{array}{l} L \text{ can be converted to a link } L' \text{ with } g_{\mathrm{alg}}(L') = \\ 0 \text{ by } p \text{ null-homologous } +1\text{-twists and } n \text{ null-} \\ \text{homologous } -1\text{-twists.} \end{array} \right\}.$$

Proof. Theorem 1.1 shows that if L is obtained from L' with $g_{\mathrm{alg}}(L') = 0$ by p null-homologous $+1$-twists and n null-homologous -1-twists, then $g_{\mathrm{alg}}(L) \leq \max\{n,p\}$. On the other hand, applying Proposition 3.1 repeatedly shows that L can be converted into a link L' with $g_{\mathrm{alg}}(L') = 0$, by $g_{\mathrm{alg}}(L)$ pairs of null-homologous $+1$-twists and -1-twists. □

4 Satellite knots

In this section we prove Theorem 1.4. First we note that null-homologous twisting is preserved under satellite operations.

Lemma 4.1 *Let K and K' be knots related by a null-homologous n-twist, then for any pattern $P \subseteq S^1 \times D^2$, the satellite knots $P(K)$ and $P(K')$ are related by a null-homologous n-twist.*

Proof. Let X_P denote the complement $X_P = S^1 \times D^2 \setminus \nu P$ which comes with a meridian μ and distinguished longitude λ in $\partial(S^1 \times D^2)$. The knot complement $S^3 \setminus \nu P(K)$ is obtained by gluing X_P to $S^3 \setminus \nu K$ so that μ and λ are glued to the meridian and null-homologous longitude of K respectively. The complement $S^3 \setminus \nu P(K')$ is constructed similarly by gluing X_P to $S^3 \setminus \nu K'$.

Since K and K' are related by a null-homologous n-twist there is a null-homologous curve $C \subset S^3 \setminus \nu K$ such that performing $1/n$ surgery on C yields $S^3 \setminus \nu K'$. Since C is null-homologous in $S^3 \setminus \nu K$, surgering C takes the meridian and null-homologous longitude of $S^3 \setminus \nu K$ to the meridian and null-homologous longitude of $S^3 \setminus \nu K'$. Thus if we consider C as a curve in $S^3 \setminus \nu P(K) = S^3 \setminus \nu K \cup X_P$, we see that $1/n$ surgery on C will produce $S^3 \setminus \nu P(K')$. Since C is null-homologous in $S^3 \setminus \nu K$ it is null-homologous in $S^3 \setminus \nu P(K)$, thus $P(K)$ and $P(K')$ are related by a null-homologous n-twist. □

Lemma 4.2 *Let K' be a knot with $\Delta_{K'}(t) = 1$, then for any pattern P, we have $g_{\text{alg}}(P(K')) = g_{\text{alg}}(P(U))$.*

Proof. We refer the reader to [12, Proof of Theorem 6.15]. In this proof, Lickorish constructs a Seifert matrix for $P(K')$ of the form $\begin{pmatrix} M & 0 \\ 0 & X \end{pmatrix}$ where M is a Seifert matrix for $P(U)$ and X is a matrix satisfying $\det(tX - X^T) = \Delta_{K'}(t^w)$, where w is the winding number of P. Since $\Delta_{K'}(t) = 1$, this shows that $P(K')$ and $P(U)$ are S-equivalent, and hence have the same algebraic genus. □

Theorem 1.4 *For a satellite knot $P(K)$, we have*

$$g_{\text{alg}}(P(K)) \leq g_{\text{alg}}(P(U)) + g_{\text{alg}}(K).$$

Proof. By Proposition 3.1, K can be converted into a knot K' with Alexander polynomial one by a sequence of at most $g_{\text{alg}}(K)$ pairs of null-homologus $+1$-twists and -1-twists. By Lemma 4.1 this shows that $P(K)$ can be converted to $P(K')$ by a similar sequence of twists. Thus we have

$$g_{\text{alg}}(P(K)) \leq g_{\text{alg}}(K) + g_{\text{alg}}(P(K')).$$

By Lemma 4.2 we have $g_{\text{alg}}(P(K')) = g_{\text{alg}}(P(U))$ so this is the desired bound. □

5 Torus knots

We now gather the ingredients to prove Theorem 1.5.

Lemma 5.1 *For any $a,b \geq 1$, we have*

$$g_{\text{alg}}(T_{2^a,2^b}) < \frac{2^{a+b}}{3}.$$

Proof. We will show that $T_{2^a,2^b}$ can be converted to the unlink using at most $\frac{2^{a+b}}{3}$ null-homologous twists. By Theorem 1.1, this shows that

$$g_{\text{alg}}(T_{2^a,2^b}) \leq \left\lfloor \frac{2^{a+b}}{3} \right\rfloor < \frac{2^{a+b}}{3},$$

where the strict inequality follows from the fact that $\frac{2^{a+b}}{3}$ is not an integer. Given a full twist on 2^{k+1} strands oriented so that all crossings are positive, we can perform a null-homologous -1-twist to produce two parallel sets of 2^k strands each with two positive full twists. This is depicted in Figure 7. Thus if we let T_k denote the number null-homologous twisting moves required to undo a full twist on 2^k strands we see that T_k satisfies the recursive bound $T_{k+1} \leq 1+4T_k$. Note that $T_1 = 1$ since a full twist on two strands can be undone by a single crossing change. Now the solution to the recursion relation $c_{k+1} = 1+4c_k$ with $c_1 = 1$ is $c_k = \frac{4^k-1}{3}$. Thus we see that $T_k \leq \frac{4^k-1}{3}$ for all k.

Without loss of generality suppose that $2^a \leq 2^b$. The link $T_{2^a,2^b}$ can be viewed as 2^{b-a} full twists on 2^a strands. Thus $T_{2^a,2^b}$ can be converted into the unlink $T_{2^a,0}$ by removing 2^{b-a} positive full twists on 2^a strands. Thus

$$g_{\text{alg}}(T_{2^a,2^b}) \leq 2^{b-a} \times \frac{4^a-1}{3} < \frac{2^{a+b}}{3}.$$

□

Lemma 5.2 *For any $a,b,c \geq 1$,*

$$g_{\text{alg}}(T_{a,b+c}) \leq g_{\text{alg}}(T_{a,b}) + g_{\text{alg}}(T_{a,c}) + a.$$

Proof. Observe that $T_{a,b+c}$ can be converted into the split link $T_{a,b} \sqcup T_{a,c}$ by performing a oriented crossing resolutions: if one considers $T_{a,b+c}$ as the closure of the braid word $(\sigma_1 \cdots \sigma_{b+c-1})^a$, then these resolutions corresponding to deleting all instances of σ_b from this braid word. Thus $T_{a,b+c}$ can be obtained from the split link $T_{a,b} \sqcup T_{a,c}$ by a oriented band moves. Thus, Lemma 2.4 implies that

$$g_{\text{alg}}(T_{a,b+c}) \leq g_{\text{alg}}(T_{a,b}) + g_{\text{alg}}(T_{a,c}) + a,$$

as required. □

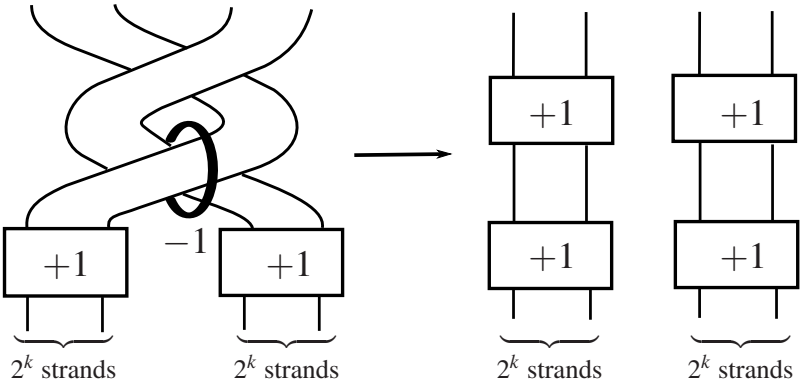


Fig. 7 Converting a full twist on 2^{k+1} strands into four full twists on 2^k strands with a single null-homologous twist. Each box contains a full twist.

Combining the bounds from Lemma 5.1 and Lemma 5.2 yields Theorem 1.5.

Theorem 1.5 *For any torus knot or link $T_{p,q}$ with $p,q>1$ we have*

$$g_4^{\text{top}}(T_{p,q}) \leq g_{\text{alg}}(T_{p,q}) < \frac{pq}{3} + p\log_2 q + q\log_2 p.$$

Proof. Suppose that p and q are represented in binary as $\sum_{i=0}^{k} 2^{a_i} = q$ and $p = \sum_{j=0}^{l} 2^{b_j}$, i.e. so that when represented in binary q has $k+1$ non-zero digits and p has $l+1$ non-zero digits when represented in binary. Notice that we have

$$k \leq \log_2 q \quad \text{and} \quad l \leq \log_2 p. \tag{3}$$

For any given any i and j, Lemma 5.1 shows that the algebraic genus of the link $L_{i,j} := T_{2^{b_j},2^{a_i}}$ satisfies

$$g_{\text{alg}}(L_{i,j}) < \frac{2^{a_i+b_j}}{3}.$$

By applying Lemma 5.2 to $L_{0,j},\dots,L_{k,j}$, we see that $T_{q,2^{b_j}}$ satisfies

$$\begin{aligned} g_{\text{alg}}(T_{q,2^{b_j}}) &< 2^{b_j}k + \sum_{i=0}^{k}\left(\frac{2^{a_i+b_j}}{3}\right) \\ &= 2^{b^j}k + \frac{2^{b_j}q}{3}. \end{aligned}$$

So by applying Lemma 5.2 to $T_{q,2^{b_0}},\dots,T_{q,2^{b_l}}$, we see that $T_{p,q}$ satisfies

$$g_{\text{alg}}(T_{p,q}) < lq + \sum_{j=0}^{l} \left(\frac{2^{b_j} q}{3} + 2^{b_j} k \right)$$
$$= \frac{pq}{3} + ql + pk.$$

By (3), this shows that

$$g_{\text{alg}}(T_{p,q}) < \frac{pq}{3} + p \log_2 q + q \log_2 p,$$

as required. □

6 Anisotropic Seifert forms

In this section we prove Proposition 1.3 which shows that shows that most pairs of null-homologous twisting operations can change the algebraic genus and the topological slice genus by two. Recall that a quadratic form $q : \mathbb{Z}^n \to \mathbb{Z}$ is *isotropic* if there is $v \neq 0$ such that $q(v) = 0$ and *anisotropic* otherwise.

Lemma 6.1 *Let K be a knot with a Seifert surface F and associated Seifert form θ. If $g_4^{\text{top}}(K) < g(F)$, then the quadratic form on $H_1(F;\mathbb{Z})$ defined by $v \mapsto \theta(v,v)$ is isotropic.*

Proof. Given a knot with a genus g Seifert surface F and Seifert form $\theta : \mathbb{Z}^{2g} \times \mathbb{Z}^{2g} \to \mathbb{Z}$, Taylor defines a knot invariant [16]:

$$t(K) := g - a(\theta),$$

where $a(\theta)$ is the rank of a maximal isotropic subgroup of $\mathbb{Z}^{2g}$ (i.e. the maximal rank of a subgroup on which θ is identically 0). As discussed in [11, Section 2], this invariant is known to be a lower bound for the topological slice genus. In particular, we have $a(\theta) \geq g(F) - g_4^{\text{top}}(K)$. Thus if $g_4^{\text{top}}(K) < g(F)$, then θ has a non-trivial isotropic subgroup, as required. □

In order to apply Lemma 6.1 we will need to show certain forms are anisotropic. This requires some elementary number theory. For a prime p, we use $\left(\frac{n}{p}\right)$ to denote the Legendre symbol of n modulo p.

Lemma 6.2 *Let p be an odd prime and let a,b,M,N be positive integers coprime to p. The quadratic form*

$$q(x_1,x_2,x_3,x_4) = ax_1^2 - bx_2^2 + p(Mx_3^2 + Nx_4^2)$$

is anisotropic if $\left(\frac{ab}{p}\right) = -1$.

Proof. We will show that if q is isotropic, then $\left(\frac{ab}{p}\right) = 1$. If q is isotropic, then there are integers y_1, y_2, y_3, y_4 such that $\gcd(y_1, y_2, y_3, y_4) = 1$ and

$$ay_2^2 - by_1^2 = p(My_3^2 + Ny_4^2). \tag{4}$$

Since p divides the right hand side, we see that y_1 and y_2 provide a solution to the equation $aX^2 \equiv bY^2 \bmod p$. Moreover, this is a non-trivial solution, that is $y_1, y_2 \not\equiv 0 \bmod p$. Assume for sake of contradiction that $y_1 \equiv y_2 \equiv 0 \bmod p$. If both sides of (4) are non-zero, then the largest power of p dividing the left hand side is even, but the largest power of p dividing the right hand side is odd. Thus both sides of (4) must be zero. This implies that $y_3 = y_4 = 0$, which would imply that $\gcd(y_1, \dots, y_4) \geq p > 1$. Thus we must have $y_1, y_2 \not\equiv 0 \bmod p$.

Since the quadratic residues form an index two subgroup in $(\mathbb{Z}/p\mathbb{Z})^\times$, the equation $aX^2 \equiv bY^2 \bmod p$ has a non-trivial solution if and only if both a and b are quadratic residues or both a and b are quadratic non-residues modulo p. In either case, this implies that if $aX^2 \equiv bY^2 \bmod p$ has a non-trivial solution, then $\left(\frac{ab}{p}\right) = 1$, as required. □

Lemma 6.3 *For any integer $n > 0$ which is not a square, there is an odd prime p such that*

$$\left(\frac{n}{p}\right) = -1.$$

Proof. This is a standard application of quadratic reciprocity and Dirichlet's theorem on primes in arithmetic progressions. Suppose that n has prime factorization $n = p_1^{a_1} \dots p_k^{a_k}$. Since n is not a square, at least one of the a_i is odd. Without loss of generality assume that a_1 is odd. Suppose first that p_1 is an odd prime. By the Chinese remainder theorem and Dirichlet's theorem on primes in arithmetic progressions, we can choose a prime p satisfying the congruences $p \equiv 1 \bmod 4$, $p \equiv 1 \bmod p_i$ for $i > 1$ and $p \equiv q \bmod p_1$, where q satisfies $\left(\frac{q}{p_1}\right) = -1$. It follows from quadratic reciprocity that such a p satisfies $\left(\frac{n}{p}\right) = -1$.

If $p_1 = 2$, then we choose p to be a prime satisfying $p \equiv 5 \bmod 8$ and $p \equiv 1 \bmod p_i$ for all $i > 1$. Using quadratic reciprocity and the fact that $\left(\frac{2}{p}\right) = -1$ for $p \equiv 5 \bmod 8$, we see that such a p satisfies $\left(\frac{n}{p}\right) = -1$. □

Proposition 1.3 *For any $m, n \in \mathbb{Z}$ such that $-mn$ is not a square, there is a knot K with $g_{\text{alg}}(K) = g_4^{\text{top}}(K) = 2$, which can be unknotted by performing a null-homologous m-twist and a null-homologous n-twist.*

Proof. For integers a, b, c, d let $K = K(a, b, c, d)$ be the knot as shown in Figure 8. With respect to the Seifert surface and basis shown in Figure 9, this has Seifert matrix

$$M = \begin{pmatrix} a & 0 & 1 & 0 \\ 0 & b & 0 & 1 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{pmatrix}. \tag{5}$$

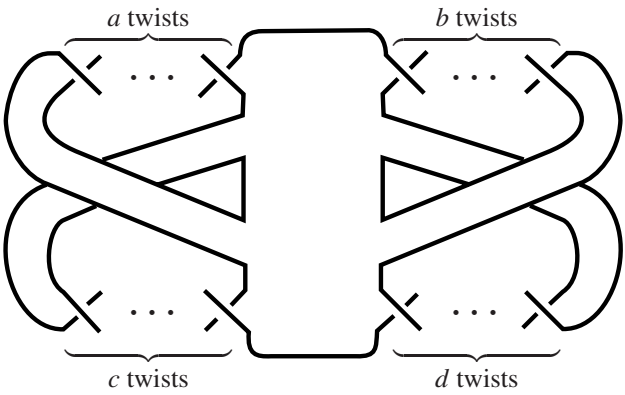


Fig. 8 The knot $K(a,b,c,d)$.

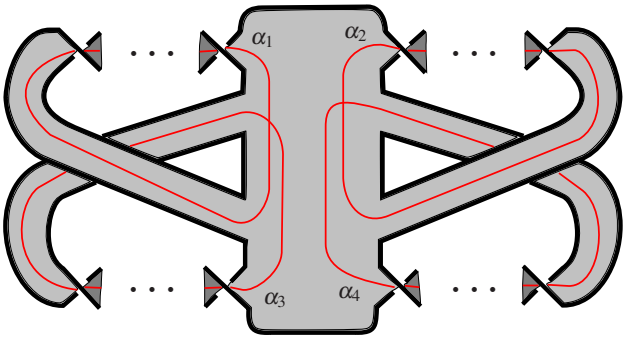


Fig. 9 A Seifert surface for $K(a,b,c,d)$.

Notice that for any c and d the knot $K(0,0,c,d)$ is the unknot. Thus we see that $K = K(a,b,c,d)$ can be unknotted by performing a null-homologous a-twist and a null-homologous b-twist. The aim is to show that for any a,b such that $-ab$ is not a square, then we can find c and d such that $g_4^{\text{top}}(K(a,b,c,d)) = 2$. Without loss of generality, assume that $a > 0$. First suppose that we also have $b > 0$. In this case, one can easily see from (5) that $\sigma(K(a,b,c,d)) = 4$ for any $c > 0$ and $d > 0$.

Thus it suffices to consider the case where $a > 0$ and $b < 0$. By Lemma 6.1 it suffices to find c,d ensuring that the Seifert form is anisotropic. That is we need to show that the quadratic form

$$q(x_1,x_2,x_3,x_4) = ax_1^2 + x_1x_3 + cx_3^2 - |b|x_2^2 + x_2x_4 + cx_4^2 \tag{6}$$

is not always isotropic. This can be diagonalized over $\mathbb{Q}$ as

$$q(x_1,x_2,x_3,x_4) = a\left(x_1+\frac{x_3}{2a}\right)^2 + a(4ac-1)\left(\frac{x_3}{2a}\right)^2 - |b|\left(x_2-\frac{x_4}{2|b|}\right)^2 + |b|(4|b|d+1)\left(\frac{x_4}{2|b|}\right)^2.$$

Since a quadratic form is isotropic over $\mathbb{Z}$ if and only if it is isotropic over $\mathbb{Q}$. It suffices to show that the form

$$\widetilde{q}(x_1,x_2,x_3,x_4) = ax_1^2 - |b|x_2^2 + a(4ac-1)x_3^2 + |b|(4|b|d+1)x_4^2$$

is anisotropic for some choice of c and d. Since we are assuming that $-ab$ is not a square, then Lemma 6.3 shows there is an odd prime p with $\left(\frac{a|b|}{p}\right) = -1$. Since p is coprime to a and b, we can find c and d such that p divides both $(4ac-1)$ and $(4|b|d+1)$ but p^2 does not divide either $(4ac-1)$ or $(4|b|d+1)$. Lemma 6.2 shows that for such c and d the Seifert form is anisotropic and hence $g_4^{\text{top}}(K(a,b,c,d)) = 2$, as required. □

Acknowledgements The statement of Theorem 1.4 was first found and proven by Peter Feller, Allison N. Miller and Juanita Pinzon-Caicedo [9]. The author is grateful to Peter Feller for bringing this result to his attention and for other illuminating conversations.

References

1. Baader, S., Feller, P., Lewark, L., Liechti, L.: On the topological 4-genus of torus knots. Trans. Amer. Math. Soc. **370**(4), 2639–2656 (2018)
2. Baader, S., Lewark, L.: The stable 4-genus of alternating knots. Asian J. Math. **21**(6), 1183–1190 (2017)
3. Borodzik, M., Friedl, S.: On the algebraic unknotting number. Trans. London Math. Soc. **1**(1), 57–84 (2014)
4. Borodzik, M., Friedl, S.: The unknotting number and classical invariants, I. Algebr. Geom. Topol. **15**(1), 85–135 (2015)
5. Feller, P.: The degree of the Alexander polynomial is an upper bound for the topological slice genus. Geom. Topol. **20**(3), 1763–1771 (2016)
6. Feller, P., Lewark, L.: On classical upper bounds for slice genera. Selecta Math. (N.S.) **24**(5), 4885–4916 (2018)
7. Feller, P., Lewark, L.: Balanced algebraic unknotting, linking forms, and surfaces in three- and four-space. arXiv:1905.08305 (2019)
8. Feller, P., McCoy, D.: On 2-bridge knots with differing smooth and topological slice genera. Proc. Amer. Math. Soc. **144**(12), 5435–5442 (2016)
9. Feller, P., Miller, A.N., Pinzon-Caicedo, J.: A note on the topological slice genus of satellite knots. arXiv:1908.03760 (2019)
10. Freedman, M.H.: The topology of four-dimensional manifolds. J. Differential Geom. **17**(3), 357–453 (1982)
11. Lewark, L., McCoy, D.: On calculating the slice genera of 11- and 12-crossing knots. Exp. Math. **28**(1), 81–94 (2019)
12. Lickorish, W.R.: An Introduction to Knot Theory. Springer (1997)
13. Liechti, L.: Positive braid knots of maximal topological 4-genus. Math. Proc. Cambridge Philos. Soc. **161**(3), 559–568 (2016)

14. Livingston, C.: Null-homologous unknottings. arXiv:1902.05405 (2019)
15. Rudolph, L.: Some topologically locally-flat surfaces in the complex projective plane. Comment. Math. Helv. **59**(4), 592–599 (1984)
16. Taylor, L.R.: On the genera of knots. In: Topology of low-dimensional manifolds (Proc. Second Sussex Conf., Chelwood Gate, 1977), *Lecture Notes in Math.*, vol. 722, pp. 144–154. Springer, Berlin (1979)

A slicing obstruction from the $10/8+4$ theorem

Linh Truong

Abstract Using the $10/8+4$ theorem of Hopkins, Lin, Shi, and Xu, we derive a smooth slicing obstruction for knots in the three-sphere using a spin 4-manifold whose boundary is 0–surgery on a knot. This improves upon the slicing obstruction bound by Vafaee and Donald that relies on Furuta's 10/8 theorem. We give an example where our obstruction is able to detect the smooth non-sliceness of a knot by using a spin 4-manifold for which the Donald-Vafaee slice obstruction fails.

1 Introduction

A knot in the three–sphere is smoothly slice if it bounds a disk that is smoothly embedded in the four–ball. Classical obstructions to sliceness include the Fox–Milnor condition [3] on the Alexander polynomial, the $\mathbb{Z}_2$–valued Arf invariant [14], and the Levine–Tristram signature [8, 15]. Furthermore, modern Floer homologies and Khovanov homology produce powerful sliceness obstructions. Heegaard Floer concordance invariants include τ of Ozsváth–Szabó [10], the $\{-1,0,+1\}$–valued invariant ε of Hom [6], the piecewise–linear function $\Upsilon(t)$ [11], the involutive Heegaard Floer homology concordance invariants $\overline{V_0}$ and $\underline{V_0}$ [5], as well as ϕ_i homomorphisms of [1]. Rasmussen [13] defined the s–invariant using Khovanov–Lee homology, and Piccirillo recently used the s–invariant to show that the Conway knot is not slice [12].

We study an obstruction to sliceness derived from handlebody theory. We call a four-manifold a two-handlebody if it can be obtained by attaching two-handles to a four-ball. In [2] Donald and Vafaee used Furuta's $10/8$ theorem [4] to obtain a slicing obstruction. This obstruction is able to detect nontrivial torsion elements

Linh Truong
University of Michigan, Ann Arbor, MI 48103, U.S.A.
e-mail: tlinh@umich.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_8

in the concordance group as well as find topologically slice knots which are not smoothly slice.

We apply the recent $10/8+4$ theorem of Hopkins, Lin, Shi, and Xu [7], which improves on Furuta's inequality, to improve the Donald–Vafaee slicing obstruction.

Theorem 1. *Let $K \subset S^3$ be a smoothly slice knot and X be a spin two-handlebody with $\partial X \cong S^3_0(X)$. If $b_2(X) \neq 1, 3,$ or 23, then*

$$b_2(X) \geq \frac{10}{8}|\sigma(X)| + 5.$$

We will give an example in Proposition 1 of a knot K and a spin two-handlebody with boundary $S^3_0(K)$ where our obstruction detects the non-sliceness of K and the Donald-Vafaee slice obstruction fails using this spin 2-handlebody.

2 The slicing obstruction

In [2] Donald and Vafaee used Furuta's 10/8 theorem to obtain a slicing obstruction.

Theorem 2 ([2]). *Let $K \subset S^3$ be a smoothly slice knot and X be a spin 2-handlebody with $\partial X \cong S^3_0(X)$. Then either $b_2(X) = 1$ or*

$$4b_2(X) \geq 5|\sigma(X)| + 12$$

Recently, Hopkins, Lin, Shi, and Xu have improved Furuta's theorem with the following 10/8+4 theorem.

Theorem 3 ([7]). *Any closed simply connected smooth spin 4-manifold M that is not homeomorphic to S^4, $S^2 \times S^2$, or $K3$ must satisfy the inequality*

$$b_2(M) \geq \frac{10}{8}|\sigma(M)| + 4.$$

Using the above theorem, we prove Theorem 1.

Proof (Proof of Theorem 1). The proof is identical to the proof of [2, Theorem 1.1], except one applies the 10/8+4 theorem of [7] instead of Furuta's 10/8 theorem.

If K is smoothly slice, then $S^3_0(K)$ embeds smoothly in S^4 (see for example [9, Theorem 1.8]). The embedding splits S^4 into two spin 4-manifolds U and V with a common boundary $S^3_0(K)$. Since $S^3_0(K)$ has the same integral homology as $S^1 \times S^2$, the Mayer-Vietoris sequence shows that U and V have the same integral homology as $S^2 \times D^2$ and $S^1 \times D^3$, respectively.

Let X be a spin 2-handlebody with $\partial X \cong \partial V \cong S^3_0(K)$ (where $\cong$ denotes orientation-preserving diffeomorphism), and let $W = X \cup_{S^3_0(K)} -V^4$. We restrict the spin structure on X to the boundary $S^3_0(K)$ and extend this spin structure on $S^3_0(K)$ over the manifold V. Then W is spin since the spin structures of X and

V agree on the boundary and spin structures behave well with respect to gluing. By Novikov additivity, W has signature $\sigma(W) = \sigma(X) + \sigma(V)$. Since $\sigma(V) = 0$, we have $\sigma(W) = \sigma(X)$. As in [2] we will show that $b_2(W) = b_2(X) - 1$. The Euler characteristic satisfies $\chi(W) = \chi(X) = 1 + b_2(X)$, where the first equality uses $\chi(V) = \chi(S^3_0(K)) = 0$ and the second equality holds since X is a 2-handlebody. Since $H_1(W,X;\mathbb{Q}) \cong H_1(V,Y;\mathbb{Q}) = 0$, it follows from the exact sequence for the pair (W,X) that $b_1(W) = b_3(W) = 0$. Therefore, $b_2(W) = b_2(X) - 1$.

If $b_2(X) \neq 1$, 3, or 23, then W cannot be homeomorphic to S^4, $S^2 \times S^2$, or $K3$. The result follows by applying the Hopkins, Lin, Shi, and Xu theorem [7]. □

Remark 1. This improves upon the slicing obstruction by Donald and Vafaee (under some restrictions on the second Betti number of the spin 2-handlebody).

We give an example of a knot and a spin 4-manifold where one can apply our obstruction. Let K' be a knot that is the closure of the braid word

$$K' = (\sigma_{12}\sigma_{11}\cdots\sigma_1)^{12}(\sigma_7\sigma_8\cdots\sigma_{12})^{-7}(\sigma_1\sigma_2\cdots\sigma_{10})^{-11}b,$$

where

$$b = (\sigma_3\sigma_2\sigma_1)(\sigma_4\sigma_3\sigma_2)(\sigma_2\sigma_1)(\sigma_3\sigma_2)(\sigma_1)^{-2}(\sigma_3\sigma_4)^{-1}\sigma_5^{-2}.$$

See Figure 1. The knot K' is presented as a generalized twisted torus knot. It is the closure of a braid formed by taking a $(13,12)$ torus knot and then adding one negative full twist on seven strands, one negative full twist on eleven strands, and the braid b. As noted in [2] the obstruction from Theorem 1 is generally easier to apply to knots like this because they can be unknotted efficiently by blowing up to remove full twists.

Proposition 1. *The knot K' is not smoothly slice.*

Proof. Add a 0-framed 2-handle to ∂D^4 along K' and blow up three times as follows. Blow up once negatively across thirteen strands on the top, then blow up positively across seven and eleven strands indicated by the two boxes labeled with -1 in Figure 1. This gives a manifold with second Betti number 4 and signature 1. The characteristic link has one component, with framing $-13^2 + 7^2 + 11^2 = 1$. Four Reidemeister I moves immediately show that this knot is isotopic to the knot in Figure 5 of [2], which in turn is isotopic to the figure eight knot.

At this point, we follow the procedure that Donald and Vafaee use to show that the figure eight knot is not slice. They apply a sequence of blow-ups, blow-downs, and handleslides until the characteristic link is empty and then apply their slice obstruction. Starting with the 1-framed figure eight knot, the same sequence of blow-ups, blow-downs, and handleslides can be applied until the characteristic link is empty.

This procedure is shown in Figure 2 and detailed below.

1. First blow up negatively twice as indicated in Figure 2(B). This gives $b_2 = 6$ and $\sigma = -1$.
2. Slide one of the two blow up curves over the other, resulting in Figure 2(C).

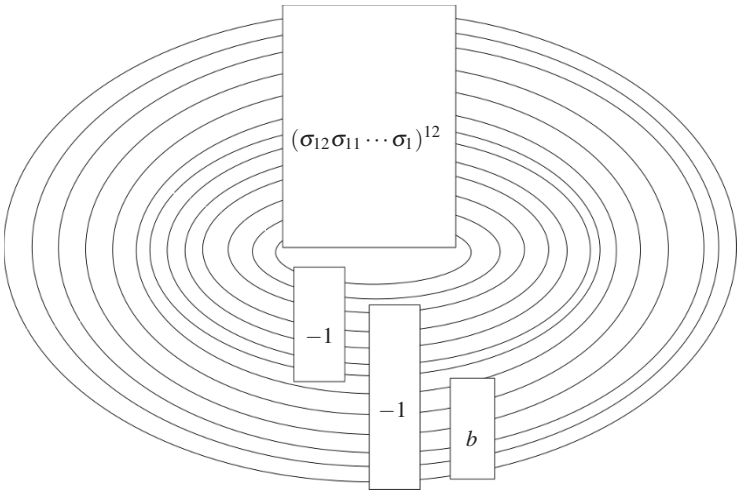


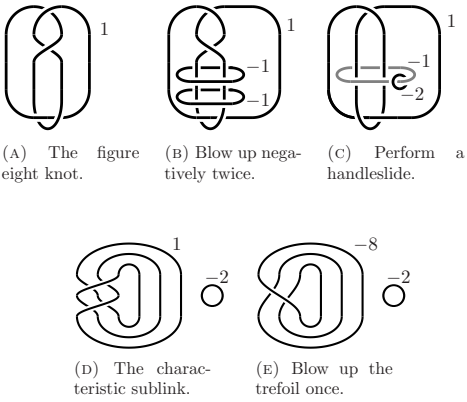
Fig. 1 The knot K' is a generalized twisted torus knot obtained from the torus knot $T_{13,12}$ by adding one negative full twist around seven adjacent strands, one negative full twist around eleven adjacent strands, and the braid b.

3. Figure 2(D) shows just the characteristic link, a split link whose components are a 1-framed trefoil and a -2-framed unknot. Blowing up negatively once around the three strands of the trefoil changes the characteristic link to a two-component unlink with framings -8 and -2 as in Figure 2(E). This is inside a four-manifold with $\sigma = -2$ and second Betti number $b_2 = 7$.
4. Positively blowing up the meridians eight times changes both framings in the characteristic link to -1 and gives $b_2 = 15$ and $\sigma = 6$.
5. Blow down negatively twice, resulting in an empty characteristic link.

The result is a spin 4-manifold X with boundary $S^3_0(K')$ with $b_2(X) = 13$ and $\sigma(X) = 8$. Thus, Theorem 1 concludes that K' is not smoothly slice. □

We observe that with the spin two-handlebody in the above proof, the Donald-Vafaee slice obstruction fails to detect the smooth non-sliceness of K'.

Acknowledgements I am grateful to John Baldwin for suggesting this topic during a problem session at the Virginia Topology Conference in December 2018. I thank Lisa Piccirillo and Maggie Miller for interesting discussions while we visited MATRIX Institute at the University of Melbourne in Creswick for the workshop Topology of manifolds: Interactions between high and low dimensions. I wish to thank the organizers of this workshop for the invitation to attend, and the NSF and MATRIX for providing funding. I would also like to thank Nathan Dunfield for his assistance with Snappy. I thank the anonymous referee for helpful comments. I was partially supported by NSF grant DMS-1606451.



(A) The figure eight knot. (B) Blow up negatively twice. (C) Perform a handleslide.

(D) The characteristic sublink. (E) Blow up the trefoil once.

Fig. 2 A sequence of blow-ups and handleslides that shows $S^3_0(K')$ bounds a spin manifold with $b_2 = 13$ and $\sigma = 8$. These diagrams come from the figure eight example in [2] with different framing coefficients.

References

1. Dai, I., Hom, J., Stoffregen, M., Truong, L.: More concordance homomorphisms from knot Floer homology (2019). To appear, Geom. Topol., arXiv:1902.03333
2. Donald, A., Vafaee, F.: A slicing obstruction from the $\frac{10}{8}$ theorem. Proc. Amer. Math. Soc. **144**(12), 5397–5405 (2016). URL https://doi.org/10.1090/proc/13056
3. Fox, R.H., Milnor, J.W.: Singularities of 2-spheres in 4-space and cobordism of knots. Osaka Math. J. **3**, 257–267 (1966). URL http://projecteuclid.org/euclid.ojm/1200691730
4. Furuta, M.: Monopole equation and the $\frac{11}{8}$-conjecture. Math. Res. Lett. **8**(3), 279–291 (2001). URL https://doi.org/10.4310/MRL.2001.v8.n3.a5
5. Hendricks, K., Manolescu, C.: Involutive Heegaard Floer homology. Duke Math. J. **166**(7), 1211–1299 (2017). URL https://doi.org/10.1215/00127094-3793141
6. Hom, J.: Bordered Heegaard Floer homology and the tau-invariant of cable knots. J. Topol. **7**(2), 287–326 (2014). URL http://dx.doi.org/10.1112/jtopol/jtt030
7. Hopkins, M.J., Lin, J., Shi, X.D., Xu, Z.: Intersection forms of spin 4-manifolds and the Pin(2)-equivariant Mahowald invariant (2018). Preprint, arXiv:1812.04052
8. Levine, J.: Knot cobordism groups in codimension two. Comment. Math. Helv. **44**, 229–244 (1969). URL https://doi.org/10.1007/BF02564525
9. Miller, A.N., Piccirillo, L.: Knot traces and concordance. J. Topol. **11**(1), 201–220 (2018). URL https://doi.org/10.1112/topo.12054
10. Ozsváth, P., Szabó, Z.: Knot Floer homology and the four-ball genus. Geom. Topol. **7**, 615–639 (2003). URL http://dx.doi.org/10.2140/gt.2003.7.615
11. Ozsváth, P.S., Stipsicz, A.I., Szabó, Z.: Concordance homomorphisms from knot Floer homology. Adv. Math. **315**, 366–426 (2017). URL https://doi.org/10.1016/j.aim.2017.05.017
12. Piccirillo, L.: The Conway knot is not slice. Ann. of Math. (2) **191**(2), 581–591 (2020). URL https://doi-org.proxy.lib.umich.edu/10.4007/annals.2020.191.2.5

13. Rasmussen, J.: Khovanov homology and the slice genus. Invent. Math. **182**(2), 419–447 (2010). URL http://dx.doi.org/10.1007/s00222-010-0275-6
14. Robertello, R.A.: An invariant of knot cobordism. Comm. Pure Appl. Math. **18**, 543–555 (1965). URL https://doi.org/10.1002/cpa.3160180309
15. Tristram, A.G.: Some cobordism invariants for links. Proc. Cambridge Philos. Soc. **66**, 251–264 (1969). URL https://doi.org/10.1017/s0305004100044947

Chapter 2

Ergodic Theory, Diophantine Approximation and Related Topics

A generalised multidimensional Jarník-Besicovitch theorem

Mumtaz Hussain

Abstract In this short note we prove a general multidimensional Jarník-Besicovitch theorem which gives the Hausdorff dimension of simultaneously approximable set of points with error of approximations dependent on continuous functions in all dimensions. Consequently, the Hausdorff dimension of the set varies along continuous functions. This resolves a problem posed by Barral-Seuret (2011).

1 Localised Jarník-Besicovitch theorem

The Jarník-Besicovitch set is of foundational nature in the theory of metric Diophantine approximation;

$$W(\tau) := \left\{ x \in [0,1) : \left| x - \frac{p}{q} \right| < \frac{1}{q^{\tau}} \quad \text{for infinitely many } (p,q) \in \mathbb{Z} \times \mathbb{N} \right\}.$$

It has been generalised in various directions such as replacing the error of approximation by an arbitrary function tending to zero and hence studying the associated metrical theory has received much attention over the years. We refer the reader to [2] for a survey of metric theory of Diophantine approximation. Staying within the scope of Jarník-Besicovtich set, for any $x \in [0,1)$, let us define the approximation order of x to be

$$\delta(x) = \sup\{\tau : x \in W(\tau)\}.$$

From the asymptotic form of Dirichlet's theorem (1842), it follows that $\delta(x) \geq 2$ for all irrational numbers x. For any $\tau \geq 2$, the classical Jarník-Besicovitch theorem (1928, 1934) states that

Mumtaz Hussain
La Trobe University, POBox199, Bendigo 3552, Australia
e-mail: m.hussain@latrobe.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_9

$$\dim_{\mathscr{H}}\{x\in[0,1):\delta(x)\geq\tau\}=\dim_{\mathscr{H}}\{x\in[0,1):\delta(x)=\tau\}=\frac{2}{\tau}.$$

Here and throughout, $\dim_{\mathscr{H}}(X)$, denotes the Hausdorff dimension of a set X. For any $s\in\mathbb{R}^+$, $\mathscr{H}^s$ denotes the s-dimensional Hausdorff measure of X. In the case $s=d$, the s-dimensional Hausdorff measure is comparable with the d-dimensional Lebesgue measure. Finally, $B(x,r)$ denotes a ball centred at x and of radius r.

In [1], Barral-Seuret investigated the structure of the set of points with their approximation order varying along a continuous function $\tau(x)\geq 2$. They called the corresponding set as the *localised Jarník-Besicovitch set*

$$W_{\rm loc}(\tau(x)):=\{x\in[0,1):\delta(x)=\tau(x)\}.$$

They proved that the Hausdorff dimension of the set $W_{\rm loc}(\tau(x))$ to be

$$\frac{2}{\min\{\tau(x):x\in\mathbb{R}\}}.$$

Roughly speaking this result gives the size of the set of real numbers with a prescribed order of approximation. For example, for real numbers $0<a<b<1$, it gives

$$\dim_{\mathscr{H}}\{x\in[a,b]:\delta(x)=2(1+x)\}=\frac{1}{1+a}.$$

The result of Barral-Seuret was further generalised to the settings of continued fractions by Wang-Wu-Xu in [8].

In the higher dimensions the analogue of Jarník-Besicovitch set has been well studied specifically by Rynne in a sequence of papers in the 90's. To state the most relevant result, we introduce a little notation first. Let $\underline{\tau}=(\tau_1,\dots,\tau_d)$ be a vector of strictly positive numbers and let $W_d(\underline{\tau})$ denote the set of all $\mathbf{x}\in[0,1)^d$ for which the system of inequalities

$$|qx_i-p_i|<q^{-\tau_i},\quad 1\leq i\leq d,$$

are satisfied for infinitely many $(p_1,\dots,p_d,q)\in\mathbb{Z}^d\times\mathbb{N}$. The Hausdorff dimension of this set was determined by Rynne [5].

Theorem 1 (Rynne, 1998). *Let* $\frac{1}{d}\leq\tau_1\leq\tau_2\dots\leq\tau_d$. *Then*

$$\dim_{\mathscr{H}}W_d(\underline{\tau})=\min_{1\leq j\leq d}\left\{\frac{1+d+j\tau_j-\sum_{i=1}^{j}\tau_i)}{1+\tau_j}\right\}.$$

In this paper, we replace the constant vector $\underline{\tau}$ in the set $W_d(\underline{\tau})$ with the function

$$\underline{\tau}(\mathbf{x}):=\{(\tau_1(x_1),\dots,\tau_d(x_d)):x_1,\dots,x_d\in[0,1]\},$$

where every function $\tau_i(x_i)$ is a continuous function on $[0,1]$. To be precise, let $W_d(\underline{\tau}(\mathbf{x}))$ denote the set of all $\mathbf{x} \in [0,1)^d$ for which the system of inequalities

$$|qx_i - p_i| < q^{-\tau_i(x_i)}, \quad 1 \le i \le d,$$

are satisfied for infinitely many $(p_1,\dots,p_d,q) \in \mathbb{Z}^d \times \mathbb{N}$. We calculate the Hausdorff dimension of this set and, thus, answer a question [1, §6] raised by Barral-Seuret of extending their one dimensional result to higher dimensions.

Theorem 2. *Let*

$$\frac{1}{d} \le \min_{x_1\in[0,1]} \tau_1(x_1) \le \min_{x_2\in[0,1]} \tau_2(x_2) \le \cdots \le \min_{x_d\in[0,1]} \tau_d(x_d).$$

Then

$$\dim_{\mathscr{H}} W_d(\underline{\tau}(\mathbf{x})) = \min_{1\le j\le d} \left\{ \frac{d+1+j \min\limits_{x_j\in[0,1]} \tau_j(x_j) - \sum_{i=1}^{j} j \min\limits_{x_i\in[0,1]} \tau_i(x_i)}{1+\min\limits_{x_j\in[0,1]} \tau_j(x_j)} \right\}.$$

2 Proof

2.1 The upper bound

The upper bound relies on the natural covering of the set $W_d(\underline{\tau}(\mathbf{x}))$. Here we prove for $d=2$ for clarity by showing that the s-dimensional Hausdorff measure of this set is zero whenever $s > \dim_{\mathscr{H}} W_2(\underline{\tau}(\mathbf{x}))$. The general case $d>2$ follows similarly.

$$\begin{aligned}
W_2(\underline{\tau}(\mathbf{x})) &= \left\{ (x_1,x_2)\in[0,1]^2 : \begin{array}{c} |qx_1-p_1| < q^{-\tau_1(x_1)},\ |qx_2-p_2| < q^{-\tau_2(x_2)} \\ \text{for infinitely many} \quad (p_1,p_2,q)\in\mathbb{Z}^2\times\mathbb{N} \end{array} \right\} \\
&\subseteq \bigcup_{q=N}^{\infty} \bigcup_{p_1,p_2\le q} \left\{ (x_1,x_2)\in[0,1]^2 : |qx_1-p_1| < q^{-\tau_1(x_1)},\ |qx_2-p_2| < q^{-\tau_2(x_2)} \right\} \\
&= \bigcup_{q=N}^{\infty} \bigcup_{p_1,p_2\le q} B\left(\frac{p_1}{q}, q^{-1-\tau_1(x_1)}\right) \times B\left(\frac{p_2}{q}, q^{-1-\tau_2(x_2)}\right) \\
&\subseteq \bigcup_{q=N}^{\infty} \bigcup_{p_1,p_2\le q}^{\infty} B\left(\frac{p_1}{q}, q^{-1-\min\limits_{x_1\in[0,1]}\tau_1(x_1)}\right) \times B\left(\frac{p_2}{q}, q^{-1-\min\limits_{x_2\in[0,1]}\tau_2(x_2)}\right)
\end{aligned}$$

So, $W_2(\underline{\tau}(\mathbf{x}))$ is a subset of a collection of rectangles and each one of them

$$R = B\left(\frac{p_1}{q}, \frac{1}{q^{1+\min\limits_{x_1\in[0,1]}\tau_1(x_1)}}\right) \times B\left(\frac{p_2}{q}, \frac{1}{q^{1+\min\limits_{x_2\in[0,1]}\tau_2(x_2)}}\right)$$

can be covered in two ways: either by collection of squares formed by shorter side lengths or by a bigger square of side length as the longer side of the rectangle.

Case I. Since $\min\limits_{x_1\in[0,1]}\tau_1(x_1) \le \min\limits_{x_2\in[0,1]}\tau_2(x_2)$, the rectangle R can be covered by at most

$$2q^{\left(\min\limits_{x_2\in[0,1]}\tau_2(x_2)-\min\limits_{x_1\in[0,1]}\tau_1(x_1)\right)}$$

squares of side length $q^{-1-\min\limits_{x_2\in[0,1]}\tau_2(x_2)}$. Hence the s-dimensional Hausdorff measure of $W_2(\underline{\tau}(\mathbf{x}))$ can be estimated as

$$\begin{aligned}\mathscr{H}^s\left(W_2(\underline{\tau}(\mathbf{x}))\right) &\le 2\liminf_{N\to\infty}\sum_{q=N}^{\infty} q^2 q^{\left(\min\limits_{x_2\in[0,1]}\tau_2(x_2)-\min\limits_{x_1\in[0,1]}\tau_1(x_1)\right)} q^{-s\left(1+\min\limits_{x_2\in[0,1]}\tau_2(x_2)\right)} \\ &\le 2\liminf_{N\to\infty}\sum_{q=N}^{\infty} q^{2+\min\limits_{x_2\in[0,1]}\tau_2(x_2)-\min\limits_{x_1\in[0,1]}\tau_1(x_1)-s\left(1+\min\limits_{x_2\in[0,1]}\tau_2(x_2)\right)}.\end{aligned}$$

Therefore, for any

$$s > \frac{3+\min\limits_{x_2\in[0,1]}\tau_2(x_2)-\min\limits_{x_1\in[0,1]}\tau_1(x_1)}{1+\min\limits_{x_2\in[0,1]}\tau_2(x_2)},$$

$\mathscr{H}^s\left(W_2(\underline{\tau}(\mathbf{x})\right) = 0$. This shows that

$$\dim_{\mathscr{H}} W_2(\underline{\tau}(\mathbf{x}) \le \frac{3+\min\limits_{x_2\in[0,1]}\tau_2(x_2)-\min\limits_{x_1\in[0,1]}\tau_1(x_1)}{1+\min\limits_{x_2\in[0,1]}\tau_2(x_2)}.$$

Case II. The second case concerns covering the rectangle R by the square formed by the longer side length $q^{-1-\min\limits_{x_1\in[0,1]}\tau_1(x_1)}$. Hence the s-dimensional Hausdorff measure of $W_2(\underline{\tau}(\mathbf{x}))$ can be estimated as

$$\mathscr{H}^s\left(W_2(\underline{\tau}(\mathbf{x}))\right) \le \liminf_{N\to\infty}\sum_{q=N}^{\infty} q^{2-s\left(1+\min\limits_{x_1\in[0,1]}\tau_1(x_1)\right)}.$$

Therefore, for any $s > \frac{3}{1+\min\limits_{x_1\in[0,1]}\tau_1(x_1)}$, $\mathscr{H}^s\left(W_2(\underline{\tau}(\mathbf{x})\right) = 0$. This shows that

$$\dim_{\mathscr{H}} W_2(\underline{\tau}(\mathbf{x})) \le \frac{3}{1+\min\limits_{x_1\in[0,1]} \tau_1(x_1)}.$$

Hence combining both the above cases, we have

$$\dim_{\mathscr{H}} W_2(\underline{\tau}(\mathbf{x})) \le \min\left(\frac{3}{1+\min\limits_{x_1\in[0,1]} \tau_1(x_1)}, \frac{3+\min\limits_{x_2\in[0,1]} \tau_2(x_2)-\min\limits_{x_1\in[0,1]} \tau_1(x_1)}{1+\min\limits_{x_2\in[0,1]} \tau_2(x_2)}\right).$$

Following similar line of covering, as above, for arbitrary d, we have

$$\dim_{\mathscr{H}} W_d(\underline{\tau}(\mathbf{x})) \le \min_{1\le j\le d}\left\{\frac{d+1+j\min\limits_{x_j\in[0,1]} \tau_j(x_j)-\sum_{i=1}^{j}\min\limits_{x_i\in[0,1]} \tau_i(x_i)}{1+\min\limits_{x_j\in[0,1]} \tau_j(x_j)}\right\}.$$

2.2 The lower bound

The main ingredient in proving the lower bound of Theorem 2 is the following mass transference principle, from balls to rectangles, proved by Wang-Wu-Xu in [7]. We refer the reader to [3] for more intricate result regarding the generalised Hausdorff measure criterion which also, of course, implies the Hausdorff dimension results. To state the Wang-Wu-Xu result we need a bit more notation. Let $\{x_n\}_{n\in\mathbb{N}} \subset [0,1]^d$ with $d \ge 1$ be a sequence of rationals and let $\{r_n\}_{n\ge1}$ be a sequence of positive numbers tending to zero. Define the limsup set generated by balls

$$W := \left\{x \in [0,1]^d : x \in B(x_n, r_n) \text{ for i.m. } n \in \mathbb{N}\right\} = \limsup_{n\to\infty} B(x_n, r_n).$$

For any $\mathbf{a} = (a_1, \ldots, a_d)$, with $1 \le a_1 \le \ldots \le a_d$, define the limsup set generated by rectangles

$$W^{\mathbf{a}} := \left\{x \in [0,1]^d : x \in B^{\mathbf{a}}(x_n, r_n) \text{ for i.m. } n \in \mathbb{N}\right\} = \limsup_{n\to\infty} B^{\mathbf{a}}(x_n, r_n)$$

where $B^{\mathbf{a}}(x,r)$ denotes a rectangle with center x and side length $(r^{a_1} \ldots, r^{a_d})$.

The main result of [7] is the following mass transference principle (see also [6, Theorem 2.4].

Theorem 3 (Wang-Wu-Xu, 2015). *Let $\{B_i : i \ge 1\}$ be a sequence of balls such that for any ball $B \subset [0,1]^d$, $\mathscr{H}^d(B \cap \limsup_{i\to\infty} B_i) = \mathscr{H}^d(B)$. Let $\mathbf{a} = (a_1, \ldots, a_d)$, with $1 \le a_1 \le \ldots \le a_d$. Then we have*

$$\dim_{\mathscr{H}} W^{\mathbf{a}} \geq \min_{1\leq j\leq d}\left\{\frac{d+ja_j-\sum_{i=1}^{j}a_i}{a_j}\right\} := s(\mathbf{a})$$

and for any ball $B \subset [0,1]^d$,

$$\mathscr{H}^{s(\mathbf{a})}(B\cap W^{\mathbf{a}}) = \mathscr{H}^{s(\mathbf{a})}(B).$$

It is worth stressing that the Lebesgue measure of the set $W^{\mathbf{a}}$ being full i.e. $\mathscr{H}^d(W^{\mathbf{a}}) = 1$, in general settings, follows from the Khintchine-Groshev type theorem proved in [4, 2]. Having Theorem 3 at our disposal, we are in a position to prove the lower bound of Theorem 2. First note that the sequence of rectangles in $[0,1]^d$ can be written as

$$B^{\mathbf{a}}(x_n, r_n) := \prod_{i=1}^{d} B\left(x_{n,i}, r_n^{a_i}\right)$$

It is also clear from the definition of Hausdorff measure that, since $s(\mathbf{a}) \leq d$, we have $\mathscr{H}^{s(\mathbf{a})}(B\cap W^{\mathbf{a}}) = \mathscr{H}^{s(\mathbf{a})}(B) > 0$.

Now consider a localised limsup set by replacing the constant exponents a_i with continuous functions $a_i(x_i)$ for all $1 \leq i \leq d$. Given a sequence of balls $\{B^{\mathbf{a}(\mathbf{x})}(x_n, r_n)\}_{n\geq 1}$ in a compact bounded cube $\mathscr{C} = C_1 \times \cdots \times C_d$ of $[0,1]^d$, where

$$\mathbf{a}(\mathbf{x}) = (a_1(x_1), \ldots, a_d(x_d))$$

is a d-dimensional continuous function with

$$1 \leq a_1(x_1) \leq \ldots \leq a_d(x_d).$$

Consider the limsup set

$$W^{\mathbf{a},L} = \left\{\mathbf{x} \in \mathscr{C} : \mathbf{x} \in B^{\mathbf{a}(\mathbf{x})}(x_n, r_n) \text{ for infinitely many } n \in \mathbb{N}\right\}.$$

Let

$$a_0 = \min_{\mathbf{x}\in\mathscr{C}}\{a_1(x_1), \ldots, a_d(x_d)\}.$$

Then there exists a ball $B := B^{\mathbf{a}(\mathbf{x})}(x_n, r_n) \subset \mathscr{C}$ such that

$$\min(a_1(x_1), \ldots, a_d(x_d)) \leq a_0 + \varepsilon \quad \forall \mathbf{x} \in B. \tag{1}$$

Then

$$W^{a_0+\varepsilon} = \limsup_{n\to\infty} B^{a_0+\varepsilon}(x_n, r_n) = \limsup_{n\to\infty} \prod_{i=1}^{d} B\left(x_{n,i}, r_n^{a_0+\varepsilon}\right).$$

Now for $\mathbf{a}(\mathbf{x})$ satisfying (1), define

$$s(\mathbf{a}(\mathbf{x})) := \min_{1\le j\le d}\left\{\frac{d+j\min\limits_{x_j\in C_j} a_j(x_j)-\sum_{i=1}^{j}\min\limits_{x_i\in C_i} a_i(x_i)}{\min\limits_{x_j\in C_j} a_j(x_j)}\right\}.$$

Then

$$\begin{aligned}\mathscr{H}^{s(\mathbf{a}(\mathbf{x}))}(W^{\mathbf{a},L}) &\ge \mathscr{H}^{s(\mathbf{a}(\mathbf{x}))}(B\cap W^{\mathbf{a},L})\\ &\ge \mathscr{H}^{s(\mathbf{a}(\mathbf{x}))}(B\cap W^{a_0+\varepsilon})\\ &\ge \mathscr{H}^{s(\mathbf{a}(\mathbf{x}))}(B) \quad \text{by letting } \varepsilon\to 0\\ &> 0.\end{aligned}$$

Hence from the definition of Hausdorff dimension, it follows that

$$\dim_{\mathscr{H}} W^{\mathbf{a},L} \ge s(\mathbf{a}(\mathbf{x})).$$

The lower bound of the proof of Theorem 2 follows by identifying

$$\mathscr{C}=[0,1]^d, \mathbf{a}(\mathbf{x})=\underline{\tau}(\mathbf{x}), B(x_i,r_i)=B\left(\frac{p_i}{q},\frac{1}{q^{1+\tau_i(x_i)}}\right)$$

to yield that

$$\dim_{\mathscr{H}} W_d(\underline{\tau}(\mathbf{x})) \ge \min_{1\le j\le d}\left\{\frac{d+1+j\min\limits_{x_j\in[0,1]}\tau_j(x_j)-\sum_{i=1}^{j}\min\limits_{x_i\in[0,1]}\tau_i(x_i)}{1+\min\limits_{x_j\in[0,1]}\tau_j(x_j)}\right\}.$$

Remark 1. It is worth stressing that if we look at the set $W^{\mathbf{a},L}$ locally, then the power functions $(a_i(x_i))$ in the above proof are almost constants. In this sense the Hausdorff measure result of Theorem 3 is applicable. In comparison, the proof of Barral-Seuret [1] is much more involved as they tackled the problem of exact approximation order and Hausdorff dimension of level sets.

Finally we would like to point out that, very recently, Wang-Wu has introduced a mass transference principle [6] from rectangles to rectangles for the linear form settings. Given this new avatar, we envisage that the main result of this paper may be extended to the dual linear forms but it would require careful synthesis of the framework introduced in their paper.

Acknowledgements The author is supported by the ARC DP200100994. Part of this work was carried out during the workshop "Ergodic Theory, Diophantine approximation and related topics" sponsored by the MATRIX Research Institute. The author thank Baowei Wang and Weiliang Wang for useful discussions on this topic.

References

1. Julien Barral and Stéphane Seuret, *A localized Jarník-Besicovitch theorem*, Adv. Math. **226** (2011), no. 4, 3191–3215. MR 2764886
2. Victor Beresnevich, Felipe Ramírez, and Sanju Velani, *Metric Diophantine approximation: aspects of recent work*, Dynamics and analytic number theory, London Math. Soc. Lecture Note Ser., vol. 437, Cambridge Univ. Press, Cambridge, 2016, pp. 1–95. MR 3618787
3. Mumtaz Hussain and David Simmons, *A general principle for Hausdorff measure*, Proc. Amer. Math. Soc. **147** (2019), no. 9, 3897–3904. MR 3993782
4. Mumtaz Hussain and Tatiana Yusupova, *A note on the weighted Khintchine-Groshev theorem*, J. Théor. Nombres Bordeaux **26** (2014), no. 2, 385–397. MR 3320485
5. Bryan P. Rynne, *Hausdorff dimension and generalized simultaneous Diophantine approximation*, Bull. London Math. Soc. **30** (1998), no. 4, 365–376. MR 1620813
6. Bao-Wei Wang and Jun Wu, *Mass transference principle from rectangles to rectangles in Diophantine approximation*, Pre-Print: arXiv:1909.00924 (2019).
7. Bao-Wei Wang, Jun Wu, and Jian Xu, *Mass transference principle for limsup sets generated by rectangles*, Math. Proc. Cambridge Philos. Soc. **158** (2015), no. 3, 419–437. MR 3335419
8. _____, *A generalization of the Jarník-Besicovitch theorem by continued fractions*, Ergodic Theory Dynam. Systems **36** (2016), no. 4, 1278–1306. MR 3492979

Chapter 3

Influencing Public Health Policy with Data-informed Mathematical Models Of Infectious Diseases

Model structures and structural identifiability: What? Why? How?

Jason M. Whyte

Abstract We may attempt to encapsulate what we know about a physical system by a model structure, S. This collection of related models is defined by parametric relationships between system features; say observables (outputs), unobservable variables (states), and applied inputs. Each parameter vector in some parameter space is associated with a completely specified model in S. Before choosing a model in S to predict system behaviour, we must estimate its parameters from system observations. Inconveniently, multiple models (associated with distinct parameter estimates) may approximate data equally well. Yet, if these equally valid alternatives produce dissimilar predictions of unobserved quantities, then we cannot confidently make predictions. Thus, our study may not yield any useful result.

We may anticipate the non-uniqueness of parameter estimates ahead of data collection by testing S for structural global identifiability (SGI). Here we will provide an overview of the importance of SGI, some essential theory and distinctions, and demonstrate these in testing some examples.

1 Introduction

A "model structure" (or simply "structure") is essentially a collection of related models of some particular class (say the linear, first-order, homogeneous, constant-coefficient ODEs in n variables), as summarised by mathematical relationships between system variables that depend on parameters. For example, in a "controlled state-space structure" we may draw on our knowledge of the system to relate time-varying quantities such as "states" ($\mathbf{x}$) that we may not be able to observe, and (typically known) controls or "inputs" ($\mathbf{u}$) which act on some part of our system,

Jason M. Whyte
ACEMS, School of Mathematics and Statistics; and CEBRA, School of BioSciences
University of Melbourne, Parkville, Victoria 3010, Australia
e-mail: jason.whyte@unimelb.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_10

to "outputs" ($\mathbf{y}$) we can observe. A structure is a useful construct when seeking to model some physical system for which our knowledge is incomplete. We choose some suitable parameter space, and each parameter vector therein is associated with a model in our structure, where we use "model" to mean a completely specified set of mathematical relationships between system variables.

In order to illustrate the concept of a structure, we will consider S_1, a controlled state-space structure of "compartmental" models, meaning that these are subject to a "conservation of mass" condition—matter is neither created nor destroyed. When we are interested in a system evolving in continuous time, a structure will employ ordinary differential equations (ODEs) to describe the time course of the states. Compartmental structures are often appropriate for the modelling of biological systems. To illustrate this, let us consider a simple biochemical system, where we consider the interconversion and consumption of chemical species, as in a cellular process. Structure S_1 has three state variables, x_1, x_2, and x_3, representing concentrations of three distinct chemical species, or "compartments". Matter may be excreted from the system, delivered into the system, or converted between the forms. We assume that the system receives some infusion of x_3 via input u.

Using standard notation for compartmental systems, a real parameter k_{ij} ($i, j = 1,2,3$, $i \neq j$) represents the rate constant for the conversion of x_j into x_i. A real parameter k_{0j} is the rate constant associated with the loss of material from x_j to the "environment" outside of the system. If reactions are governed by "first-order mass-action kinetics", the rate of conversion (or excretion) of some species at time t depends linearly on the amount of that species at time t.

Given our physical system and modelling paradigm, (and understanding that an expression such as $\dot{x}$ represents $\mathrm{d}x/\mathrm{d}t$) we may write the "representative model" of S_1 as

$$\begin{aligned} \dot{x}_1(t) &= -(k_{01}+k_{21})x_1(t) - k_{12}x_2(t)\,, \\ \dot{x}_2(t) &= k_{21}x_1(t) - (k_{12}+k_{32})x_2(t) + k_{23}x_3(t)\,, \\ \dot{x}_3(t) &= k_{32}x_2(t) - k_{23}x_3(t) + u(t)\,, \end{aligned} \tag{1}$$

where we set initial conditions for our state variables (where $\top$ denotes transpose)

$$\big(x_1(0)\ x_2(0)\ x_3(0)\big)^\top = \big(0\ x_{2_0}\ 0\big)^\top\,. \tag{2}$$

Supposing that x_1 is the only variable we can observe over time, our output is

$$y(t) = x_1(t)\,. \tag{3}$$

We may represent this "single-input single-output" (SISO) structure by a compartmental diagram, as in Fig. 1. Squares represent distinct chemical species, thin arrows show the conversion of mass to other forms, or excretion from the system. The rates of conversion or excretion are determined by the product of the associated parameter and the state variable at the source of the arrow. The thick arrow shows an input, and the circle linked to x_1 indicates that this compartment is observed. More specifically, Fig. 1 and (1)–(3) illustrate the representative model of a controlled (due to the input u) compartmental (mass is conserved) linear (describing the man-

ner in which the states and input appear) time-invariant (coefficients of the input and state variables are constants) state-space structure.[1]

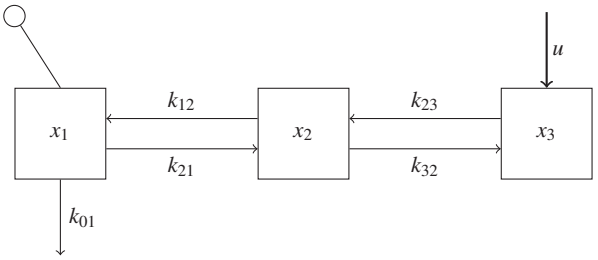


Fig. 1 A compartmental diagram of the chemical system as modelled by the representative model shown in (1)–(3). Matter in the compartments representing the chemical species x_1, x_2, and x_3, is transferred between compartments. Matter is lost from the x_1 compartment to the environment and this compartment is observed. Input u delivers mass to the x_3 compartment.

At this juncture, establishing some conventions will aid our further discussion of structures.

Convention 1. When discussing features of a structure M, we represent its associated parameter space with Θ, which we may specify more particularly as necessary. Given arbitrary parameter vector $\boldsymbol{\theta} \in \Theta$, we shall always use $M(\boldsymbol{\theta})$ to represent M's representative system. When considering some specific parameter vector, say $\boldsymbol{\alpha}$, we shall represent the associated model by $M(\boldsymbol{\alpha})$, which we will understand to be completely specified.

Convention 2. When we apply some descriptors (e.g. controlled compartmental linear time-invariant state-space) to either a structure's representative system (as in the example above) or to a structure, these descriptors transfer to the other. The descriptors also apply to all systems in the structure, except for possibly degenerate systems associated with a subset of parameter space of measure zero.

Convention 2 foreshadows a case where some small number of models in a structure may have properties different from those of other models. We may account for this complication in a manner that assists our intended analysis of structures.

Property 1. Given structure M with parameter set Θ, a property of M is generic if it holds "almost everywhere" in Θ. That is, we allow that the property may not hold on some subset(s) of Θ of measure zero.

Having specified a structure for a physical system, we may expect it to contain some model which will encapsulate the system's features of interest, and provide insights into aspects of the system's behaviour. For example, we may hope to achieve objectives, such as to accurately:

[1] We will treat classes of structures more formally in Sect. 2.

O1 predict system outputs at unobserved times within the time range for which we have data,
O2 estimate the time course of states,
O3 anticipate system behaviour in situations for which we do not have data, such as under a proposed change in experimental or environmental conditions,
O4 compare the effects of a range of proposed actions on the system, allowing us to discern which actions have the potential to produce beneficial results.

We can only hope to consistently gain such insights if our modelling effort provides reliable predictions. Yet, features of an assumed structure may make this challenging, or impossible. As such, we can benefit from interrogating structures in advance of their use to ascertain their suitability.

To explain further, we may expect to arrive at a particular model in M that we can use for prediction after using data to estimate our parameter vector in a process of "parameter identification" (PI). In essence, PI uses some objective function to quantify the goodness-of-fit of predictions made for some $\boldsymbol{\alpha} \in \Theta$ to data, and an algorithm that searches through Θ to improve upon this as much as possible. The goal is to determine those parameter vectors which optimise the objective function. Suppose that there is a "true" (unknown) parameter vector $\boldsymbol{\theta}^* \in \Theta$ such that $M(\boldsymbol{\theta}^*)$ reproduces the actual dynamics of our physical system, including that relating to any unobservable states. As data is typically sparse and subject to noise, whilst we expect that we cannot exactly recover $\boldsymbol{\theta}^*$, we intend that PI can obtain a good approximation to it.

This ambition is frustrated when the value of the objective function is virtually constant over some region of parameter space. Upon encountering such a region, a search algorithm is unable to find a search direction that will improve the objective function's value. This may lead to an unsatisfactory result. For example, the PI process may terminate without returning any parameter estimate.

Alternatively, PI's results may defy interpretation. Suppose PI returns multiple feasible, equally valid estimates of $\boldsymbol{\theta}^*$. If we lack further constraints on the elements of $\boldsymbol{\theta}^*$ (e.g. relative sizes), we cannot discern which of the alternative estimates to use as our approximation.

This state of affairs may not matter if our only concern is O1, or we do not need to specifically know $\boldsymbol{\theta}^*$. However, suppose that using M with alternative parameter estimates yields substantially different results for outcomes O2–O4. Then, we cannot confidently use M for prediction.

Cox and Huber [9] provided one example of such an unsatisfactory outcome. The authors showed that two parameter vectors returned by PI lead to equally good predictions of the observed time series of counts of malignant cancer cells in a patient, yet produce substantially different counts for the time after an "intervention"—a reduction in the carcinogenic components to which the patient is exposed.

PI may fail to uniquely estimate a parameter vector due an inherent property of M. As such, our non-uniqueness problem is independent of the amount and quality of data we have. That is, improvements in the volume of data or accuracy of its measurement cannot resolve the problem.

We expect to anticipate the non-uniqueness of parameter estimates when scrutiny of our structure shows that it is not structurally globally identifiable (SGI).[2] The concept was first formalised for state-space structures in Bellman and Åström [4] with reference to compartmental structures similar to that shown in Fig. 1.

One tests a structure to determine whether or not it is SGI in an idealised framework.

Convention 3. The framework employed in testing a structure M for SGI is defined by assumptions including:

- the structure correctly represents our physical system,
- a record of error-free data that is infinite in extent is available,
- and others that may be particular to the assumed class of structure, or testing method.

Some methods, e.g. those employing similarity transforms [21] or Markov and initial parameters, [14], are only applicable when M is "generically minimal". That is, for almost all $\boldsymbol{\theta} \in \Theta$ we cannot reduce $M(\boldsymbol{\theta})$ to a system of fewer states that produces an identical output.

The test aims to discern whether or not it is possible for PI applied to idealised data to only return the true vector $\boldsymbol{\theta}^*$, for almost all $\boldsymbol{\theta}^* \in \Theta$. The test result is definitive in this case.

Suppose that structure M is classified as SGI. Then, it may be possible for PI applied to actual (limited in extent, noisy) data to return a unique estimate for $\boldsymbol{\theta}^*$, but this is not guaranteed. As such, we can only consider an SGI model as possibly useful for prediction. Still, the value of knowing that M is SGI is the assurance that we are not almost certain to fail in our objective before we commence our study. Alternatively, it is extremely unlikely that PI applied to a non-SGI model and actual data will return a unique estimate of $\boldsymbol{\theta}^*$. In this case, we should not immediately proceed to make predictions following PI. Instead, we may seek to propagate parameter uncertainty through our structure so as to produce a range of predictions, allowing us to quantify prediction uncertainty. From this we may judge whether or not we can obtain sufficiently useful predictions for our purposes.

Aside from merely encouraging caution, the result of testing structure M for structural global identifiability[3] can deliver useful insights. The test result allows us to distinguish between individual parameters we may estimate uniquely, and those we cannot.

---

[2] The literature has various alternative terms for SGI, some of which may be equivalent only under particular conditions. For two examples, Audoly et al. [3], used "structurally *a priori* identifiable", where *a priori* emphasises that one can test a structure in advance of data collection. Godfrey [12] favoured "deterministic identifiability" in discussing compartmental models, for reasons relating to the degree of *a priori* knowledge of a system and the dependence of the result of testing on the combination of inputs. We will consider this second matter in Sect. 4.

[3] In the interests of brevity, henceforth we use SGI as a shorthand for this noun, in addition to the adjective used earlier, expecting that the reader can infer the meaning from context.

Further, awareness that a structure is not SGI can assist in correcting the problem. The test may allow us to recognise those parameter combinations which PI may return uniquely. This knowledge may guide reparameterisation of $M(\theta)$ so as to produce the representative system of a new structure that is SGI. Additionally, having learned that M is not SGI, one can examine whether it is possible that modifying M (e.g. holding some parameters constant), or the combination of M and planned data collection (e.g. supposing that an additional variable is measured, and rewriting M to include this as another output), will remedy this. Thus, we can treat the process of testing a structure for SGI as an iterative process. We can detect a structure's undesirable features ahead of data collection, address them, test the revised structure, and continue this process until the structure is satisfactory.

Analytical inspection of (in particular, more complex) structures to anticipate the uniqueness or otherwise of parameter estimates is often not straightforward. The difficulties of testing a structure for SGI, as well as how the results of PI applied to real data can be worse than that predicted by theory, have encouraged numerical approaches to the task. (See [12, Chapter 8] for an introduction.) Broadly, approaches seeking to demonstrate "numerical" (or "practical") identifiability are based on assuming some number of parameter vectors; using each of these with the structure to simulate data at a limited number of observation times, or under a limited number of conditions (e.g. applied inputs or values of experimental variables), or subject to noise, or some combination of these; conducting PI; and investigating the features of parameter estimates to determine if these adequately approximate assumed values.

Testing a structure for numerical identifiability may determine when PI is unlikely to yield accurate results. However, unlike analytical scrutiny, these investigations may not provide clear guidance on how to remedy the problem.

In this paper we will provide an introduction to the testing of (state-space) structures for SGI. There are a variety of testing methods available (see, for example, [11]) although many are not an ideal means of introducing the field of identifiability analysis. As such, we intend that our choices of testing method and examples will allow us to illustrate some important issues without having to encounter unnecessary algebraic and conceptual complexity.

In choosing example structures, we have limited ourselves to a class which are linear in the state variables, as demonstrated in the representative model given in (1)–(3) . We further restrict these to compartmental structures. Given these choices, the "Transfer Function Approach" (TFA, see for example [8]), which makes use of features of the Laplace transform of a structure's output function,[4] is appropriate for our purposes. Although one of the older testing methods, it is still included in relatively recent texts presenting a range of methods (e.g. [11]), and:

1. is conceptually rather more straightforward than other methods,
2. has the unusual distinction of being applicable to a structure that is not generically minimal, and
3. is unambiguously appropriate for compartmental structures.

[4] For this reason, the approach is also known as the "Laplace transform method", as seen in [12, Chapter 6].

To explain the significance of Points 2 and 3, we note that a general linear state-space structure may be judged as generically minimal as a consequence of having the generic properties of controllability and observability. The conditions used in deciding this are appropriate for linear systems—these have a state space which is a vector space. However, the state space of a positive linear system is a polyhedral cone, and so it does not seem appropriate to treat these as we would a general linear system.

Certain authors have sought to highlight differences between features of linear systems and linear positive systems. In the context of discrete-time systems, Benvenuti and Farina sought to show

> ... that the minimality problem for positive linear systems is inherently different from that of ordinary linear systems ..." ([5, Page 219]).

Whyte [26, Chapter 3, Section 5.2] considered some of the literature's perspectives on controllability of linear state-space systems. Briefly, the origins of the area related to linear "structured" systems (see Poljak [17]) which are generally distinct from linear compartmental systems (a type of "descriptor" system; see Yamada and Luenberger [27]). This lead to suspicions that it may not always be inappropriate to test a linear compartmental structure for generic minimality using the machinery designed for general linear structures. By choosing to use the TFA in analysing a structure, Point 2 allows us to avoid this potential issue.

Further, the TFA has shown promise in the analysis of structures of linear switching systems (LSSs) (Whyte [25, 26]). Structures of switching systems (especially those which evolve in continuous time) are largely neglected in the literature. Yet methods under development may assist in the scrutiny of structures used to model epidemics, such as where an intervention causes an abrupt change in some parameter values.

Discussions at a recent workshop "Identifiability problems in systems biology" held at the American Institute of Mathematics ([1]) highlighted a degree of inconsistency in certain key definitions used in the field of identifiability analysis. As such, here we will draw on efforts to propose transparent and coherent definitions in the analysis of uncontrolled structures (Whyte [25, 26]) in suggesting equivalent definitions for controlled structures.

The remainder of this paper is organised as follows. In Sect. 2 we present some preliminary material and introduce certain classes of structures that aid us in presenting the TFA. In Sect. 3 we outline the general theory of testing an uncontrolled structure for SGI, particularise this to uncontrolled linear time-invariant (LTI) state-space structures, and consider an example. Section 4 proceeds similarly for controlled LTI state-space structures, where we draw an important distinction between testing approaches based on how much information we are able to elicit from our structure. Finally, in Sect. 5 we summarise some concepts in the testing of structures and offer some concluding remarks.

We conclude this section by establishing notation.

1.1 Notation

The field of real numbers is denoted by $\mathbb{R}$. The subset of $\mathbb{R}$ containing only positive (non-negative) values is denoted by $\mathbb{R}_+$ ($\bar{\mathbb{R}}_+$). The natural numbers $\{1,2,3,\ldots\}$ are denoted by $\mathbb{N}$, and we define $\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$.

The field of complex numbers is denoted by $\mathbb{C}$. The real part of $z \in \mathbb{C}$ is denoted by $\mathrm{Re}(z)$. Given some $a \in \mathbb{R}$, a useful set for the following discussion is

$$\mathrm{H}_a \triangleq \{s \in \mathbb{C} \big| \mathrm{Re}(s) > a\}. \tag{4}$$

We use a bold lower-case (upper-case) symbol such as $\mathbf{a}$ ($\mathbf{A}$) to denote a vector (matrix), and a superscript $\top$ associated with any such object indicates its transpose. Given vector $\mathbf{a}$, $\dot{\mathbf{a}}$ denotes its derivative with respect to time. To specify the (i,j)-th element of $\mathbf{A}$ we may use $a_{i,j}$, or prefer the simplicity of $(\mathbf{A})_{\mathbf{i,j}}$ when $\mathbf{A}$ is a product of terms. For $n \in \mathbb{N}$, we use $\mathrm{diag}(a_1, a_2, \ldots, a_n)$ to denote the square diagonal matrix having $a_1, \ldots, a_n$ on the main diagonal and zeros elsewhere. A special diagonal matrix is the $(n \times n)$ identity matrix $\mathbf{I_n} \in \mathbb{R}^{\mathbf{n}\times\mathbf{n}}$, having a main diagonal of n 1s.

Given field $\mathbb{F}$ and some indeterminate w, $\mathbb{F}(w)$ denotes the field of rational functions in w over $\mathbb{F}$. Given $a, b \in \mathbb{N}_0$ and $\mathbb{F}$, we use $\mathbb{F}^{a\times b}$ to denote the set of matrices of a rows and b columns having elements in $\mathbb{F}$. When at least one of a or b is zero, it is convenient to have $\mathbb{F}^{a\times b}$ represent a set of "empty matrices", and we can disregard any matrix in this set as it arises.

2 Preliminaries

In this section we will define certain classes of structures, and present an overview of some useful properties, in preparation for a discussion of how we may test these structures for SGI.

We will aim to illustrate the features of systems by introducing sufficient systems theory, beginning with some conventions. Suppose we have a set of input values U, a set of output values Y, and a time set $T \subseteq \bar{\mathbb{R}}_+$. Let $\mathscr{U}$ denote a set of input functions such that for $u \in \mathscr{U}, u : T \to U^T : t \mapsto u(t) \in U$. That is, $\mathscr{U}$ is a set of input functions taking values in the set U. Similarly, let $\mathscr{Y}$ denote a set of functions such that for $y \in \mathscr{Y}, y : T \to Y^T : t \mapsto y(t) \in Y$. That is, $\mathscr{Y}$ is a set of output functions taking values in a set Y. Finally, let ζ denote an "input-output" map from $\mathscr{U}$ to $\mathscr{Y}$. We use these definitions in presenting a general type of system in Definition 1. From this we may obtain other system types by imposing suitable conditions.

Definition 1. An **input-output system** on time set T is a triple $(\mathscr{U}, \mathscr{Y}, \zeta)$.

Contained within the input-output systems are the state-space systems, which are of particular interest to us here. To aid our discussion of these, given some time set T we define the set

$$T_+^2 \triangleq \left\{(t_2, t_1);\ t_2 \geq t_1,\ t_1, t_2 \in T\right\}. \tag{5}$$

2.1 State-space structures

In the following definitions and discussion we draw on Whyte [26, Section 3.4], which was informed by Caines [6, Appendix 2]).

Definition 2 (Adapted from Whyte [26, Definition 3.8]). A **state-space system** Σ is a quintuple $(\mathscr{U}, X, \mathscr{Y}, \Phi, \eta)$ where

- $\mathscr{U}$ is a set of input functions.
- X is a set, called the state-space of Σ, with elements called states.
- $\mathscr{Y}$ is a set of output functions.
- $\Phi(\cdot,\cdot,\cdot,\cdot)$ is the state transition function, which maps $T_+^2 \times X \times \mathscr{U}$ into X.
  To illustrate this, consider time interval $T \subseteq \bar{\mathbb{R}}_+$ with $t_0 \triangleq \inf T$. Suppose Σ is subject to input function $u \in \mathscr{U}$. Further, suppose that at $t = t_0$ we have that $x_0 \in X$ is the initial state of Σ. Then, for $(t,t_0) \in T_+^2$, $\Phi(t,t_0,x_0,u)$ determines the state of Σ as a consequence of time t, x_0, and u. Under these conditions, we may concisely refer to $\Phi(t,t_0,x_0,u)$ as the state of Σ at time t.
- $\eta(\cdot,\cdot,\cdot)$ is the output map, which maps $T \times X \times \mathscr{U}$ into Y.
  That is, at some time $t \in T$, η determines the output vector that results from three inputs: t, the state of Σ at that time, and the input u.

Further, the following four properties hold:

SS1: The Identity Property of Φ

$$\Phi(t,t,x,u) = x, \text{ for all } t \in T,\ x \in X \text{ and } u \in \mathscr{U}\ .$$

That is, suppose the state of Σ at time t is x. Then, if no time has elapsed from t, Φ does not move the state away from x.

SS2: The Nonanticipative Property of Φ
Suppose we have any $u_1, u_2 \in \mathscr{U}$ such that these functions are identical on time interval $[t_0,t_1]$, where $(t_1,t_0) \in T_+^2 \subset \mathbb{R}_+^2$. Then, for all $x \in X$ we have

$$\Phi(t_1,t_0,x,u_1) = \Phi(t_1,t_0,x,u_2)\ .$$

To explain this, suppose the state of Σ at time t_0 is some $x \in X$. The Nonanticipative Property of Φ means that Σ reaches the same state at time t_1 for Φ subject to either u_1 or u_2. Equivalently, differences between u_1 and u_2 for any time greater than t_1 do not influence the evolution of the state of Σ on $[t_0,t_1]$ under Φ.

SS3: The Semigroup Property of Φ
For all $(t_1,t_0),(t_2,t_1) \in T_+^2$, $x \in X$, and $u \in \mathscr{U}$,

$$\Phi(t_2,t_0,x,u) = \Phi\big(t_2,t_1,\Phi(t_1,t_0,x,u),u\big)\ .$$

To explain, suppose we have system Σ with initial state x at time t_0 and input u. Suppose Φ acts on time interval $[t_0,t_1]$ resulting in some particular state (say $x_1 \triangleq \Phi(t_1,t_0,x,u)$) at t_1. Suppose then Φ uses x_1 as an initial state for evolving the state of Σ on $[t_1,t_2]$, resulting in a particular state (say $x_2 \triangleq \Phi\big(t_2,t_1,\Phi(t_1,t_0,x,u),u\big)$) at t_2. Due to the Semigroup Property of Φ, system Σ also reaches state x_2 at t_2 if Φ is used to evolve the state on $[t_0,t_2]$.

SS4: The Instantaneous Output Map η

For all $x \in X$, $u \in \mathscr{U}$, $(t,t_0) \in T_+^2$, the function $y : T \to Y$ defined via

$$y(t) = \eta\big(t, \Phi(t,t_0,x,u), u(t)\big)$$

is a segment of a function in $\mathscr{Y}$.

That is, we can use η to define the instantaneous output of Σ at current time t through t, the state of Σ at time t ($\Phi(t,t_0,x,u)$) and the value of the input at time t ($u(t)$). This property is useful as y provides a simpler means of illustrating the output of Σ than does η when we wish to introduce particular system types.

We will now illustrate some useful classes of continuous-time state-space structures, beginning with a general type. Henceforth we consider spaces for states, inputs, and outputs of $X \subseteq \mathbb{R}^n$, $U \subseteq \mathbb{R}^m$, and $Y \subseteq \mathbb{R}^k$, respectively, where accordingly indices $n,m,k \in \mathbb{N}$ determine the dimensions of our state, input, and output vectors. For arbitrary parameter vector $\boldsymbol{\theta} \in \Theta$, and input $\mathbf{u} \in \mathscr{U}$, at time $t \in T$ a controlled state-space structure M has representative system $M(\boldsymbol{\theta})$ of the general form:

$$\begin{aligned} \dot{\mathbf{x}}(t;\boldsymbol{\theta}) = \mathbf{f}(\mathbf{x},\mathbf{u},t;\boldsymbol{\theta}), \quad \mathbf{x}(0;\boldsymbol{\theta}) = \mathbf{x}_0(\boldsymbol{\theta}) \,, \\ \mathbf{y}(t;\boldsymbol{\theta}) = \mathbf{g}(\mathbf{x},\mathbf{u},t;\boldsymbol{\theta}) \,, \end{aligned} \tag{6}$$

where $\mathbf{f}$ and $\mathbf{g}$ satisfy the relevant properties SS1–SS4 of Definition 2.

A subtype of the controlled state-space structures are an uncontrolled class, lacking inputs. If an uncontrolled state-space structure has indices for the state and output spaces of n and k respectively, then a representative model is similar to (6):

$$\begin{aligned} \dot{\mathbf{x}}(t;\boldsymbol{\theta}) = \mathbf{f}(\mathbf{x},t;\boldsymbol{\theta}), \quad \mathbf{x}(0;\boldsymbol{\theta}) = \mathbf{x}_0(\boldsymbol{\theta}) \,, \\ \dot{\mathbf{y}}(t;\boldsymbol{\theta}) = \mathbf{g}(\mathbf{x},t;\boldsymbol{\theta}) \,. \end{aligned} \tag{7}$$

We will now introduce a particular class of the general state-space structures described above—that of linear time-invariant (LTI) structures. An LTI structure has a representative system that is particular form of (6). We will use specific examples of LTI structures to illustrate the testing of a structure for SGI in Sects. 3 and 4.

2.2 Continuous-time linear, time-invariant structures

The following definitions are adapted from Whyte [26, Definition 3.21], which drew on concepts from van den Hof [14].

Definition 3. Given indices $n,m,k \in \mathbb{N}$, a **controlled continuous-time linear time-invariant state-space structure** (or, more briefly, **an LTI structure**) M has state, input, and output spaces $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, and $Y = \mathbb{R}^k$, respectively. For parameter set $\Theta \subseteq \mathbb{R}^p$ ($p \in \mathbb{N}$), M has mappings

$$\mathbf{A} : \Theta \to \mathbb{R}^{n\times n}\,, \quad \mathbf{B} : \Theta \to \mathbb{R}^{n\times m}\,, \quad \mathbf{C} : \Theta \to \mathbb{R}^{k\times n}\,, \quad \mathbf{x_0} : \Theta \to \mathbb{R}^n\,, \tag{8}$$

where the particular pattern of non-zero elements in the "system matrices" shown in (8) defines M. More specifically, mappings in (8) dictate the relationships between state variables $\mathbf{x}$, inputs $\mathbf{u}$, and outputs $\mathbf{y}$ for all times $t \in T \subseteq \mathbb{R}_+$. Thus, for arbitrary $\boldsymbol{\theta} \in \Theta$, M's representative system $M(\boldsymbol{\theta})$ has the form

$$\dot{\mathbf{x}}(t,\mathbf{u};\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\,\mathbf{x}(t,\mathbf{u};\boldsymbol{\theta}) + \mathbf{B}(\boldsymbol{\theta})\,\mathbf{u}(t)\,, \quad \mathbf{x}(0;\boldsymbol{\theta}) = \mathbf{x}_0(\boldsymbol{\theta})\,, \tag{9}$$

$$y(t;\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})\,\mathbf{x}(t;\boldsymbol{\theta})\,. \tag{10}$$

Defining

$$L\Sigma P(n,m,k) \triangleq \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{k\times n} \times \mathbb{R}^n\,, \tag{11}$$

then

$$SL\Sigma P(n,m,k) \triangleq \left\{ \left(\mathbf{A}(\Theta),\mathbf{B}(\boldsymbol{\theta}),\mathbf{C}(\boldsymbol{\theta}),\mathbf{x_0}(\boldsymbol{\theta})\right) \in L\Sigma P(n,m,k) \,\middle|\, \boldsymbol{\theta} \in \Theta \right\} \tag{12}$$

is the set of system matrices associated with systems in M. Thus, we may consider the matrices of a particular system in M as obtained by the parameterisation map $f : \Theta \to SL\Sigma P(n,m,k)$ such that

$$f(\boldsymbol{\theta}) = \Big(\mathbf{A}(\boldsymbol{\theta}),\mathbf{B}(\boldsymbol{\theta}),\mathbf{C}(\boldsymbol{\theta}),\mathbf{x_0}(\boldsymbol{\theta})\Big).$$

Together, the matrices and vector defined by (8), and the indices n, m, and k, are the **system parameters** of $M(\boldsymbol{\theta})$.

We may consider an **uncontrolled LTI structure** having indices $n,k \in \mathbb{N}$ as a form of controlled LTI structure having $n,m,k \in \mathbb{N}_0$ by setting $m = 0$. As such, systems in the uncontrolled structure have $X = \mathbb{R}^n$ and $Y = \mathbb{R}^k$. By omitting the empty matrix $\mathbf{B}$ from (9) we obtain the form of the uncontrolled structure's representative system:

$$\dot{\mathbf{x}}(t;\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\,\mathbf{x}(t;\boldsymbol{\theta}), \quad \mathbf{x}(0;\boldsymbol{\theta}) = \mathbf{x}_0(\boldsymbol{\theta})\,, \tag{13}$$

$$y(t;\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})\,\mathbf{x}(t;\boldsymbol{\theta})\,, \tag{14}$$

where the system matrices are $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{C} \in \mathbb{R}^{k\times n}$, and $\mathbf{x_0} \in \mathbb{R}^{\mathbf{n}}$.

As a notational convenience, we allow sets defined in (11) and (12) to apply to this context, where $L\Sigma P(n,0,k)$ and $SL\Sigma P(n,0,k)$ are understood as neglecting the irrelevant $\mathbf{B}$.

In modelling biological systems, we may employ a subclass of the LTI state-space structures in which systems have states, inputs, and outputs subject to constraints informed by physical considerations. This, in turn, imposes conditions on

the structure's system matrices. Our summary of the conditions in the following definition is informed by the treatment of compartmental LTI systems given in van den Hof [14].

Definition 4 (Classes of LTI state-space structures). A **positive LTI state-space structure** with indices $n,m,k \in \mathbb{N}$ is an LTI state-space structure after Definition 3, having representative system of the form given in (9) and (10), where states, outputs, and inputs are restricted to non-negative values. That is, the structure has $X = \bar{\mathbb{R}}^n_+$, $U = \bar{\mathbb{R}}^m_+$, and $Y = \bar{\mathbb{R}}^k_+$.

A **compartmental LTI structure** with indices $n,m,k \in \mathbb{N}$ is a positive LTI state-space structure for which systems in the structure have system matrices subject to "conservation of mass" conditions:

- all elements of $\mathbf{B}$ and $\mathbf{C}$ are non-negative, and
- for $\mathbf{A} = (a_{i,j})_{i,j=1,\ldots,n}$,

$$\begin{aligned} a_{ij} &\geq 0\,, \qquad\qquad i,j \in \{1,\ldots,n\}\,,\ i \neq j\,, \\ a_{ii} &\leq -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ji}\,, \quad i \in \{1,\ldots,n\}\,. \end{aligned} \tag{15}$$

An **uncontrolled positive LTI structure** or an **uncontrolled compartmental LTI structure** with indices n,k belongs to a subclass of the corresponding class of controlled LTI structures with indices n,k,m. The relationship between the controlled and uncontrolled forms is as for that between LTI structures and uncontrolled LTI structures presented in Definition 3. The representative system of any such uncontrolled structure has the form outlined in (13) and (14), subject to appropriate restrictions on state and output spaces, X and Y, respectively.

We shall now consider some properties of controlled LTI structures which will inform our testing of these structures for SGI subsequently.

2.3 Features of the states and outputs of a controlled LTI structure

A consideration of some features of the states and outputs of LTI structures here will allow us to appreciate the utility of the TFA in testing such a structure for SGI in Sect. 3.

2.3.1 The time course of states and outputs

In this discussion we adapt the treatment of uncontrolled LTI systems given in Whyte [26, Chapter 3] and combine this with insights from Seber and Wild [18, Chapter 8]. In this subsection, in the interests of brevity, we we will neglect the dependence of systems on $\boldsymbol{\theta}$.

Let us consider a structure defined by system matrices in $SL\Sigma P(n,m,k)$ (recall (11)), where we assume the structure is defined on time set $T = \bar{\mathbb{R}}_+$. Recall that states evolve according to an ODE system as in (9). Given state space $X = \mathbb{R}^n$, the solution for state vector $\mathbf{x}(\mathbf{t})$ depends on the matrix exponential $e^{\mathbf{At}} \in \mathbb{R}^{n\times n}$ through

$$\mathbf{x}(\mathbf{t}) = \mathrm{e}^{\mathbf{At}}\mathbf{x_0} + \int_{\mathbf{0}}^{\mathbf{t}} \mathrm{e}^{\mathbf{A(t-t')}}\mathbf{B}\boldsymbol{u}(\mathbf{t'})\mathrm{d}\mathbf{t'} \,, \tag{16}$$

provided that the integral exists. Assuming this existence, we may use (14) and the convolution operator $*$ to express response as

$$\boldsymbol{y}(t) = \mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{x_0} + \mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{B} * \mathbf{u}(\mathbf{t}) \,. \tag{17}$$

Let us presume a situation typical in the modelling of physical systems—that the elements of $\mathbf{A}$ are finite. Let us suppose that the n (finite and not necessarily distinct) eigenvalues of $\mathbf{A}$ are ordered from largest to smallest and labelled as λ_i, $i = 1,\ldots,n$. In the interests of simplicity, we also assume that $\mathbf{A}$ has n linearly independent right eigenvectors $\mathbf{s_i}$, $i = 1,\ldots,n$, where each is associated with the appropriate λ_i. We define $\mathbf{S} \in \mathbb{R}^{n\times n}$ as the matrix for which the i-th column is $\mathbf{s_i}$. We may then employ a spectral decomposition $\mathbf{A} \equiv \mathbf{S\Lambda S^{-1}}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. As a result, we may rewrite our matrix exponential:

$$\mathrm{e}^{\mathbf{At}} \equiv \mathbf{S}\mathrm{e}^{\mathbf{\Lambda t}}\mathbf{S^{-1}} \,, \tag{18}$$

noting that each element is a sum of (up to n) exponentials, with exponents drawn from λ_i ($i = 1,\ldots,n$).

With this in mind, let us turn our attention towards the terms $\mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{x_0} \in \mathbb{R}^{\mathbf{k}\times\mathbf{1}}$ and $\mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{B} \in \mathbb{R}^{\mathbf{k}\times\mathbf{m}}$ on the the right-hand side of (17). As $\mathbf{x_0}$ is a constant vector, and $\mathbf{B}$ and $\mathbf{C}$ are constant matrices, then each element of $\mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{x_0}$ and $\mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{B}$ is also a sum of exponentials in λ_i ($i = 1,\ldots,n$).

Suppose λ_1 has multiplicity $\mu \geq 1$. Hence, the largest possible dominant term in any of our sums of exponentials involves $t^\mu e^{\lambda_1 t}$. Hence, there exist real constants $K > 0$ and $\lambda > \lambda_1$ such that for all $t \in \bar{\mathbb{R}}_+$ we have

$$K\mathrm{e}^{\lambda t} \geq \begin{cases} \left|\left(\boldsymbol{C}\mathrm{e}^{\mathbf{At}}\mathbf{x_0}\right)_{i,1}\right| & i = 1,\ldots,k \,, \\ \left|\left(\mathbf{C}\mathrm{e}^{\mathbf{At}}\mathbf{B}\right)_{i,j}\right| & \begin{matrix} i = 1,\ldots,k \,, \\ j = 1,\ldots,m \,. \end{matrix} \end{cases} \tag{19}$$

The existence of these bounds will prove important when we consider the application of the TFA to an LTI structure. Towards this, we shall consider some features of the Laplace transform of the output of LTI structures.

2.3.2 The Laplace transform of an LTI structure output function

We recall the definition of the Laplace transform of a real-valued function.

Definition 5. Suppose some real-valued function f is defined for all non-negative time. (That is, $f : \bar{\mathbb{R}}_+ \mapsto \mathbb{R},\ t \mapsto f(t)$.) We represent the (unilateral) Laplace transform of f with respect to the transform variable $s \in \mathbb{C}$ by

$$\mathscr{L}\{f\}(s) \triangleq \int_0^\infty f(t) \cdot e^{-st} \mathrm{d}t \ ,$$

if this exists on some domain of convergence $\mathscr{D} \subset \mathbb{C}$.

Let us consider a controlled LTI structure S with parameter set Θ, with a representative system $S(\boldsymbol{\theta})$, having the form shown in (9) and (10). We assume system matrices belong to $SL\Sigma P(n,m,k)$ (recall (12)). Suppose that given input $\mathbf{u}$, $\mathscr{L}\{\mathbf{u}\}(\mathbf{s})$ exists. In this case the Laplace transform of output $\mathbf{y}$ given $\mathbf{u}$ is[5]

$$\mathscr{L}\{\mathbf{y}(\cdot,\mathbf{u};\boldsymbol{\theta})\}(s;\boldsymbol{\theta}) = \mathbf{V}(s;\boldsymbol{\theta}) + \mathbf{W}(s;\boldsymbol{\theta})\mathscr{L}\{\mathbf{u}\}(s) \in \mathbb{R}(s)^{k\times 1} \ , \tag{20}$$

where

$$\mathbf{V}(s;\boldsymbol{\theta}) \triangleq \mathbf{C}(\boldsymbol{\theta})\big(s\mathbf{I}_n - \mathbf{A}(\boldsymbol{\theta})\big)^{-1}\mathbf{x_0}(\boldsymbol{\theta}) \in \mathbb{R}(s)^{k\times 1} \ , \tag{21}$$

$$\mathbf{W}(s;\boldsymbol{\theta}) \triangleq \mathbf{C}(\boldsymbol{\theta})\big(s\mathbf{I}_n - \mathbf{A}(\boldsymbol{\theta})\big)^{-1}\mathbf{B}(\boldsymbol{\theta}) \in \mathbb{R}(s)^{k\times m} \ , \tag{22}$$

and, owing to (19), each element of $\mathbf{V}$ and $\mathbf{W}$ is defined for all $s \in H_\lambda$.

Definition 6. We refer to $\mathbf{V}$ and $\mathbf{W}$ as "transfer matrices", and each element of these is a transfer function—specifically, a rational function in s. We term any such element an *unprocessed transfer function*.

Property 2. The degree of the denominator of any unprocessed transfer function in $\mathbf{V}$ or $\mathbf{W}$ is at most n. Similarly, if S is a compartmental structure, the degree of the numerator of any transfer function is at most $n-1$. If we can cancel any factors in s between the numerator and denominator of the transfer function (pole-zero cancellation), then we will obtain a degree for each of the numerator and denominator which is lower than previously.

Suppose that pole-zero cancellation occurs in each unprocessed transfer function in $\mathbf{V}$ and $\mathbf{W}$. Then, S is not generically minimal (recall Convention 3).

When we have an uncontrolled LTI structure, (20) reduces to

$$\mathscr{L}\{\mathbf{y}(\cdot;\boldsymbol{\theta})\}(s) = \mathbf{V}(s;\boldsymbol{\theta}) \in \mathbb{R}(s)^{k\times 1} \ , \tag{23}$$

with $\mathbf{V}$ as in (21), and the discussion of matrix elements given above also applies.

[5] We note that others, such as Walter and Pronzato [22, Chapter 2, Page 22], have considered such expressions. However, the notation employed may make the description of transfer functions in testing a structure for SGI unnecessarily complicated. As such, we employ a simpler notation here. We also include $\mathbf{x_0}$ in $\mathbf{V}$ (unlike say in the equivalent matrix $\mathbf{H_2}$ in [22]), as otherwise the initial conditions do not feature in the test equations.

We may now proceed to consider definitions and processes relating to structures and structural global identifiability, informed by Convention 3. By way of introduction, we begin with the rather more straightforward matter of the testing of uncontrolled structures.

3 Testing an uncontrolled structure for structural global identifiability

We will consider the testing of an uncontrolled structure for SGI following what we may call the "classical" approach originally outlined by Bellman and Åström [4]. We follow the treatment of [26] which drew on aspects of Denis-Vidal and Joly-Blanchard [10]. In essence, we judge a structure as SGI (or otherwise) with reference to the solution set of test equations.

Definition 7. Suppose we have a structure of uncontrolled state-space systems M, having parameter set Θ (an open subset of $\mathbb{R}^p$, $p \in \mathbb{N}$), and time set $T \subseteq [0,\infty)$. For some unspecified $\boldsymbol{\theta} \in \Theta$, M has representative model $M(\boldsymbol{\theta})$, which has state function $\mathbf{x}(\cdot;\boldsymbol{\theta}) \in \mathbb{R}^{\mathbf{n}}$ and output $\mathbf{y}(\cdot;\boldsymbol{\theta}) \in \mathbb{R}^k$ (recall (7)). Suppose that systems in M satisfy conditions:

1. The functions $\mathbf{f}(\mathbf{x},\cdot;\boldsymbol{\theta})$ and $\mathbf{g}(\mathbf{x},\cdot;\boldsymbol{\theta})$ are real and analytic for every $\boldsymbol{\theta} \in \Theta$ on $\mathscr{S}$ (a connected open subset of $\mathbb{R}^n$ such that $\boldsymbol{x}(t;\boldsymbol{\theta}) \in \mathscr{S}$ for every $t \in [0,\tau]$, $\tau > 0$).
2. $\mathbf{f}(\mathbf{x_0}(\boldsymbol{\theta});\boldsymbol{\theta}) \neq \mathbf{0}$ for almost all $\boldsymbol{\theta} \in \Theta$.

Then, for some finite time $\tau > 0$, we consider the set

$$\mathscr{I}(M) \triangleq \left\{ \boldsymbol{\theta}' \in \Theta : \mathbf{y}(t;\boldsymbol{\theta}') = \mathbf{y}(t;\boldsymbol{\theta}) \quad \forall t \in [0,\tau] \right\} . \tag{24}$$

If, for almost all $\boldsymbol{\theta} \in \Theta$:

$\mathscr{I}(M) = \{\boldsymbol{\theta}\}$, M is structurally globally identifiable (SGI);
the elements of $\mathscr{I}(M)$ are denumerable, M is structurally locally identifiable (SLI);
the elements of $\mathscr{I}(M)$ are not denumerable, M is structurally unidentifiable (SU).

We note that some care is needed in the application of Definition 7, as it is not appropriate in all cases. Condition 1 ensures that the definition is not applicable to all classes of systems, including switching systems. Condition 2 indicates that the initial state cannot be an equilibrium point, as otherwise response is constant for all time. Such a response cannot provide information on system dynamics. If the constant response is atypical, it does not provide an appropriate idealisation of real data. Thus, it is inappropriate to use a constant response in testing the structure for SGI.

Remark 1. Instead of the test described above, one may test a structure for the property of structural local identifiability ([20]). This is able to judge a structure as either

SLI (the structure may actually be SGI, but we cannot discern this), or SU. Discerning that a structure is SLI may be adequate in some circumstances, and the tests tend to be easier to apply than tests for SGI.

In general, the output of system $M(\boldsymbol{\theta})$ features "(structural) invariants" [19] (or "observational parameters" [15]) $\phi(\boldsymbol{\theta})$ which define the time course of output. We may use these to summarise the properties of the whole structure.[6]

Thus, invariants allow us to test a structure for SGI using algebraic conditions that are addressed more easily than a functional relationship as in (24). Here we formalise this property by rewriting Definition 7 in terms of invariants. This leads to a test of a structure for SGI that is easier to apply than its predecessor.

Definition 8. Suppose that structure M satisfies Conditions 1 and 2 of Definition 7. Then, for some arbitrary $\boldsymbol{\theta} \in \Theta$, we define the set

$$\mathscr{I}(M,\phi) \triangleq \left\{ \boldsymbol{\theta}' \in \Theta : \boldsymbol{\phi(\theta')} = \phi(\boldsymbol{\theta}) \right\} \equiv \mathscr{I}(M) \,. \tag{25}$$

It follows that determination of $\mathscr{I}(M,\phi)$ allows classification of M according to Definition 7.

Given Definition 8, we may propose a process for testing a structure for SGI.

Proposition 1.

Step 1 *Obtain invariants $\phi(\boldsymbol{\theta})$: there are various approaches, but some have requirements (e.g. that the structure is generically minimal) that may be difficult to check.*

Step 2 *Form alternative invariants $\phi(\boldsymbol{\theta}')$ by substituting $\boldsymbol{\theta}'$ for $\boldsymbol{\theta}$ in $\phi(\boldsymbol{\theta})$.*

Step 3 *Form equations $\phi(\boldsymbol{\theta}') = \phi(\boldsymbol{\theta})$.*

Step 4 *Solve equations.*

Step 5 *Scrutinise solution set to make a judgement on M according to Definition 8.*

Step 1 poses a key problem : how may we obtain some suitable ϕ? When considering an LTI structure, the TFA is appropriate. We will now introduce the approach, proceeding to illustrate its application to an uncontrolled LTI structure in Sect. 3.2.

3.1 The Transfer Function Approach

Consider a compartmental LTI structure S with indices $n,k \in \mathbb{N}$ and $m \in \mathbb{N}_0$, having system matrices belonging to $SL\Sigma P(n,m,k)$ (recalling that $m = 0$ indicates an uncontrolled structure). Recall the idealised framework employed in the testing of a structure for SGI shown in Convention 3. As such, we consider S defined for time set

[6] We can conceive of invariants most directly when a structure is defined by one set of mathematical relations for all time. Otherwise, say for structures of switching systems, we require a more flexible approach ([23, 24]). Such structures are beyond the introductory intentions of this chapter.

$T = \bar{\mathbb{R}}_+$. Recall (20), and the discussion of Sect. 2.3.1 which guarantees that there exists some λ such that the Laplace transform of $\boldsymbol{y}$ has a domain of convergence. Then, given transfer matrices $\mathbf{V}$ and $\mathbf{W}$ (as appropriate), we may extract invariants for use in testing S for SGI. First, we must place the transfer functions into a specific form.

Definition 9 (Canonical form of a transfer function). Given compartmental LTI structure S of $n \in \mathbb{N}$ states, suppose that associated with $S(\boldsymbol{\theta})$ is a transfer matrix (as in (20)) $\mathbf{Z}$, composed of unprocessed transfer functions. Given element $z_{i,j}(s;\boldsymbol{\theta}) \in \mathbb{C}(s)$, we obtain the associated *transfer function in canonical form* by cancelling any common factors between the numerator and denominator, and rewriting to ensure that the denominator polynomial is monic. The result is an expression of the form:

$$z_{i,j}(s;\boldsymbol{\theta}) = \frac{\omega_{i,j,r+p}(\boldsymbol{\theta})s^p + \cdots + \omega_{i,j,r}(\boldsymbol{\theta})}{s^r + \omega_{i,j,r-1}(\boldsymbol{\theta})s^{r-1} + \cdots + \omega_{i,j,0}(\boldsymbol{\theta})}, \quad \forall s \in \mathbb{C}_0 \supseteq H_\lambda \,, \\ r \in \{1,\ldots,n\} \,, \quad p \in \{0,\ldots,r-1\} \,. \tag{26}$$

The coefficients $\omega_{i,j,0},\ldots,\omega_{i,j,r+p}$ in (26) contribute invariants towards $\phi(\boldsymbol{\theta})$.

3.2 A demonstration of the testing of an uncontrolled LTI structure for SGI

Recalling the general form of systems in an uncontrolled compartmental LTI structure from (13) and (14), let us consider a particular example S_0, with representative system:

$$\dot{\mathbf{x}}_{\mathbf{0}}(\mathbf{t};\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\,\mathbf{x}_{\mathbf{0}}(\mathbf{t};\boldsymbol{\theta}) \qquad \mathbf{x}_{\mathbf{0}}(\mathbf{0};\boldsymbol{\theta}) = \mathbf{x}_{\mathbf{0}_0}(\boldsymbol{\theta}) \,, \tag{27}$$

$$y_0(t;\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})\,\mathbf{x}_{\mathbf{0}}(t;\boldsymbol{\theta}) \,, \tag{28}$$

where the state vector is $\mathbf{x}_{\mathbf{0}}(t;\boldsymbol{\theta}) = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top$, and the system matrices belong to $SL\Sigma P(3,0,1)$. These have the form:

$$\mathbf{x}_{\mathbf{0}_0}(\boldsymbol{\theta}) = \begin{bmatrix} 0 \\ x_{20} \\ 0 \end{bmatrix}, \quad \mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} -k_{01}-k_{21} & k_{12} & 0 \\ k_{21} & -k_{12}-k_{32} & k_{23} \\ 0 & k_{32} & -k_{23} \end{bmatrix}, \quad \mathbf{C}(\boldsymbol{\theta}) = \begin{bmatrix} 1\ 0\ 0 \end{bmatrix}, \tag{29}$$

and we have parameter vector

$$\boldsymbol{\theta} = (k_{01}, k_{12}, k_{21}, k_{23}, k_{32}, x_{20})^\top \in \mathbb{R}^5_+ \,. \tag{30}$$

Condition 1 of Definition 7 is satisfied for linear systems. To test whether S_0 satisfies Condition 2 of Definition 7, we note that

$$\dot{\mathbf{x}}_{\mathbf{0}}(0,\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\mathbf{x}_{\mathbf{0}_0}(\boldsymbol{\theta}) = \begin{bmatrix} k_{12}x_{20} \\ -(k_{12}+k_{32})x_{20} \\ k_{32}x_{20} \end{bmatrix} \neq \mathbf{0} \quad \text{(as all parameters are strictly positive)}, \tag{31}$$

and thus the condition is satisfied for all $\boldsymbol{\theta} \in \Theta$. As the conditions of Definition 7 are satisfied, we may proceed in testing S_0 for SGI following Proposition 1 and Definition 8.

Recall that in this uncontrolled case, the Laplace transform of the output function has the form of (23). Following the notation introduced earlier, we write the transform for $y_0(\cdot;\boldsymbol{\theta})$ as ${}^{S_0}V(s;\boldsymbol{\theta})$, which is a scalar, and the only source of invariants for S_0. Deriving the expression (and neglecting the matrix indices of Definition 9 for simplicity) yields

$${}^{S_0}V(s;\boldsymbol{\theta}) = \frac{\phi_4(\boldsymbol{\theta})s + \phi_3(\boldsymbol{\theta})}{s^3 + \phi_2(\boldsymbol{\theta})s^2 + \phi_1(\boldsymbol{\theta})s + \phi_0(\boldsymbol{\theta})}, \quad \forall s \in \mathbb{C}_0, \tag{32}$$

where

$$\begin{aligned} \phi_0(\boldsymbol{\theta}) &= k_{01}k_{12}k_{23}\,, \\ \phi_1(\boldsymbol{\theta}) &= k_{01}k_{12} + k_{01}k_{23} + k_{01}k_{32} + k_{12}k_{23} + k_{21}k_{23} + k_{21}k_{32}\,, \\ \phi_2(\boldsymbol{\theta}) &= k_{01} + k_{12} + k_{21} + k_{23} + k_{32}\,, \\ \phi_3(\boldsymbol{\theta}) &= k_{12}k_{23}x_{20}\,, \\ \phi_4(\boldsymbol{\theta}) &= k_{12}x_{20}\,. \end{aligned} \tag{33}$$

We set

$$\boldsymbol{\phi}_{\mathbf{0}}(\boldsymbol{\theta}) \triangleq \Big(\phi_0(\boldsymbol{\theta}), \phi_1(\boldsymbol{\theta}), \phi_2(\boldsymbol{\theta}), \phi_3(\boldsymbol{\theta}), \phi_4(\boldsymbol{\theta})\Big)^{\top}, \tag{34}$$

and defining

$$\boldsymbol{\theta}' \triangleq \big(k'_{01}, k'_{12}, k'_{21}, k'_{23}, k'_{32}, x'_{20}\big)^{\top} \in \mathbb{R}^5_+ \tag{35}$$

allows us to form the test equations

$$\boldsymbol{\phi}_{\mathbf{0}}(\boldsymbol{\theta}') = \boldsymbol{\phi}_{\mathbf{0}}(\boldsymbol{\theta})\,. \tag{36}$$

We have six parameters, and merely five conditions. As such, we expect that S_0 is not SGI. Solving System (36) for feasible $\boldsymbol{\theta}'$ yields the solution set:

$$\mathscr{I}(S_0, \boldsymbol{\phi}_{\mathbf{0}}) = \left\{ \boldsymbol{\theta}' \in \mathbb{R}^5_+ \;\middle|\; \begin{array}{l} \left\{ \frac{x'_{20}k_{01}}{x_{20}},\ \frac{k_{12}x_{20}}{x'_{20}},\ \Psi - \frac{\sqrt{\Pi}}{2x'_{20}},\ k_{23},\ \frac{\chi+\sqrt{\Pi}}{2x'_{20}},\ x'_{20} \right\}, \\ \left\{ \frac{x'_{20}k_{01}}{x_{20}},\ \frac{k_{12}x_{20}}{x'_{20}},\ \Psi + \frac{\sqrt{\Pi}}{2x'_{20}},\ k_{23},\ \frac{\chi-\sqrt{\Pi}}{2x'_{20}},\ x'_{20} \right\} \end{array} \right\}, \tag{37}$$

where we interpret x'_{20} as a free parameter,

$$\Psi \triangleq \frac{\phi_1(\boldsymbol{\theta})}{2} - \frac{k_{01}x'_{20}}{x_{20}} - \frac{k_{12}x_{20}}{2x'_{20}},$$

and setting $\Xi \triangleq k_{01} + k_{21} - k_{23}$ allows us to write

$$\begin{aligned}\chi &\triangleq (\Xi + k_{12} + k_{32})x'_{20} - k_{12}x_{20},\\ \Pi &\triangleq \left(\Xi^2 + 2(k_{12} - k_{32})\Xi + (k_{12} + k_{32})^2\right)x'^2_{20}\\ &\quad - 2k_{12}x_{20}(\Xi - k_{12} + k_{32})x'_{20} + k^2_{12}x^2_{20}.\end{aligned} \tag{38}$$

By substituting $x'_{20} = x_{20}$ into either of the solution families given in (37) we see that the trivial solution $\boldsymbol{\theta}' = \boldsymbol{\theta}$ is also valid, as we would expect. We note that the parameter k_{23} is SGI.

Even though structure S_0 contains relatively simple models, (37) with (38) show that the solutions for $\boldsymbol{\theta}'$ in terms of $\boldsymbol{\theta}$ are somewhat complicated, and not particularly easy to categorise. However, we see in (37) that there are two distinct families of solutions. As x'_{20} is free in each, there are uncountably infinitely-many feasible vectors $\boldsymbol{\theta}'$ that reproduce the structure's output for a nominated $\boldsymbol{\theta}$. As such, we judge S_0 as SU.

4 Testing a controlled structure for structural global identifiability

In considering the properties of a controlled state-space structure, we must account for the effects of inputs. Returning to the testing overview outlined in Proposition 1, it is appropriate to precede Step 1 with a new step:

Step 0 Specify the set of inputs which may be applied to the structure.

It is also appropriate for us to adapt the definitions that suit uncontrolled structures for this setting.

Definition 10. Suppose we have controlled state-space model structure M having parameter set Θ and set of input functions $\mathscr{U}$, and time set $T \subseteq [0,\infty)$. For some unspecified parameter vector and input, $\boldsymbol{\theta} \in \Theta$ and $\mathbf{u} \in \mathscr{U}$ respectively, we illustrate M with representative model $M(\boldsymbol{\theta})$ (say, as in (6)), having state function $\mathbf{x}(\cdot, \mathbf{u}; \boldsymbol{\theta}) \in \mathbb{R}^{\mathbf{n}}$ and output function $\mathbf{y}(\cdot, \mathbf{u}; \boldsymbol{\theta}) \in \mathbb{R}^{\mathbf{k}}$.

Suppose that for each $\mathbf{u} \in \mathscr{U}$ systems in M satisfy conditions:

1. Functions $\mathbf{f}(\mathbf{x}, \mathbf{u}, \cdot; \boldsymbol{\theta})$ and $\mathbf{g}(\mathbf{x}, \mathbf{u}, \cdot; \boldsymbol{\theta})$ are real and analytic for every $\boldsymbol{\theta} \in \Theta$ on $\mathscr{S}$ (a connected open subset of $\mathbb{R}^n$ such that $\mathbf{x}(\mathbf{t}, \mathbf{u}; \boldsymbol{\theta}) \in \mathscr{S}$ for every $t \in [0, \tau]$, $\tau > 0$).
2. For t belonging to (at least) some subinterval of $[0, \tau]$, $\mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{t}; \boldsymbol{\theta}) \neq \mathbf{0}$ for almost all $\boldsymbol{\theta} \in \Theta$.

Given finite time $\tau > 0$, we define

$$\mathscr{I}(M,\mathscr{U}) \triangleq \left\{ \boldsymbol{\theta}' \in \Theta : \mathbf{y}(t,\boldsymbol{u};\boldsymbol{\theta}') = \mathbf{y}(t,\boldsymbol{u};\boldsymbol{\theta}) \quad \forall t \in [0,\tau],\ \forall \boldsymbol{u} \in \mathscr{U} \right\}. \tag{39}$$

If, for almost all $\boldsymbol{\theta} \in \Theta$:

$\mathscr{I}(M,\mathscr{U}) = \{\boldsymbol{\theta}\}$: M is structurally globally identifiable for input set $\mathscr{U}$ ($\mathscr{U}$-SGI);
the elements of $\mathscr{I}(M,\mathscr{U})$ are denumerable: M is structurally locally identifiable for input set $\mathscr{U}$ ($\mathscr{U}$-SLI);
the elements of $\mathscr{I}(M,\mathscr{U})$ are not denumerable: M is structurally unidentifiable for input set $\mathscr{U}$ ($\mathscr{U}$-SU).

Remark 2. Conditions 1 and 2 of Definition 10 play similar roles to the corresponding conditions of Definition 7. Condition 1 excludes from consideration structures subject to discontinuities in the state or output functions, for which we cannot readily define invariants. Condition 2 relates to conditions which allow us to elicit informative input from a system in M. This loosens the condition of the uncontrolled case, where a system at equilibrium at $t = 0$ remains there. The controlled case is different; a system at an equilibrium state may be displaced by the action of an input. However, this alone does not guarantee that the output of a controlled system is informative for any input in $\mathscr{U}$. As such, Condition 2 seeks to preclude the case where the system's state is largely constant, possibly changing only at isolated points on $[0,\tau]$. By doing so, we expect to obtain useful (non-degenerate) output, and possibly, invariants subsequently, depending on the nature of $\mathscr{U}$.

Should Conditions 1 and 2 not hold for any $\mathbf{u} \in \mathscr{U}$, it is appropriate to remove these from the input set.

Suppose M satisfies Conditions 1 and 2 of Definition 10, and we may observe M's outputs for $\mathscr{U}$ containing a sufficiently broad range of inputs (e.g. the set of piecewise continuous functions defined on T, [19]). Then, within our idealised testing framework (Convention 3) we can access the structure's invariants, say ϕ. In such a case, rather than making a judgement on M using Definition 10, we may use ϕ with the more convenient Definition 8.

Let us turn our attention to the application of Definition 10 when M is a controlled compartmental LTI structure. By physical reasoning ($\mathbf{x}$ is real and does not exhibit jumps, and these properties are transferred to $\mathbf{y}$) we expect that Condition 1 is satisfied. Checking Condition 2 may not be trivial in general, and so it may be easier to verify an alternative condition, even if this is stricter than necessary. For example, if we were to show that $\dot{\mathbf{x}}(\mathbf{t};\boldsymbol{\theta}) \neq \mathbf{0}$ for almost all $\boldsymbol{\theta} \in \Theta$ and any $t \in [0,\tau]$ for finite τ, then Condition 2 is satisfied.

In practice, conditions such as those of Definition 10 do not typically feature in discussions of the testing of controlled LTI structures for SGI. This is likely due to the expectation that one can access a structure's invariants if the input set meets only modest requirements: that $\mathscr{U}$ is sufficiently diverse, and that the Laplace transform of any input in $\mathscr{U}$ exists. Satisfying these conditions allows us to derive transfer matrices $\mathbf{W}$ and $\mathbf{V}$ as in (20), place transfer functions contained therein in canonical form (recall Definition 9), and obtain ϕ from their coefficients.

In various situations, for practical or ethical reasons, one is limited in the nature and number of inputs that one can apply to some physical system. In such a case, it is not appropriate to assume that we may access ϕ from M. As such, the testing framework seen in Definition 8 is an inappropriate idealisation. However, we may consider the result of such a test as a "best case scenario"—we would not expect to obtain a more favourable result from a limited set of inputs. As such, if a test using ϕ shows that M is SU, we can be almost certain that PI applied to the output from our physical system resulting from a limited set of inputs will not obtain unique parameter estimates. Inconveniently, when the test classifies M as SGI or SLI, we cannot necessarily ascertain whether this judgement will also apply when we know that limited inputs are available. As such, it is appropriate to return to Definition 10 and consider a test for generic uniqueness of parameter vectors that takes into account the set of available inputs, and which does not require invariants.

Some authors have noted situations where—unlike in the testing of a structure for SGI based on invariants—we may not consider inputs as being applied sequentially to yield separate output time courses. For example, in considering LTI compartmental structures, Godfrey [12, Page 95] cautioned:

> However, when more than one input is applied simultaneously, identifiability may depend on the shape of the two inputs, and it is then essential to examine the form of the observations $\mathbf{Y}(\mathbf{s})$ [the Laplace transform of $\mathbf{y}$] rather than individual transfer functions.

In noting the importance of the available set of inputs, Jacquez and Greif [15, Page 201] sought to distinguish "system identifiability" (which we understand as SGI) from "model identifiability" which depends on some particular inputs (as we have allowed for in Definition 10). The authors noted the confusion caused by failing to distinguish between these different properties. To the best of our knowledge, the literature does not have consistent terminology to distinguish these concepts, which may be a consequence of how infrequently it is explicitly considered.

We will seek to reuse the TFA machinery in considering what parameter information we may glean from the idealised output of a compartmental LTI structure subject to a single input. Let us consider such a structure S having system matrices in $SL\Sigma P(n,m,k)$. Suppose that we can observe idealised output for a single input $\mathbf{u}$, that is $\mathscr{U} = \{\mathbf{u}\}$, and that $\mathscr{L}\{\mathbf{u}\}(\mathbf{s})$ exists. Then, we may obtain parameter information for testing S for SGI given $\boldsymbol{u}$ from

$$\mathscr{L}\{y\}(s;\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})\big(s\mathbf{I} - \mathbf{A}(\boldsymbol{\theta})\big)^{-1}\big(\mathbf{x_0}(\boldsymbol{\theta}) + \mathbf{B}(\boldsymbol{\theta})\mathscr{L}\{\mathbf{u}\}(s)\big)\,. \tag{40}$$

In order to demonstrate the difference between the testing of a controlled structure when invariants are and are not obtainable, we shall consider an example structure for which different input sets are available. Recall the SISO structure S_1 from Sect. 1. Following definitions from Sect. 2, we rewrite the representative system in state-space form as

$$\dot{\mathbf{x}}_1(t;\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\,\mathbf{x_1}(t;\boldsymbol{\theta}) + \mathbf{B}(\boldsymbol{\theta})u(t)\,, \quad \mathbf{x_1}(0;\boldsymbol{\theta}) = \mathbf{x_{1_0}}(\boldsymbol{\theta})\,, \tag{41}$$

$$y_1(t;\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})\,\mathbf{x_1}(t;\boldsymbol{\theta})\,, \tag{42}$$

where the state vector is $\mathbf{x_1}(t;\boldsymbol{\theta}) = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top$, and system matrices belong to $SL\Sigma P(3,1,1)$. Specifically we have

$$\mathbf{x_1}(0;\boldsymbol{\theta}) = \begin{bmatrix} 0 \\ x_{20} \\ 0 \end{bmatrix}, \quad \mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} -k_{01}-k_{21} & k_{12} & 0 \\ k_{21} & -k_{12}-k_{32} & k_{23} \\ 0 & k_{32} & -k_{23} \end{bmatrix}, \\ \mathbf{B}(\boldsymbol{\theta}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{C}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}. \tag{43}$$

Recalling (21) and (22), the transfer matrices here are scalars, which henceforth we denote by ${}^{S_1}W$ and ${}^{S_1}V$. We note that by neglecting $\mathbf{B}$ we obtain the uncontrolled LTI structure S_0 (recall (27) and (28)). Structure S_1 has the same parameter vector as S_0, shown in (30).

Below we proceed to test S_1 for SGI under the assumption that we can obtain its invariants.

4.1 A demonstration of the testing of a controlled LTI structure for SGI when invariants are accessible from outputs

Let us assume that we have the idealised outputs of S_1 for a sufficiently large input set $\mathscr{U}$ such that we can obtain ${}^{S_1}W$ and ${}^{S_1}V$. By converting each of these rational functions into the canonical form, we may obtain each coefficient of s. The collection of these specifies a vector of invariants. We shall recall the steps of Proposition 1 in testing S_1 for SGI.

Towards Step 1, those invariants relating to the response due to the initial conditions reside in ${}^{S_1}V \equiv {}^{S_0}V$. We collected these invariants in (34).

The behaviour of S_1 differs from that of S_0 due to the invariants relating to inputs, held in ${}^{S_1}W$. Following (22), we see that ${}^{S_1}W \triangleq \mathbf{C}(s\mathbf{I}_3 - \mathbf{A})^{-1}\mathbf{B}$, from which we obtain the transfer function in canonical form:

$${}^{S_1}W(\boldsymbol{\theta}) = \frac{\omega_0(\boldsymbol{\theta})}{s^3 + \phi_2(\boldsymbol{\theta})s^2 + \phi_1(\boldsymbol{\theta})s + \phi_0(\boldsymbol{\theta})}, \tag{44}$$

where the denominator invariants repeat the corresponding coefficients in $\mathscr{L}\{y_0\}(s;\boldsymbol{\theta})$ (recall (33)), and

$$\omega_0(\boldsymbol{\theta}) = k_{12}k_{23}. \tag{45}$$

Thus, only $\omega_0(\boldsymbol{\theta})$ provides an invariant that is novel compared to those from ${}^{S_0}V(\boldsymbol{\theta})$.

Drawing on (34) and (45), we complete Step 1 by forming the vector of distinct invariants associated with S_1:

$$\boldsymbol{\phi_1}(\boldsymbol{\theta}) \triangleq (\underbrace{\phi_0(\boldsymbol{\theta}), \phi_1(\boldsymbol{\theta}), \phi_2(\boldsymbol{\theta})}_{\substack{\text{common to} \\ {}^{S_1}V(\boldsymbol{\theta}),\, {}^{S_1}W(\boldsymbol{\theta}) \\ \text{denominators}}}, \underbrace{\phi_3(\boldsymbol{\theta}), \phi_4(\boldsymbol{\theta})}_{\substack{\text{from numerator} \\ \text{of } {}^{S_1}V(\boldsymbol{\theta})}}, \underbrace{\omega_0(\boldsymbol{\theta})}_{\substack{\text{from numerator} \\ \text{of } {}^{S_1}W(\boldsymbol{\theta})}})^\top . \tag{46}$$

Following Step 2 we use $\boldsymbol{\phi_1}(\boldsymbol{\theta})$ from (46) to form the invariants dependent on our alternative parameter $\boldsymbol{\theta}'$, (as in (35)), $\boldsymbol{\phi_1}(\boldsymbol{\theta}')$. Step 3 directs us to form the test equations $\boldsymbol{\phi_1}(\boldsymbol{\theta}') = \boldsymbol{\phi_1}(\boldsymbol{\theta})$. Upon solving for feasible $\boldsymbol{\theta}'$ we obtain

$$\mathscr{I}(S_1, \boldsymbol{\phi_1}) = \left\{ \boldsymbol{\theta}' \in \mathbb{R}^5_+ \;\middle|\; \begin{matrix} \{k_{01},\, k_{12},\, k_{21}, & & k_{23},\, k_{32}, & & x_{20}\}, \\ \{k_{01},\, k_{12},\, -k_{01}+k_{23}+k_{32},\, k_{23},\, k_{01}+k_{21}-k_{23},\, x_{20}\} \end{matrix} \right\}. \tag{47}$$

Equation (47) shows that we can obtain unique estimates for k'_{01}, k'_{12}, k'_{23}, and x'_{20} (i.e. the corresponding true values in $\boldsymbol{\theta}$) for any $\boldsymbol{\theta} \in \mathbb{R}^6_+$. However, for each of k'_{21} and k'_{32} we see there are two distinct solutions whenever $-k_{01}+k_{23}+k_{32} > 0$ and $k_{01}+k_{21}-k_{23} > 0$. That is, the structure is SLI.

Inspection of the second solution family in (47) reveals $k'_{21}+k'_{32} = k_{21}+k_{32}$. This may hint that a reparameterisation of S_1 so as to replace occurrences of $k_{21}+k_{32}$ (which may occur in combination with other parameters) with appropriate new parameters would produce a new structure which is SGI. Whilst there are techniques for generating alternative structures that produce the same output (e.g. [21]), in general, finding a suitable reparameterisation amongst these is not a trivial undertaking. Given this, we may have to find some means of managing an SU structure. For example, we may determine bounds on the values of parameters by testing for "interval identifiability". If the bounds are sufficiently narrow, we may tolerate an SU structure (see [13] for examples).

We shall now consider S_1 in the more restrictive setting where our idealised output results from the application of one specific input.

4.2 A demonstration of the testing of a controlled LTI structure for SGI when invariants are not accessible from outputs

Suppose that we can only observe the idealised output of S_1 for the single input $u = \delta(t-0)$—the impulsive input at time zero. Noting that $\mathscr{L}\{\delta(t-0)\}(s) = 1$, and recalling (20), we may write

$$\mathscr{L}\{y_2(\cdot, \boldsymbol{\theta})\}(s) = {}^{S_1}V(s; \boldsymbol{\theta}) + {}^{S_1}W(s; \boldsymbol{\theta}), \tag{48}$$

where the terms on the right-hand side are given by (32) (recalling ${}^{S_1}V(s;\boldsymbol{\theta}) \equiv {}^{S_0}V(s;\boldsymbol{\theta})$) and (44), respectively.

The sum of the two transfer functions on the right-hand side of (48) is also a rational function in s, and hence is analogous to a transfer function. As such, it is

convenient to process this in a manner similar to that shown in Sect. 4.1. Thus, ensuring that the right-hand side of (48) is in the canonical form, and simplifying, yields an expression (which is similar to the canonical form of $\mathscr{L}\{y_0(\cdot,\boldsymbol{\theta})\}(s)$, recall (32)):

$$\mathscr{L}\{y_2\}(s;\boldsymbol{\theta}) = \frac{\phi_4(\boldsymbol{\theta})s+\beta(\boldsymbol{\theta})}{s^3+\phi_2(\boldsymbol{\theta})s^2+\phi_1(\boldsymbol{\theta})s+\phi_0(\boldsymbol{\theta})}, \quad \forall s \in \mathbb{C}_0, \tag{49}$$

where, recalling (33) and (45),

$$\beta(\boldsymbol{\theta}) \triangleq \phi_3(\boldsymbol{\theta}) + \omega_0(\boldsymbol{\theta}) = k_{12}k_{23}(x_{20}+1)\,.$$

Remark 3. Given the input $u = \delta(t-0)$ and that S_1 is an open system, mass present in the system due to the input and initial conditions is lost to the environment over time. As $t \to \infty$, the system approaches its steady state $\mathbf{x}^* = \mathbf{0}$. We note that (49) is the Laplace transform of an output function that is a sum of exponentials in t (recall Sect. 2.3.1) as a result of being a linear combination of the individual state variables. As all $\boldsymbol{\theta}$ are positive, all invariants in (49) are also positive. As such, we see that $\mathbf{y}$ is not constant. We infer that the state function $\mathbf{x}$ is time-varying, and that it leads to an informative output function. Thus, S_1 for this u satisfies Condition 2 of Definition 10.

We note that $\mathscr{L}\{y_2\}(s;\boldsymbol{\theta})$ and $\mathscr{L}\{y_0\}(s;\boldsymbol{\theta})$ differ only in the constant term of their numerators. The coefficients in (49) play a similar role to invariants as they determine the output. As a further conceptual and notational convenience, we write

$$\boldsymbol{\phi_2}(\boldsymbol{\theta}) \triangleq (\phi_0(\boldsymbol{\theta}), \phi_1(\boldsymbol{\theta}), \phi_2(\boldsymbol{\theta}), \beta(\boldsymbol{\theta}), \phi_4(\boldsymbol{\theta}))^\top.$$

Following Steps 2 and 3 of Proposition 1 leads to a system of test equations $\boldsymbol{\phi_2}(\boldsymbol{\theta}') = \boldsymbol{\phi_2}(\boldsymbol{\theta})$, containing four of the five equations used in testing S_0 for SGI.

Let us consider the difference between the systems of equations which follow from $\boldsymbol{\phi_0}$ and $\boldsymbol{\phi_2}$. The analysis of S_0 produces a novel equation involving ϕ_3. In analysing S_1 output due to a single input here, the novel equation is due to $\beta(\boldsymbol{\theta})$. This allows k_{12}, k_{23}, and x_{20} more freedom than that permitted by the ϕ_3 equation. Thus, solving $\boldsymbol{\phi_2}(\boldsymbol{\theta}') = \boldsymbol{\phi_2}(\boldsymbol{\theta})$ yields an even more complicated solution set than that seen for S_0 in (37) and (38). As a kindness to the reader, we shall not present the solution sets here. However, classification of the structure is straightforward as $\boldsymbol{\phi_2}(\boldsymbol{\theta})$ provides five equations, yet we have six parameters. Thus, when the input set is $\mathscr{U} = \{\delta(t-0)\}$, we classify S_1 as $\mathscr{U}$-SU.

This is a less-favourable result than the classification of S_1 as SLI (recall the the assumption that outputs are available for a broad enough range of inputs) as demonstrated in (47). This result reinforces the claim that, when intending to test a structure for SGI, it is appropriate to specify the inputs which will be applied to physical system. Thence, we may judge whether or not the associated idealised output allows determination of invariants, and use this knowledge in choosing an appropriate testing method.

5 Concluding remarks

This overview has aimed to highlight the benefits of testing model structures for the property of structural global identifiability (SGI). Moreover, by assembling crucial definitions, drawing important distinctions, and providing test examples, we have sought to illuminate some important concepts in the field of identifiability analysis. We hope that this will encourage and assist interrogation of proposed structures so as to recognise those that are not SGI. This will allow researchers to anticipate the frustrations almost certain to accompany the use of a non-SGI structure (especially, an unidentifiable one) in modelling and parameter estimation.

Progress in the field of identifiability analysis is ongoing through the development of new methods of testing structures for SGI or SLI, and refinements to their implementation. However, certain practical matters are yet to receive widespread consideration. We conclude with brief comments on a selection of these.

Competition—or collaboration—between testing methods? Over a period of time, the literature has reported that one cannot generally anticipate which method will be easiest to apply to a given case, (e.g. [12, Page 96]), or that testing methods may suit some problems more than others (e.g. [7]). Consequently, when considering software implementations of testing methods, we may not be able to anticipate which method will produce a result in the shortest time, or at all. This uncertainty has prompted various comparisons aimed at evaluating the utility of alternative methods for testing structures for SGI.

One may wonder if a competitive treatment of methods is a limiting one. That is, might there be benefits in combining methods so as to draw upon their strengths? For example, in considering controlled compartmental LTI structures, the TFA provides a means of ascertaining whether or not a structure is generically minimal. If the conclusion is positive, we may then change our approach and apply a suitable testing method that uses a type of invariant expected to be simpler than those used in the TFA. For example, we may choose Markov and initial parameters as invariants, expecting these polynomials in the parameters to have a lower degree than those seen in transfer function coefficients. Given such simpler invariants, the resulting test equations will have a reduced algebraic complexity. We could reasonably expect to solve these more quickly than equations obtained from the TFA.

Reproducibility of analysis There is a growing concern over the reproducibility of studies in computational biology ([16]). We expect a greater awareness of identifiability analysis to encourage the asking of questions that will contribute to a rigorous and defensible modelling practice. Beyond this, we may also ponder how to promote reproducibility through the processes by which identifiability analysis is undertaken.

For all but the simplest cases, testing a structure for SGI requires the use of a computer algebra system (CAS). Often this is a commercial product, such as Maple, Mathematica, or MATLAB. However, as for all complex computer code, one cannot necessarily guarantee that results produced by a CAS will be correct in all situations (see, for example, [2] noting a limitation of certain versions of Maple). As such, it

is good practice for us to check that results obtained from one CAS agree with those from another.

Performing such a comparison might not be straightforward. Recall that the classical approach to testing a structure for SGI requires the solution of a system of algebraic equations. If two CASs employ differing methods in solving a given system, the solution sets may appear quite dissimilar, even if they are, in fact, the same. This complicates the task of determining whether or not the solution sets are equivalent.

We may be able to make choices that can reduce the complexity of the comparison problem. One approach is to seek to direct the output of CASs by specifying similar options in their commands where this is possible. For example, the "solve" command in Maple allows the user to specify various options, including some relating to how any solutions are displayed. Another Maple option allows some variables to be specified as "functions of a free variable". We may be encouraged to use this given the form of solutions obtained from another CAS which we would like to emulate.

The seeming dissimilarity of solutions may be due to features of CAS solution algorithms that we cannot directly control. As such, we may seek to manage these by further scrutinising our equations (or more fundamentally our invariants $\phi(\boldsymbol{\theta})$), before we attempt to solve them.

Suppose that each (multivariate polynomial) element of $\phi(\boldsymbol{\theta})$ is some combination of simpler polynomials in parameters $\boldsymbol{\theta}$. We may determine these new polynomials by calculating a Gröbner basis for $\phi(\boldsymbol{\theta})$.[7] This requires an "ordering" of parameters, which determines how terms are arranged within a polynomial, and how monomials are arranged within terms.[8] We may obtain differing bases depending on the chosen ordering.

In certain solution methods (such as "nonlinsolve" in version 1.4 of Python package SymPy) a CAS may (effectively) calculate a Gröbner basis for invariants, choosing an ordering without user input. In such cases, should different CASs employ differing orderings, the solutions of test equations may appear quite different. As such, it may be useful for the user to obtain a Gröbner basis for a specified ordering, and use this in formulating test equations for each CAS.

We shall illustrate the importance of the choice of ordering by returning to our example structure S_1. In Maple 2019 (version 1) we used the "Basis" command (from the "Groebner" package) to compute Gröbner bases for $\phi_{\mathbf{1}}(\boldsymbol{\theta})$ under different orderings. We varied the ordering of parameters, as specified by the "plex()" option (pure lexicographical ordering). The ordering indicates a decreasing preference for eliminating parameters from our input polynomials (here, our invariants) as we proceed from the start of the list, with the aim of forming a "triangular" system in $\boldsymbol{\theta}$. As

[7] We may consider a Gröbner basis for a list of polynomials as analogous to the reduced row-echelon form of a system of linear equations.

[8] For example, the polynomial $x^2y + 2xy^3 - 4x + y$ employs "pure lexicographical ordering" with "$x > y$"—terms are arranged by decreasing degree of monomials in x, and within each term any monomial in x appears before one in y. Changing the ordering to "$y > x$" yields an alternative form: $2y^3x + yx^2 + y - 4x$.

such, those parameters occurring earlier in the list are more likely to be eliminated than those occurring later.

Using the ordering $k_{21} > k_{32} > k_{01} > k_{12} > k_{23} > x_{20}$ yields the Gröbner basis:

$$\mathbf{b_1}(\boldsymbol{\theta}) \triangleq \begin{bmatrix} k_{12}x_{20}, \\ k_{12}k_{23}, \\ -k_{01}k_{12} + k_{12}k_{32} + k_{23}^2 + 2k_{23}k_{32} + k_{32}^2, \\ k_{01} + k_{12} + k_{21} + k_{23} + k_{32} \end{bmatrix}. \tag{50}$$

Alternatively, with the ordering $k_{23} > k_{32} > x_{20} > k_{21} > k_{12} > k_{01}$, Maple produces the Gröbner basis:

$$\mathbf{b_2}(\boldsymbol{\theta}) \triangleq \begin{bmatrix} k_{01}^2 + 2k_{01}k_{21} + k_{21}k_{12} + k_{21}^2 \\ k_{12}x_{20} \\ k_{01}k_{12} + k_{12}^2 + k_{21}k_{12} + k_{12}k_{32} \\ k_{01} + k_{12} + k_{21} + k_{23} + k_{32}\,. \end{bmatrix} \tag{51}$$

The Gröbner bases $\mathbf{b_1}(\boldsymbol{\theta})$ and $\mathbf{b_2}(\boldsymbol{\theta})$ are not identical, having only two components (the first and fourth components of $\mathbf{b_1}(\boldsymbol{\theta})$) in common. (We also note that although (46) shows $\phi_1(\boldsymbol{\theta})$ as comprised of six invariants, (50) (or (51)) shows that in the testing of S_1 for SGI, $\boldsymbol{\theta}$ is subject to only four independent conditions.)

Suppose now that—in a similar manner as we did for $\phi_1(\boldsymbol{\theta})$—we use $\mathbf{b_1}(\boldsymbol{\theta})$ and $\mathbf{b_2}(\boldsymbol{\theta})$ in turn to define two distinct systems of four SGI test equations. The associated solution sets for $\boldsymbol{\theta}'$, $\mathscr{I}(S_1, \mathbf{b_1})$ and $\mathscr{I}(S_1, \mathbf{b_2})$ respectively, determined by Maple appear to be quite different. For example, $\mathscr{I}(S_1, \mathbf{b_1})$ shows k'_{12} and k'_{32} as free parameters, whereas $\mathscr{I}(S_1, \mathbf{b_2})$ has k'_{01} and k'_{21} free. This result suggests that using a Gröbner basis of our invariants to define SGI test conditions may remove one cause of unwanted variation between results obtained by different CASs.

When faced with (potential or actual) disparities between CAS results, access to the source code may illuminate the cause of the divergence, and contribute to its resolution. However, certain CAS do not permit such access to the source. In light of this, we are currently developing open-source code using the programming language Python, making particular use of the SymPy (symbolic algebra) package. By implementing this in the Jupyter notebook environment, we intend to develop implementations of testing algorithms (as we have for the TFA approach) that are readily accessible to the scientific community, and permit user customisation.

Acknowledgements The author is grateful to the organisers of the programme "Influencing public health policy with data-informed mathematical models of infectious diseases" at MATRIX (Creswick, Victoria, July 1–12 2019) for the invitation to present, and exposure to aspects of infectious disease modelling. Appreciation goes also to the organisers of the programme "Identifiability problems in systems biology" held at the American Institute of Mathematics, San Jose, California (August 19–23 2019) and its participants, for useful discussions on contemporary problems.

References

1. American Institute of Mathematics: Identifiability problems in systems biology (2019). URL https://aimath.org/workshops/upcoming/identbio/
2. Armando, A., Ballarin, C.: A reconstruction and extension of Maple's assume facility via constraint contextual rewriting. Journal of Symbolic Computation **39**, 503–521 (2005)
3. Audoly, S., Bellu, G., D'Angiò, L., Saccomani, M.P., Cobelli, C.: Global identifiability of nonlinear models of biological systems. IEEE Transactions on Biomedical Engineering **48**(1), 55–65 (2001)
4. Bellman, R., Åström, K.J.: On structural identifiability. Mathematical Biosciences **7**, 329–339 (1970). URL https://doi.org/10.1016/0025-5564(70)90132-X
5. Benvenuti, L., Farina, L.: Minimal positive realizations: a survey of recent results and open problems. Kybernetika **39**(2), 217–228 (2003)
6. Caines, P.E.: Linear Stochastic Systems. John Wiley & Sons, Inc. (1988)
7. Chis, O.T., Banga, J.R., Balsa-Canto, E.: Structural identifiability of systems biology models: a critical comparison of methods. PloS One **6**(11), e27755 (2011)
8. Cobelli, C., DiStefano III, J.J.: Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology **239**(1), R7–R24 (1980). URL https://doi.org/10.1152/ajpregu.1980.239.1.R7. PMID: 7396041
9. Cox Jr., L.A., Huber, W.A.: Symmetry, Identifiability, and Prediction Uncertainties in Multistage Clonal Expansion (MSCE) Models of Carcinogenesis. Risk Analysis: An International Journal **27**(6), 1441–1453 (2007)
10. Denis-Vidal, L., Joly-Blanchard, G.: Equivalence and identifiability analysis of uncontrolled nonlinear dynamical systems. Automatica **40**(2), 287–292 (2004)
11. DiStefano III, J.: Dynamic systems biology modeling and simulation. Academic Press (2015)
12. Godfrey, K.: Compartmental Models and Their Application. Academic Press Inc. (1983)
13. Godfrey, K., DiStefano III, J.: Identifiability of model parameters. IFAC Proceedings Volumes **18**(5), 89–114 (1985)
14. van den Hof, J.M.: Structural identifiability from input-output observations. Tech. Rep. BS-9514, Centrum voor Wiskunde en Informatica (1995)
15. Jacquez, J.A., Greif, P.: Numerical parameter identifiability and estimability: Integrating identifiability, estimability and optimal sampling design. Mathematical Biosciences **77**(1-2), 201–227 (1985)
16. Laubenbacher, R., Hastings, A.: Editorial. Bulletin of Mathematical Biology **80**(12), 3069–3070 (2018). URL https://doi.org/10.1007/s11538-018-0501-8
17. Poljak, S.: On the gap between the structural controllability of time-varying and time-invariant systems. IEEE Transactions on Automatic Control **37**(12), 1961–1965 (1992)
18. Seber, G.A.F., Wild, C.J.: Nonlinear Regression. Wiley series in probability and statistics. Wiley (2003)
19. Vajda, S.: Structural equivalence of linear systems and compartmental models. Mathematical Biosciences **55**(1-2), 39–64 (1981)
20. Villaverde, A.F., Barreiro, A., Papachristodoulou, A.: Structural Identifiability of Dynamic Systems Biology Models. PLoS Computational Biology **12**(10), e1005153 (2016)
21. Walter, E., Lecourtier, Y.: Unidentifiable compartmental models: what to do? Mathematical Biosciences **56**(1-2), 1–25 (1981)
22. Walter, É., Pronzato, L.: Identification of Parametric Models from Experimental Data. Communication and Control Engineering. Springer (1997)
23. Whyte, J.M.: On deterministic identifiability of uncontrolled linear switching systems. WSEAS Transactions on Systems **6**(5), 1028–1036 (2007)
24. Whyte, J.M.: A preliminary approach to deterministic identifiability of uncontrolled linear switching systems. In: 3rd WSEAS International Conference on Mathematical Biology and Ecology (MABE'07), Proceedings of the WSEAS International Conferences. Gold Coast, Queensland, Australia (2007)

25. Whyte, J.M.: Inferring global *a priori* identifiability of optical biosensor experiment models. In: G.Z. Li, X. Hu, S. Kim, H. Ressom, M. Hughes, B. Liu, G. McLachlan, M. Liebman, H. Sun (eds.) IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2013), pp. 17–22. Shanghai, China (2013). https://doi.org/10.1109/BIBM.2013.6732453
26. Whyte, J.M.: Global *a priori* identifiability of models of flow-cell optical biosensor experiments. Ph.D. thesis, School of Mathematics and Statistics, University of Melbourne, Victoria, Australia (2016)
27. Yamada, T., Luenberger, D.G.: Generic Controllability Theorems for Descriptor Systems. IEEE Transactions on Automatic Control **AC-30**(2), 144–152 (1985)

Chapter 4

International Workshop on Spatial Statistics

Spatial modelling of linear regression coefficients for gauge measurements against satellite estimates

Benjamin Hines, Yuriy Kuleshov and Guoqi Qian

Abstract Satellite imagery provides estimates for the amount of precipitation that has occurred in a region, these estimates are then used in models for predicting future precipitation trends. As these satellite images only provide an estimate for the amount of precipitation that has occurred, it is important that they be accurate estimates. If we assume that a rain gauge correctly measures the amount of precipitation that has occurred in some location over a specified time interval, then we can compare the satellite precipitation estimate to the gauge measurement for the same time interval. By expressing the relationship between the gauge measurement and the satellite precipitation estimate for the same time interval as a linear equation we can then spatially map the coefficients of this linear relationship to inspect the spatial trends of the regression coefficients. We then model the coefficients of the linear equations of each location by a spatial linear model and then use this model to predict the coefficients in location where there are no rain gauges available.

1 Introduction

The ability to measure precipitation accurately is very important for many reasons. Knowing how much precipitation has occurred in a given region will help us improve our knowledge about the seasonal patterns and climate trends in said region and the regions around it. Rain gauges are used to measure the amount of precipitation that has occurred in a given time period, where precipitation is captured and then measured at equally spaced intervals. Figure 1 shows the locations of all the

Benjamin Hines and Guoqi Qian
School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3010, Australia, e-mail: benjamin.hines@unimelb.edu.au, e-mail: g.qian@ms.unimelb.edu.au.

Yuriy Kuleshov
Australian Bureau of Meterology, Docklands Melbourne VIC 3008 Australia; School of Science, RMIT University, e-mail: yuriy.kuleshov@bom.gov.au.

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_11

Bureau of Meteorology's monthly rain gauges around Australia that are operational as of August 2018 in which there are 865 of. Clearly, the gauges are not uniformly placed around Australia, but are distributed based on the population concentration. If we wanted to know how much precipitation has occurred in a region where there

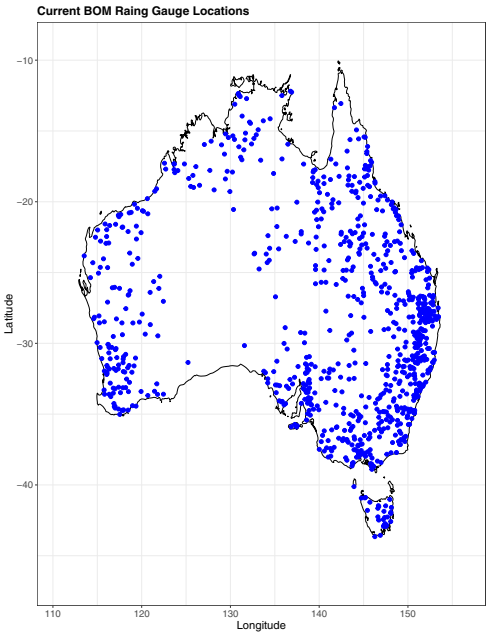


Fig. 1: Bureau of Meteorology's Australian monthly rain gauge locations as of August 2018.

are no rain gauges, we would have to rely on satellite precipitation estimates. Satellite precipitation estimation is done by taking an image of a cloud from above and then estimating the amount of precipitation that will come from that cloud based on its physical characteristics [9]. As these satellite images only provide an estimate for how much precipitation has occurred in a region, it is important to know if these estimates are close to what a rain gauge would measure. If the satellite images do indeed provide us with good estimates for the amount of precipitation, then there is no issue with using these estimates as measurements for regions in which there are no rain gauges located. However, if the satellite images do not provide us with a good estimate for how much precipitation a rain gauge has measured for some time interval for a given location, then we can look at the difference between the gauge measurement and the satellite estimate and how there may be some relationship between the two, which may depend on the location of interest.

2 Methodology

Let $\mathbf{x}_i \in \mathbb{R}^2$ for $i = 1, \ldots, m$ be the two dimensional coordinates description of the i^{th} location where x_{i1} and x_{i2} are the longitude and latitude for location i respectively. We then define $Y_{ij}^{[g]}$ and $Y_{ij}^{[s]}$ to be the gauge measurement and satellite estimate for the i^{th} location respectively for the j^{th} time period where $j = 1, \ldots, n_i$. We can then consider $Y_{ij}^{[g]}$ and $Y_{ij}^{[s]}$ to have a linear relationship, i.e.

$$Y_{i,j}^{[g]} = \beta_0^{[i,j]} + \beta_1^{[i,j]} Y_{i,j}^{[s]}.$$

Now if we consider the relationship between the gauge measurements and satellite estimates to be temporally stationary (coefficients are the same regardless of time) [5], then we can express a linear equation for location i as

$$\mathbf{Y}_i^{[g]} = \beta_0^{[i]} \mathbf{1}_{n_i} + \beta_1^{[i]} \mathbf{Y}_i^{[s]} + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\mathbf{Y}_i^{[g]} = \left(Y_{i,1}^{[g]}, \ldots, Y_{i,n_i}^{[g]}\right)^T$, $\mathbf{Y}_i^{[s]} = \left(Y_{i,1}^{[s]}, \ldots, Y_{i,n_i}^{[s]}\right)^T$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \sigma_i^2 I_{n_i}\right)$ is the noise term and $\mathbf{1}_{n_i}$ is an $n_i \times 1$ vector with every entry being 1. We would expect to estimate $\beta_0^{[i]} = 0$ and $\beta_1^{[i]} = 1$ for $i = 1, \ldots, m$ as we expect the satellite precipitation estimate to be equal to the gauge measurement. The conventional method for estimating the coefficients $\beta_0^{[i]}$ and $\beta_1^{[i]}$ in this situation is by minimising the ordinary least squares equation

$$\hat{\beta}^{[i]} = \underset{\beta_0^{[i]}, \beta_1^{[i]}}{\arg\min} ||\mathbf{Y}_i^{[g]} - \beta_0^{[i]} \mathbf{1}_{n_i} - \beta_1^{[i]} \mathbf{Y}_i^{[s]}||_2^2 \tag{2}$$

where $\hat{\beta}^{[i]} = \left(\hat{\beta}_0^{[i]}, \hat{\beta}_1^{[i]}\right)^T$ and $||\mathbf{t}||_p = \left(\sum_{i=1}^n |t_i|^p\right)^{1/p}$. The solution to equation (2) can be shown to be

$$\hat{\beta}^{[i]} = \left(V_i^T V_i\right)^{-1} V_i^T \mathbf{Y}_i^{[g]}$$

with $V_i = \left(\mathbf{1}_{n_i}, \mathbf{Y}_i^{[s]}\right)$ an $n_i \times 2$ matrix [3]. In doing this we can obtain estimates for the coefficients $\beta_0^{[i]}$ and $\beta_1^{[i]}$ for all locations which we define by

$$\beta_\ell = \left[\hat{\beta}_\ell^{[1]} \ \hat{\beta}_\ell^{[2]} \ \cdots \ \hat{\beta}_\ell^{[m]}\right]^T \tag{3}$$

an $m \times 1$ vector for $\ell = 0, 1$.

Once we have our estimates for the coefficients of each location as in equation (3), we can think of these spatially specified coefficients as a spatial process of a geostatistical random field [14], where we can construct a model to test if there is some sort of spatial dependency structure. Consider the following spatial linear model:

$$\beta_\ell = \alpha_\ell \mathbf{1}_m + \lambda_\ell W_{m,\ell}\beta_\ell + \varepsilon(\beta_\ell) \tag{4}$$

for $\ell = 0, 1$, where λ_ℓ is the autocorrelation parameter, α_ℓ is the intercept coefficient of the model, $W_{m,\ell}$ is a given $m \times m$ weight matrix representing the spatial distances of the observations and $\varepsilon(\beta_\ell) \sim \mathcal{N}\left(\mathbf{0}, \sigma^2(\beta_\ell) I_m\right)$ for $\ell = 0, 1$ [13]. Furthermore, $W_{m,\ell} = \left\{ w_{ij}^{[\ell]} \right\}$, where $w_{ij}^{[\ell]}$ is given by some function of a known distance metric $d(\mathbf{x}_i, \mathbf{x}_j)$ where locations $\mathbf{x}_i$ and $\mathbf{x}_j$ are the locations for the coefficients $\hat{\beta}_\ell^{[i]}$ and $\hat{\beta}_\ell^{[j]}$ respectively. Note that $w_{ii}^{[\ell]} = 0$ for all $i = 1, \ldots, m$ as the observation cannot depend on itself. The choice of which metric and function we use to define out weight matrix $W_{m,\ell}$ is vital to being able to model the spatial data well. If the weight matrix does not give a good representation of the true nature of spatial region of interest, then the estimates calculated for the parameters are likely to be biased and inconsistent [2, 6]. Clearly, good selection of the weight matrix is essential for providing unbiased estimates of parameters and should be done in a way such that every entry is consistent to some rule.

Rearranging equation (4)

$$\begin{aligned}
\beta_\ell &= \alpha_\ell \mathbf{1}_m + \lambda_\ell W_{m,\ell}\beta_\ell + \varepsilon(\beta_\ell) \\
\beta_\ell - \lambda_\ell W_{m,\ell}\beta_\ell &= \alpha_\ell \mathbf{1}_m + \varepsilon(\beta_\ell) \\
(I_m - \lambda_\ell W_{m,\ell})\beta_\ell &= \alpha_\ell \mathbf{1}_m + \varepsilon(\beta_\ell) \\
\beta_\ell &= \alpha_\ell (I_m - \lambda_\ell W_{m,\ell})^{-1}\mathbf{1}_m + (I_m - \lambda_\ell W_{m,\ell})^{-1}\varepsilon(\beta_\ell)
\end{aligned}$$

where I_m is the $m \times m$ identity matrix. It can then be easily shown that the mean and the variance of β_ℓ is

$$\alpha_\ell (I_m - \lambda_\ell W_{m,\ell})^{-1}\mathbf{1}_m$$

and

$$(I_m - \lambda_\ell W_{m,\ell})^{-1}(I_m - \lambda_\ell W_{m,\ell}^T)^{-1}\sigma^2(\beta_\ell)$$

respectively, and due to linearity, we have

$$\beta_\ell \sim \mathcal{N}\left(\alpha_\ell (I_m - \lambda_\ell W_{m,\ell})^{-1}\mathbf{1}_m, (I_m - \lambda_\ell W_{m,\ell})^{-1}(I_m - \lambda_\ell W_{m,\ell}^T)^{-1}\sigma^2(\beta_\ell)\right)$$

for $\ell = 0, 1$. We can estimate λ_ℓ and α_ℓ by minimising the ordinary least squares equation in both λ_ℓ and α_ℓ. The ordinary least squares equation for β_ℓ is given by

$$\begin{aligned}
\left\{\hat{\lambda}_\ell, \hat{\alpha}_\ell\right\} &= \underset{\lambda_\ell, \alpha_\ell}{\arg\min} \, ||\beta_\ell - \alpha_\ell \mathbf{1}_m - \lambda W_{m,\ell}\beta_\ell||_2^2 \\
&= \underset{\lambda_\ell, \alpha_\ell}{\arg\min} \left(\beta_\ell - \alpha_\ell \mathbf{1}_m - \lambda_\ell W_{m,\ell}\beta_\ell\right)^T \left(\beta_\ell - \alpha_\ell \mathbf{1}_m - \lambda_\ell W_{m,\ell}\beta_\ell\right).
\end{aligned} \tag{5}$$

We then minimise equation (5) with respect to λ_ℓ and α_ℓ by taking the derivatives, setting to zero and then solving simultaneously. This yields

$$\lambda_\ell = \frac{\beta_\ell^T (W_{m,\ell} + W_{m,\ell}^T)\beta_\ell - 2\alpha_\ell \beta_\ell^T W_{m,\ell}^T \mathbf{1}_m}{2\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell}. \tag{6}$$

and

$$\alpha_\ell = \frac{1}{m}\mathbf{1}_m^T (I_m - \lambda_\ell W_{m,\ell})\beta_\ell \tag{7}$$

Then solving simultaneously gives estimators for λ_ℓ and α_ℓ as

$$\hat{\lambda}_\ell = \frac{\beta_\ell^T (W_{m,\ell} + W_{m,\ell}^T)\beta_\ell - \frac{2}{m}\beta_\ell^T W_{m,\ell}^T \mathbf{1}_m}{2\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell - \frac{2}{m}\beta_\ell^T W_{m,\ell}^T \mathbf{1}_m \mathbf{1}_m^T W_{m,\ell}\beta_\ell} \tag{8}$$

and

$$\hat{\alpha}_\ell = \frac{2\mathbf{1}_m^T \beta_\ell \beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell - \beta_\ell^T (W_{m,\ell} + W_{m,\ell}^T)\beta_\ell \mathbf{1}_m^T W_{m,\ell}\beta_\ell}{2m\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell - \beta_\ell^T W_{m,\ell}^T \mathbf{1}_m \mathbf{1}_m^T W_{m,\ell}\beta_\ell} \tag{9}$$

respectively. These estimators given in equations (8) and (9) have been shown through testing to be biased and inconsistent especially for λ when $||\beta||_2$ is large [10, 11], and is often accepted to be true [12], thus we should not use them. We can however use equations (6) and (7) in an iterative algorithm which updates at each step as following, where for this case $X = \mathbf{1}_m$,

Algorithm 1 Spatial Parameters Estimation (SPE) Algorithm

1: **procedure** SPE($\beta, W_{m,\ell}, X$)
2: Initialise $\lambda_\ell^{[0]} = 0$ and calculate for $i = 1, 2, \ldots$
3: **while** $|\lambda^{[i]} - \lambda^{[i-1]}| > \varepsilon$ **do**
4: $\alpha_\ell^{[i]} = \left(X^T X\right)^{-1} X^T \left(I - \lambda_\ell^{[i-1]} W_{m,\ell}\right)\beta_\ell$
5: $\lambda_\ell^{[i]} = \frac{\beta_\ell^T \left(W_{m,\ell}^T + W_{m,\ell}\right)\beta_\ell - \left(\beta_\ell^T W_{m,\ell}^T X \alpha_\ell^{[i]} + \alpha_\ell^{[i]^T} X^T W_{m,\ell}\beta_\ell\right)}{2\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell}$
6: $i = i + 1$
7: **return** $\lambda^{[i]}$ and $\beta^{[i]}$

Once we have created these models with estimates for λ_ℓ and α_ℓ for $\ell = 0, 1$, we can then use these models to predict what the coefficients would be for a given location in Australia by spatial interpolation, i.e. let $\mathbf{x}_h$ be a location where there is no rain gauge and let $\mathbf{w}_{h,\ell} = \left(w_{h,1}^{[\ell]}, \ldots, w_{h,m}^{[\ell]}\right)^T$ be the spatial weight vector with each entry being a function of the distance metric from $\mathbf{x}_h$ to all other known coefficient locations (entries corresponding to locations with unknown coefficients are set to zero). Thus our estimate for the coefficients at location $\mathbf{x}_h$ are given by

$$\tilde{\beta}_\ell^{[h]} = \hat{\alpha}_\ell + \hat{\lambda}_\ell \mathbf{w}_{h,\ell}\beta_\ell \tag{10}$$

for $\ell = 0, 1$. From this we can then predict what the corresponding gauge measurement of location $\mathbf{x}_h$ for the j^{th} time interval would be from the corresponding satellite estimate by

$$\tilde{Y}_{hj}^{[g]} = \tilde{\beta}_0^{[h]} + \tilde{\beta}_1^{[h]} Y_{hj}^{[s]} \tag{11}$$

3 Results

The bureau of meteorology between January 2003 and August 2018 has had over 3000 rain gauge stations taking monthly precipitation measurements be in operation, 3368 of which we are using for this study ($m = 3368$). Figure 2 shows us the locations of the 3368 monthly rain gauge stations as well as the correlation between the precipitation measured by these gauges and what was estimated by the satellite imagery. As we can see, majority of these locations have reasonably high correlation between the gauge measurement and the satellite estimate, with 1921 out of the 3368 stations having a correlation factor greater than 0.7. The correlation factor also appears to be following a spatial trend, indicating that there may be a spatial trend in the relationship between gauge measurements and Satellite estimates. While gauge

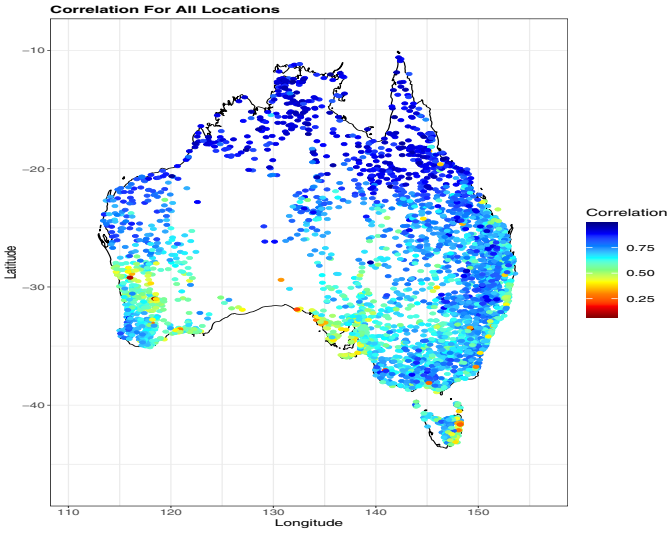


Fig. 2: Sample pearson correlation between the gauge measurement and the satellite estimate of the available 3368 Bureau of Meteorology rain gauge stations.

measurements and satellite estimates are highly positively correlated for the majority of locations we need to be able to justify that the relationship between the two is linear and also temporally stationary. Naturally we would assume that the

relationship between the gauge measurements and the satellite estimates is linear as in equation (1) as we would assume that the satellites give somewhat accurate estimates for the amount of precipitation, we have $\mathbb{E}\left[Y_{i,j}^{[g]}\right] = Y_{i,j}^{[s]}$. To justify this assumption, consider figure 3. We can see in the top left and top right plots the time

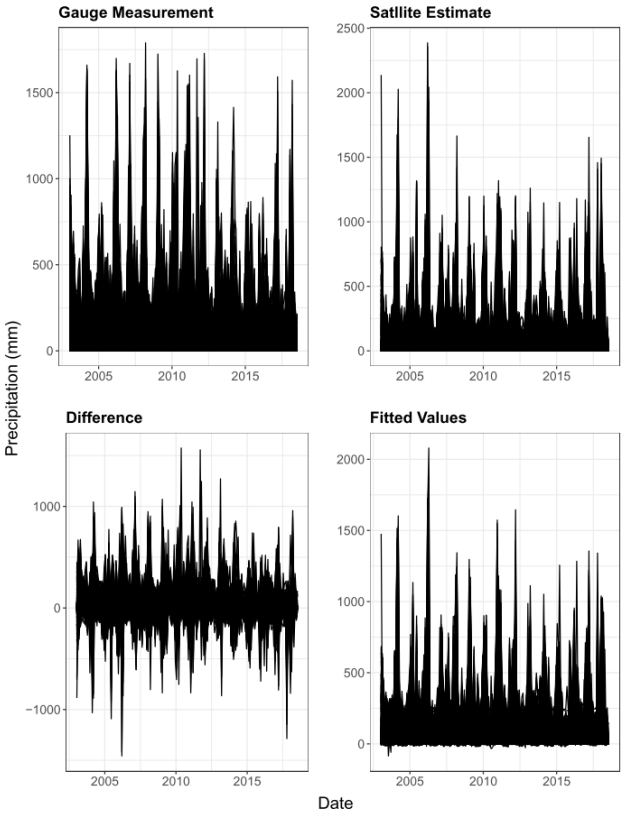


Fig. 3: Time series of all locations for gauge measurements (top left), satellite estimates (top right), difference of the gauge measurement and satellite estimate (bottom left) and linear regression fitted values (bottom right) in millimetres starting at January 2003 and ending at august 2018.

series for both the gauge measurements and satellite estimates respectively for all 3368 locations, where the satellite estimates are provided by the Japan Aerospace Exploration Agency (JAXA). In the bottom left plot of figure 3 we have the difference between the gauge measurements and the satellite estimates ($Y_{i,j}^{[g]} - Y_{i,j}^{[s]}$ for all $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$) and we can see that on average the satellite imagery tends to overestimate the amount of precipitation that a gauge has measured. We can also see that there are points in the time series that there are very large differ-

ences between the gauge measurement and the satellite estimate in both directions. In the bottom right plot we fit the simple linear model to every location as described in equation (1) and recreate the time series in the top left plot using the fitted values of these models. We can see that it tends to recreate the rain gauge time series fairly well which gives evidence to the assumption of a linear relationship. However, the points where the gauge measurement and satellite estimate are significantly different suggests that there a significant outliers in the data. To justify the temporal stationarity assumption we can consider the difference between the gauge and satellite measurements at each location ($Y_{i,j}^{[g]} - Y_{i,j}^{[s]}$) as a time series then we can perform the augmented Dickey-Fuller test. We are testing the null hypothesis that there is a unit root present in the time series against the alternative that the time series is stationary [4]. The result of this test gives no locations with a p-value above 0.05 meaning the time series at each location are stationary. Meaning that the mean, variance and autocorrelation of the time series does not change with time [5], and thus we can justify using one model for each location with one set of coefficients that do not change with time.

In figure 4, we look at the total amount of precipitation recorded at each rain gauge station for the entire time that it was operational and compare it to the total estimates for that location over the same time period. We can see that whilst

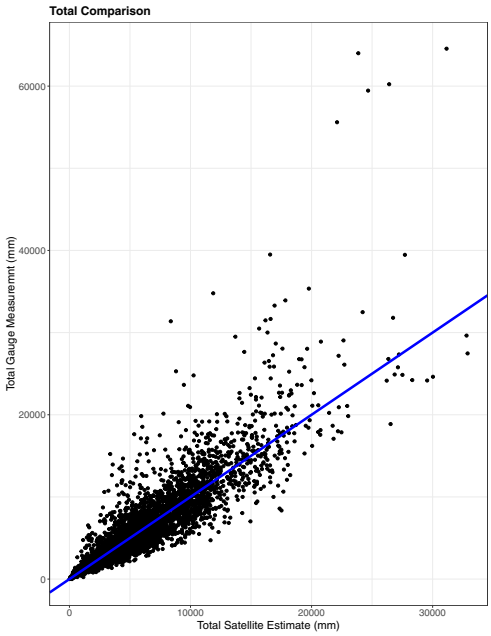


Fig. 4: Total gauge measurement of precipitation versus total satellite estimates of precipitation for each location over the same time interval compared to the line $y = x$.

the trend follows the red $y = x$ line quiet closely, giving more evidence to the assumption of a linear relationship, the variance in the difference between the total gauge measurements and the total satellite estimates increases as the measurements increases. If the satellite consistently only overestimates or only underestimates the gauge measurement in locations where there is a lot of precipitation, this problem will only be exacerbated when summing up the entire series. However, this issue could also be due to a single point in which the satellite estimate is significantly different to the gauge measurement as was seen in the bottom left plot of figure 3, which then skews the overall difference in the total comparison. In figure 5, we have the monthly gauge measurements plotted against their corresponding satellite estimates for rain gauge station 031030 (located on the eastern coast of northern Queensland, north of Cairns). The black $y = x$ line is what we would expect to see given that the satellite imagery gives an accurate estimate. We can see that this data has a significant outlier where the satellite estimates there to have been over 2200 millimetres of precipitation in a month compared to the gauge which only measured there to be just under 1000 millimetres of precipitation in that same month. Due to

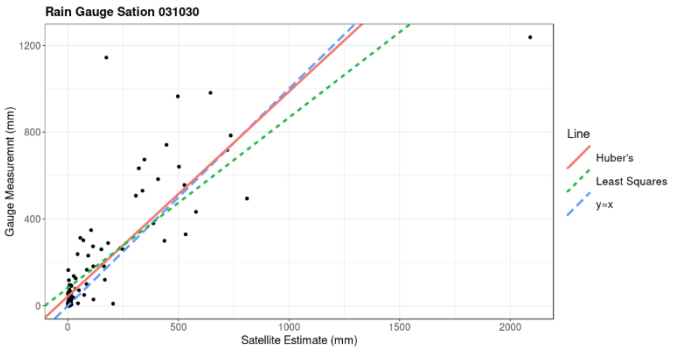


Fig. 5: Monthly gauge measurements of precipitation versus its corresponding monthly satellite estimates of precipitation for Station 031030, with the least squares regression line and the Huber's robust regression line.

the squared nature of the ordinary least squares in equation (2), these outliers can significantly influence the fit of the regression model (blue line) and thus causes the model to poorly capture the true trend of the relationship between the gauge measurement and the satellite estimate. We can reduce the influence of the significant outliers by using a robust regression method to model the relationship between the gauge measurements and the satellite estimates. Estimating parameters by robust regression requires us to minimise a different loss function as to what was shown in equation (2). We will be using Huber's loss to estimate $\beta^{[i]}$ in our robust regression model. Parameter estimation by Huber's loss is given by

$$\hat{\beta}_{\delta}^{[i]} = \underset{\beta_0^{[i]},\beta_1^{[i]}}{\arg\min} \begin{cases} \frac{1}{2}||\mathbf{Y}_i^{[g]} - \beta_0^{[i]}\mathbf{1}_{n_i} - \beta_1^{[i]}\mathbf{Y}_i^{[s]}||_2^2 & ,|Y_{ij}^{[g]} - \beta_0^{[i]} - \beta_1^{[i]}Y_{ij}^{[s]}| \leq \delta \\ ||\mathbf{Y}_i^{[g]} - \beta_0^{[i]}\mathbf{1}_{n_i} - \beta_1^{[i]}\mathbf{Y}_i^{[s]}||_1 - \frac{1}{2}\delta^2 & ,\text{otherwise} \end{cases} \tag{12}$$

where δ is a tuning parameter found by cross validation. In other words, we minimise the squared error for fitted values when they are within δ of the observed value and then minimise the absolute error for when fitted values are further than δ away from the observed value [8]. Figure 5 shows how the green robust regression line compares to the ordinary least squares regression line and we can see that by using Huber's loss, the significant outlier has much less influence on the fit of the regression line and which is a lot closer to the line $y = x$ than the ordinary least squares regression line.

In figure 6 we can see a clear trend in how the coefficients behave based on their location. As can be expected, higher values of the intercept coefficient, are

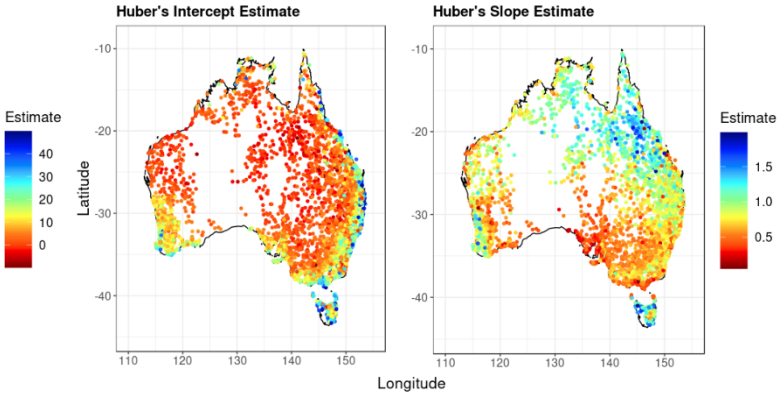


Fig. 6: Mappings of the Huber's regression coefficient estimates (intercept:left, slope:right) for the relationship between the gauge measurement and the satellite estimate for every location.

associated with lower values of the slope coefficient. There appears to be spatial trends in the relationship between the satellite estimate and the gauge measurement. Whilst a lot of coefficients are not giving values that we would expect with the intercept having a range of around -10 to 50 and the slope having a range of 0.05 to 1.8, this is due to the amount of noise that is present in data for some locations. While the noise is giving values for coefficients far off what we would expect to see for some locations, majority of locations have coefficients a lot closer to what we would expect to see. Figure 7 shows histograms of the coefficients estimated by Huber's loss in equation (12) we can see that the estimated values for the intercept are positively skewed with a mean of about 10 while the estimated values for the slope are quite symmetric with a mean of about 0.8.

To use a spatial linear model, we need to decide on what the weight matrices $W_{m,\ell}$ will be. As mentioned earlier, selection of a weight matrix is essential in creating

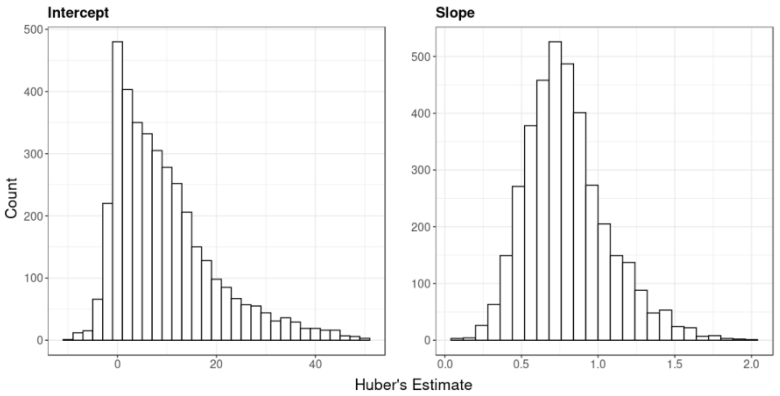


Fig. 7: Histograms of the estimated values of β_0 (left) and β_1 (right) by Huber's loss.

a good model. There are many different methods that could be used with different distance metrics. In this paper we use a mixture of k-nearest neighbours and inverse distance weighting (IDW) with an addition of the distance to the coast of the locations. That is,

$$w_{i,j} = \begin{cases} \frac{1/\left(d(\mathbf{x}_i,\mathbf{x}_j)^{\gamma}+dc(\mathbf{x}_i)+dc(\mathbf{x}_j)\right)}{\sum_{\mathbf{x}_t \in ne(\mathbf{x}_i)} 1/(d(\mathbf{x}_i,\mathbf{x}_t)^{\gamma}+dc(\mathbf{x}_i)+dc(\mathbf{x}_t))} & \text{, if } \mathbf{x}_j \in ne(\mathbf{x}_i) \\ 0 & \text{, otherwise} \end{cases} \tag{13}$$

where $ne(\mathbf{x}_i)$ is the neighbourhood of the location $\mathbf{x}_i$ which is determined by which other observed location are the k closest to it and $dc(\mathbf{x}_i)$ is the distance from the location $\mathbf{x}_i$ to the nearest coastal point. We use the distance to the coast as a factor in creating the weight matrices as we are working with rainfall data, the nature of the spatial relationships may change when a location is a further away from the coast, where there is significantly less rain. As the Surface of Australia lays on the surface of an approximate sphere, we should use a distance metric that considers the spherical nature of the domain. Thus we use a cosine distance metric, given by

$$d(\mathbf{x}_i,\mathbf{x}_j) = r\Delta\sigma$$

for

$$\Delta\sigma = \arccos\left(\sin x_{i,1}\sin x_{j,1} + \cos x_{i,1}\cos x_{j,1}\cos|x_{i,2}-x_{j,2}|\right)$$

with $r \approx 6371$ *km* (radius of the Earth). Whilst using the cosine distance metric doesn't significantly change the results compared to euclidean distance both in terms of the tuning parameters (k and γ) and also the end result, it is better to use a more accurate representation of the distance between locations.

Another well known neighbourhood defining method is the d-nearest neighbours method, where a neighbourhood for location $\mathbf{x}_i$ is defined to be the other locations

within a distance d of it. However, this method is not ideal for spatial processes where the observed locations are inconsistently placed as is gives a high amount of variance in the number of neighbours locations can have. For example, we can set our d to be a distance of 167 kilometres, which results in rain gauge station 13043 (South Karlamilyi National Park, Western Australia) having no neighbours, and rain gauge station 41103 (Toowoomba, Queensland) having 297 neighbours. The power coefficient γ is a non-negative value that represents the smoothness of the weight matrix. By cross-validation we can get the optimal values for creating the weight matrices for the intercept and slope as $(k = 7, \gamma = 0.92)$ and $(k = 8, \gamma = 0.79)$ respectively.

$$\begin{aligned} \hat{\beta}_0 &= 0.122 \times \mathbf{1}_m + 0.983 \times W_{m,0}\beta_0 \\ \hat{\beta}_1 &= 0.021 \times \mathbf{1}_m + 0.972 \times W_{m,1}\beta_1 \end{aligned} \tag{14}$$

respectively. The estimates for λ_0 and λ_1 as 0.983 and 0.972 respectively have associated likelihood ratio p-values that are significantly small ($< 10^{-100}$) which indicate that there is a high degree of spatial dependency. Figure 8 shows us how the

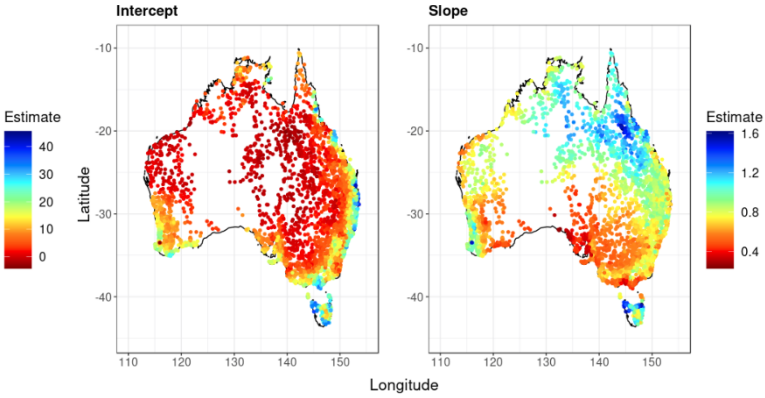


Fig. 8: Spatially modelled fitted values for the intercept (left) and slope (right) with estimation done by the SPE algorithm.

equations in (14) model the coefficients and as we can see, our models fit the data quite well as they appear to have captured the spatial trends present in the data. We can also see that locations with their coefficients significantly different to the surrounding locations coefficients' are not estimated as well by the model.

The fitted values that are not estimated as well as their surrounding coefficients may be able to be modelled better with the inclusion of a confounding factor. Thus far we have only looked at longitude and latitude contributing to how our coefficients are estimated by model (4), but it is possible that the elevation dimension is significant in modelling a coefficient's behaviour. Elevation could help us to explain a coefficient's behaviour as the satellite's image in which the precipitation is

estimated from is taken from above, and rain gauge stations that are at a higher elevation will be closer to the cloud and the satellite. Whilst the difference in elevation is marginal compared the altitude of the satellite, it is not marginal compared the altitude of the clouds. The elevation range of the rain gauge stations is 1,868 meters, with multiple at sea level and the highest located in Kosciuszko National Park, New South Wales. We can see in figure 9 an elevation mapping of Australia with locations of interest being labelled. Precipitation causing clouds can occur anywhere between ground level and 6000 metres [7], therefore the altitude of the rain gauge station may help us better model the coefficients' behaviour. This new model with

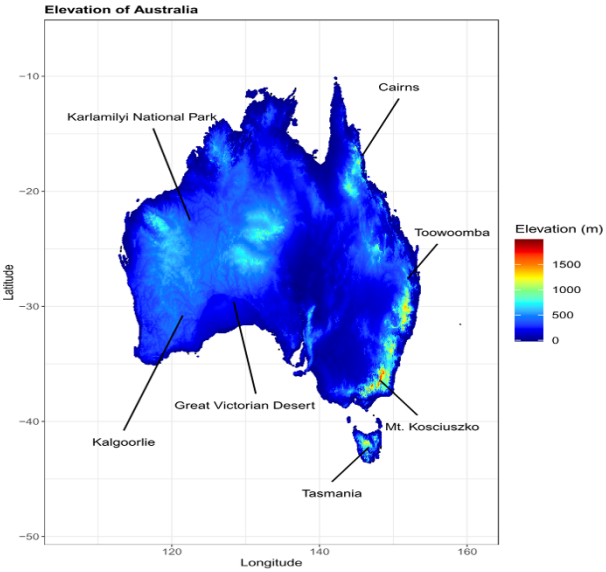


Fig. 9: Map of Australia showing the elevation (m).

the elevation confounding factor included can be expressed as

$$\beta_\ell = \alpha_\ell \mathbf{1}_m + \eta_\ell \mathbf{e} + \lambda_\ell W_{m,\ell} \beta_\ell + \varepsilon(\beta_\ell) \tag{15}$$

for $\ell = 0, 1$, with $\mathbf{e}$ being an $m \times 1$ vector where e_i is the elevation of the rain gauge station at location $\mathbf{x}_i$ for $i = 1, \ldots, m$. We can again use the SPE algorithm to find estimates for the unknown parameters with $X = (\mathbf{1}_m, \mathbf{e})$. We find here that the optimal values for defining the weight matrices for the intercept is the same, but the slope is now $(k = 8, \gamma = 0.8)$. For the spatial linear model of the intercept, the coefficient η_0 of the elevation confounding factor is estimated to be 0.0016 with an associated p-value of 5×10^{-5}, lowering the residual error from 5.429 to 5.417. For the slope coefficient spatial model, the elevation parameter is significant with the coefficient η_1 being estimated to be 6.3×10^{-5} with an associated p-value of 4.6×10^{-7}, lowering the residual error from 0.1502 to 0.1496. Therefore, the equations for the

intercept and slope models are now given by

$$\begin{aligned}\hat{\beta}_0 &= -0.310 \times \mathbf{1}_m + 0.0016 \times \mathbf{e} + 0.986 \times W_{m,0}\beta_0 \\ \hat{\beta}_1 &= 0.012 \times \mathbf{1}_m + 6.3 \times 10^{-5} \times \mathbf{e} + 0.964 \times W_{m,1}\beta_1\end{aligned}. \tag{16}$$

Now that we have estimates for the parameters α_ℓ, η_ℓ and λ_ℓ, we can recreate the

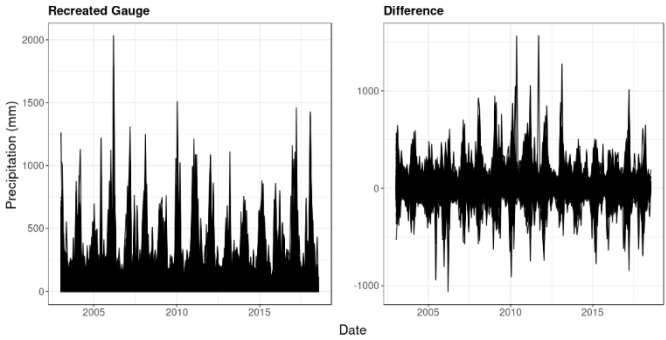


Fig. 10: Recreated time series using the fitted values from equations (16) for all observed locations (left) and the difference between the gauge measurements and the recreated gauge measurements (right).

time series of observations and compare this to the original gauge measurements time series as in figure 3. In the left plot of figure 10 we recreated the time series using the fitted values from the models in equation (16) for all observed locations over all time intervals. We can see that this method does yield a similar result to what the simple linear model used for the bottom right plot of figure 3 indicating that the spatial linear model does a good job of recreating the coefficients estimated by the linear model in equation (1). We can also see in the right plot of figure 10 that difference between the gauge measurements and the recreated gauge measurements is for the most part, around zero. There are many points where the gauge measurements and recreated gauge measurements are significantly different, however we must remember that we are using Huber's estimator as a loss function instead of the ordinary least squares and therefore less weight is given to 'outliers' so these points are not estimated as well by the model.

We can proceed to predict how the intercept and the slope would behave in regions where there are very few or even no rain gauges. We generate 50000 grid points over all of Australia. We then define the weight matrix $\mathbf{w}_{h,\ell}$ for each of the new points from their neighbourhoods as defined in equation (13) with $(k = 7, \gamma = 0.92)$ and $(k = 8, \gamma = 0.8)$ for the intercept and slope respectively. Recall that $\mathbf{w}_{h,\ell}$ will only depend on the existing rain gauge locations. We then use equation (10) with the addition of the elevation confounding factor

$$\tilde{\beta}_\ell^{[h]} = \hat{\alpha}_\ell + \hat{\eta}_\ell e_h + \hat{\lambda}_\ell \mathbf{w}_{h,\ell}\beta_\ell$$

to estimate the intercept and slope for location $\mathbf{x}_h$, where e_h is the elevation at $\mathbf{x}_h$ for all new locations ($h = m+1, \ldots, m+50000$). Note that due to the sparsity of existing rain gauges in some areas many of these new points will be defined to have the same neighbourhoods which results in their coefficients to be predicted as the same. Figure 11 shows how the models would predict these new points and as we can see, these added points follow the spatial dependency we had expected to see. Following from here, the gauge measurement time series at those new locations can be estimated. For example, consider map coordinates longitude 126.284 and latitude -29.068, located in Western Australia just east of the Great Victoria Desert Nature Reserve and about 500 kilometres from Kalgoorli at an elevation of 232 metres. The models in equation (16) give the estimated intercept and slope for this location as 3.42 and 0.46 respectively. The JAXA satellite estimate for the month of August 2018 is 10.3mm, substituting 10.3 into equation (11) gives 8.2mm as the gauge replicate for that location for month of August 2018. There are other loss functions

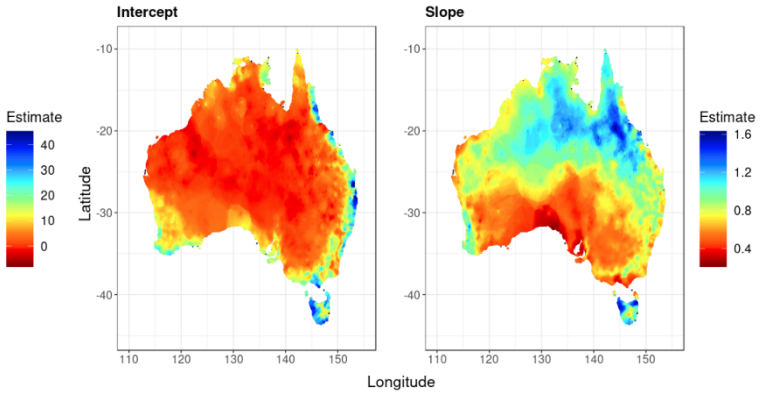


Fig. 11: Spatially Predicted value plots for the intercept (left) and slope (right) of the Huber's regression coefficients of the relationship of the gauge measurement against the satellite estimate of precipitation using the SPE algorithm.

that we could use besides the ordinary least squares to develop an algorithm for estimating the spatial parameters of a model as in equations (4) and (15). We could use a robust regression but as can be seen in by the histograms in figure 7 there are no significant outliers that will skew the estimation of the parameters by having too much leverage, thus there is no need to use robust regression in this case. However, at the moment we are estimating the spatial parameters of the intercept and slope separately when the intercept and slope are obviously dependent. We can also use the observed values from the gauge measurements and the satellite estimates in our new minimisation

$$\sum_{i=1}^{m} \left|\left| \mathbf{Y}_i^{[g]} - \mathbf{1}_{n_i} \left(\mathbf{x}_i^T \alpha_0 + \lambda_0 \mathbf{w}_{0i}^T \beta_0 \right) - \mathbf{Y}_i^{[s]} \left(\mathbf{x}_i^T \alpha_1 + \lambda_1 \mathbf{w}_{1i}^T \beta_1 \right) \right|\right|_2^2 \qquad (17)$$

where $\alpha_\ell = (\alpha_\ell, \eta_\ell)$. By taking the derivative of equation (17) with respect to the spatial parameters $\lambda_0, \lambda_1, \alpha_0$ and α_1, setting to zero and rearranging, we can obtain 4 equations for each of the parameters which all depend on the other parameters. We can then create another algorithm similar to the SPE algorithm that can give estimates for the spatial parameters. This algorithm gives parameter estimates as the following

$$\begin{array}{lll} \hat{\lambda}_0 = 1.17 & \hat{\alpha}_0 = 0.97 & \hat{\eta}_0 = -0.02 \\ \hat{\lambda}_1 = 0.9 & \hat{\alpha}_1 = 0.06 & \hat{\eta}_1 = 0.0003 \end{array}.$$

Note that in some literature there exists the restriction of $|\lambda_\ell| < 1$ [1]. We can see in figure 12 how this new loss function in equation (17), where $\mathbf{w}_{0i}$ and $\mathbf{w}_{1i}$ are defined the same as above, alters the prediction of the coefficients, especially in locations where the number of gauge locations is sparse. Interestingly, the range of the

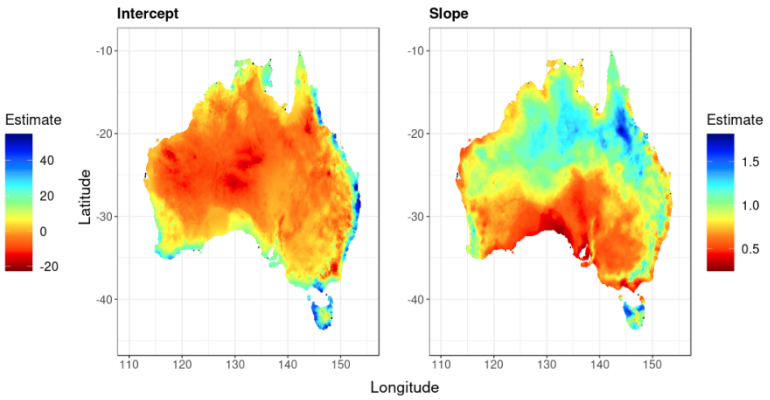


Fig. 12: Spatially Predicted value plots for the intercept (left) and slope (right) of the Huber's regression coefficients of the relationship of the gauge measurement against the satellite estimate of precipitation using the loss function in equation (17).

predicted slope values has slightly decreased where as the range for the predicted intercept values has slightly increased. Again we can give an estimate for what the gauge measurement would read in a location where there are no rain gauges, in the same location as used above, the new estimates for the intercept and slope are now given as -0.487 and 0.538 respectively which gives the recreated gauge measurement as 5.05, significantly lower than the previously estimated gauge measurement.

4 Conclusion

There is clear evidence that there is some difference between the gauge measurements and the satellite estimates. For the majority of locations, the satellite estimates on average that slightly more precipitation has occurred than what the gauge has measured for the same time interval. There is also clear evidence of spatial dependency for the relationship between the gauge measurement and satellite estimate.

The spatial linear model appears to be able to model the spatial dependencies of the robust regression coefficients well. A problem with attempting to predict how the intercept and slope coefficients would behave for locations where there are no rain gauges (as shown in figure 11), is that there is potential for areas to have very localised spatial behaviour. This localised spatial behaviour of the intercept and slope coefficients may not have been captured as the neighbours for these locations are far away and may be behaving differently based on their own localised spatial behaviour. We can see in the plots of figure 6, in locations such as Tasmania and north-east Australia that there appears to be very localised behaviour.

In comparing each of the spatial models used, they each have their own pros and cons. The first model in equation (4) while good as it easy to compute and easy to interpret, the model is probably overly simple as it does not use any confounding factors, the estimation is for the coefficients is done independently of each other and it does not depend on the gauge measurements or the satellite estimates. The second model in equation (15) is also easy to compute and it uses elevation which was found to be statistically significant for both the intercept and the slope. However, similarly to the first model, it is not good that this model also does estimation of the coefficients independently and does not depend of the gauge measurements of the satellite estimates. The last model using the loss function in equation (17) performs better than the other two models in fitting the known gauge measurements due to the parameters being estimated dependently and including the gauge measurements and satellite estimates. However, due to the parameters being estimated dependently there is a larger residual error of the coefficients and it more computationally complex than the other two models.

The amount of noise with the readings does present a problem in itself. The large amount of noise makes for the intercept's coefficient in some locations to be so large, that even if there was no precipitation estimated by the satellite images for the month, the model could give that a rain gauge would have recorded 50mm. One way to adjust for this would be to have each observation be a measured/estimated over a greater time interval such as 3 months or even a year instead of only a month. An issue with increasing the length of the time interval is that there are some rain gauge stations that were only in operation for a couple years or even less and increasing the length of the time interval would decrease what little amount of observations they had making for a less accurate estimation of the coefficients. Another approach to reduce the influence of the amount of noise on the fit of the model is to use an even more robust loss function to estimate the intercept and slope coefficients.

There is potential for more confounding factors to be included in the spatial linear models that may help explain the spatial relationship between the gauge mea-

surements and the satellite estimates. However, we must keep in mind that with spatial modelling that including too many explanatory variables can result in spatial over-fitting where the explanatory variables explain the spatial dependencies in the process, i.e. as the number of explanatory variables increases, $|\hat{\lambda}_\ell|$ decreases [14]. While in this paper we have modelled the relationship between the gauge measurements and the satellite estimates for each location by a strictly linear relationship, this may not be the case, the relationship may be more complex and to properly model this relationship we may need to consider transformations of the variables. We could also explore more complicated weight functions, that could include more spatially descriptive factors, or have the distance to the coast factor be included in a different way.

The loss function in equation (17), which has parameter estimation depending of the intercept and the slope depending on each other and also the gauge measurements and the satellite estimates, gives interesting results for the coefficient predicted as in figure 12. There are alterations that could be made to this loss function, such as the minimisations at each location are given weights. We could do this in many ways such as the weights being dependent on how many observations are at a given location, giving more weight to locations with more observations, or we could define weights depending on how isolated the observed location is.

References

1. Arbia, G.: A primer for spatial econometrics: with applications in R. Springer (2014)
2. Beenstock, M., Felsenstein, D., et al.: The Econometric Analysis of Non-Stationary Spatial Panel Data. Springer (2019)
3. Freedman, D.A.: Statistical models: theory and practice. cambridge university press (2009)
4. Fuller, W.A.: Introduction to statistical time series, vol. 428. John Wiley & Sons (2009)
5. Hamilton, J.D.: Time series analysis, vol. 2. Princeton university press Princeton, NJ (1994)
6. Herrera, M., Mur, J., Ruiz Marin, M.: Selecting the most adequate spatial weighting matrix: A study on criteria. Tech. rep., University Library of Munich, Germany (2012)
7. Houze Jr, R.A.: Cloud dynamics, vol. 104. Academic press (2014)
8. Huber, P.J., et al.: Robust estimation of a location parameter. The annals of mathematical statistics **35**(1), 73–101 (1964)
9. Kachi, M., Kubota, T., Ushio, T., Shige, S., Kida, S., Aonashi, K., Okamoto, K., Oki, R.: Development and utilization of"jaxa global rainfall watch"system based on combined microwave and infrared radiometers aboard satellites. IEEJ Transactions on Fundamentals and Materials **131**, 729–737 (2011)
10. Lee, L.F.: Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica **72**(6), 1899–1925 (2004)
11. Lee, L.f., Yu, J.: Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics **154**(2), 165–185 (2010)
12. Li, H., Calder, C.A., Cressie, N.: Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geographical Analysis **39**(4), 357–375 (2007)
13. Ord, K.: Estimation methods for models of spatial interaction. Journal of the American Statistical Association **70**(349), 120–126 (1975)
14. Schabenberger, O., Gotway, C.A.: Statistical methods for spatial data analysis. Chapman and Hall/CRC (2017)

Chapter 5

Mathematics of Physiological Rhythms

The new frontier of Network Physiology: Emerging physiologic states in health and disease from integrated organ network interactions

Plamen Ch. Ivanov, Jilin W.J.L. Wang, Xiyun Zhang, and Bolun Chen

Abstract An intriguing question in the new field of Network Physiology is how organ systems in the human body dynamically interact to coordinate functions, to maintain healthy homeostasis, and to generate distinct physiological states and behaviors at the organism level. Physiological systems exhibit complex dynamics, operate at different time scales and are regulated by multi-component mechanisms, which poses challenges to studying physiologic coupling and network interactions among systems with diverse dynamics. We present a conceptual framework and a method based on the concept of time delay stability to probe transient physiologic network interactions in a group of healthy subjects during sleep. We investigate the multi-layer network structure and dynamics of interactions among (i) physiologically relevant brain rhythms within and across cortical locations, (ii) brain rhythms and key peripheral organ systems, and (iii) the network structure and dynamics among peripheral organ systems across distinct physiological states. We demonstrate that each physiologic state (sleep stage) is characterized by a specific network structure and link strength distribution. The entire physiological network undergoes hierarchical reorganization across layers with the transition from one stage to another. Our findings are consistent across subjects and indicate a robust association of organ network structure and dynamics with physiologic state and function. The presented Network Physiology approach provides a new framework to explore physiologic states under health and disease through networks of organ interactions.

Plamen Ch. Ivanov · Jilin W.J.L. Wang · Bolun Chen
Keck Laboratory for Network Physiology, Department of Physics, Boston University, Boston 02215, USA
e-mail: plamen@buphy.bu.edu

Xiyun Zhang
Keck Laboratory for Network Physiology, Department of Physics, Boston University, Boston 02215, USA
Department of Physics, Jinan University, Guangzhou 510632, China

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_12

1 Introduction

The human organism consists of various physiological systems, each with its structural organization and exhibits complex dynamics with nonlinear and transient characteristics. States and functions at the organism level are traditionally defined by the dynamics of individual organ systems, and their modulation in response to internal and external perturbations. However, coordinated network interactions among organs are essential to generating distinct physiological states and maintaining health at the organism level. Manifested as synchronized bursting activities with certain time delays, these interactions occur through different coupling forms [1, 3], stochastic and nonlinear feedback across spatial-temporal scales and at multiple levels of integration to optimize and coordinate organ functions. Disrupting network communications can lead to dysfunction of individual systems or the collapse of the entire organism. Currently, there is no established theoretical framework, computational and analytic formalism to probe interactions between diverse systems in the human organism.

Here we present a new methodology adequate to identify and quantify the coupling of systems with different temporal characteristics and signal outputs. We apply Network Physiology approach [4, 14, 15] and the novel concept of time delay stability [2], and we demonstrate their utility to study transient synchronous bursts in systems dynamics as a fundamental form of physiologic network communications. We investigate new aspects of network interactions among brain rhythms across and within cortical locations, and their relation to neural plasticity in response to changes in autonomic regulation underlying different physiologic states. Further, we uncover dynamical features of brain-organ and organ-organ networks as a new signature of physiologic control and establish an association of network structure and dynamics with physiologic state and function. The presented methodology is an initial step in developing novel signal processing and computational tools and reported findings to establish building blocks of an atlas of dynamical interactions among key organ systems in the human body.

2 Method

2.1 Data

The data used in this work are multi-channel signals synchronously recorded from key physiological systems during night-time sleep with an average duration of 7.9h (EU SIESTA databases [18]). We analyze two subsets of the database: (i) 52 healthy subjects (26 female, 26 male, ages 20-34 years); (ii) 34 healthy subjects (17 female, 17 male, ages 20-40 years). All participants provided written informed consent. The research protocol was approved by the Institutional Review Boards of Boston Uni-

versity; data collection conducted according to the principles expressed in the Declaration of Helsinki; sleep stages scored in 30s epochs by certified technicians.

Standard polysomnogram recordings follow the American Academy of Sleep Medicine Manual [5]. Signals include EEG (channels Fp1, Fp2, C3, C4, O1 and O2), ECG, respiratory waves, EOG, EMG from chin and leg. From the raw signals we extract: spectral power in windows of 2s with 1s overlap for all physiologically relevant cortical rhythms (EEG frequency bands): δ (0.5-3.5Hz), θ (4-7.5Hz), α (8-11.5Hz), σ (12-15.5Hz), β (16-19.5Hz), γ_1 (20-33.5 Hz), γ_2 (34-99.5 Hz); variance of EOG and EMG in 2s windows with 1s overlap; heartbeat RR intervals and interbreath intervals are re-sampled to 1Hz after which values are inverted to obtain heart rate and respiratory rate. Thus, all time series have time resolution of 1s prior to analysis.

2.2 Time Delay Stability (TDS) Method

Physiological systems exhibit complex time-varying dynamics characterized by coherent bursts in activation across systems in response to modulation in physiologic state and condition (Fig. 1 Top left). We develop a new approach to (i) quantify pair-wise coupling and network interactions among diverse systems with bursting dynamics, and (ii) track the evolution of networks of organ interactions across states and conditions. We introduce a novel concept, Time Delay Stability (TDS), and a TDS method (Fig. 1) to study the time delay with which bursts of activity in a given system are consistently followed by corresponding bursts in the signal output of other systems. Within this framework, periods of TDS, i.e., constant time delay between bursts in the activation of two systems, indicate coupling.

To probe the interaction between two physiologic systems X and Y, we consider their output signals $\{x\}$ and $\{y\}$, each of length N. We divide signals $\{x\}$ and $\{y\}$ into N_L overlapping segments of equal length $L = 60$s. We chose an overlap of $L/2 = 30$s which corresponds to the time resolution of the conventional sleep-stage scoring epochs, and thus $N_L = [2N/L]$. Prior to analysis, each segment is normalized separately to zero mean and unit standard deviation to remove constant trends so that the estimated coupling between signals is not affected by relative amplitudes.

Next, we calculate the cross-correlation function,

$$C_{xy}^{\nu}(\tau) = \frac{1}{L}\sum_{i=1}^{L} x_{i+(\nu-1)\frac{L}{2}}^{\nu}\, y_{i+(\nu-1)\frac{L}{2}+\tau}^{\nu}, \tag{1}$$

within each segment $\nu = 1, \ldots, (N_L - 1)$ by applying periodic boundary conditions. For each segment ν we define the time delay τ_0^{ν} corresponding to the maximum in the absolute value of $C_{xy}^{\nu}(\tau)$ in this segment:

$$\tau_0^{\nu} = \tau|_{|C_{xy}^{\nu}(\tau)| \geq |C_{xy}^{\nu}(\tau')|} \quad \forall \tau'. \tag{2}$$

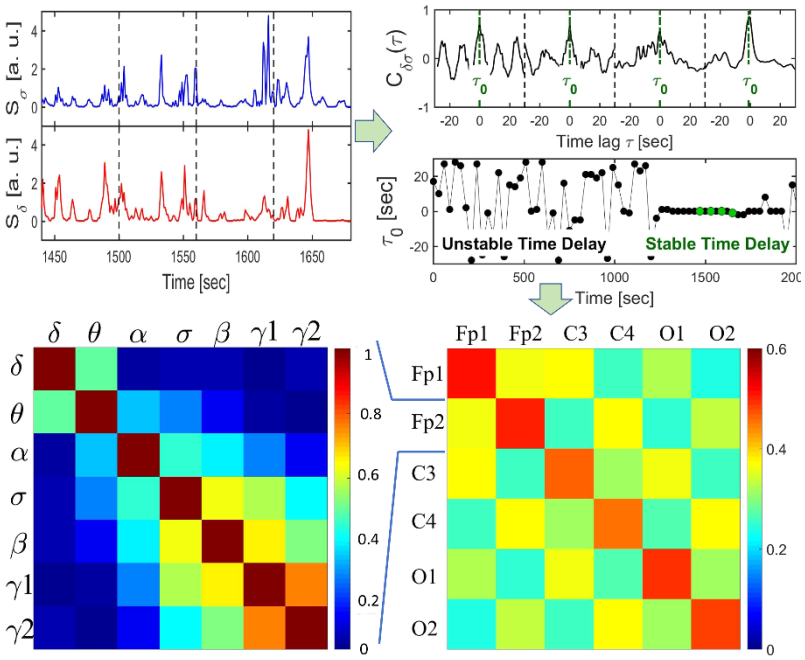


Fig. 1 Degree of coupling between brain rhythms quantified by Time Delay Stability (TDS). Schematic presentation of the TDS method. (Top left) Segments of time series representing EEG spectral power $S(\sigma)$ and $S(\delta)$ of the σ and δ cortical rhythms shown for consecutive 60s windows (vertical dashed lines). (Top right) Coordinated synchronous bursts in $S(\sigma)$ and $S(\delta)$ lead to pronounced cross-correlation $C_{\delta\sigma}$ with dominant peak within each time window located at time lag τ_0, representing the time delay between the two signals. Time delay τ_0 between $S(\sigma)$ and $S(\delta)$ is plotted for consecutive 60s windows with step of 30s (green dots mark τ_0 for the windows shown in the $C_{\delta\sigma}$ plot). Note the transition at $\sim$ 1200s from a segment with strongly fluctuating τ_0 to a stable time delay regime with $\tau_0 \approx$ constant. Such regime of time delay stability (TDS) indicates the onset of physiological coupling. The fraction of time (%) in the EEG recording when TDS is observed quantifies the degree of coupling strength (%TDS). (Bottom left) TDS matrix representing the degree of coupling between different brain rhythms (δ, θ, α, σ, β, γ_1, and γ_2) derived from two cortical locations (Fp1 and Fp2 EEG channels). Matrix elements represent the coupling strength, measured as %TDS, for each pair of brain rhythms. (Bottom right) TDS matrix representing the average coupling of all brain rhythms across each pair of EEG channels (Fp1, Fp2, C3, C4, O1, O2). Matrix elements show cortical rhythms interactions for one representative healthy young subject during Wake. Color code indicates the average coupling strength.

Time periods of stable interrelation between two signals are represented by segments of approximately constant τ_0 in the newly defined series of time delays, $\{\tau_0^\nu\}_{\nu=1,\ldots,N_L-1}$. In contrast, absence of stable coupling between the signals corresponds to large fluctuations in τ_0 (Fig. 1 Top right).

Third, we identify two systems as linked if their corresponding signals exhibit a time delay that does not change by more than $\pm$1s for several consecutive segments ν. We track the values of τ_0 along the series $\{\tau_0^\nu\}$: when for at least four out of

five consecutive segments ν (corresponding to a window of 5×30s) the time delay remains in the interval $[\tau_0 - 1, \tau_0 + 1]$ these segments are labeled as stable. This procedure is repeated for a sliding window with a step size one along the entire series $\{\tau_0^\nu\}$. The TDS value is finally calculated as the fraction (%TDS) of stable points in the time series $\{\tau_0^\nu\}$. Thus, longer periods of TDS between the output signals of two systems reflect more stable interaction and stronger coupling between these systems, and the links strength in physiologic networks is determined by the percentage of time when TDS is observed: higher %TDS means stronger links.

We have tested several different values for the window size L, i.e., $L = 30$, 60, 120, and 180s with non-overlapping windows as well as window overlaps $L/2$ and $L/4$. The overall TDS results were not significantly different for the different combinations of L and overlap, however, there was a tendency to noisier τ_0 vs t signals for shorter windows and less overlap (Fig. 1 Top left). On the other hand larger windows reduce the time resolution of the TDS.

The TDS method is general, and can be applied to diverse systems with bursting dynamics. It is more reliable in identifying physiological coupling compared with traditional cross-correlation, cross-coherence, and classical Granger causality approaches, which are not suitable for heterogeneous non-stationary signals with time varying coupling, and are affected by the degree of auto-correlations and irregular bursts embedded in these signals [4, 23]. Several relevant signal processing techniques have been developed for automated pattern discovery (e.g., dynamic time wrapping method used for machine learning and information retrieval), which may perform well when time-series are at a similar scale with low noise. They are not tailored for high-frequency bursting signals in multiple-channel polysomnogram recordings that exhibit transient dynamics and strong stochastic fluctuations.

2.3 Averaging procedure for assessing links strength in physiological networks

We utilize a specific procedure to quantify the group average strength of a particular network link for a given physiological state (sleep stage). A standard averaging procedure, where the strength of a network link during a given sleep stage is first calculated for one subject and is then averaged for all subjects, would give equal weight for all subjects in the group average. However, we note that the total duration of each sleep stage (sum of all episodes of a given stage) during night-time sleep varies from subject to subject. Thus, we perform a weighted averaging procedure where the contribution of each subject in the group average link strength for a sleep stage is weighted proportionally to the total duration of that sleep stage during the night.

Specifically, links in our network analysis are obtained by quantifying TDS for each pair of physiological systems after calculating the weighted average for all subjects during a given physiological state (sleep stage): $\%\text{TDS} = (\Sigma_i s_i / \Sigma_i S_i) \times 100$ where S_i indicates the total duration of a given sleep stage for subject i, and s_i is the total duration of TDS within S_i for the considered pair of physiological signals.

Artifacts related to specific behaviors of individual subjects (excessive movement, respiratory perturbations, etc.) or to the quality of recording of specific channels (due to loose lead contact) may lead to outliers in the estimate of some links strength in the network for a given subject. Further, links that are outliers in the physiological network of one subject may not be outliers in the network of another subject (same artifacts may not repeat for different subjects). To address this problem, for each pair of physiological signals (specific network link) we obtain the distribution and standard deviation of %TDS values (link strength) derived from all subjects in the group. Subjects for whom the considered network link has %TDS value above the group average + 2 are then removed, and a weighted average for the link is obtained based on the remaining subjects in the group, thus removing outliers in the calculation of the group-averaged link strength. This procedure is repeated for each link in the network. Considering all network links for all subjects in our database during a given physiologic state, this procedure led to $< 3\%$ of links removed as outliers in the calculation of the reported group average results for the different physiological networks.

To avoid unreal couplings due to small cross-correlation peaks, we only look at stable periods of TDS – only when four out of five consecutive segments with maximal correlation appear at approximately the same delay do we consider them to be stable. A network link between two systems is defined when their interaction is characterized by TDS value above a significance threshold determined by a surrogate analysis test. For each link in a given sleep stage, 200 surrogates are generated considering signals from two distinct and randomly chosen subjects, and a surrogate average link strength (%TDS) is obtained. The procedure is repeated for each network link to obtain a distribution of surrogate link strengths in each sleep stage. For each distribution, the mean μ_{surr} and standard deviation σ_{surr} are estimated. The significance threshold at 95% confidence level for network links strength is defined as $\mu_{\text{surr}} + 2\sigma_{\text{surr}}$ for each sleep stage. The %TDS thresholds in surrogate tests are around 2.5% which are much lower than those in real empirical data (around 50-60%), validating our method's effectiveness.

3 Results

We focus on physiological systems network dynamics during sleep because sleep stages are well-defined physiological states with specific neuroautonomic regulation, and external influences due to physical activity or sensory inputs are reduced. The structure of our database, comprising of multi-channel synchronously recorded signals from different organ systems, allows to investigate the dynamics of interactions among organ systems and their network organization during different physiological states (sleep stages). Using the TDS method, specially tailored to probe interactions among systems with diverse dynamics, we aim to quantify coupling between organ systems and their network characteristics. It is essential to understand how physiologic regulation underlying a given state influences the dynamics of or-

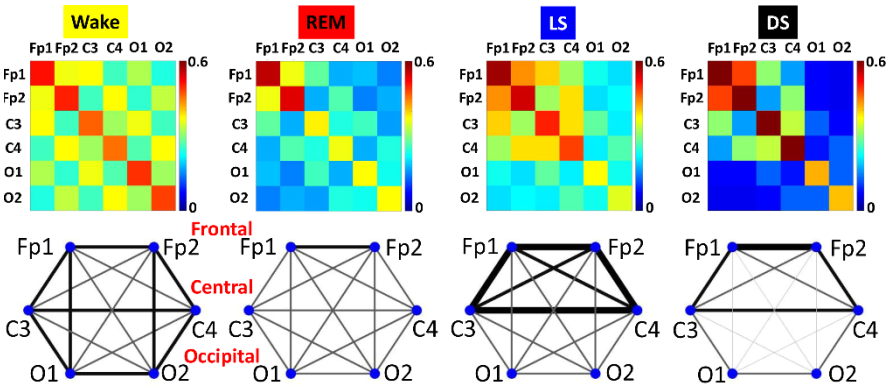


Fig. 2 Dynamic networks of brain rhythms interactions across cortical locations and transitions with physiological states. (Top) Time Delay Stability (TDS) matrices representing the average strength of coupling among all cortical rhythms (EEG frequency bands) across EEG channels, obtained from continuous overnight recordings for different sleep stages and averaged for a group of healthy subjects. Color code for matrix elements marks the coupling strength (%TDS). Transitions across sleep stages lead to changes in the average coupling strength of brain rhythm interactions across cortical locations and associated reorganization in TDS matrix structure characterized by stronger interactions during Wake and LS (warmer colors) compared to REM and DS (colder colors). (Bottom) Network representation of the group average TDS matrices for different sleep stages. Network nodes indicate cortical areas: Frontal (Fp1 and Fp2), Central (C3 and C4) and Occipital (O1 and O2). Each network link represents the coupling strength averaged over all pairs of rhythms from two different cortical areas, where wider and darker links indicate stronger coupling. Links are separated in four modules (with %TDS $<$12%; 12%-30%; 30%-38%; $>$38%). Dramatic reorganization in network structure is observed with transition from one sleep stage to another, with more homogeneous links (coupling strength) distribution during Wake and REM, heterogeneous and modularized links during LS and DS. Reorganization in network links heterogeneity is paralleled by a pronounced sleep-stage stratification pattern – average network links strength is significantly different comparing all four sleep stages (one-way ANOVA rank test $p \leq 0.001$), and pairwise comparisons of Wake vs REM and LS vs DS both show significant difference (Mann-Whitney test $p \leq 0.001$).

gan network communications, and how integration of organ systems as a network leads to emergent behaviors and physiological functions at the organism level [16].

3.1 Networks of brain rhythms interactions across cortical locations

We first investigate the network of interactions among different brain rhythms. Sleep stages are traditionally defined by the presence of dominant brain rhythms in cortical EEG dynamics. However, little is known whether and how brain rhythms across cortical locations interact as a network to generate sleep stages [22]. We consider seven distinct cortical rhythms from six cortical areas (EEG channels) that are traditionally used in sleep-stage scoring. Our TDS analysis shows pronounced coupling

for all pairs of rhythms, well above the significance threshold at 2.5%TDS, indicating physiologically relevant network interactions. Further, we find that the complex network of brain rhythms interactions across locations changes with transition from one sleep stage to another. A clear sleep-stage stratification is observed when we coarse-grain the network by averaging the coupling strength over all pairs of rhythms for each two cortical areas – globally the network is characterized by much stronger coupling among brain rhythms during Wake and LS compared to REM and DS, as demonstrated by the coarse-grained TDS matrix in Fig. 2. Moreover, there is a pronounced reorganization in network topology with transition across physiologic states, where each sleep state is characterized by specific modules of cortical locations with strong or weak interactions (Fig. 2).

3.2 Network interactions among brain rhythms within cortical areas

We next investigate the network of brain rhythms interactions within each of the six cortical locations separately. We find that higher frequency brain rhythms exhibit stronger coupling (i.e. more synchronous bursting activity) – a behaviour which is consistently observed for all six cortical locations and sleep stages, as shown by the TDS matrices in Fig.3. With transition from one sleep stage to another, there is a significant reorganization in both links strength and topology for all local networks of brain rhythms interactions: while Wake is characterized by similar network link strength and topology for all six cortical areas, local networks of brain rhythms interactions during REM, LS and DS exhibit different structure with higher connectivity and link strength in the Frontal areas compared to the Central and Occipital areas.

The existing literature focuses on how a given rhythm (such as δ) interacts with itself across different brain locations. Since neuronal populations in six cortical layers generate different brain rhythms that project onto the scalp, cross-frequency coupling naturally occurs at the same location, manifesting neuronal populations' synchronous activities and quantifying inter-layer coordination between cortical neurons. Moreover, in the next section, we will discuss specific functional forms of couplings among cortical rhythms in the same brain area [21].

3.3 Coexisting networks of brain rhythm interactions represent different types of physiologic coupling

In contrast to Fig. 3 where network links represent %TDS (Method), network links in Fig. 4 [21] represent the degree of synchronous or asynchronous modulation in the spectral power amplitude of different (dominant and non-dominant) brain rhythms. To probe the collective behavior of brain rhythms in relation to physiologic states, we construct networks of positive and anti-correlated interactions from equal-

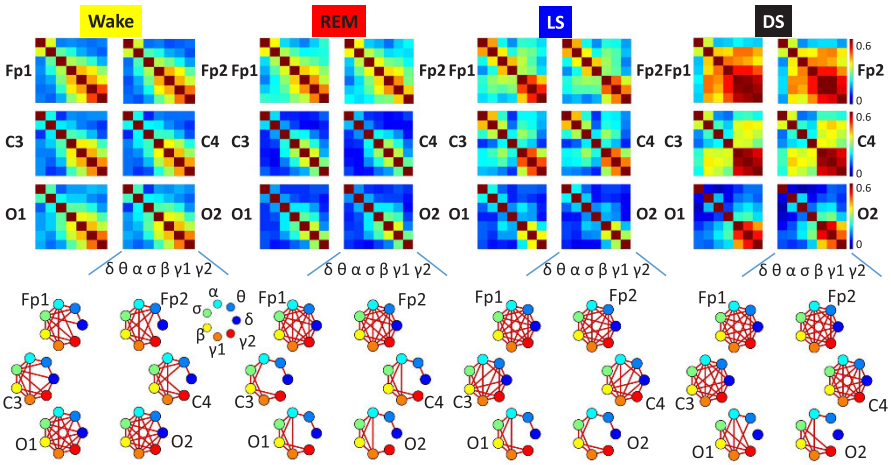


Fig. 3 Plasticity in network interactions among brain rhythms at specific cortical locations as function of physiologic state. (Top) Time Delay Stability (TDS) matrices quantify interactions for specific pairs of cortical rhythms (EEG frequency bands δ, θ, α, σ, β, γ_1, and γ_2) within six cortical areas (EEG channels: Frontal Fp1 and Fp2, Central C3 and C4, Occipital O1 and O2). Color code of matrix elements marks the coupling strength for each pair of brain rhythms averaged over a group of healthy subjects. Changes in TDS matrix structure for different brain areas and sleep stages indicates plasticity of brain rhythms interactions as function of physiologic state. (Bottom) Network presentation of the TDS matrices at six cortical locations for different sleep stages. Network nodes in color mark cortical rhythms derived from a given EEG channel representing a cortical location. Network links (in red) represent the strength of interaction for each pair of brain rhythms at a given EEG channel location (only links with %TDS$\geq$ 25% are shown). Network connectivity significantly changes at cortical locations during a given sleep stage, as well as with transition across stages (one-way ANOVA tests $p \leq 0.001$), indicating a complex reorganization and plasticity in brain rhythm interactions necessary to facilitate physiologic functions associated with distinct physiologic states.

time cross-correlation among brain rhythms, and we track their evolution across sleep stages. This network approach helps to visualize and dissect brain wave interactions where positive- and anti-correlated behaviors coexist [21]. It also provides a first demonstration of how brain rhythms coordinate collectively as a network to generate distinct physiologic states.

During DS, we observe a pronounced network cluster of anti-correlated interactions between the δ wave and all other brain waves (Fig. 4). We also identify a coexisting complementary network during DS comprised of only positively-correlated interactions between all brain waves except δ (Fig. 4). With the transition from DS to LS, REM, and wake, the links strength in the anti-correlated cluster between δ and all other brain waves decreases. In contrast, new positively-correlated links emerge, indicating a complex reorganization among brain rhythms across physiologic states. We note that links in the positively-correlated networks represent parallel coordination of brain wave activation, whereas links in the anti-correlated networks correspond to brain wave interactions of reciprocal and complementary

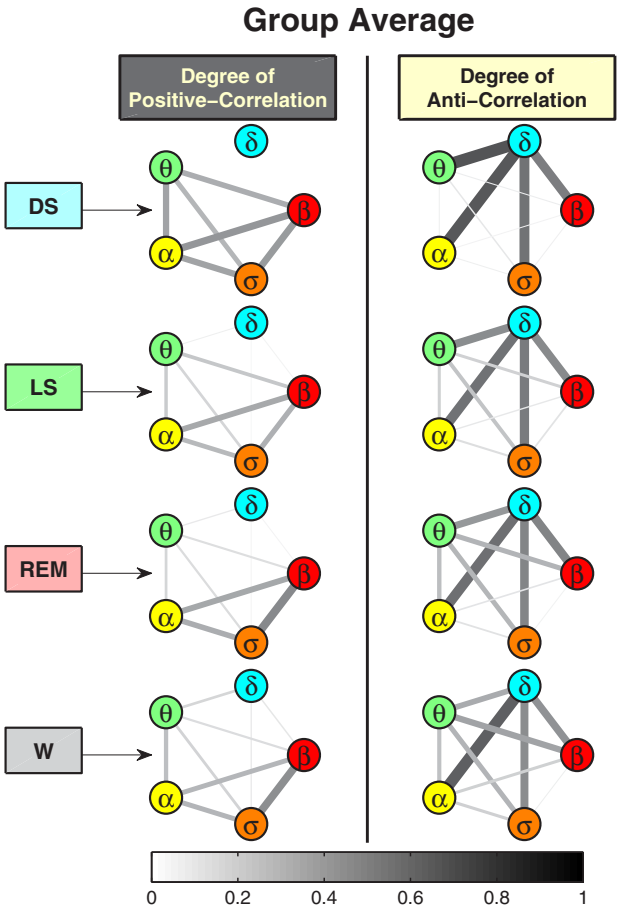


Fig. 4 Coexisting networks of brain rhythm interactions at channel C3 represent different types of physiologic coupling and exhibit distinct topology across sleep stages. Network nodes represent brain waves (EEG frequency bands) and network links indicate the degree of equal-time cross-correlations for each pair of brain waves (line thickness and darkness correspond linearly to link strength). Two types of networks are shown: left column, where links strength reflects the fraction of time when significant positive correlation (with $C > 0.5$) is found between a given pair of brain waves; right column, where links strength corresponds to the fraction of time when significant anti-correlation (with $C < -0.5$) is observed. These networks of interactions evolve across sleep stages – the links strength in the anti-correlated cluster between δ and all other brain waves decrease, while new positively-correlated links emerge. Remarkably, the coexistence of both positively- and anti-correlated networks of brain waves interactions within each physiologic state indicates a transient on/off nature of brain-wave communications, where links of different nature can emerge during different periods of time within the same physiologic state. The specific topology and clustering of brain wave networks during different sleep stages demonstrate a direct association between brain wave communications and physiologic state and function.

nature (opposite direction of modulation). Specifically, the δ-α interaction is always characterized by strong anti-correlation during all sleep stages, and there is

no δ-α link in the positively-correlated networks (Fig. 4). This observation is consistent with the traditional understanding of δ and α waves as the predominant brain rhythms for two opposite physiologic states, i.e., sleep vs. wake. However, the classical description of these physiologic states does not address the nature of δ-α interaction. Our analyses reveal the complex dynamics of reciprocal and competing nature in the coupling between δ and α waves, which transcends all physiologic states. In contrast to δ-α interactions, links associated with the θ wave that show positive correlations during DS become increasingly anti-correlated during LS and REM, indicating a very different role of θ-wave interactions compared to α- and δ-wave. Note that network links represent the fraction of time when a specific type of cross-correlation (positive or negative) is observed. Thus, the coexistence of both positively- and anti-correlated networks of brain-wave interactions within each physiologic state indicates a transient on/off nature of brain rhythms communications, where links of different nature can emerge during different times in the same physiologic state.

The traditional paradigm in brain research focuses on exploring the temporal dynamics and role of individual brain rhythms, and their association with specific physiologic states and functions [10, 26]. It is motivated by observations of quasi-steady-state behavior of brain rhythms at large time scales within a given physiologic state (e.g., sleep or wake, sleep stages) [6, 28], and changes in the amplitude (i.e., spectral power) of individual brain rhythms, their synchrony and coherence across cortical areas with the transition from one physiologic state to another [7, 11, 12, 31]. Our study aims to address the question of how dominant and non-dominant brain rhythms dynamically interact. We demonstrate that synchronous short-term modulations in the amplitude of brain rhythms that occur on top of their quasi-steady-state behavior at large time scales carry key information about the coupling among brain rhythms that are essential characteristics of a physiologic state. The presented here approach can detect higher-order interactions among both dominant and non-dominant brain rhythms embedded in their fine temporal structure at small time scales. It can quantify the change in brain rhythms network communications with transition across distinct sleep stages (Fig. 4). The uncovered coupling forms and network coordination among brain rhythms provide new insights into intrinsic physiologic interactions.

3.4 Dynamics of brain-organs interactions

Brain dynamics play an important role in the neuroautonomic regulation of organ systems. However, it remains unknown how brain rhythms simultaneously coordinate the function of different organs. We analyze the coupling of all seven brain rhythms from all six cortical locations with five key organ systems: heart, lungs, chin, eye and leg.

There are several key questions related to the nature of brain-organ interactions: (i) how different brain areas (EEG-channel locations) are involved in the commu-

nications and control of each organ system, (ii) which brain-wave frequency bands mediate the brain-organ communications, and (iii) how the networks of brain-organ interactions evolve with transitions across physiologic states. To this end, we apply the TDS method to identify and quantify dynamical links in the networks, which would serve as unique physiological maps of brain-organ interactions.

3.4.1 Brain-heart interactions

To demonstrate the rich dynamical features in brain-organ communications, let us first examine the network of brain-heart interactions. As indicated by the radar chart inside the heart hexagon in Fig. 5 [2], the network exhibits a relatively symmetric distribution of the average links strength for different brain areas, with a slight prevalence in strength for the links between the heart and the Central brain areas (C3 and C4). The spatial symmetry in the average brain-heart link strength holds for all sleep stages. Systematically investigating the links strength in the brain-heart network for all seven frequency bands and different sleep stages, we find that the average link strength for the entire network of brain-heart interactions is highest during W and LS, lower during REM and lowest during DS (Fig. 5). Further, this sleep-stage stratification pattern is consistently observed for all three sub-networks representing the Frontal-heart, Central-heart and Occipital-heart links across all frequency bands. Thus, our analysis shows that the strength of all links in the brain-heart network, regardless of brain areas or frequency bands, is modulated in the same way with transitions across sleep stages.

3.4.2 Sub-networks of brain-organ interactions

We find that sub-networks of brain rhythms interacting with distinct organ systems exhibit different average links strength, indicating a more synchronous activity and stronger coupling of brain rhythms with the dynamics of some organ systems compared to others, as shown by different size of network nodes in Fig. 6.

Further, we find that while all brain rhythms play certain role in the network of brain-organs interactions, a particular rhythm serves as the main mediator of network communications for a given organ system. Thus, a very structured dynamic network of brain-organs interaction emerges, where different brain rhythms are involved as main mediators of the function of different organ systems during a given physiological state (marked by different node circumference color in Fig.6). With transition from one sleep stage to another, a different brain rhythm may take the role as the main mediator in network interaction with a given organ system – e.g., brain-heart network interactions are mediated by γ_2 rhythms during Wake, γ_1 and β rhythms during REM and LS, and δ rhythms during DS, reflecting previously unrecognized aspects in the autonomic regulation of organ systems (Fig. 6) [2, 20].

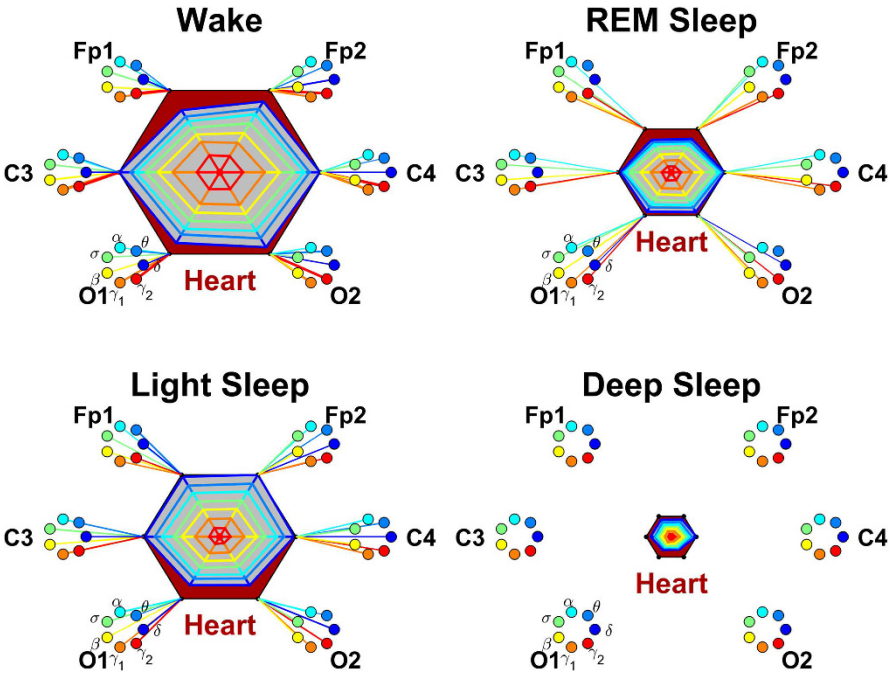


Fig. 5 Networks of brain-heart interactions during different physiologic states. Brain areas are represented by Frontal (Fp1 and Fp2), Central (C3 and C4) and Occipital (O1 and O2) EEG channels. Network nodes with different colors represent seven frequency bands ($\delta, \theta, \alpha, \delta, \beta, \gamma_1, \gamma_2$) in the spectral power of each EEG channel. Network links between the heart (red hexagon) and EEG frequency nodes at different locations are determined based on the TDS measure, and links strength is illustrated by the line thickness. Shown are links with strength 5%TDS. Radar-charts centered in each hexagon represent the relative contribution of brain control from different brain areas to the strength of network links during different sleep stages. The length of each segment along each radius in the radar-charts represents TDS coupling strength between the heart and each frequency band at each EEG channel location. These segments are shown in the same color as the corresponding EEG frequency nodes. During W and REM, the brain-heart network interactions are mediated mainly through high-frequency γ_1 and γ_2 bands (orange and red links), while during LS and DS, the interactions are mediated uniformly through all frequency bands. The brain-heart network is characterized by relatively symmetric links strength to all six brain areas, as shown by the symmetric radar-charts in each hexagon. A pronounced stratification pattern is observed for the overall strength of network links—stronger links during W and LS (larger hexagons) and weaker links during REM and DS (smaller hexagons). Notably, there are no links in the brain-heart network during DS (all links $< 5\%$TDS).

3.4.3 Networks of organ interactions

Finally, we apply our TDS analysis to probe interactions among organ systems. We find that pairs of organ systems are characterized by different coupling and correspondingly by different group average network links strength. As in the cases of brain-brain and brain-organs interactions, our analyses show that each sleep stage is characterized by a specific network topology of organ interactions (Fig. 6). The re-

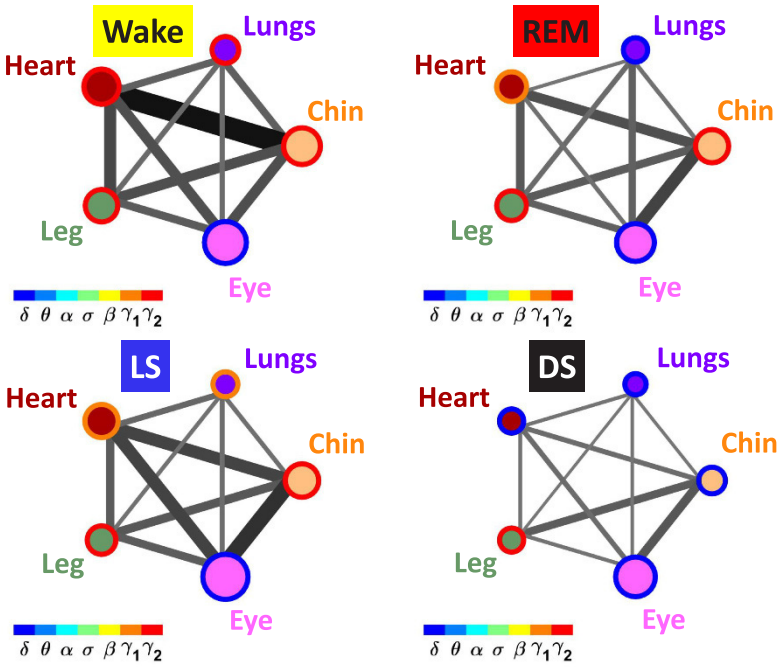


Fig. 6 Dynamic networks of organ interactions across sleep stages. Interactions among organ systems are represented by weighted undirected graphs, where network links between organ systems reflect the strength of dynamic coupling measured as %TDS and averaged for a group of healthy subjects. Darker and thicker links correspond to stronger interaction with higher %TDS. Network nodes represent key organ systems. The size of each organ node is proportional to the strength of the overall interaction of the organ with all brain rhythms at six cortical EEG channel locations (see Fig. 3). Color bars represent physiologically relevant cortical rhythms (EEG frequency bands). The circumference color of each organ node corresponds to the cortical rhythm exhibiting dominant coupling with the organ system when averaged over all cortical locations. Significant reorganization in network topology (links strength) for different sleep stages (all stages comparison one-way ANOVA rank test $p \leq 0.001$, and pairwise comparisons of Wake vs REM and LS vs DS with Mann-Whitney test $p \leq 0.003$) indicates an association between organs network interactions and physiologic function.

sults for the group average network characteristics (topology and link strength) are consistent with results obtained for individual subjects in our database, indicating a robust association of organ network interactions with physiologic state and function.

3.5 Network integration of interactions between the brain and peripheral organ systems

After separately investigating the networks of interactions between the brain and different organ systems, we integrate all brain-organ interactions into a single net-

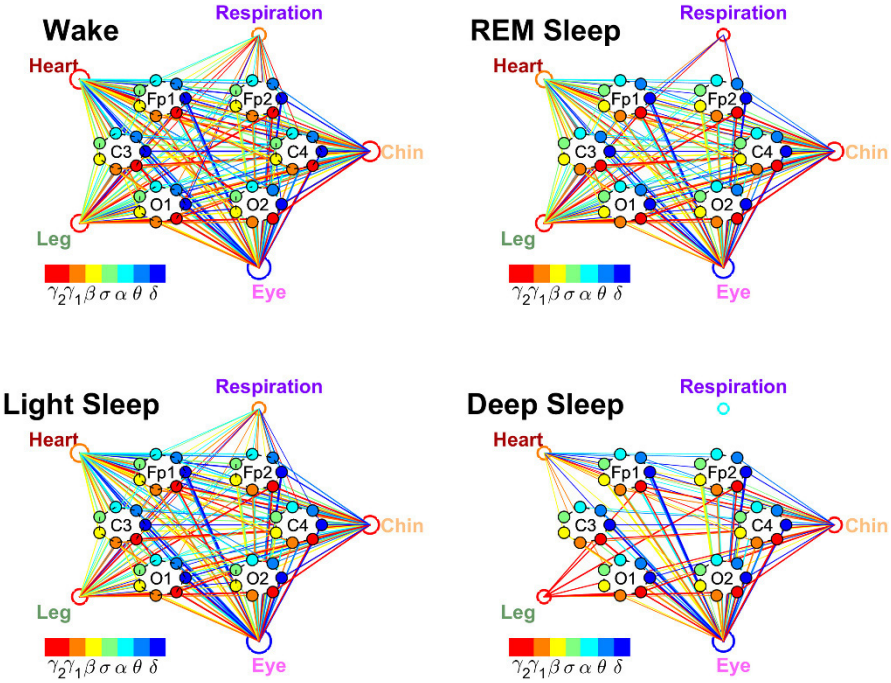


Fig. 7 Networks of physiologic interactions between brain areas and key organ systems during different physiologic states. Brain areas are represented by Frontal (Fp1 and Fp2), Central (C3 and C4) and Occipital (O1 and O2) EEG channels. Interactions between brain channels and organ systems are represented by weighted undirected graphs. The size of each organ node in the network is proportional to the strength of the overall brain-organ interaction as measured by the summation of the TDS links strength for all frequency bands and EEG channel locations. The color of each organ node corresponds to the dominant frequency band in the coupling of the organ system with the brain. The width of each link reflects the strength of dynamic coupling as measured by %TDS, and colors of the links correspond to the colors of the nodes representing the different frequency bands (color bars). Plotted are only links with strength 3%TDS. Thicker links correspond to stronger coupling and higher time delay stability. The physiological network exhibits transitions across sleep stages – lowest number of links during DS, higher during REM, and highest during LS and W. For different organs, brain-organ interactions are mediated through different dominant frequency bands, e.g., the chin and the leg are predominantly coupled to the brain through the high frequency γ_2 band during all sleep stages whereas brain-eye network interactions are mediated mainly through low-frequency δ band. The complex networks of dynamic interactions between key organ systems and the brain undergoes a hierarchical reorganization across different sleep stages, indicating a previously unknown mechanism of regulation.

work. It allows us to simultaneously compare several essential characteristics of the global network. Specifically, we track the number of links, their strength, the brain areas and frequency bands involved in the interactions between the brain and the group of organ systems and how this global brain-organs network evolves across physiologic states (Fig. 7) [2].

This integrative approach makes it possible to compare the predominant frequency band through which the interaction between the brain and different organs is mediated for several organ systems simultaneously during a given physiologic state. We find that the heart, leg and chin always interact with the brain mainly through the high-frequency γ_1 and γ_2 bands (red-colored links in Fig. 7, whereas the brain-eye interactions are mediated through lowest-frequency δ band (blue colored links in Fig. 7). There is no single dominating frequency for the brain-respiration interaction. The interaction between brain and the respiratory system is always weaker than other brain-organ interactions, indicating a relatively weak physiologic coupling between brain and respiration compared to other organs at the time scales (> 2.5 min) over which the TDS analysis is performed. Further, with transitions across sleep stages, we observe a complex hierarchical reorganization in both the number and the strength of links in the integrated brain-organs network – lowest number of links during DS (sparse network), higher during REM, and highest number of links involving most of the frequency bands during LS and W. Remarkably, this structural reorganization of the integrated brain-organs network is consistent with the sleep-stage stratification patterns observed for each organ system, indicating a previously unknown rule for neural regulation of organ systems.

4 Summary

We show that the concept of time delay stability and the TDS method we developed can be successfully employed to quantify the coupling and network interactions of systems with complex time-varying and diverse dynamics. Utilizing continuous recording during sleep from healthy young subjects, we demonstrate that each sleep stage is uniquely characterized by a network of physiologic interactions across scales in the human organism – from coupling among brain rhythms within and across cortical locations to networks of organ interactions. We find that with the transition from one state to another, physiologic network structure undergoes a consistent reorganization that occurs across scales. The introduced here method to infer interactions among diverse dynamic systems and the reported empirical findings provide new insights into the mechanisms of autonomic regulation underlying physiologic states, and represent first building blocks in the emerging field of Network Physiology [4, 14], where recent studies have uncovered robust associations between physiological states and networks of physiologic interactions [8, 13, 17, 27, 32] and under various clinical conditions [9, 19, 24, 25, 30]. Thus, the proposed here Network Physiology approach [2, 15, 16, 29], reveals fundamental new laws of physiologic regulation and can enhance our understanding of how behaviors and functions emerge at the organism level out of integrated network interaction among diverse systems.

Acknowledgements This work was supported by the W. M. Keck Foundation, the National Institutes of Health (Grant 1R01-HL098437), the US-Israel Binational Science Foundation (Grant 2012219), the Office of Naval Research (Grant 000141010078).

References

1. Bartsch, R.P., Ivanov, P.C.: Coexisting forms of coupling and phase-transitions in physiological networks. In: International Conference on Nonlinear Dynamics of Electronic Systems, pp. 270–287. Springer (2014)
2. Bartsch, R.P., Liu, K.K., Bashan, A., Ivanov, P.C.: Network physiology: how organ systems dynamically interact. PloS one **10**(11), e0142,143 (2015)
3. Bartsch, R.P., Liu, K.K., Ma, Q.D., Ivanov, P.C.: Three independent forms of cardio-respiratory coupling: transitions across sleep stages. In: Computing in Cardiology 2014, pp. 781–784. IEEE (2014)
4. Bashan, A., Bartsch, R.P., Kantelhardt, J.W., Havlin, S., Ivanov, P.C.: Network physiology reveals relations between network topology and physiological function. Nature communications **3**(1), 1–9 (2012)
5. Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C., Vaughn, B.V., et al.: The aasm manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine **176**, 2012 (2012)
6. Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C., Vaughn, B.V., et al.: The aasm manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine **176**, 2012 (2012)
7. Bian, Z., Li, Q., Wang, L., Lu, C., Yin, S., Li, X.: Relative power and coherence of eeg series are related to amnestic mild cognitive impairment in diabetes. Frontiers in aging neuroscience **6**, 11 (2014)
8. Bogdan, P.: Taming the unknown unknowns in complex systems: challenges and opportunities for modeling, analysis and control of complex (biological) collectives. Frontiers in Physiology **10** (2019)
9. Bolton, T.A., Wotruba, D., Buechler, R., Theodoridou, A., Michels, L., Kollias, S., Rössler, W., Heekeren, K., Van De Ville, D.: Triple network model dynamically revisited: lower salience network state switching in pre-psychosis. Frontiers in physiology **11**, 66 (2020)
10. Carskadon, M.A., Dement, W.C., et al.: Normal human sleep: an overview. Principles and practice of sleep medicine **4**, 13–23 (2005)
11. Chorlian, D.B., Rangaswamy, M., Porjesz, B.: Eeg coherence: topography and frequency structure. Experimental brain research **198**(1), 59 (2009)
12. Cimenser, A., Purdon, P.L., Pierce, E.T., Walsh, J.L., Salazar-Gomez, A.F., Harrell, P.G., Tavares-Stoeckel, C., Habeeb, K., Brown, E.N.: Tracking brain states under general anesthesia by using global coherence analysis. Proceedings of the National Academy of Sciences **108**(21), 8832–8837 (2011)
13. Faes, L., Nollo, G., Jurysta, F., Marinazzo, D.: Information dynamics of brain–heart physiological networks during sleep. New Journal of Physics **16**(10), 105,005 (2014)
14. Ivanov, P.C., Bartsch, R.P.: Network physiology: mapping interactions between networks of physiologic networks. In: Networks of Networks: the last Frontier of Complexity, pp. 203–222. Springer (2014)
15. Ivanov, P.C., Liu, K.K., Bartsch, R.P.: Focus on the emerging new fields of network physiology and network medicine. New journal of physics **18**(10), 100,201 (2016)
16. Ivanov, P.C., Liu, K.K., Lin, A., Bartsch, R.P.: Network physiology: From neural plasticity to organ network interactions. In: Emergent Complexity from Nonlinearity, in Physics, Engineering and the Life Sciences, pp. 145–165. Springer (2017)
17. Kerkman, J.N., Bekius, A., Boonstra, T.W., Daffertshofer, A., Dominici, N.: Muscle synergies and coherence networks reflect different modes of coordination during walking. Frontiers in Physiology **11** (2020)
18. Klosh, G., Kemp, B., Penzel, T., Schlogl, A., Rappelsberger, P., Trenker, E., Gruber, G., Zeithofer, J., Saletu, B., Herrmann, W., et al.: The siesta project polygraphic and clinical database. IEEE Engineering in Medicine and Biology Magazine **20**(3), 51–57 (2001)
19. Lavanga, M., Bollen, B., Jansen, K., Ortibus, E., Naulaers, G., Van Huffel, S., Caicedo, A.: A bradycardia-based stress calculator for the neonatal intensive care unit: a multisystem approach. Frontiers in Physiology **11** (2020)

20. Lin, A., Liu, K.K., Bartsch, R.P., Ivanov, P.C.: Delay-correlation landscape reveals characteristic time delays of brain rhythms and heart interactions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**(2067), 20150,182 (2016)
21. Lin, A., Liu, K.K., Bartsch, R.P., Ivanov, P.C.: Dynamic network interactions among distinct brain rhythms as a hallmark of physiologic state and function. Communications Biology **3**(1), 1–11 (2020)
22. Liu, K.K., Bartsch, R.P., Lin, A., Mantegna, R.N., Ivanov, P.C.: Plasticity of brain wave network interactions and evolution across physiologic states. Frontiers in neural circuits **9**, 62 (2015)
23. Liu, K.K., Bartsch, R.P., Ma, Q.D., Ivanov, P.C.: Major component analysis of dynamic networks of physiologic organ interactions. In: Journal of Physics: Conference Series, vol. 640, p. 012013 (2015)
24. Liu, L., Shao, Z., Lv, J., Xu, F., Ren, S., Jin, Q., Yang, J., Ma, W., Xie, H., Zhang, D., et al.: Identification of early warning signals at the critical transition point of colorectal cancer based on dynamic network analysis. Frontiers in bioengineering and biotechnology **8**, 530 (2020)
25. Moorman, J.R., Lake, D.E., Ivanov, P.C.: Early detection of sepsis—a role for network physiology? Critical care medicine **44**(5), e312–e313 (2016)
26. Niedermeyer, E., da Silva, F.L.: Electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins (2005)
27. Pereira-Ferrero, V.H., Lewis, T.G., Pereira Ferrero, L.G., Duarte, L.T.: Complex networks models and spectral decomposition in the analysis of swimming athletes' performance at olympic games. Frontiers in physiology **10**, 1134 (2019)
28. Rechtschaffen, A., Kales, A.: A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. bethesda, md: Us dept of health, education and welfare. Public Health Service (1968)
29. Rizzo, R., Zhang, X., Wang, J.W.J.L., Lombardi, F., Ivanov, P.Ch.: Network Physiology of Cortico–Muscular Interactions. Frontiers in Physiology **11**, 558070 (2020)
30. Tan, Y.Y., Montagnese, S., Mani, A.R.: Organ system network disruption is associated with poor prognosis in patients with chronic liver failure. Frontiers in Physiology **11**, 983 (2020)
31. Tanaka, H., Hayashi, M., Hori, T.: Topographic mapping of electroencephalography coherence in hypnagogic state. Psychiatry and clinical neurosciences **52**(2), 147–148 (1998)
32. Wang, Z., Liu, Z.: A brief review of chimera state in empirical brain networks. Frontiers in Physiology **11** (2020)

Modelling oscillating living systems: Cell energy metabolism as weighted networks of nonautonomous oscillators

Joe Rowland Adams and Aneta Stefanovska

Abstract Oscillations are a common feature throughout life, forming a key mechanism by which living systems can regulate their internal processes and exchange information. To understand the functions and behaviours of these processes, we must understand the nature of their oscillations. Studying oscillations can be difficult within existing physical models that simulate the changes in a system's masses through autonomous differential equations. We discuss an alternative approach that focuses on the phases of oscillating processes and incorporates time as a key consideration. We will also consider the application of these theories to the cell energy metabolic system, and present a novel model using weighted nonautonomous Kuramoto oscillator networks in this context.

1 Introduction

It is increasingly clear that a wide variety of biological processes are rhythmic in nature, from glycolysis within a cell to the heart pumping blood throughout the body [6, 31]. Replicating this fluctuating behaviour poses a challenge to many traditional modelling methods, which can rely upon approximations of the system as thermodynamically closed and linear, and which examine the system asymptotically in time. Such models may only generate oscillations at particular parameter selections and modulations, oscillate with a high degree of stability, and exist in a steady state within most of their parameter space. This is in contrast to much of what has been observed of oscillating living systems, where the oscillations continually fluc-

Joe Rowland Adams
Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK
e-mail: j.adams6@lancaster.ac.uk

Aneta Stefanovska
Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK
e-mail: aneta@lancaster.ac.uk

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_13

tuate in their frequency and amplitude, and continue until the death of the system itself [2,5,6,9,13,14,18,19,24–26,28–30,35–37]. We will present and discuss a different approach that rethinks how fluctuating biological systems are best modelled, applied to the cell energy metabolic system.

2 Principles of an alternative approach

In table 2 we outline the key principles that form our method for modelling oscillating living systems, which we discuss further in this section.

Table 1 Summary of the principles informing our modelling approach contrasted to those of mainstream approaches

Mainstream principles	Our principles
Open systems can be modelled as perturbed closed systems	Open systems can only be fully represented by open models
Oscillations result from instability of a dynamical system	Oscillations are inherent to the dynamics of open systems. Living systems continuously exchange energy and matter with the environment and each process is characterized by self-sustained oscillations on a certain time-scale
Nonlinear systems can be recombined from linear systems	Nonlinear systems are best understood by nonlinear models
Time variation in living systems is often due to noise, and can be averaged out over asymptotic time	Time variation in living systems is often deterministic, and must be modelled as nonautonomous to reflect the full system dynamics

It is easy to see that biological systems are open: without being able to exchange mass and energy across its boundary a cell would die, the blood would not be oxygenated by the lungs, and neurons would not receive the energy they need to fire [8, 23, 31, 40]. While it can be mathematically simpler to treat these systems as closed off to their environment, doing so is not modelling them in their healthy, existing state, but instead a dead or dying one. The first principle of our approach is therefore to allow the modelled system to be open. Attempting to model transfers of mass in an open system can be distinctly difficult. Tracking each unit of mass throughout the entire system necessitates the inclusion of processes that may otherwise not need to be considered, and are often challenging, if not impossible, to measure experimentally in their living states.

Oscillations can often be considered as a perturbation of a system away from its 'natural' steady state. However, an attempt to remove oscillations from an otherwise oscillatory system would be equivalent to destroying the system itself: oscillations not only allow a compartmentalisation of otherwise conflicting processes, but play a significant role in the exchange of information and regulation throughout living systems [37]. Therefore we instead treat them as an intrinsic result of the openness of living systems.

While modelling systems' interactions linearly also simplifies the mathematics, it does not reflect the biological reality. Biological systems endemically exhibit transitions in behaviour disproportionate to environmental changes [7], and so we propose to model them as nonlinearly interacting phase oscillators [32].

The fourth key principle of our approach is that living systems should be studied according to the time scales in which they actually exist and function. Analysing the properties of a system in an asymptotic time frame can erase dynamics that exist for only short times. Lucas et al., for example, demonstrated that nonautonomous phase oscillators may synchronise intermittently, and that this is missed when using asymptotic methods [21].

This variation of frequency of oscillation is seen throughout biology [2, 19, 25, 37], and hence our model considers nonlinearly interacting phase oscillations with nonautonomous frequencies, analysed on finite time scales.

3 Modelling a cell's energy metabolism

Our model brings together these four principles to examine the oscillations of the energy metabolism of a single cell. The focus of this model is the production of ATP, a key molecule in maintaining cellular functions, by glycolysis, consuming glucose, and mitochondrial oxidative phosphorylation (OXPHOS), consuming oxygen [8, 40]. We build on the work of Lancaster et al. [20], who modelled each metabolic process as a singular nonautonomous phase oscillator. This model is based on the theory of chronotaxic systems, which characterises nonautonomous oscillations as a method for stabilising against external perturbations [34]. We extend this to include multiple oscillators of each process, transforming the glycolytic and OXPHOS processes into weighted networks of Kuramoto oscillators [17]. We also incorporate the findings of Lucas et al. [21], deterministically varying the frequencies of the oscillations.

This model is represented diagrammatically in figure 1. It consists of four main elements – two weighted Kuramoto networks of phase oscillators representing glycolysis and OXPHOS, and two sets of phase oscillators driving these networks, representing glucose and oxygen.

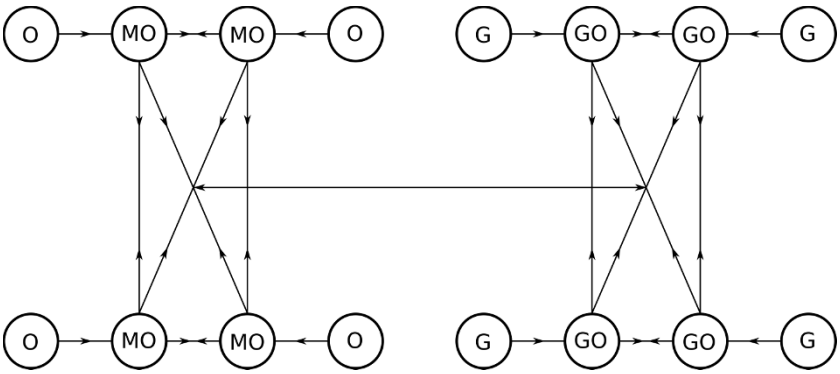


Fig. 1 Oscillator model diagram, where each circle represents a glycolysis (GO), glucose (G), mitochondrial OXPHOS (MO) or oxygen (O) oscillator, and each line a coupling.

That these processes are oscillatory has been extensively established by experimentation, and further, that they may do so nonautonomously [2, 5, 6, 9, 13, 14, 18–20, 24–26, 28–30, 35–37]. The networks of the model reflect the fact that glycolysis occurs in a cell distributed throughout the cytosol, undergoing multiple different reactions simultaneously, and that these reactions appear to communicate through the exchange of acetaldehyde molecules [10, 16, 22, 29, 39]. Similarly, cells contain multiple mitochondria, each undergoing OXPHOS, communicating through molecular exchange, common regulation and inter-mitochondrial nano tunnels [3, 4, 12, 18, 30, 38]. The weighting of these networks, such that neighbouring oscillators influence one another more strongly than those more separated, reflects the spatial distances between these individual processes, and the diffusive nature of their molecular-exchange-driven communications.

We now introduce the mathematical formulation of these elements, beginning with the concept of phase oscillators. These are derived from ordinary differential equations that exhibit self-sustaining oscillations in their state dynamics. Phase, in this circumstance, is defined as the position of the equation along its oscillatory cycle at a given time. The frequency here refers to the velocity of this phase, which we allow to vary in time. We choose to focus on phase as the building blocks of our model, initially discarding the amplitude of the oscillations. This is because at a microscopic level the oscillator is a unit defined with a phase only, while the amplitude is built at a mesoscopic level, resulting from the mean field of the network.

The oscillators' phase can be further defined in the immediate region around its oscillations in state space through the use of isochrons. Isochrons connect all points in the region adjacent to a stable cycle with the one point on the cycle that, after a time, will first meet the perturbed points back on the cycle as the perturbation decays. Thus all these points are defined by the same phase [27, 32].

For nonautonomous oscillators we may also make this extension of definition, by considering each state in time as an autonomous system of slightly different frequency to the ones proceeding and following it. So long as the cycle of each autonomous system exists in the region of attraction of the system proceeding it, we

may define the former's phase via the isochrons of the latter system. This assumption hence requires that the change in the oscillator's frequency over time remains small in comparison to the frequency itself [15].

Having defined phase in the region of nonautonomous cycles, we can consider methods of coupling oscillators. Because our approach focuses on the frequencies and phases of the systems involved, phase coupling is used to model the effects of the biological processes on one another. Through this form of coupling, oscillatory systems perturb one another's phase in a backwards or forwards direction, depending on the comparative directions of oscillation of the two systems. Too strong coupling, however, can perturb the phase beyond the region defined by isochrons. Therefore in order for the perturbed system to remain in the region of its original cycle, where phase is defined, we must further require that that the coupling generating the perturbation is only weak [11, 27, 32].

We may now consider the equations of the model. First, the glycolysis and OXPHOS intra-network connections are defined as

$$\begin{aligned}\dot{\theta}_{GONi} &= \frac{K_{GO}}{N}\sum_{j=1}^{N} W_{ij}\sin\left(\theta_{GOj}-\theta_{GOi}\right)\\ \dot{\theta}_{MONi} &= \frac{K_{MO}}{M}\sum_{j=1}^{M} W_{ij}\sin\left(\theta_{MOj}-\theta_{MOi}\right),\end{aligned} \tag{1}$$

where the subscript GO represents the glycolytic network and MO the OXPHOS, N the number of glycolytic oscillators, M the number of OXPHOS oscillators, K_X the relevant network coupling strength and θ_X the phase.

The weighting of edges within the glycolytic and mitochondrial networks consists of more heavily weighting shorter edges, where the nodes are positioned equidistantly around a ring. Mathematically, for $i \leq \frac{N}{2}$

$$W_{ij} = \begin{cases} \dfrac{W}{|i-j|}, & \text{for } j \in \left[1, i+\frac{N}{2}-1\right] \\ \dfrac{W}{|j-N-i|}, & \text{for } j \in \left[i+\frac{N}{2}, N\right], \end{cases} \tag{2}$$

and for $N \geq i > \frac{N}{2}$

$$W_{ij} = \begin{cases} \dfrac{W}{|i-j|}, & \text{for } j \in \left[i-\frac{N}{2}+1, N\right] \\ \dfrac{W}{|j+N-i|}, & \text{for } j \in \left[1, i-\frac{N}{2}\right], \end{cases} \tag{3}$$

where i denotes the index of the node under consideration, j the index of the node at the other end of the corresponding edge, N the number of nodes in the network, W a constant to be chosen, and W_{ij} the resulting weighting of the edge connecting nodes i and j.

Next, the glucose and oxygen driving are defined as,

$$\begin{aligned}\dot{\theta}_{GOGi} &= \varepsilon_G \sin\left(\theta_{GOi} - \theta_{Gi}\right)\\ \dot{\theta}_{MOOi} &= \varepsilon_O \sin\left(\theta_{MOi} - \theta_{Oi}\right),\end{aligned} \tag{4}$$

where the subscript G represents the glucose driving and O the oxygen, and ε_X represents the coupling strength of the relevant driving.

Finally, the inter-network interactions arise through coupling each network to the mean field of the other [33], such that

$$\begin{aligned}\dot{\theta}_{GOMOi} &= F_{GO} r_{MO} \sin\left(\Psi_{MO} - \theta_{GOi}\right)\\ \dot{\theta}_{MOGOi} &= F_{MO} r_{GO} \sin\left(\Psi_{GO} - \theta_{MOi}\right).\end{aligned} \tag{5}$$

Here F_X is the intra-network coupling strength, r_X the Kuramoto order parameter, where $r_X e^{i\phi} = \frac{1}{N}\sum_{k=1}^{N} e^{i\theta_{Xk}}$ and ϕ is the phase of the mean field arising from the network, such that $r_X = 1$ indicates a totally ordered network, while $r_X = 0$ a totally disordered one. Further, the average phase of network X is $\Psi_X = \frac{1}{N}\sum_{i=1}^{N} \theta_{Xi}$.

The four governing differential phase equations therefore are,

$$\begin{aligned}\dot{\theta}_{Gi} &= \omega_{Gi}(t)\\ \dot{\theta}_{Oi} &= \omega_{Oi}(t)\\ \dot{\theta}_{GOi} &= \omega_{GOi}(t) + \dot{\theta}_{GONi} - \dot{\theta}_{GOGi} + \dot{\theta}_{GOMOi}\\ \dot{\theta}_{MOi} &= \omega_{MOi}(t) + \dot{\theta}_{MONi} - \dot{\theta}_{MOOi} - \dot{\theta}_{MOGOi},\end{aligned} \tag{6}$$

where $\omega_X(t)$ is the time-varying natural frequency of oscillator X. The signs of the inter-network coupling terms are opposite to represent the inhibitory effects of OXPHOS on glycolysis, and the excitatory effects of glycolysis on OXPHOS [20].

A comparison between an output of this model and an experimental observation of cellular glycolysis in shown figure 2. The experimental data were obtained by Amemiya et al. [2], who optically measured the NADH fluorescence, a by-product of glycolysis, of batches of HeLa cells cultured under a variety of glucose starvation conditions. The model output is the combined Kuramoto order parameter of the glycolytic and OXPHOS networks, defined as

$$\Psi_{GOMO} = \frac{1}{(N+M)}\left(\sum_{i=1}^{N}\theta_{GOi} + \sum_{j=1}^{M}\theta_{MOj}\right).$$

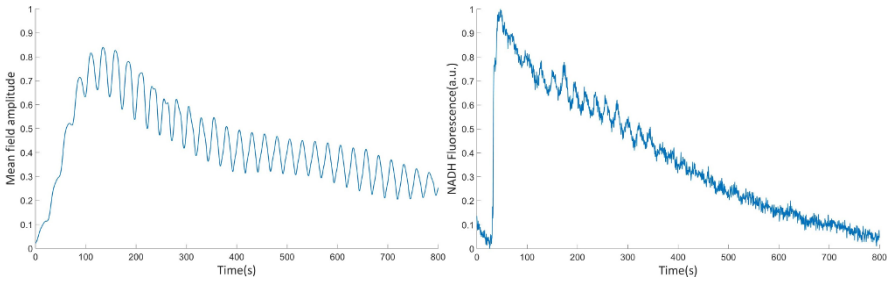


Fig. 2 Sample output of the model (left) and the NADH fluorescence of a single HeLa cell from the Amemiya et al experiment [2], normalised to within the range [0, 1] (right). The model output is represented by the combined Kuramoto order parameter of both the glycolytic and OXPHOS networks.

The parameter values are given in table 2.

Table 2 Parameters used in the simulation to generate the output displayed in figure 2

Parameter	Value(s)
ε_G	[0.1, 0.26]
ε_O	0.01
K_{GO}	1
K_{MO}	1
F_{GO}	0.05
F_{MO}	0.05
ω_G	[0.015, 0.065] Hz
ω_{GO}	[0.02, 0.04] Hz
ω_{MO}	[0.025, 0.075] Hz
ω_O	[0.02, 0.04] Hz
N	100
M	100
W	1

These results can be compared to the model of the same experiment by Amemiya et al. [1] who constructed a classical autonomous model of just the glycolytic process of a HeLa cell, in which mass was assumed to be conserved. Figure 2 in [1] presents an analogous output to what we have shown here. The model by Amemiya et al. involved 22 parameters in 7 governing equations, while our model relies on the 13 parameters of table 2 in the 4 governing equations shown in equation 6.

While the overall trend and oscillating nature of the model output in figure 2 are represented in the experimental data, we are undoing more analysis of the model to better replicate the oscillation death and frequency seen in the experiment. Further details of this simulation and analysis will be presented elsewhere.

4 Outlook

Modelling oscillating biological systems in their living state is a complex task. In order to reproduce every oscillation, variation of frequency, and different regime of stability a system offers, oscillations and nonautonomicity must be built in to the foundations of a model.

Using this approach, we can replicate oscillatory biological data in all its variety with only small changes to model parameters, that can themselves be matched to experimental measurements. Investigating the parameters at which various combinations of the oscillators of the model synchronise, and the transitions between these relationships, can also reveal a significant amount about a biological system. Each of these regimes can be understood as a healthy or pathological state of the system, revealing the breakdown of which mechanisms can be identified with which diseases [20].

Further, analysing the synchronisation of nonautonomous oscillator networks in finite time has already uncovered the new phenomenon of intermittent synchronisation [21]. Investigation of the metabolic model we have presented here, which introduces multiple networks and more complex forms of coupling, promises yet more unseen stabilisation behaviours.

References

1. Amemiya, T., Shibata, K., Du, Y., Nakata, S., Yamaguchi, T.: Modeling studies of heterogeneities in glycolytic oscillations in hela cervical cancer cells. Chaos **29**, 033,132 (2019)
2. Amemiya, T., Shibata, K., Itoh, Y., Itoh, K., Watanabe, M., Yamaguchi, T.: Primordial oscillations in life: Direct observation of glycolytic oscillations in individual HeLa cervical cancer cells. Chaos **27**, 104,602 (2017)
3. Aon, M.A., Cortassa, S., Marbán, E., O'Rourke, B.: Synchronized whole cell oscillations in mitochondrial metabolism triggered by a local release of reactive oxygen species in cardiac myocytes. J. Biol. Chem. **278**, 44,735–44,744 (2003)
4. Aon, M.A., Cortassa, S., O'Rourke, B.: Percolation and criticality in a mitochondrial network. PNAS **101**, 4447–4452 (2004)
5. Bechtel, W., Abrahamsen, A.: Complex Biological Mechanisms: Cyclic, Oscillatory, and Autonomous. In: C. Hooker (ed.) Philosophy of Complex Systems, *Handbook of the Philosophy of Science*, vol. 10, pp. 257–285. North-Holland, Amsterdam (2011)
6. Betz, A., Chance, B.: Phase relationship of glycolytic intermediates in yeast cells with oscillatory metabolic control. Archives of Biochemistry and Biophysics **109**, 585–594 (1965)
7. Carballido-Landeira, J., Escribano, B. (eds.): Nonlinear Dynamics in Biological Systems. Springer International Publishing, New York (2016)
8. Chaudhry, R., Varacallo, M.: Biochemistry, Glycolysis. StatPearls Publishing, Treasure Island (FL) (2020)
9. Ganitkevich, V., Mattea, V., Benndorf, K.: Glycolytic oscillations in single ischemic cardiomyocytes at near anoxia. J. Gen. Physiol. **135**, 307–319 (2010)
10. Gustavsson, A.K., Niekerk, D.D.v., Adiels, C.B., Preez, F.B.d., Goksör, M., Snoep, J.L.: Sustained glycolytic oscillations in individual isolated yeast cells. The FEBS Journal **279**, 2837–2847 (2012)

11. Hagos, Z., Stankovski, T., Newman, J., Pereira, T., McClintock, P.V.E., Stefanovska, A.: Synchronization transitions caused by time-varying coupling functions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **377**, 20190,275 (2019)
12. Iotti, S., Borsari, M., Bendahan, D.: Oscillations in energy metabolism. Biochim. Biophys. Acta **1797**, 1353–1361 (2010)
13. Jung, S.K., Kauri, L.M., Qian, W.J., Kennedy, R.T.: Correlated Oscillations in Glucose Consumption, Oxygen Consumption, and Intracellular Free Ca2+ in Single Islets of Langerhans. J. Biol. Chem. **275**, 6642–6650 (2000)
14. Kennedy, R.T., Kauri, L.M., Dahlgren, G.M., Jung, S.K.: Metabolic Oscillations in beta-Cells. Diabetes **51**, S152–S161 (2002)
15. Kloeden, P., Rasmussen, M.: Nonautonomous Dynamical Systems. American Mathematical Society, Providence (2011)
16. Kohnhorst, C.L., Kyoung, M., Jeon, M., Schmitt, D.L., Kennedy, E.L., Ramirez, J., Bracey, S.M., Luu, B.T., Russell, S.J., An, S.: Identification of a multienzyme complex for glucose metabolism in living cells. J. Biol. Chem. **292**, 9191–9203 (2017)
17. Kuramoto, Y.: Chemical Oscillations, Waves, and Turbulence. Springer, Berlin, Heidelberg (1984)
18. Kurz, F.T., Aon, M.A., O'Rourke, B., Armoundas, A.A.: Spatio-temporal oscillations of individual mitochondria in cardiac myocytes reveal modulation of synchronized mitochondrial clusters. Proc. Natl. Acad. Sci. U.S.A. **107**, 14,315–14,320 (2010)
19. Kurz, F.T., Aon, M.A., O'Rourke, B., Armoundas, A.A.: Wavelet analysis reveals heterogeneous time-dependent oscillations of individual mitochondria. Am J Physiol Heart Circ Physiol **299**, H1736–H1740 (2010)
20. Lancaster, G., Suprunenko, Y.F., Jenkins, K., Stefanovska, A.: Modelling chronotaxicity of cellular energy metabolism to facilitate the identification of altered metabolic states. Sci Rep **6**, 29,584 (2016)
21. Lucas, M., Fanelli, D., Stefanovska, A.: Nonautonomous driving induces stability in network of identical oscillators. Phys. Rev. E **99**, 012,309 (2019)
22. Madsen, M.F., Danø, S., Sørensen, P.G.: On the mechanisms of glycolytic oscillations in yeast: Mechanisms of glycolytic oscillations. FEBS Journal **272**, 2648–2660 (2005)
23. Muoio, V., Persson, P.B., Sendeski, M.M.: The neurovascular unit – concept review. Acta Physiologica **210**, 790–798 (2014)
24. Olsen, L.F., Andersen, A.Z., Lunding, A., Brasen, J.C., Poulsen, A.K.: Regulation of Glycolytic Oscillations by Mitochondrial and Plasma Membrane H+-ATPases. Biophys J **96**, 3850–3861 (2009)
25. O'Rourke, B., Ramza, B.M., Marban, E.: Oscillations of membrane current and excitability driven by metabolic oscillations in heart cells. Science **265**, 962–966 (1994)
26. Ozalp, V.C., Pedersen, T.R., Nielsen, L.J., Olsen, L.F.: Time-resolved measurements of intracellular ATP in the yeast Saccharomyces cerevisiae using a new type of nanobiosensor. J. Biol. Chem. **285**, 37,579–37,588 (2010)
27. Pikovsky, A., Rosenblum, M., Kurths, J.: Synchronization | Nonlinear science and fluid dynamics. Cambridge University Press, New York (2003)
28. Porat-Shliom, N., Chen, Y., Tora, M., Shitara, A., Masedunskas, A., Weigert, R.: In Vivo Tissue-wide Synchronization of Mitochondrial Metabolic Oscillations. Cell Reports **9**, 514–521 (2014)
29. Richard, P.: The rhythm of yeast. FEMS Microbiol Rev **27**, 547–557 (2003)
30. Saleet Jafri, M., Kotulska, M.: Modeling the mechanism of metabolic oscillations in ischemic cardiac myocytes. Journal of Theoretical Biology **242**, 801–817 (2006)
31. Stefanovska, A., Bračič, M.: Physics of the human cardiovascular system. Contemp. Phys. **40**(1), 31–55 (1999)
32. Strogatz, S.H.: Nonlinear Dynamics and Chaos. CRC Press, Boca Raton (2015)
33. Strogatz, S.H., Mirollo, R.E.: Stability of incoherence in a population of coupled oscillators. J Stat Phys **63**, 613–635 (1991)

34. Suprunenko, Y.F., Clemson, P.T., Stefanovska, A.: Chronotaxic systems: A new class of self-sustained nonautonomous oscillators. Phys. Rev. Lett. **111**(2), 024,101 (2013)
35. Thoke, H.S., Tobiesen, A., Brewer, J., Hansen, P.L., Stock, R.P., Olsen, L.F., Bagatolli, L.A.: Tight coupling of metabolic oscillations and intracellular water dynamics in Saccharomyces cerevisiae. PLoS ONE **10**, e0117,308 (2015)
36. Tu, B.P., Kudlicki, A., Rowicka, M., McKnight, S.L.: Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. Science **310**, 1152–1158 (2005)
37. Tu, B.P., McKnight, S.L.: Metabolic cycles as an underlying basis of biological oscillations. Nat Rev Mol Cell Biol **7**, 696–701 (2006)
38. Vincent, A.E., Turnbull, D.M., Eisner, V., Hajnóczky, G., Picard, M.: Mitochondrial Nanotunnels. Trends Cell Biol. **27**, 787–799 (2017)
39. Weber, A., Prokazov, Y., Zuschratter, W., Hauser, M.J.B.: Desynchronisation of Glycolytic Oscillations in Yeast Cell Populations. PLoS ONE **7**, e43,276 (2012)
40. Wilson, D.F.: Oxidative phosphorylation: regulation and role in cellular and tissue metabolism. The Journal of Physiology **595**, 7023–7038 (2017)

A time-series approach to assess physiological and biomechanical regulatory mechanisms

Ruben Fossion, Ana Leonor Rivera, Lesli Alvarez-Millán, Lorena García-Iglesias, Octavio Lecona, Adriana Robles-Cabrera, Bruno Estañol

Abstract In various areas of Medicine there is interest to incorporate information on homeostasis and regulation to increase the predictive power of prognostic scales. This has proven to be difficult in practice because of an uncomplete understanding of how regulation works dynamically and because a common methodology does not exist to quantify the quality of regulation independent from the specific mechanism. In the present contribution, it is shown that time series of *regulated* and *effector variables* from different regulatory mechanisms show universal features that may be used to assess the underlying regulation.

Ruben Fossion
Instituto de Ciencias Nucleares & Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico
e-mail: ruben.fossion@nucleares.unam.mx

Ana Leonor Rivera
Instituto de Ciencias Nucleares & Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico

Lesli Alvarez-Millán
Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico

Lorena García-Iglesias
Doctorado en Ciencias, Universidad Autóoma del Estado de Morelos, Avenida Universidad 1001, Chamilpa, 62209 Cuernavaca, Morelos, Mexico

Octavio Lecona
Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico

Adriana Robles-Cabrera
Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico

Bruno Estañol
Laboratorio de Neurofisiología Clínica, Departamento de Neurología y Psiquiatría, Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán", Mexico City, Mexico

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_14

1 Introduction

Apart from diagnosis and treatment, a third important task in medicine is *prognosis*. Short-term prognostic scales of the order of days and weeks as for use in critical and intensive care are based on vital signs, in particular point (one-time) measurements of heart rate, respiration rate, blood pressure, blood oxygen saturation, body temperature, state of consciousness, etc. Longer-term scales of the order of months and years as used in geriatrics and palliative care include different aspects of the patient's functionality (physical, emotional, social, etc.) because average values of physiological parameters tend to be within their normative ranges [4]. Attemps are being made to incorporate also measures related to regulation and physiological reserves, in particular in geriatric frailty scales, but these attempts have remained mostly theoretical because of problems related to how to assess and measure regulation [26, 38].

Physiological regulation was first advanced by Claude Bernard in the 2nd half of the 19th century as the approximate constancy of the internal environment ("milieu intérieur") of the human body in the face of perturbations received from the external environment. At the beginning of the 20th century, Walter Cannon included also adaptive physiological responses to these external perturbations in an extended concept which he coined *homeostasis*. Although homeostasis is one of the core concepts of physiology, several sticky points remain in its understanding, e.g., how constant is the approximate constancy of the internal environment, whether physiological regulation works as an on-off switch or whether it is active continuously, whether the setpoint of a homeostatic mechanism is stable or whether it may change over time, and how various homeostatic mechanisms interconnect hierarchically [23].

At least 3 different strategies exist to quantify homeostasis in clinical practice, but no strategy is 100% satisfactory [9]. First, point (one-time) measurements of vital parameters allow to check whether these are within their normal ranges. The most obvious example is measuring body temperature to detect fever. Unfortunately, only information on the end result of regulation is obtained but not on the dynamical adaptation that underlies regulation. Second, the stimulus-response approach where responses to standardized stimuli are measured, allowing to distinguish between adaptive and non-adaptive physiological responses. A typical example is the glucose tolerance test in the diagnosis of diabetes. The major drawback is that not all experimental stimuli are applicable in vulnerable populations such as elderly adults or critical care patients. Third, realizing that the human body is never really in rest but continuously responding to a wide variety of internal and external stimuli, the time series of the spontaneous fluctuations of specific physiological variables can be analyzed statistically in order to quantify the activity of the corresponding regulatory mechanisms. The most studied physiological time series is heart rate variability (HRV) which offers a non-invasive proxy to assess the autonomous nervous system [20, 32]. The disadvantage of this approach using physiological time series is that having no information on the perturbations that generate these fluctuations makes it difficult to compare between different subjects. Also, it is not obvious how the statistics of these fluctuations evolve from ideal conditions of youth and health to adverse

conditions of ageing or disease when physiological regulation becomes suboptimal. Two different hypotheses exist to interpret these fluctuations but appear to mutually contradict each other: the *loss of complexity* paradigm of Lipsitz and Goldberger which predicts that complexity and variability decrease [19], whereas the *critical transitions* paradigm of Scheffer et al. states that variability and non-gaussianity increase in adverse conditions [33]. Moreover, West argues that homeostasis and traditional gaussian statistics constitute conceptual barriers to understand the spontaneous fluctuations of physiological variables which should rather be studied within the context of the new field of *fractal physiology* [39].

In order to solve this paradox, we point to the fact that different variables may play different roles in physiological regulation. Indeed, recent advances in physiology education distinguish between on the one hand *regulated variables* such as core temperature and blood pressure that represent Bernard's internal environment and that are supposed to remain constant, and on the other hand *effector variables* such as skin temperature and heart rate that are responsible for Cannon's adaptive responses [23, 24]. We reasoned that these very distinct roles generate different statistics for the corresponding time series [9, 10, 11]: (i) in optimal conditions, regulated variables are characterized by a small variability around their respective setpoints which reflects the characteristic constancy of the internal environment, whereas effector variables have a large variability reflecting their adaptive capacity, and (ii) in adverse conditions, adaptive capacity and variability decrease for effector variables with as a consequence a loss of the constancy of the internal environment and therefore an increase of the variability of the regulated variables. In order to compare the variability from variables that often are measured in different units, we rescaled fluctuations to percentages around the median value,

$$\Delta X = 100 \times \left(\frac{X - \text{median}(X)}{\text{median}(X)} \right), \tag{1}$$

where X is the variable of interest, and we introduced a *homeostatic parameter*,

$$\alpha = \text{SD}(\Delta X_e)/\text{SD}(\Delta X_r), \tag{2}$$

which compares the relative variability of the regulated variable X_r of a specific regulatory mechanism and X_e a corresponding effector variable and where the standard deviation SD may be used as a measure of variability [10, 28]. It has been suggested that this approach may constitute a "bridge" between the loss of complexity and critical transition paradigms [25]. In the present contribution, we will focus on variables that are measurable continuously and in a non-invasive way, and we will compare examples from our previous publications (body temperature and cardiovascular variables) with new examples from physiology and biomechanics (ventilatory variables and gait), see Table 1.

Table 1 Regulatory mechanisms typically consist of one variable that is to be regulated and maintained constant and various effector variables that are responsible for adaptive responses to perturbations. Specific homeostatic mechanisms are often studied separately as if they function independently from each other which of course is an approximation; instead, homeostasis is known to work in a *hierarchical* or *nested* way, where a regulated variable such as blood pressure at the systemic scale may function as an effector variable at a local scale. Time series of regulated and effector variables may show universal behaviour independent from the specific field of study, from physiology to biomechanics.

regulated variable	effector variables
core temperature	skin temperature, vasomotor effects (vasoconstriction, vasodilatation), shivering, sweating, etc.
blood oxygen saturation	breathing rate, breathing amplitude, etc.
blood pressure[a]	heart rate, ejection fraction, cardiac output, vasomotor effects (vasoconstriction, vasodilatation), shivering, sweating, etc.
blood flow[b]	blood pressure, heart rate, ejection fraction, cardiac output, vasomotor effects (vasoconstriction, vasodilatation), etc.
average walking speed	step length, cadence, etc.

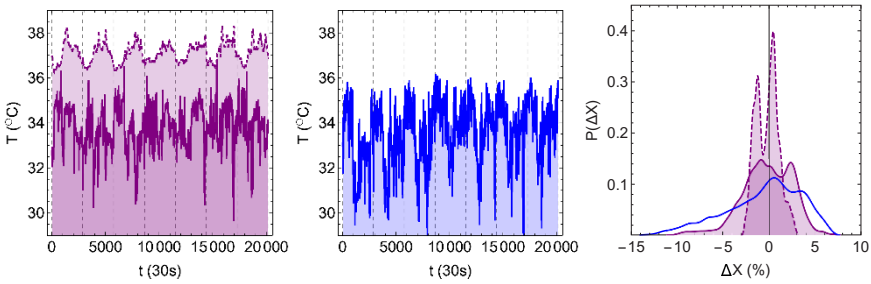
[a] systemic/extrinsic regulation
[b] local/intrinsic regulation or autoregulation

2 Selected physiological and biomechanical regulatory mechanisms

2.1 Physiological regulation of body temperature

The temperature of the human body depends on where it is measured [31]. It is the core body temperature which represents the internal environment and which is to be maintained constant in the face of changes of the external environment. One of the most important effector variables allowing adaptation to these external changes is skin temperature, which is modulated by limiting (vasoconstriction) or stimulating blood flow (vasodilatation) through the capillaries below the skin. Skin temperature depends on where on the body surface it is measured, in part because of variations in the local surface-to-volume ratio. Heat transfer is limited and a higher temperature is maintained proximally (on the trunk), whereas distally (on the extremities) heat transfer is enhanced and temperature tends to be lower. Temperature variations are slow and need to be measured over long time intervals, hours to days, to be studied as a time series, see Fig. 1. Core temperature T_{core} is a difficult variable to measure, because a sensor needs to be introduced in a body orifice and maintained there for the whole duration of the experiment. We explored body temperature regulation previously [8, 10], here we also aimed at exploring whether skin temperature when measured proximally, e.g., at the clavicula fossa T_{clav}, might function as a proxy for T_{core} to illustrate the dynamics of body temperature homeostasis in a more accessible way and we contrasted with distal skin temperature measured at the wrist, T_{wrist}. We compared the probability distribution functions (PDF) of all variables using the

dimensionless fluctuations of eq. (1) and calculated the homeostatic parameter α of eq. (2). It can be observed that the PDF of T_{core} is a superposition of 2 gaussian distributions corresponding to small variations of $\approx 1\%$ around a day and a night setpoint. In the PDF of T_{clav} the 2 local maxima of the circadian cycle are still present but in a less prominent way and variability is larger $\approx 5\%$. In the PDF of T_{wrist} the circadian cycle almost has become invisible, variability has increased dramatically $\approx 10\%$ and the distribution is non-gaussian and skewed to the left. We indeed observe a larger variability for the effector variable T_{wrist} than for the regulated variable T_{core} ($\alpha = 3.18$) or T_{clav} ($\alpha = 2.15$), which confirms our working hypothesis and indicates that indeed T_{clav} may possibly serve as a proxy for T_{core} to assess body temperature regulation. Previously, we also found that variability of T_{wrist} decreases with adverse conditions of being overweight and obesity which may indicate that adaptive capacity is lost [8, 10].



(a) core temperature and proxy (b) skin temperature at wrist (c) variability

Fig. 1 Homeostasis of body temperature. Shown are (a) the regulated variable of core temperature T_{core} and its possible proxy of skin temperature at the clavicula fossa T_{clav}, (b) effector variable of skin temperature at the wrist T_{wrist}, and (c) probability distribution functions (PDF) of fluctuations ΔX of eq. (1) for T_{core}, T_{clav} and T_{wrist} comparing their variability. Time series are measured continuously using thermochron iButtons at 30s sample intervals over 7 successive days (vertical gridlines indicate midnight) and the prominent 24h periodic pattern is the circadian cycle. All panels use the same color and style coding for the different variables, T_{core} (dashed and shaded purple curve), T_{clav} (continuous and shaded purple curve) and T_{wrist} (continuous blue curve). Data are from a healthy male adult.

2.2 Physiological regulation of blood oxygen saturation

Blood oxygen saturation, the percentage of arterial red blood cells carrying oxygen, is one of the only regulated variables that can be measured continuously and in a non-invasive way using a digital oximeter which can be developed in a well-equiped university lab at the undergraduate level [14, 21]. In optimal conditions, blood oxygen saturation is above 95% at sea level (above 90% at higher altitudes such as Mexico City located at 2000m), below 80% organ functionality is compro-

mised and below 70% consciousness is lost. Corresponding efector variables include breathing rate and breathing amplitude which can be measured using a chest strap. Fig. 2 compares conditions of rest and exercise (2km walking). It can be seen that during physical effort, muscles consume oxygen at an increased rate, breathing rate is increased but nevertheless blood oxygen saturation is slightly lower than during rest. The occasional large peaks in breathing amplitude during rest before the effort correspond to sighs; breathing amplitude is increased drastically during rest after the effort. Variability is larger for breathing amplitude (50-100%) than for breathing rate (20-30%), and variability of both effector variables is larger than in the case of the regulated variable (1%), with a homeostastic parameter of breathing rate with respect to blood oxygen saturation of $\alpha = 18.27$ (rest pre), $\alpha = 6.27$ (walk) and $\alpha = 10.71$ (rest post), and a homeostastic parameter of breathing amplitude with respect to blood oxygen saturation of $\alpha = 61.41$ (rest pre), $\alpha = 38.57$ (walk) and $\alpha = 37.59$ (rest post), which confirms our working hypothesis.

2.3 Physiological regulation of blood pressure

It is straightforward to realize point (one time) measurements of arterial blood pressure using a sphygmomanometer. It is much more difficult to continuously monitor blood pressure. Non-invasive devices exist, using volume-clamp techniques based on control theory, such as the Finapres of Finapres Medical Systems and the CNAP Monitor from CNSystems (CNAP stands for continuous non-invasive arterial pressure), but these very expensive [1]. Heart rate on the other hand can be measured easily using an electrocardiogram on the chest (ECG) or photoplethysmography (PPG) on the finger or earlobes, and which is the principle of measurement used by commercial smartwatches that monitor hear rate. There is a clear consensus that a high heart rate variability (HRV) constitutes a protective factor for health [20, 32]. The significance of blood pressure variability (BPV) is less clear, although there are indications that a high BPV represents a risk factor for negative health outcomes, which raises the question whether in the specific case of hypertension treatment should only focus on lowering high blood pressure levels or should try to reduce BPV as well [27]. In previous publications, we found evidence for a higher variability for heart rate than for systolic blood pressure in health [11], and a decrease of HRV and an increase of systolic BPV in the adverse condition of type-2 diabetes mellitus, in correspondence with our working hypothesis [9, 10, 28, 29]. In contrast, here heart rate and systolic blood pressure would seem to have similar variabilities, see Fig. 3, panels (a)-(c). A difference between both variables is that the distribution for systolic blood pressure behaves symmetrical and gaussian, whereas the distribution for heart rate is asymmetrical and right-skewed. Panels (d)-(f) compare average distributions over 5 min segments with the distribution for the whole time series of 2 hr for heart rate, systolic and diastolic blood pressure. It can be seen that variability increases only slightly for heart rate with time scale but that it increases much more importantly for systolic blood pressure. Diastolic blood pressure results to be much

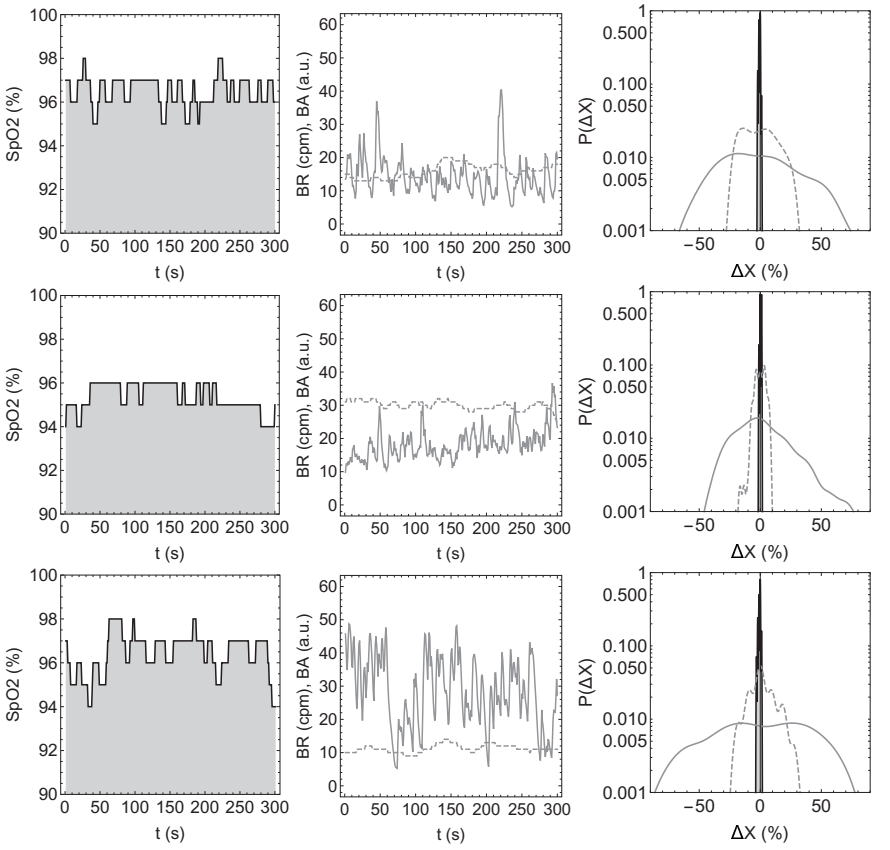
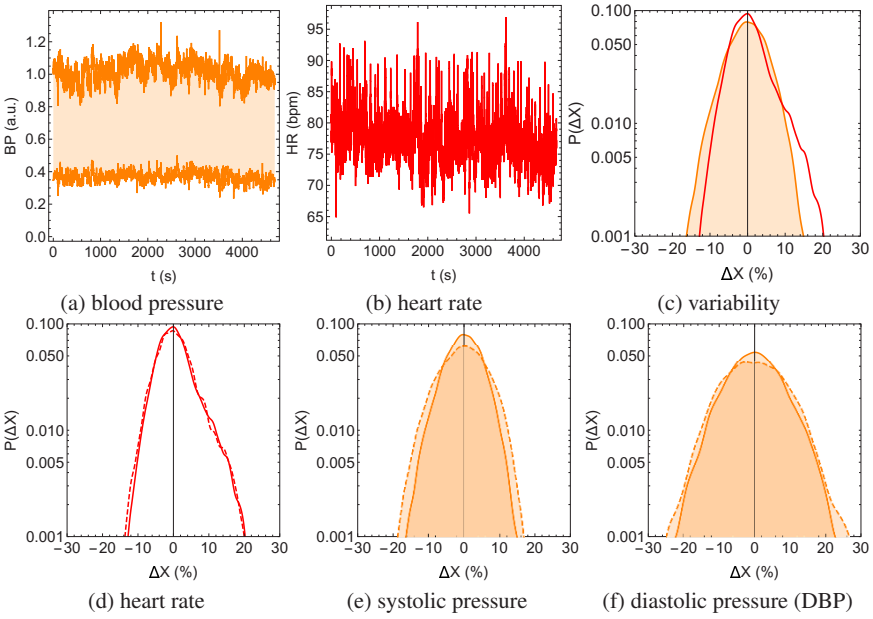


Fig. 2 Homeostasis of blood oxygen saturation in rest before (top row), after (bottom row) and during the physical effort of a 2km walk (middle row). Shown are (a) blood oxygen saturation SpO2 in percent, (b) breathing rate BR in cycles per minute and breathing amplitude BA in arbitrary units, and (c) probability distribution functions (PDF) of fluctuations ΔX of eq. (1) for SpO2, BR and BA comparing their variability. All time series are fragments of 5 min, measured at 1s sample intervals using a Masimo MightySAT oximeter (SpO2) and a Zephyr Bioharness (BA and BR). All panels use the same color and style coding for the different variables, SpO2 (continuous and shaded black curve), BR (dashed gray curve) and BA (continuous gray curve). Data are from a healthy female undergraduate student.

more variable than both heart rate and systolic blood pressure. Homeostatic parameters for heart rate with respect to systolic blood pressure are $\alpha = 1.07$ (average over fragments of 5min) and $\alpha = 0.83$ (over whole time series of 2hrs) and for heart rate with respect to diastolic blood pressure $\alpha = 0.66$ (average over fragments of 5min) and $\alpha = 0.61$ (over whole time series of 2hrs). Some physiological considerations may be able to explain these values. Systolic blood pressure depends mostly on cardiac output and in lesser degree also on arterial elasticity, whereas diastolic

blood pressure depends in the first place on arterial elasticity [36]. Therefore, it may be expected that systolic blood pressure participates actively in regulation whereas diastolic blood pressure possible plays a more passive role. Another consideration is that blood pressure may be a regulated variable at the systemic scale but plays an effector role at the local scale of specific organs, see Table 1, and the discussion section. This more passive role for diastolic blood pressure may be reflected by the homeostatic parameter α which appears to be independent from the time scale, whereas the variation in α with time for systolic blood pressure may indicate an alternation between the different roles of effector and regulator.



(a) blood pressure (b) heart rate (c) variability

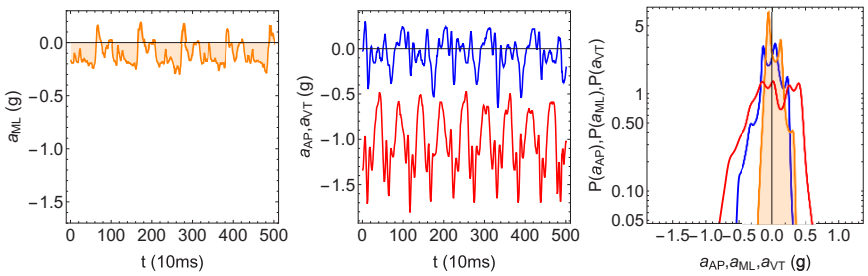
(d) heart rate (e) systolic pressure (f) diastolic pressure (DBP)

Fig. 3 Homeostasis of blood pressure. Shown are (a) systolic SBP and diastolic blood pressure DBP in arbitrary units, (b) heart rate HR in beats per minute and (c) average probability distribution functions (PDF) over 5min fragments of fluctuations ΔX of eq. (1) for SBP and HR comparing their variability. Average PDFs over 5min fragments are compared with PDFs of the complete 2hr time series for (d) heart rate HR, (e) systolic blood pressure SBP and (f) diastolic blood pressure DBP. Time series are on beat-to-beat basis with a length of 2hrs in supine rest. Data from a healthy young adult from the Physionet-Fantasia database [17, 13].

2.4 Biomechanical regulation of gait

The specific term of "homeostasis" is reserved for physiological regulation. Regulation also appears in other disciplines of medicine, e.g., the dynamics of biped

gait in biomechanics. Average walking speed v, possibly a regulated variable, is considered as the 6th vital sign because physical funcionality and independence are compromised when it drops below approx. 1m/s, e.g., in the case of age-associated frailty [12]. Elderly adults are often described to walk with a "cautious" or "senile gait", i.e., with small steps and a high step frequency or cadence [41, 18], both probably effector variables. When balance is altered, because of external factors such as walking in a moving train [2] or on a ship at sea [37] or because of internal factors such as pregnancy [22], obesity [5] or ageing [18], gait becomes similar to a "waddle", i.e., with an increased step width and larger associated mediolateral acceleration, also an effector variable. Fig. 4 shows fragments of time series of the effector variables of mediolateral a_{ML} and anterioposterior acceleration a_{AP} which oscillate around 0 and where each oscillation corresponds to an individual step, and vertical acceleration a_{VT} which oscillates around -1 because of the constant contribution of gravity. The PDFs of variability around the corresponding medians show that variability is smaller for a_{ML} than for a_{AP} and a_{VT}, in particular $\alpha = 1.53$ in the former case and $\alpha = 2.86$ in the latter case (comparing here different effector variables), which makes sense because the former component does not contribute to the forward movement whereas the latter two components do (this is obvious for a_{AP} but also applies to a_{VT} where especially for running vertical acceleration must be large enough to suspend both feet in the air simultaneously for each step). We have preliminary results that the variability of a_{AP} and a_{VT} decreases and that the variability of a_{ML} increases with age-associated frailty, constituting a suboptimal regulation of gait, which is also the reason that these variables are presented in different panels in Fig. 4.



(a) mediolateral (b) anterioposterior and vertical (c) variability

Fig. 4 Regulation of gait. Shown are (a) mediolateral acceleration a_{ML} in units of the Earth gravitational constant $g = 9.81\mathrm{m/s}^2$, (b) anterioposterior a_{AP} and vertical acceleration a_{VT} in units of g and (c) probability distribution functions (PDF) of fluctuations ΔX of eq. (1) for $a_{\mathrm{ML}}, a_{\mathrm{AP}}$ and a_{VT} comparing their variability. Time series fragments of 5s are shown and the PDFs are for the complete walk exercise of 2km (approx. 20min) of Fig. 2. Measured with the triaxial accelerometer of the Zephyr Bioharness with a sampling interval of 10ms. All panels use the same color and style coding for the different variables, a_{ML} (shaded orange curve), a_{AP} (blue curve) and a_{VT} (red curve). Data from a healthy female undergraduate student.

3 Discussion

Homeostatic principles of approximate constancy and therefore small variability of *regulated variables* associated to the internal environment and large variability because of adaptation to perturbations by *effector variables* are very clearly observed in time series related to body temperature homeostasis and the homeostasis of blood oxygen saturation, where skin temperature is much more variable than core temperature (Fig. 1), and breathing rate and breathing amplitude are much more variable than blood oxygen saturation (Fig. 2). In previous contributions, we have shown that in adverse conditions of ageing and/or disease the variability of regulated and effector variables deviates in opposite directions: decreasing for effector variables reflecting their diminished adaptive capacity and consequently increasing for regulated variables reflecting the more instable internal environment [9, 10]. This approach may offer a "bridge" between the paradigms of *loss of complexity* [19] and *critical transitions* [33].

This time-series approach can help to solve some of the "sticky points" mentioned in the introduction. The spontaneous fluctuations of physiological time series at all time scales clearly contradict the notion that physiological regulation would work as an on-off switch (on at some times and off at other times) and suggest that regulation is working continuously. Circadian cycles are a clear example of how the setpoint of a homeostatic mechanism may change over time, e.g., core body temperature is lower during the night than during the day, probably with the objective to save energy.

Another "sticky question" is how constant precisely the approximate constancy of the internal environment is. The answer may be that it depends on the specific homeostatic regulatory mechanism in question. In the cases of core temperature and blood oxygen saturation, we saw that they are maintained within a few percent of their median values. It may be interesting as well to consider the relative variability of specific regulated variables with respect to a corresponding effector variable. Since the work of Schrödinger where he interpreted the phenomenon of life from the perspective of physics, people have wondered about the order of internal structures and processes of the human body whereas according to the 2nd law of thermodynamics entropy should increase with time [30, 34]. Regulatory processes may function as an *entropy pump*, creating order in the internal environment by pumping excess entropy towards the external environment and creating extra entropy in the process. The homeostatic parameter α may express quantitatively how effective specific regulatory mechanisms are as an entropy pump. If true, then body temperature regulation and in particular blood oxygen saturation regulation would appear to work as good entropy pumps.

Standard textbooks on physiology discuss different homeostatic regulatory mechanisms one for one and separately, e.g., those listed in Table 1, as if they were independent from each other. This is not the case, different homeostatic regulatory mechanisms are interconnected as conveyed by the concept of *hierarchical* or *nested homeostasis* [6, 24]. One example may be blood pressure homeostasis. Although heart rate variability is one of the best studied physiological time series, its

statistics and that of the corresponding systolic and diastolic blood pressure is difficult to interpret in terms of regulatory mechanisms (Fig. 3). A possible reason is that blood pressure may play the role of a regulated variable at the systemic scale of the whole body but may function as an effector variable at the local scale of specific organs. Systemic blood pressure must be kept at a level which is convenient at average for all organs and tissues in the body, whereas some specific organs are so vital for survival, e.g., brain, heart, kidney, and possibly also eye [16], skin [40], etc., that their blood flow is very closely regulated and maintained constant by compensating between blood pressure on the one hand and vasomotor effects on the other hand [7, 15].

Regulated variables, such as core temperature and aslo blood pressure, tend to be more difficult to monitor continuously and non-invasively than corresponding effector variables, such as skin temperature and heart rate. We saw that regulated variables tend to be more symmetrical and gaussian, whereas effector variables appear to be characterized by more asymmetrical and non-gaussian distributions with a tail. Perhaps the focus on heart rate variability which is easy to monitor and which is well studied, has led West to conclude that homeostasis and gaussian distribution are obstacles to understand physiological time series [39]. Our results seem to indicate that gaussian statistics describes well regulated variables associated to the approximately constant internal environment, whereas fractal statistics and power laws may be better suited to describe the associated effector variables [11]. Homeostasis then is a concept which allows to combine the dynamics of regulated and associated effector variables within a same regulatory mechanism.

This time-series approach appears to capture general features of regulatory mechanisms and is applicable not only to physiological mechanisms, such as respiration and body temperature, but also biomechanical mechanisms, such as gait, and therefore is promising to incorporate information from regulation in prognostic scales in various medical disciplines.

4 Conclusions

Time series of physiological and biomechanical variables appear to reflect key aspects of the underlying regulatory mechanisms. The relative variability of a regulated variable and the corresponding effector variables would seem to offer a means to quantify the quality of the regulation. An advantage of focusing on such common and universal features would be that using the same methodology various regulatory mechanisms can be incorporated into prognostic scales.

Acknowledgements Funding for this work was supplied by Dirección General de Asuntos del Personal Académico (DGAPA) from Universidad Nacional Autónoma de México (UNAM) with Grant IA102619.

References

1. Bartels, K., Thiele, R.H.: Advances in photoplethysmography: Beyond arterial oxygen saturation. Can J Anesth **62** 1313–1328 (2015)
2. Baumgart, C., Wilhelm Hoppe, M., Freiwald, J.: Long-term adaptations to unexpected surface perturbations: Postural control during stance and gait in train conductors. J. Motor Behav. **48** 341–347 (2016)
3. Branco, M., Santos-Rocha, R., Vieira, F.: Biomechanics of gait during pregnancy. Sci. World J.
4. Brandan, M.E., Avila, M.A., Fossion, R., Zapata-Fonseca, L.: Una mirada a la investigación futura en Física Médica en México. In: Lucío-Martínez, J.L., Torres-Labansat, M. (eds.) Presente y futuro de la Ciencia en México: Retos y perspectivas de la Física, pp. 175-188.
5. Browning, R.C.: Locomotion mechanics in obese adults and children. Curr. Obes. Rep. **1** 152–159 (2012)
6. Carpenter, R.H.S.: Homeostasis: A plea for a unified approach. Adv. Physiol. Educ. **28** S180–S187 (2004)
7. Estañol, B., Rivera, A.L., Martínez Memije, R., Fossion, R., Gómez, F., Bernal, K., Murúa Beltrán, S., Delgado-García, G., Frank, A.: From supine to standing: In vivo segregation of myogenic and baroreceptor vasoconstriction in humans. Physiol. Rep. **4** e13053 (2016)
8. Fossion, R., Stephens, C.R., García-Pelagio, K.P., García-Iglesias, L.: Data mining and time-series analysis as two complementary approaches to study body temperature in obesity. Proceedings of Digital Health Dh'2017, London, UK, July 02-05, pp. 190-194 (2017).
9. Fossion, R., Rivera, A.L., Estañol, B.: A physicist's view of homeostasis: How time series of continuous monitoring reflect the function of physiological variables in regulatory mechanisms. Physiol. Meas. **39**, 084007 (2018)
10. Fossion, R., Fossion, J.P.J., Rivera, A.L., Lecona, O.A., Toledo-Roy, J.C., García-Pelagio, K.P., García-Iglesias, L., Estañol, B.: Homeostasis from a time-series perspective: An intuitive interpretation of the variability of physiological variables. In: Olivares-Quiroz, L., Resendis-Antonio, O. (eds.) Quantitative models for microscopic to macroscopic biological macromolecules and tissues, pp. 87–109, Springer Int. Publishing AG, Cham (2018)
11. Fossion, R., Sáenz-Burrola, A., Zapata-Fonseca, L.: On the stability and adaptability of human physiology: Gaussians meet heavy-tailed distributions. INTERdisciplina **8** 55–81 (2020)
12. Fritz, S., Lusardi, M.: White paper: Walking speed, the sixth vital sign. J. Geriatr. Phys. Ther. **32**, 2–5 (2014)
13. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M.,, Ivanov, P.Ch., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation **101** e215-e220 (2000)
14. González-Villa, E.A.: Desarrollo y caracterización de un oxímetro de pulso para el análisis de series de tiempo. Undergraduate thesis, UNAM (2015).
15. Hall, J.E.: Guyton and Hall textbook on medical physiology. 12th ed. Saunders Elsevier, Philadelphia (2011)
16. He, Z., Vingrys, A.J., Armitage, J.A., Bui, B.V.: The role of blood pressure in glaucoma. Clin. Exp. Optometry **94** 133–149 (2011)
17. Iyengar, N., Peng, C.-K., Morin, R., Goldberger, A.L., Lipsitz, L.A.: Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. Am. J. Physiol. **271** 1078–1084 (1996)
18. Lindemann, U.: Spatiotemporal gait analysis of older persons in clinical practice and research. Z Gerontol Geriat **53** 171–178 (2020)
19. Lipsitz, L.A., Goldberger, A.L.: Loss of "complexity" and aging: Potential applications of fractals and chaos theory to senescence. J. Am. Med. Assoc. **267** 1806–1809 (1992)
20. Malik, M.: Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Eur. Heart J. **17** 354–381 (1996)

21. Margerito-Segundo, D.: Desarrollo de un dispositivo medidor de oxígeno y su correlación con los patrones de actividad física. Undergraduate thesis, UNAM (2020)
22. McCrory, J.L., Chambers, A.J., Daftary, A., Redfern, M.S.: The pregnant "waddle": An evaluation of torso kinematics in pregnancy. J. Biomech. **47** 2964-–2968 (2014)
23. Michael, J., Cliff, W., McFarland, J., Modell, H., Wright, A.: The core concepts of physiology: A new paradigm for teaching physiology. Springer Nature, NY (2017)
24. Modell, H., Cliff, W., Michael, J., McFarland, J., Wenderoth, M.P., Wright, A.: A physiologist's view of homeostasis. Adv. Physiol. Educ. **39**, 259–266 (2015)
25. Nakazato, Y., Sugiyama, T., Ohno, R., Shimoyama, H., Leung, D.L., Cohen, A.A., Kurane, R., Hirose, S., Watanabe, A., Shimoyama, H.: Estimation of homeostatic dysregulation and frailty using biomarker variability: A principal component analysis of hemodialysis patients. Sci. Rep. **10** 10314 (2020)
26. Olde Rikkert, M.G.M., Melis, R.J.F.: Rerouting geriatric medicine by complementing static frailty measures with dynamic resilience indicators of recovery potential. Front. Physiol. **10** 723 (2019)
27. Parati, G., Ochoa, J.E., Lombardi, C., Bilo, G.: Assessment and management of blood-pressure variability. Nat. Rev. Cardiol. **10**, 143–155 (2013)
28. Rivera, A.L., Estañol, B., Sentíes-Madrid, H., Fossion, R., Toledo-Roy, J.C., Mendoza-Temis, J., Morales, I.O., Landa, E., Robles-Cabrera, A., Moreno, R., Frank, A.: Heart rate and systolic blood pressure variability in the time domain in patients with recent and long-standing diabetes mellitus. PLoS ONE **11** e0148378 (2016)
29. Rivera, A.L., Estañol, B., Fossion, R., Toledo-Roy, J.C., Callejas-Rojas, J.A., Gien-López, J.A., Delgado-García, G.R., Frank, A.: Loss of breathing modulation of heart rate variability in patients with recent and long standing diabetes mellitus type II. PLoS ONE **11** e0165904 (2016)
30. Sagan, D., Whiteside, J.H.: Gradient reduction theory: Thermodynamics and the purpose of life. In: Schneider, S.H., Miller, J.R., Crist, E., Boston, P.J. (eds.) Scientists debate Gaia: The next century. MIT Press, pp. 173–186 (2004)
31. Savastano, D.M., Gorbach, A.M., Eden, H.S., Brady, S.M., Reynolds, J.C., Yanovski, J.A.: Adiposity and human regional body temperature. Am J Clin Nutr **90** 1124–1131 (2009)
32. Schaffer, F., Ginsberg, J.P.: An Overview of Heart Rate variability Metrics and Norms. Front. Public Health **5** 258 (2017)
33. Scheffer, M.: Critical transitions in Nature and society. Princeton Univ. Press, Princeton (2009)
34. Schneider, E.D., Sagan, D.: Into the cool. Energy flow. Thermodynamics and life. University of Chicago Press, Chicago (2006)
35. Schr odinger, E., : What is Life? Dublin Institute for Advanced Studies at Trinity College, Dublin (1944)
36. Stephens, C.R., Easton, J.F., Robles-Cabrera, A., Fossion, R., de la Cruz, L., Martínez-Tapia, R., Barajas-Martínez, A., Hernández-Chávez, A., López-Rivera, J.A., Rivera, A.L.: The impact of education and age on metabolic disorders. Front. Public Health **8** 180 (2020)
37. Stoffregen, T.A.: Functional control of stance in older adults. Kinesiol. Rev. **5** 23–29 (2016)
38. Varadhan, R., Seplaki, C.S., Xue, Q.L., Bandeen-Roche, K., Fried, L.P.: Stimulus-response paradigm for characterizing the loss of resilience in homeostatic regulation associated with frailty. Mech. Ageing Dev. **129** 666–670 (2008)
39. West, B.: Homeostasis and Gaussian statistics: Barriers to understand natural variability. J. Eval. Clin. Pract. **16** 403–408 (2013).
40. Wilson, T.E., Zhang, R., Levine, B.D., Crandall, C.G.: Dynamic autoregulation of cutaneous circulation: Differential control in glabrous versus nonglabrous skin. Am. J. Physiol. **289** H385–H391 (2005)
41. Zijlstra, A., de Bruin, E.D., Bruins, N., Zijlstra, W.: The step length-frequency relationship in physically active community-dwelling older women. Eur. J. Appl. Physiol. **104** 427–434 (2008)

InterCriteria Analysis Approach as a Tool for Promising Decision Making in Physiological Rhythms

Krassimir Atanassov and Tania Pencheva

Abstract Recently developed InterCriteria Analysis (ICrA) approach has being intensively gained popularity as quite promising approach to support decision making process in biomedical informatics studies, and in particular – in physiological rhythms. ICrA has been elaborated to discern possible similarities in the behaviour of pairs of criteria when multiple objects are considered. The approach is based on the theories of intuitionistic fuzzy sets and index matrices. Up to now, ICrA has been successfully applied in economics, different industry fields, ecology, artifficial intelligence, e-learning, etc. ICrA has been demonstrated as promising tool also in studies related to medicine and bioinformatics, which are in the focus of this investigation.

1 Introduction

The idea of InterCriteria Analysis (ICrA) has been originally developed in the period 2014-2015 [4, 5]. In the coming years, the interest to the concept increased significantly. The approach has become a subject of theoretical studies as well as of applications in various fields, e.g. industry, economics, education, medical and biotechnological processes, artificial intelligence, including neural networks, expert systems, bio-inspired and metaheuristics algorithms, etc. Recently, a survey on theory and applications of ICrA approach [7] has been spread to the scientific community.

Krassimir Atanassov
Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, 105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria e-mail: krat@bas.bg

Tania Pencheva
Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, 105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria e-mail: tania.pencheva@biomed.bas.bg

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_15

The concept of ICrA is based on two mathematical formalisms, developed by Atanassov, namely the theories of Index Matrices (IM) [1, 3] and Intuitionistic Fuzzy Sets (IFSs) [2], thus relying on methodology different from the classical correlation analysis.

ICrA has been designed as a novel method for detecting the levels of pairwise correlations for a set of criteria and a set of objects (measurements or evaluations) against these criteria. The ultimate goal of ICrA is to detect if some of the criteria exhibit high enough correlations with others, so that skipping them from the further decision making process would not affect the whole process [8]. The motivation behind this method is for an eventual elimination of some of the criteria, when measurement against these comes at a higher cost, consumes more time or other resources, or is considered undesirable in any other reason. Selecting these high enough correlations requires either an expert decision or an algorithm for the precise establishment of the thresholds, beyond which the top-correlating criteria are selected in order to yield certain problem-specific conclusions.

2 Brief description of ICrA

Here we will briefly repeat the theoretical framework of the proposed approach, firstly proposed in [5]. The approach employs an index matrix M of m rows $\{O_1, ..., O_m\}$ and n columns $\{C_1, ..., C_n\}$, where for every i, j, k, l $(1 \leq i \leq j \leq m, 1 \leq k \leq l \leq n)$, O_i is an evaluated object, C_k is an evaluation criterion, and $e_{O_iC_k}$ is the evaluation of the i-th object against the k-th criterion, defined as a real number or another object that is comparable according to relation R with all the rest elements of the index matrix M:

$$M = \begin{array}{c|ccccccc} & C_1 & \dots & C_k & \dots & C_l & \dots & C_n \\ \hline O_1 & e_{O_1C_1} & \dots & e_{O_1C_k} & \dots & e_{O_1C_k} & \dots & e_{O_1C_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & & \\ O_i & e_{O_iC_1} & \dots & e_{O_iC_k} & \dots & e_{O_iC_k} & \dots & e_{O_iC_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & & \\ O_j & e_{O_jC_1} & \dots & e_{O_jC_k} & \dots & e_{O_jC_k} & \dots & e_{O_jC_n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & & \\ O_m & e_{O_mC_1} & \dots & e_{O_mC_k} & \dots & e_{O_mC_k} & \dots & e_{O_mC_n} \end{array}. \quad (1)$$

From the above requirement for comparability follows the relation $R(e_{O_iC_k}, e_{O_jC_k})$ for each i, j, k. The relation R has dual relation $\overline{R}$, which is true in the cases when relation R is false, and vice versa.

For the needs of this decision making method, pairwise comparisons between every two different criteria are made along all evaluated objects. During the comparison, it is maintained one counter of the number of times when the relation R holds, and another counter for the dual relation.

Let $S^{\mu}_{k,l}$ be the number of cases in which the relations $R(e_{O_iC_k}, e_{O_jC_k})$ and $R(e_{O_iC_l}, e_{O_jC_l})$ are simultaneously satisfied. Let also $S^{\nu}_{k,l}$ be the number of cases in which the relations $R(e_{O_iC_k}, e_{O_jC_k})$ and its dual $\overline{R}(e_{O_iC_l}, e_{O_jC_l})$ are simultaneously satisfied. As the total number of pairwise comparisons between the object is $m(m-1)/2$, it is seen that there hold the inequalities:

$$0 \leq S^{\mu}_{k,l} + S^{\nu}_{k,l} \leq m(m-1)/2. \tag{2}$$

For every k, l, such that $1 \leq k \leq l \leq m$, and for $m \geq 2$ the following two numbers are defined:

$$\mu_{C_k,C_l} = 2\frac{S^{\mu}_{k,l}}{m(m-1)}, \nu_{C_k,C_l} = 2\frac{S^{\nu}_{k,l}}{m(m-1)}. \tag{3}$$

In the terms of ICrA, μ_{C_k,C_l} is a *degree of agreement*, while ν_{C_k,C_l} – a *degree of disagreement*. Obviously, both μ_{C_k,C_l} and ν_{C_k,C_l}, are numbers in the [0, 1]-interval, and their sum is also a number in this interval. What is complement to their sum to 1 is the number π_{C_k,C_l}, which corresponds to a *degree of uncertainty*.

The pair, constructed from these two numbers, plays the role of the intuitionistic fuzzy evaluation of the relations that can be established between any two criteria C_k and C_l. In this way, the index matrix M that relates evaluated objects with evaluating criteria can be transformed to another index matrix M^* that gives the relations among the criteria:

$$M^* = \begin{array}{c|cccc} & C_1 & C_2 & \dots & C_n \\ \hline C_1 & \langle 1,0\rangle & \langle \mu_{C_1,C_2}, \nu_{C_1,C_2}\rangle & \dots & \langle \mu_{C_1,C_n}, \nu_{C_1,C_n}\rangle \\ C_2 & \langle \mu_{C_2,C_1}, \nu_{C_2,C_1}\rangle & \langle 1,0\rangle & \dots & \langle \mu_{C_2,C_n}, \nu_{C_2,C_n}\rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_n & \langle \mu_{C_n,C_1}, \nu_{C_n,C_1}\rangle & \langle \mu_{C_n,C_2}, \nu_{C_n,C_2}\rangle & \dots & \langle 1,0\rangle \end{array}. \tag{4}$$

From practical considerations, it has been more flexible to work with two index matrices M^{μ} and M^{ν}, rather than with the index matrix M^* of intuitionistic fuzzy pairs (IFPs).

The final step of the algorithm is to determine the degrees of correlation between the criteria. Let $\alpha, \beta \in [0; 1]$ be the threshold values (which are of the user's choice), against which we compare the values of μ_{C_k,C_l} and ν_{C_k,C_l}. We call that criteria C_k and C_l are in:

- *positive consonance*, if $\mu_{C_k,C_l} > \alpha$ and $\nu_{C_k,C_l} < \beta$;
- *negative consonance*, if $\mu_{C_k,C_l} < \beta$ and $\nu_{C_k,C_l} > \alpha$;
- *dissonance*, otherwise.

In a completely identical way, it is possible (though not always meaningful) to build a matrix giving the correlations between the objects. The only difference is that the input index matrix M has to be transposed, and the resultant matrix, e.g., M^{**}, is with dimensions $m \times m$.

3 ICrAData software package description

Here we provide a quick overview of ICrAData – the software implementing the ICrA approach [9]. It is written in the Java programming language and requires the installation of Java Virtual Machine. This makes it usable under Linux and Windows environment. ICrAData is freely available for use and its latest version ICrAData v2.3 can be downloaded from http://intercriteria.net/software/ (Last access August 24, 2020).

In order to easily load data from other software products, the capability to load *csv* (comma separated values) files with headers (row and column) which are taken as names for objects and criteria, was added to the software. This allows loading of tables from *MS Excel/LibreOffice Calc*.

The user interface consists of a left panel for the input data, the central panel for the result of ICrA in a coloured table view, and the rightmost panel showing the graphical interpretation of the results.

For better visualization of the results, table cell colours were added, according to the following rules, depending on the user defined α and β thresholds:

- The results are displayed in dark-green colour in case of *positive consonance*;
- The results are displayed in red colour in case of *negative consonance*;
- Otherwise, in case of *dissonance* – violet colour.

The default values used by the software ICrAData are $\alpha = 0.75$ and $\beta = 0.25$, respectively.

ICrAData saves a draft automatically each 15 minutes and on program exit in order to prevent accidental loss or overwriting of data.

The features outlined above allow for better automation in working process with program and additional improvements in that regard are also planned in the future.

4 ICrA applications in biomedical research and physiological rhythms

Going slightly beyond physiological rhythms, ICrA approach has encountered numerous applications aiming to support decision making in different areas, connected to medical investigations and bioinformatics. In [11], ICrA is applied on a dataset of thermo-dynamic parameters derived from thermograms of blood plasma proteome of patients with colorectal cancer recorded by differential scanning calorimetry. The goal of the study was to establish interdependences between the derived calorimetric parameters that were not inferred so far from the calorimetric data and to discuss their importance for the clinical application of differential scanning calorimetry.

In [10], ICrA, combined with Pearson's and Spearman's correlation analysis, is applied to a large dataset of calorimetric and biochemical parameters derived for the serum proteome of patients diagnosed with multiple myeloma. As a result, intercriteria dependences have been identified that are general for the various types

of multiple myeloma and thus can be regarded as a characteristic of this largely heterogeneous disease: strong contribution of the monoclonal protein concentration to the excess heat capacity of the immunoglobulins-assigned thermal transition; shift of the albumin assigned calorimetric transition to allocation where it overlaps with the globulins assigned transition and strong shift of the globulins assigned transition temperature attributable to monoclonal proteins conformational changes.

In [12], ICrA is applied to real data connected with health-related quality of life (HrQoL). The EQ-5D-3L questionnaire for measuring HrQoL for a representative sample of 1050 residents of Burgas (the fourth-largest in Bulgaria) is used. The data was analyzed to identify the best correlations between the indicators, to discover dependent and independent indicators and the relationships between them. The comparison can help to describe the behavior of the used indicators and their assessment. The increase of the coefficient of consonance and the entry in the zone of strong positive consonance means strong correlation between the respective pair of criteria, which may justify the removal of one of the criteria in the pair on the basis that its informational values is lesser. Removal of indicators leads to simplification of the process of evaluation.

In [13, 14] a dataset of Behterev's disease patients is analyzed applying ICrA, aiming at approbation of this novel approach to medical data with the goal to discover correlations between important health indicators based on available patients' data. The selected set of health indicators comprises: physical functioning; role functioning based by physical conditional; bodily pain; general health status; vitality; social functioning; role functioning based by emotional conditional; and mental health. Results obtained confirm once again that the health condition depends on the emotional condition and determines the social functioning of the patients under observation.

When looking for possible application of ICrA toward the physiological rhythms, in this investigation a novel idea of ICrA application in a totally new direction is proposed here, namely to adapt ICrA assessments in a way to compare two curves, which – in particular case, might represent physiological rhythms.

Let us have two lines L_1 (see Fig. 1) and L_2 (see Fig. 2).

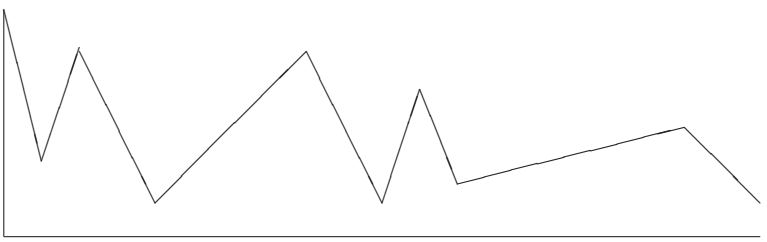


Fig. 1 Line L_1

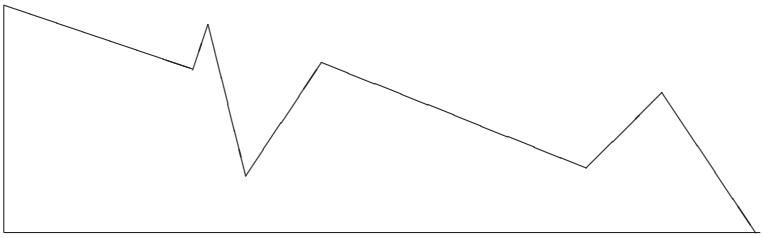


Fig. 2 Line L_2

Let M, N ad P be the degrees of coincidence, of difference and of uncertainty of between both lines. Then the three degrees can be evaluated as it is shown on Fig. 3, where the area of the part of the figure that is in white corresponds to M, the area of the part marked with horizontal lines corresponds to N and the part marked with vertical lines corresponds to P (see Fig. 3).

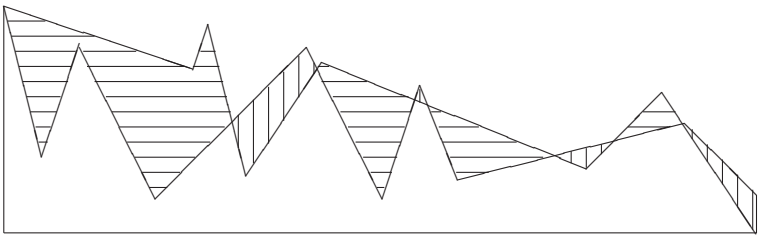


Fig. 3 An example for evaluation of degrees of coincidence, of difference and of uncertainty between lines L_1 and L_2

Further, we can assess the nearness between the two lines based on the intuitionistic evaluations μ and ν, where

$$\mu = \frac{M}{M+N+P}, \nu = \frac{N}{M+N+P}. \tag{5}$$

In fact, this is an IFP with a degree of uncertainty $\pi = 1 - \mu - \nu = \frac{P}{M+N+P}$.

Thereby, in case of n lines, we can search the nearness between them using ICrA and working with IFPs.

This idea might be further used in analyses of different types of physiological rhythms, including ECGs, as well as to even harder from a mathematical point of view electromyography (EMG) signals.

Some first steps have been done in the ICrA application in analysing the features in a database of electrocardiography signal (ECGs). The investigation is carried out over the training set of the Computing in Cardiology Challenge 2017 Database from PhysioNet (https://physionet.org). Some very promising results have been obtained, but in this investigation the researchers were faced to some limits of ICrAData, namely working with a big data. This research is in a fast progress now, both on the software improvement, as well as in data analysis.

5 Conclusions

In all applications so far, ICrA shows prerequisites to assist in decision making processes in order to guide the selection of the most appropriate choice among many. In this investigation, ICrA has been demonstrated as a quite promising tool to assist decision making in such challenging field as physiological rhythms. The novel idea for a comparison of curves presenting physiological rhythms based on ICrA approach has been introduced, which may be in benefit in other fields of research as well.

Acknowledgements This investigation has been partially supported by the National Science Fund of Bulgaria under the Grant Ref. No. DN02/10 "New Instruments for Knowledge Discovery from Data, and Their Modelling".

References

1. Atanassov, K.: Generalized index matrices. Compt. Rend. l'Acad. Bulg. Scie. **40**(11), 15–18 (1987)
2. Atanassov, K.: On Intuitionistic Fuzzy Sets Theory. Springer, Berlin (2012)
3. Atanassov, K.: Index Matrices: Towards an Augmented Matrix Calculus. Springer, Cham (2014)
4. Atanassov, K., Atanassova, V., Gluhchev, G.: Intercriteria analysis: ideas and problems. Notes Intuit Fuzzy Sets **21**(1), 81–88 (2015)
5. Atanassov, K., Mavrov, D., Atanassova, V.: Intercriteria decision making: a new approach for multicriteria decision making, based on index matrices and intuitionistic fuzzy sets. Issues Intuitionistic Fuzzy Sets Generalized Nets **11**, 1–8 (2014)
6. Atanassov, K., Szmidt, E., Kacprzyk, J.: On intuitionistic fuzzy pairs. Notes Intuit Fuzzy Sets **19**(3), 1–13 (2013)
7. Chorukova, E., Marinov P., Umlenski, I., Survey on Theory and Applications of InterCriteria Analysis Approach. Studies in Computational Intelligence, 2020, in press.
8. Doukovska, L., Atanassova, V., Mavrov, D., Radeva, I.: Intercriteria analysis of EU competitiveness using the level operator Nγ. In: Kacprzyk, J., Szmidt, E., Zadrozny, S., Atanassov, K., Krawczak, M. (eds.) Advances in Fuzzy Logic and Technology, Vol. 641 of Advances in Intelligent Systems and Computing, pp. 631–647 (2018)
9. Ikonomov, N., Vassilev, P., Roeva, O.: ICrAData – Software for InterCriteria Analysis, Int. J. Bioautomation **22**(1), 1–10 (2018).
10. Krumova, S., Todinova, S., Mavrov, D., Marinov, P., Atanassova, V., Atanassov, K., Taneva, S.: Intercriteria analysis of calorimetric data of blood serum proteome. Biochimica et Biophysica Acta — General Subjects **1861**(2), 409–417 (2017)
11. Todinova, S., Mavrov, D., Krumova, S., Marinov, P., Atanassova, V., Atanassov, K., Taneva. S. G.: Blood plasma thermograms dataset analysis by means of intercriteria and correlation analyses for the case of colorectal cancer. Int. J. Bioautomation **20**(1), 115–124 (2016)
12. Vankova, D., Sotirova, E., Bureva, V. An application of the intercriteria analysis approach to health-related quality of life, Notes on Intuit Fuzzy Sets **21**(5), 40–48 (2015)
13. Zaharieva, B., Doukovska, L., Ribagin, S., Michalikova, A., Radeva, I.: Intercriteria analysis of Behterev's kinesitherapy program, Notes on Intuit Fuzzy Sets **23**(3), 69–80 (2017)
14. Zaharieva, B., Doukovska, L., Ribagin, S., Radeva, I.: Intercriteria approach to Behterev's disease analysis, Notes on Intuit Fuzzy Sets **23**(2), 119–127 (2017)

“Ome” sweet “ome”: From the Genome to the Conductome

Christopher R. Stephens

Abstract The last few decades have seen science both changed and confronted by the appearance of enormous quantities of data, that have arisen from the development of multiple new technologies. The impact of this “data revolution” has been particularly acute in the biological sciences, where bioinformatics has made great strides in integrating such data into new theoretical frameworks and adopting new computational tools. One framework that has prospered is that of the “ome”, which adopts a more holistic view of the physical structures that make up a cell, tissue or organism and their mutual interactions. The structures associated with the principal “omes” - genome, proteome, transcriptome and metabolome - are all microscopic, being associated with different biological molecules. Recently, however, the omic approach has been applied to more “mesoscopic” structures, such as organs and tissues, with the resulting totality of structures conforming the physiolome. However, all these omes are associated with particular spatial and temporal scales, and are therefore inadequate for addressing the real complexity of living systems, which are both multi-scale and highly multi-factorial with respect to those scales. We argue that a “disease-ome”, for example, as the totality of factors associated with a given disease, requires the integration all the current omes, and more. Thus, a holistic description of an important disease, such as obesity, requires all micro, meso and macro factors, as well as an understanding of both their upstream and downstream causal relations. This is particularly challenging when the relevant factors are distant in scale. Thus, the causality between overeating and obesity at the individual level is clear. However, the link between a certain genotype and obesity or the link between food production and obesity is much less clear. In spite of this, all of these factors can, in principle, be collected and included in a prediction model,

Christopher R. Stephens
C3 - Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico, CDMX 04510
Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico, CDMX 04510
e-mail: stephens@nucleares.unam.mx

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_16

using present technology and computational tools. We argue that the fundamental concept that most naturally links the micro, meso and macro is that of behaviour, as it is influenced by both micro (nature) and macro (nurture) factors and, in turn, influences them. We discuss the concept of the Conductome - the totality of factors that influence behaviour, using as an example food consumption and obesity, and emphasise its utility as an unifying concept that allows for a truly systemic view of a living organism.

Key words: Conductome, behaviour, genome, physiolome, complexity, obesity

1 The Micro-omes

In the last few decades there has been a trend towards a more global, systemic approach to the study of biological systems. A manifestation of this more holistic approach is the proliferation of different "-omes", and their corresponding fields - "-omics" [1], where the emphasis is on identifying as complete a set of relevant "structures" as possible that belong to the corresponding "ome". The original "omes", such as the genome, originated in molecular biology, where the relevant structures, such as genes, proteins, RNA transcripts and metabolites, are all molecules, and as such we can think of the corresponding omes as all being micro-omes. In the context of these micro-omes, the natural mathematical framework for understanding structure and interactions has been that of a network, where the nodes are structures and the links represent interactions. The information needed to construct such an "ome", as representing a totality of structures, is enormous, and has required the development of advanced technologies, such as high-throughput sequencing and capillary electrophoresis mass spectrometry in the case of bioinformatics [2].

Through a physicist's eyes, this "omic" approach seems no different in spirit to the traditional approaches to be found in physics and chemistry. The Periodic Table is, in that sense, the "atom-ome", the complete set of relevant structures at the atomic level. There is also an "elementary particle-ome" and a "moleculome". Indeed, the "omes" of molecular biology should be subsets of this moleculome. Although it is important to be able to identify the set of relevant structures, or the "building blocks", that form an "ome", this difficulty pales in comparison to identifying the complete set of interactions between these structures. The Periodic Table gives us the totality of atomic structures but certainly not the totality of atomic interactions. Thus, the set of possible molecules is much larger than the set of possible atoms. Additionally, interactions are manifestly contextual, meaning that the interaction between two structures is intimately dependent on the context of the environment they are in. For example, the interactions between two carbon atoms are quite different if they are in different DNA molecules versus the same one, or close versus distant in the DNA molecule, or within a chromatin complex or not.

Another way to put this is that: understanding interactions at one scale is no guarantee that we can understand interactions at another. Thus, just as an understanding

of atomic physics does not a priori allow for a quantitative, predictive understanding of molecular physics, as molecules are emergent structures relative to atoms, so an understanding of the genome "gene by gene" does not a priori allow for a quantitative, predictive understanding of what the genome does or, indeed, even what a small set of genes do, due to the presence of genetic interactions (epistasis) [4]. The difficulty of relating structures at one scale to emergent structures at another is one of the most difficult problems in science. Exceedingly difficult in physics, while in the biological sciences it is almost overwhelming. In physics there are several theoretical frameworks that encompass passing from the micro to the macro, the link between statistical mechanics and thermodynamics being the most developed. At a more general level, synergistics [3] has been applied to both physical and biological systems and attempts to delineate generic features and general principles, such as self-organisation and the existence of a relatively small number of order parameters, that lead to a description of the macro from the micro. The standard omic approach, in contrast, is more directly phenomenological in nature. Moreover, the current molecular omes are also associated with a particular type of structure: genes with the genome and proteins with the proteome. However, genes and proteins also interact with each other in a complex fashion, from the production of proteins by the transcription of DNA through proteins as transcription factors that control gene expression.

2 Disease-omes: relating the micro to the macro

Although the "omic" approach has its origins in the micro, recently the idea has been extended to more macro "omes", such as the physiolome [5, 6, 7, 8], where the relevant structures, such as organs or tissues, are much fewer in number. In this case, a network-based approach can naturally be applied and different interaction measures introduced, such as the degree of correlation in the time series of the different organs such as lung and heart. Like the genome however, the physiolome is associated with particular spatial and temporal scales that stem from the physical structures it considers. Thus, although a goal of the omic approach is to be less reductionist, the current omes are all very much linked to a certain scale. However, a true hallmark of complexity - of living systems - is its multi-scale nature, with relevant structures at many different scales. Thus, for instance, the heart, thought of as a physiological unit, has multiple associated spatial and temporal scales: the cellular scale, where pacemaker cells set the underlying heart rate, to circadian variations in the functioning of the heart at the cellular level [9], and on to the long term irreversible changes across a lifetime that are associated with heart disease. In this case, what we do at the macro scale, such as eating a lot of high-cholesterol foods, has an effect at the cellular level, leading, for example, to atherosclerosis which, in turn, has an effect at the macro level, where an artery becomes blocked, leading to a myocardial infarction. However, myocardial infarction itself has also been confirmed to have a genetic component [10], which then introduces a scale below the cellular

level, that of a single nucleotide. Unfortunately, as with many genetic studies, the causal chain that links the observed correlation between a micro property, such as a particular Single-nucleotide polymorphism (SNP), and a macro property, such as atherosclerosis and a subsequent myocardial infarction, is very poorly understood. Of course, the cellular level must enter as a relevant scale that links the two.

Thus, we are faced with one of the principal challenges of describing truly complex phenomena: the ability to incorporate structures, and their interactions, that exist at multiple scales. In other words, a micro-level molecular "ome" is in no way sufficient to encompass an important phenomenon such as atherosclerosis. However, neither is a "meso" ome, such as the physiolome, due to the need to incorporate micro factors such as SNPs. The challenge does not stop there, however. If we consider long term changes in heart structure and function, then two other principal categories of factor enter: aging and "lifestyle", which are the remit of macro-disciplines, such as epidemiology. Studying the disease can be done "bottom up" - trying to link macro effects, such as the clinical manifestations of disease, to the micro [11]. Indeed, much of the focus of the micro-omes has been to link macroscopic disease states to genomic, proteomic, transcriptomic and metabolomic data. In this case, the abstract mathematical framework is that of a conditional probability, $P(disease\ state|state\ of\ the\ genome,\ proteome,\ metabolome\ etc)$. The more conventional approach to disease prediction however, has been "top down", linking the disease state to macro-variables such as age, sex, socio-economic status etc. A disease state however, is a complex multi-scale phenomenon, requiring a unification of the bottom-up and top down approaches, as in biological systems the micro and macro are linked and influence each other in a much deeper way than in physical systems. For instance, aging sounds simple enough to account for, but, as is known, chronological age and biological age are not the same, with the latter also having a genetic component. They are also linked by lifestyle, by which we mean the universe of external factors that affect the organism at multiple scales, from the genetic via, for example, environment-induced mutations, to the truly macro, such as the degree to which the environment itself favours the development of heart disease through, for example, diet. In contrast, in physics, with a nucleome, atomome, moleculome etc. we don't need the nucleome to understand the atomome. "Atomomics" can be developed in terms of structures - atoms - and their interactions without reference to the nucleome and its constituents - nucleons - and their interactions - the strong nuclear force.

We see then that if we wish to understand a phenomenon, such as heart disease, "omically", i.e., in the sense of a more holistic, non-reductionist perspective, it is necessary to go beyond an "ome" that is linked to a particular range of scales, as each only offers at most a partial view of the phenomenon. So, should we introduce the concept of a "myocardial infarction-ome"? where it comprehends all the factors that influence that outcome? This potentially involves the genome, proteome, metabolome, physiolome and several other omes that are still to be characterized, such as a "sociolome" or a "psycholome". However, at the same time the genome is linked to many more macro phenomenon than that of a myocardial infarction. Many, if not all diseases, have a genetic component. To try and be more precise:

imagine a set of diseases, $(C_1, C_2, \ldots, C_m)$, and a set of factors, $(X_1, X_2, \ldots, X_N)$, that are potentially related to those diseases. These factors include genetic factors, epigenetic factors, physiological factors, social factors etc. However, considering the effect of all possible causative factors on the set of all known diseases is not a recipe for success. We can group the totality of factors in a particular disease-ome in groups - genes/SNPs, proteins, metabolites, cell structures, tissue and organ changes, lifestyle factors etc. In this way however, we are faced with the perennial question of causation versus correlation. Is a SNP a "direct" cause of a disease or a correlative, indirect risk factor? Is socio-economic status a direct causal factor in the development of heart disease or a proxy for, potentially, many other more directly causal factors? What about diet? More directly causal? Then we have the fact that the impact of diet is influenced by the genome - nutrigenetics - while diet, in turn, affects gene expression - nutrigenomics. Life is complex. Literally. Nature affects nurture and nurture affects nature.

We argue then that the standard "omic" approach is still too reductionist to comprehend a complex phenomenon such as a disease, which is associated with structures and their interactions at multiple scales and where the interactions can be between structures that are naturally described at quite different scales. Although a network-based approach can be applied at the level of the disease-ome, by considering multiple diseases, for example, the natural framework for a given disease is, again, a conditional probability: $P(C = disease|\mathbf{X} = disease\ causes\ and\ risk\ factors)$ where, for example, C could be a disease state and $\mathbf{X}$ the set of factors that we wish to consider as conditioning factors on the probability to be in the disease state. Naturally, the data requirements to construct $P(C|\mathbf{X})$ are far greater than those of the micro-omes, where the latter are just one component of the disease-ome and, generically, not even the most predictive part. Furthermore, due to the multi-scale nature of the disease-ome, its construction through data has to transcend the disciplinarity that still exists as the principle foundation of scientific research.

Mathematically speaking, the disease-ome, $P(C|\mathbf{X})$, is a prediction model. Such a prediction model may be transverse or longitudinal, depending on whether or not C and $\mathbf{X}$ can be identified as states in time. For example, that C represents the development of a disease in a certain time interval and $\mathbf{X}$ represents the set of predictor variables identified in that time interval (transverse), or that they represent histories up to the beginning of that time interval (longitudinal). Of course, when $\mathbf{X}$ is high dimensional, a direct estimate of $P(C|\mathbf{X})$ is impossible, as $P(C|\mathbf{X}) = 0,\ 1$; i.e., every element is unique and either exists in a single element or doesn't. For example, no two genomes are completely identical, and the vast majority of potential genetic sequences of length ℓ - 4^ℓ - have never existed and probably never will. Thus, $P(C|\mathbf{X})$ must be estimated indirectly. There are, for instance, many machine learning based methodologies that can help in this regard. Seen abstractly, $P(C|\mathbf{X})$ represents a Bayesian belief network, where, in principle, if one could deduce its structure as a directed acylic graph would reveal the probabilistic relations between the different variables X_i, both among themselves and with the disease state itself. The goal would be to determine that graph that is most in accord with data. Unfortunately, computationally, this is an NP hard problem. Rather than search through a large

space of potential graphs, an alternative is to restrict the topology of the graphs. A particularly useful approximation in this regard is the Naive Bayes approximation, or generalisations thereof [12], that use Bayes theorem to relate the posterior probability $P(C|\mathbf{X})$ to a likelihood $P(\mathbf{X}|C)$, then assume independence of the features, X_i. In this approximation $P(\mathbf{X}|C) = \prod_{i=1}^{N} P(X_i|C)$ and so the contribution, $P(X_i|C)$, to the disease from each factor can be calculated observationally and studied separately.

3 Omes from an Ecological perspective

An ecological analogy may help intuit the difference between the two types of "ome". The traditional molecular "omes" are akin to an ecological community, where one is interested in the mutual interactions between all the structures in the system. In ecology these are typically species. The construction of a disease-ome, on the other hand, is more akin to the construction of an ecological niche, where, now, the disease itself is seen as a "species", as those factors that favour a high value for $P(C|\mathbf{X})$, relative, say, to a null hypothesis, $P(C)$, can be viewed as being niche-like, favouring the presence of the disease, C, while low values relative to the null hypothesis are anti-niche-like, favouring the absence of the disease. In this context, taking type 2 diabetes as an example, a niche factor may be the presence of a disease-related SNP, such as *rs*8050136 [13], as may be the consumption of carbonated drinks, or the price of carbonated drinks, or educational status, or hours of exercise, or age, or knowledge of the health consequences of diabetes or the health consequences of consumption of sugary foods, or a seemingly endless array of other factors. Unlike the human genome project there is, to our knowledge, no diabetes-ome project, where the goal is to obtain and integrate the multi-scale, multi-discipline data that begins to represent the totality of factors that affect the development of type 2 diabetes. Project 42, developed at the Centro de Ciencias de la Complejidad of the UNAM is a step in that direction in the context of obesity and metabolic disease. With over 3000 participants and several thousand variables, from a spectrum of previously identified SNPs for risk of obesity and metabolic disease, through demographic data, personal and family history, an ample set of biomarkers, anthropometric measurements, health knowledge, psychometrics, social characteristics, actigraphy and habits; all in a publicly available platform for analysis. The challenge of such data sets is to go beyond a static, statistical description to a process-oriented causal characterisation. Besides the right mathematical tools, this also requires domain-specific knowledge that spans multiple disciplines. The example of genetics affecting the impact of diet and diet affecting the expression of genes, while diet itself is a result of consumption and the consequence of a large number of other factors, from family environment to culture and mass marketing campaigns, speaks to the huge challenges of constructing a more process-oriented framework. Even just discovering the true underlying causal connections between, say, obesity and a single proxy variable such as educational level presents enormous

challenges. Thus, the construction of $P(C|\mathbf{X})$ via some suitable algorithm is just a first necessary step.

An advantage of a Bayesian framework for developing the disease-ome is that, based on a set of factors - "niche" dimensions - $\mathbf{X}$, it can be naturally extended by incorporating new information, such as new variables $\mathbf{X}'$. In this case, the posterior probability, $P(C|\mathbf{X})$, relative to the prior probability, $P(C)$, in the absence of the information $\mathbf{X}$ can, in its turn, be taken as a new prior probability and a new posterior probability, $P(C|\mathbf{X}'\mathbf{X})$, that incorporates information from both $\mathbf{X}$ and $\mathbf{X}'$ constructed using Bayes theorem: $P(C|\mathbf{X}'\mathbf{X}) = P(\mathbf{X}'|C\mathbf{X})P(C|\mathbf{X})/P(\mathbf{X}'|\mathbf{X})$. Again, $P(\mathbf{X}'|C\mathbf{X})$ can then be estimated using one of several Machine Learning methodologies. Dynamics can be incorporated by considering $\mathbf{X} \equiv \mathbf{X}(t)$, where the state vector $\mathbf{X}(t)$ may also contain historical information. For instance, $\mathbf{X}(t)$ may contain information about someone's historical activity level such as: actual activity level, activity level one year ago, activity level two years ago etc. [14]. Prediction in time follows naturally from considering the estimation of $P(C(t)|\mathbf{X}(t'))$, where $t' < t$. Thus, we may predict the probability for a disease state to occur at time t given the disease niche at t'. An example would be predicting if someone would become diabetic in a certain year given their disease-ome in previous years.

4 The Conductome

A process-oriented perspective requires us to think in terms of temporal development - of change. In the case of living systems, change is most naturally thought of in terms of *behaviour*. It is behaviour that naturally links cause and effect, with behaviour being the natural response - effect - to external or internal stimuli - causes. Indeed, one may argue that it is the sole medium by which organisms interact with their environment, including with other organisms. It is clear that our genome codes not only the physical structures of an organism but also what they do, up to the collective behaviour of the organism as a whole, and beyond, to the collectivity of groups of individuals. However, it should be equally clear that behaviour leaves an imprint on the genome. To a large degree, the survival of an organism in its environment is associated with what it does. A behaviour that is apt for a given environment will be propagated, genetically, if it has at least a partial genetic origin, or culturally.

The majority of, if not all, biologically relevant behaviours, such as sleeping, eating, reproducing, evading predators, are linked to clocks and underlying physical rhythms, such as the circadian rhythm, and are biological responses to fundamental properties of the earth's motion in space and time. These rhythms are a fundamental part of the niche of almost all living organisms. Thus, we would argue, that it is behaviour that is the most natural link between the micro-omes, such as the genome, and more macro-omes, that are proxied by those variables that are the area of interest of disciplines such as epidemiology, sociology and psychology. Behaviour is both caused by genetic structures and functions and, in turn, leads to changes in those structures and functions. In the omic spirit we have posited the "Conduc-

tome"[1] - the totality of behaviours of an organism and the causative factors associated with them - as the most comprehensive link between micro-omes, such as the genome, and macro-omes. As with the micro-omes, the Conductome can be approached using Complex Networks by, for instance, considering the "interactions" between different behaviours, as well as their links to other factors. Similarly, they may be considered, in analogy with a disease-ome, in terms of a probability function, $P(C|\mathbf{X})$, where C is the behaviour of interest and $\mathbf{X}$ the set of corresponding factors linked to or predictive of C.

Like many deep concepts, behaviour is difficult to characterise precisely. Take as an example, thermoregulation in mammals [15]. Mammals have different responses to external temperature. A human may sweat or may take off a coat. An elephant may flap its ears while a dog may pant. Should we classify ear flapping, panting and coat removal as behavioural adaptations and sweating as a physiological adaptation? Sweating and panting are both controlled by the autonomic nervous system [16, 17]. What about taking off a coat? We would argue that a more general and appropriate characterisation of behaviour, such as "the internally coordinated responses (actions or inactions) of whole living organisms (individuals or groups) to internal and/or external stimuli" [18] would naturally classify them all as behaviours. What about eating? Is eating more like sweating or taking off a coat? Both, of course. There is a basal mechanism for generating the urge to eat that comes from the autonomic nervous system [19]. This is a natural view from a biological perspective. However, there is much debate about just how "automatic" eating is [20], that is linked to polemical issues such as "fat shaming" and free will [21]. Clearly, as a fundamental necessity of life, the desire to seek and obtain food is pre-programed genetically. However, what about: what we eat? how much we eat? when we eat? where we eat? etc. When we eat, for instance, is associated now with an entire field of study - chrononutrition [22], while there is ample evidence that certain food preferences, such as fatty and sugary foods [23] have a neurobiological link [24]. How much we eat - portion size - is another dimension that has a strong psychological and social component, if not a direct biological one.

Although there exist ontologies of behaviour, such as the Neuro Behaviour Ontology [25], food consumption offers a good example of the complexity of classifying behaviour. We may consider food consumption as a behaviour as a class to be predicted. Obviously the simple class $C = food\ consumption = YES/NO$ is not useful, as all humans must consume food. Indeed there are a set of underlying basal behaviours that are intimately associated with the fundamental properties of life, such as homeostasis, metabolism, reproduction and adaptation to the environment. Food consumption is vital for metabolic processes and to maintain homeostasis. However, beyond the pure classification of consumption = YES/NO, we may construct a multitude of classes of interest following the discussion above. For instance, C may represent overconsumption, as defined with respect to some baseline null hypothesis, consumption of a certain food type, consumption as classified through

[1] The Conductome was introduced in the international workshop "The Human Conductome: A New Paradigm for Understanding Obesity? in the C3 ? Centro de Ciencias de la Complejidad, UNAM 29-30th November 2018.

portion size, consumption by eating times, consumption by frequency, consumption by place or, indeed, any and all combinations of the above, and more. By considering different classes we may determine the degree of heterogeneity associated with these classes/behaviours. The set of predictors, **X** - genetic, epigenetic, physiological, psychological, social - then represent the Conductome for that behaviour. Project 42, alluded to above, is, in this sense, an attempt to construct a set of variables across multiple scales that may begin to approximate in certain dimensions a Conductome for those behaviours - overconsumption and sedentariness - that are particularly related to obesity and metabolic disease. Additionally, if we do not have direct observations of a particular behaviour, we may imagine constructing a Conductome indirectly, by taking as a class C a physiological state, such as obesity or hypertriglyceridemia, that we hypothesise that is correlated with a behaviour of interest, such as overconsumption of food. Of course, when speaking of behaviour, there is a natural structural element - the central nervous system and its description at multiple spatial scales, from the cellular to the cortical - that must be included as an intermediary between the causes of behaviour and the consequences of behaviour. In this sense a "neurome" will be an essential element in understanding the link between cause and consequence and back again.

As a final reflection, although the workshop focused on physiological rhythms - periodic and periodic-like processes in human physiology - many life processes and corresponding behaviours are not periodic. Such periodicity is particularly relevant when we consider those short time intervals that are dominated by the earth's principal spatio-temporal rhythms: day, month or season, and is naturally linked to reversibility, if one thinks of returning to an initial state. However, life and its associated behaviours have a cost. Although one may analyse such costs in much more sophisticated terms - of information, entropy and free energy [26] - for present purposes one can think of the costs in terms of "wear and tear". Life inevitably leads to wear and tear, and this can be measured along two principle dimensions - temporal extent and rate. All else being equal, a human of 50 years of age will exhibit more wear and tear than a human of 20 years of age [27, 28]. Similarly, twenty years of chronic stress and inflammation due to morbid obesity will be associated with a much higher rate of wear and tear than twenty years of abstemious living. Wear and tear at the physiological level is a result of behaviour. Organisms must feed, organisms must reproduce, organisms must survive in an uncertain environment.

Acknowledgements This work has been supported by DGAPA PAPIIT grants IG101520 and IV100520, CONACyT Fronteras grant FC-2015-2-1093, SECITI grant 093/2018 and a donation from Microsoft Academic Relations. The author is grateful for very fruitful discussions with Peter Gollwitzer, Rolando Diaz-Loving, Ruud Buijs, Carolina Escobar, Sachin Panda, Per Sodersten, Cecilia Bergh, Lucia Ledesma and other participants at the first Human Conductome Workshop and to participants at the Matrix workshop Mathematics of Physiological Rhythms.

References

1. Ballereau S. et al. (2013) “Functional Genomics, Proteomics, Metabolomics and Bioinformatics for Systems Biology”. In: Prokop A., Csukás B. (eds) Systems Biology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6803-1_1
2. Horgan, R.P. and Kenny, L.C. (2011) ?Omic? technologies: genomics, transcriptomics, proteomics and metabolomics. The Obstetrician & Gynaecologist, 13: 189-195. doi:10.1576/toag.13.3.189.27672
3. Haken, H. (2004) Synergetics. Introduction and Advanced Topics, Springer, Berlin.
4. Elena Kuzmin *et al* “Systematic analysis of complex genetic interactions” (2018) Science, Vol. 360, Issue 6386, eaao1729 DOI: 10.1126/science.aao1729.
5. Ivanov, P.C. and Bartsch, R.P. (2014) “Network physiology: mapping interactions between networks of physiologic networks.” In Networks of Networks: the last Frontier of Complexity (pp. 203-222). Springer.
6. Ivanov, Plamen Ch, Kang KL Liu, Aijing Lin, and Ronny P. Bartsch. (2017) “Network physiology: From neural plasticity to organ network interactions.” In Emergent Complexity from Nonlinearity, in Physics, Engineering and the Life Sciences, pp. 145-165. Springer
7. Ivanov, P., Zhang, X., & Lombardi, F. (2019) APS March Meeting Abstracts.
8. Lin, A., Liu, K.K.L., Bartsch, R.P. et al. Dynamic network interactions among distinct brain rhythms as a hallmark of physiologic state and function. Commun Biol 3, 197 (2020).
9. Black N, D’Souza A, Wang Y, et al. (2019) “Circadian rhythm of cardiac electrophysiology, arrhythmogenesis, and the underlying mechanisms”. Heart Rhythm; 16(2): 298-307. doi:10.1016/j.hrthm.2018.08.026
10. Erdmann, Jeanette et al. (2010) “Genetic causes of myocardial infarction: new insights from genome-wide association studies.” Deutsches Arzteblatt international vol. 107, 40: 694-9. doi:10.3238/arztebl.2010.0694
11. Zhe Wang, Bing Yu (2019) “Chapter 15 - Metabolomics, Proteomics, and Genomics: An Introduction to a Clinician”, Editor(s): Vijay Nambi, Biomarkers in Cardiovascular Disease, Elsevier, Pages 159-170, ISBN 9780323548359, https://doi.org/10.1016/B978-0-323-54835-9.00015-6.
12. Christopher R. Stephens, Hugo Flores Huerta and Ana Ruíz Linares (2018) “When is the Naive Bayes approximation not so naive?” Mach. Learn. 107:397?441 https://doi.org/10.1007/s10994-017-5658-0
13. van Hoek M, Dehghan A, Witteman JC, et al. (2008) “Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study”. Diabetes; 57(11): 3122-3128. doi:10.2337/db08-0425
14. Stephens C.R., Gutiérrez J.A.B., Flores H. (2020) Bayesian Classification of Personal Histories - An application to the Obesity Epidemic. In: Hassanien A., Azar A., Gaber T., Bhatnagar R., F. Tolba M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019). AMLTA 2019. Advances in Intelligent Systems and Computing, vol 921.
15. Jeremy Terrien, Martine Perret, Fabienne Aujard (2011) “Behavioral thermoregulation in mammals: a review”, Frontiers in Bioscience 16, 1428-1444.
16. Nisha Charkoudian, “Chapter 60 - Hypothermia and Hyperthermia” (2012) Editor(s): David Robertson, Italo Biaggioni, Geoffrey Burnstock, Phillip A. Low, Julian F.R. Paton, Primer on the Autonomic Nervous System (Third Edition), Academic Press, Pages 287-289.
17. Krönert H, Pleschka K (1976) “Lingual blood flow and its hypothalamic control in the dog during panting”. Pflugers Arch. 1976; 367(1):25-31. doi:10.1007/BF00583652
18. Levitis, Daniel; William Z. Lidicker, Jr; Glenn Freund (2009) “Behavioural biologists do not agree on what constitutes behaviour”. Animal Behaviour. 78 (1): 103-10. doi:10.1016/j.anbehav.2009.03.018
19. Guarino D, Nannipieri M, Iervasi G, Taddei S and Bruno RM (2017) “The Role of the Autonomic Nervous System in the Pathophysiology of Obesity”. Front. Physiol. 8:665. doi: 10.3389/fphys.2017.00665

20. Adela R. Moldovan, Daniel David (2012) "Features of automaticity in eating behavior", Eating Behaviors, Volume 13, Issue 1, Pages 46-48.
21. David A Levitsky and Carly R Pacanowski (2012) "Free will and the obesity epidemic", Public Health Nutrition: 15(1), 126-141 doi:10.1017/S1368980011002187
22. Johnston, J. D., Ordovás, J. M., Scheer, F. A., & Turek, F. W. (2016) "Circadian Rhythms, Metabolism, and Chrononutrition in Rodents and Humans". Advances in nutrition (Bethesda, Md.), 7(2), 399-406. https://doi.org/10.3945/an.115.010777.
23. Adam Drewnowski, M.R.C. Greenwood (1983) "Cream and sugar: Human preferences for high-fat foods", Physiology & Behavior, Volume 30, Issue 4, Pages 629-633.
24. Allen S. Levine, Catherine M. Kotz, Blake A. Gosnell (2003) "Sugars and Fats: The Neurobiology of Preference", The Journal of Nutrition, Volume 133, Issue 3, Pages 831S?834S.
25. Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B; NCBO team (2012) "The National Center for Biomedical Ontology". J Am Med Inform Assoc. Mar-Apr;19(2):190-5. Epub 2011 Nov 10.
26. Erwin Schrödinger (1944) "What Is Life? : The Physical Aspect of the Living Cell". Based on lectures delivered under the auspices of the Dublin Institute for Advanced Studies at Trinity College, Dublin, in February 1943.
27. Stephens Christopher R., Easton Jonathan F., Robles-Cabrera Adriana, Fossion Ruben, de la Cruz Lizbeth, Martínez-Tapia Ricardo, Barajas-Martínez Antonio, Hernández-Chávez Alejandro, López-Rivera Juan Antonio, Rivera Ana Leonor (2020) "The Impact of Education and Age on Metabolic Disorders", Frontiers in Public Health 8, page 180; URL=https://www.frontiersin.org/article/10.3389/fpubh.2020.00180.
28. Antonio Barajas Martinez, Jonathan F. Easton, Ana Leonor Rivera, Ricardo Jesus Martinez Tapia, Lizbeth De la Cruz, Adriana Robles Cabrera, Christopher R. Stephens (2020) "Metabolic Physiological Networks: The Impact of Age", medRxiv 2020.08.05.20168997; doi: https://doi.org/10.1101/2020.08.05.20168997

Delay-differential equations for glucose-insulin regulation

Maia Angelova, Sergiy Shelyag

Abstract In this work, a model based on a system of delay differential equations, describing a process of glucose-insulin regulation in the human body, is studied numerically. For simplicity, the system is based on a single delay due to the practical importance of one of the two delays appearing in more complex models. The stability of the system is investigated numerically. The regions, where the solutions demonstrate periodicity and asymptotic stability, are explicitly calculated. The sensitivity of the solutions to the parameters of the model, which describes the insulin production in the system, is analysed.

1 Introduction

Delay differential (and, generally, functional differential) equations (DDEs) and their systems appear in natural and artificial phenomena, when the behaviour of a system explicitly depends both on its current state and its history in some functional form. Among such systems are communication networks, systems of biological and physiological regulations, population growth, infection spread, epidemics and pandemics, devices with actuators and delayed feedback, business cycle models in economics, decision making [1]. Unlike ordinary differential equations and their systems, which are finite-dimensional in phase space, DDEs are infinitely-dimensional. Inclusion of a delay in a dynamical system can lead to rather complicated dynamics, (sometimes unwanted) oscillations and even chaos. Analysis of DDEs is generally more involved, in part due to the structure of the corresponding characteristic equations, and often not allowing for an analytical treatment. Numerical solution of such equations is also not trivial due to propagating discontinuities and strict require-

Maia Angelova
Deakin University, Geelong, Victoria, Australia e-mail: maia.a@deakin.edu.au

Sergiy Shelyag
Deakin University, Geelong, Victoria, Australia, e-mail: sergiy.shelyag@deakin.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_17

ments for interpolation techniques. Nevertheless, in the recent years advances in understanding of DDEs, and analytical and computational approaches to their solution have been achieved.

It is well-known that the chemo-biological process of conversion of insulin into glucose necessarily involves a delay, which also depends on a number of physiological parameters. Only a limited number of these directly involved in glucose-insulin regulation system can be observed directly, and mathematical modelling would help estimate these parameters [2]. Furthermore, the interactions between the sub-systems of the glucose-insulin regulation may be affected by a variety of disorders and diseases, such as diabetes of multiple types [3]. Therefore, further detailed study of the system of the glucose-insulin regulation and its mathematical counterparts is warranted for better understanding of the human physiology.

In this paper, we numerically analyse the behaviour of a system of delay-differential equations, which aims to simulate the interactions in a model of glucose-insulin regulation in the human body. We study only one of the interactive terms of the system in detail (namely, the term, which describes the glucose-sensitive insulin production) and demonstrate the presence of periodic and asymptotically-stable solutions.

2 The Model

The system of delay-differential equations, which describes the glucose-insulin system regulation in human body, was introduced by [4, 5]. Its mathematical properties have been extensively studied for one- or two-delay modifications [3, 4, 6, 7, 8, 9]. We are using one-delay system given below:

$$\begin{aligned} I'(t) &= f_1(G(t)) - \frac{1}{\tau_0} I(t) \\ G'(t) &= G_{in} - f_2(G(t)) - qG(t) f_4(I(t)) + f_5(I(t-\tau)), \end{aligned} \tag{1}$$

Here, I and G are the insulin and glucose blood concentrations, respectively. The first term f_1 in the insulin equation is the insulin secretion caused by glucose intake (the effect of this term on the solution of the system will be studied in a greater detail as an example), the second term is the insulin degradation with the time scale τ_0. In the glucose equation, G_{in} is the constant glucose intake, f_2 is the constant glucose utilisation by the organism, the third term is the glucose utilisation dependent on the insulin concentration, and the fifth term f_5 is the glucose production from insulin, which includes a positive delay τ. The flowchart of the model is given in Fig. 1

The system does not allow for exact analytical solutions, therefore we produce numerical solutions for System (2). Also, we perform solution scans over the parameter ranges to determine whether the System 2 exhibits periodic or asymptotically stable behaviour.

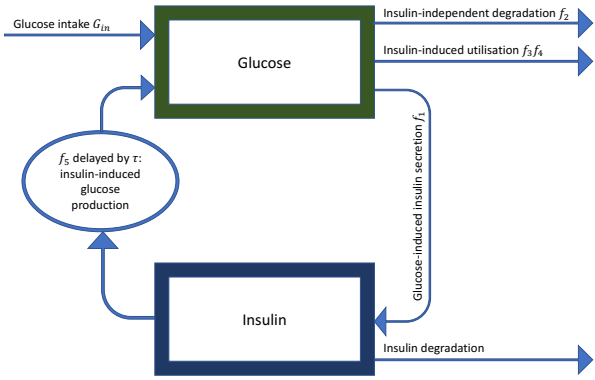


Fig. 1 The flow chart of the model. There are two processes related to production of insulin dependent on glucose concentration and production of glucose dependent on the insulin concentration. These processes create a loop between glucose and insulin compartments in the flow chart. All other processes remove glucose or insulin from the system, or add external glucose through the glucose intake term.

3 Periodic and asymptotically stable behaviour in glucose-insulin regulation system

The IVP system (2) is solved using a 4-th order Runge-Kutta method with an adaptive time step and 4-th order barycentric Lagrange interpolation of the delay term [10]. The numerical solution is therefore 4-th order precise on time. The functions f_1-f_5 are chosen as continuously differentiable, non-negative and Lipschitz-bounded on $\mathbb{R}$, and f_5 also satisfies the negative feedback condition. The functions are chosen as follows (see figure 2): $f_1(u) = a_0 + aH(u)$, $f_2(u) = bH(u)$, $f_3(u) = -qH(u)$, $f_4(u) = d + eH(u)$, and $f_5(u) = h(1 - H(u))$, where

$$H(u) = \frac{u^{N_\mathrm{H}}}{u^{N_\mathrm{H}} + 1} \tag{2}$$

is Hill function. The initial conditions are $I(0) = G(0) = 0$. Also, we set $I(t < \tau) = 0$. The latter choice, as numerical experiments demonstrate, does not change the character of the solution.

Figs. 3 and 4 show examples of the obtained periodic and asymptotically stable solutions, respectively. The parameters chosen to produce the numerical solutions are as follows: $N_\mathrm{H} = 2$, $G_{in} = 1$, $\tau_0 = 1$, $\tau = 5$, $q = 1$, $a_0 = 1$, $b = 1$, $d = 10$, $e = 10$, $h = 100$. Setting the parameter $a = 1$ led to an asymptotically stable solution, while with $a = 10$ the system demonstrated periodic behaviour. Automated differentiation between the periodic and stable solution types represents some difficulty due to

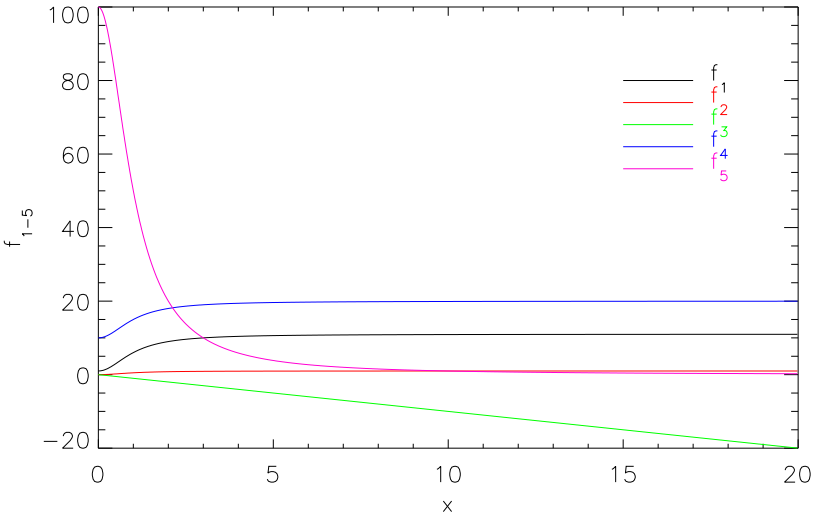


Fig. 2 Functions f_1 - f_5 as used in the numerical solution of System (2).

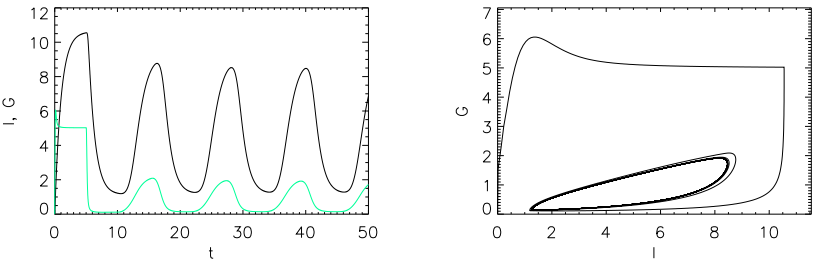


Fig. 3 An example of periodic solution for the system (2). The time evolution of I (black) and G (green) is shown in the left panel. The right panel shows the corresponding phase portrait for the system, plotted for a larger time interval $0 < t < 200$.

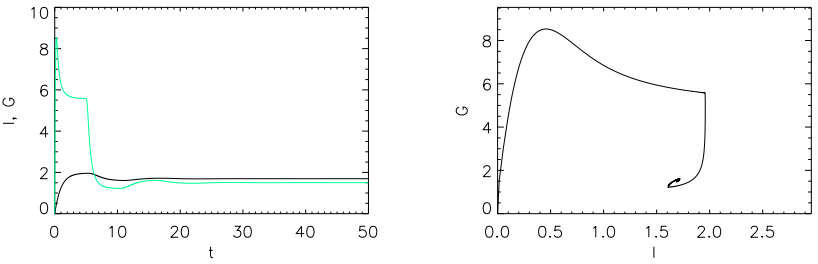


Fig. 4 Same as in Fig. 3, but for an asymptotically stable solution for the System (2).

the *a priori* unknown solution period and the time scale of amplitude decrease of the solution oscillations. This is done as follows. First, all the local extrema of the solution are located. If the number of local extrema is less than 4 (corresponding to two periods, if solution is periodic), then the solution is assumed to be stable. Otherwise, solution values are found at the positions of the solution extrema. Then, the even and odd pairs of these values are compared, and, if the difference between them is greater than some value, the solution is assumed to be stable. All other solutions are assumed periodic.

This is further demonstrated in Fig. 5, which shows solution types for System (2) on a range of scanned parameters. The scan was performed over a number of values for the Hill parameter N_{H}, a range of the delays $\tau = 0.1 - 10$ and the parameter $a = 0 - 10$. All other parameters are as used above for solution of the system (2). We also note that the system exhibits asymptotic stability over the whole range of the used parameters for the smaller values of $N_{\mathrm{H}} = 0.5,\ 1$. Only a part of the scanned parameter space is shown in the figure. The figure clearly demonstrated that the stronger the non-linearity of the Hill functions in $f_1 - f_5$ is, the wider the range of delays τ and parameter a leads to oscillatory behaviour of the solution.

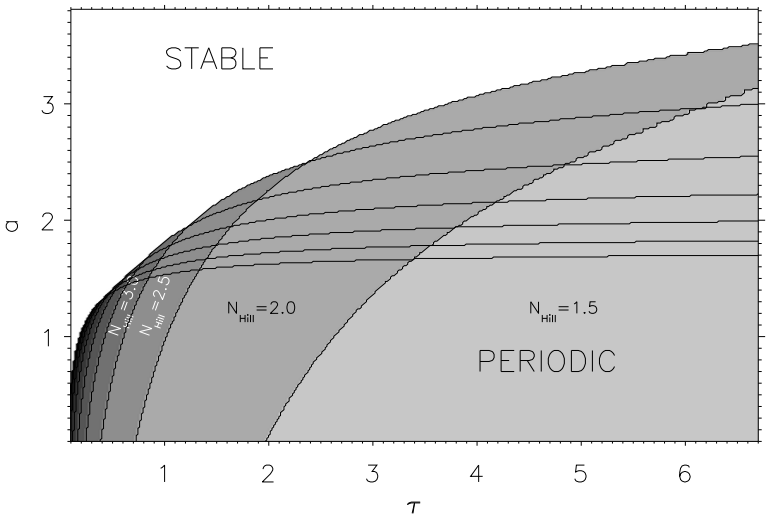


Fig. 5 Solution types obtained from numerical solution of the system (2) over a range of τ and a parameters for different values of Hill parameters N_{H}.

A scan over a range of values of the glucose intake G_{in} has also been performed for the different non-linearity indexes of the Hill function. Two-dimensional scan of the solution types, which also includes the parameter a, is shown in Fig. 3. The figure shows that, again, the stronger non-linearity (steepness) in the Hill functions provokes oscillatory behaviour in System (2) solutions. However, with the increase

of the amplitude of the non-linear part in $f_1 = a_0 + aH(u)$, the region, where periodic solution occur, shrinks.

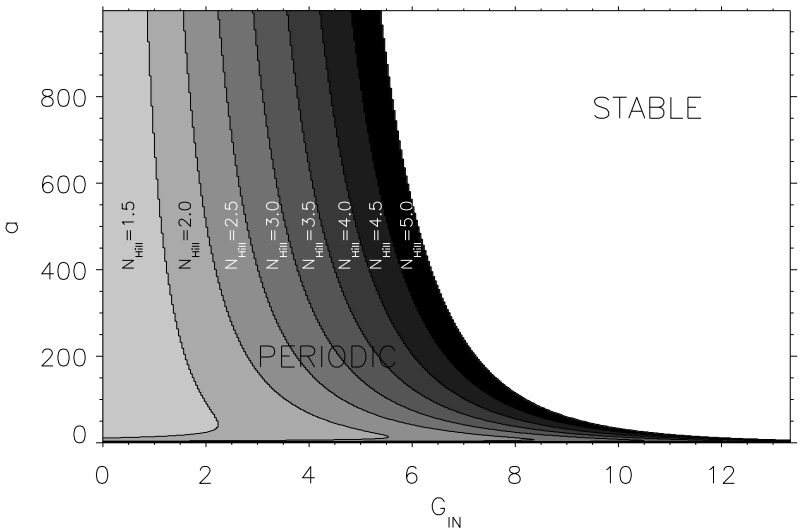


Fig. 6 Solution scan of System (2) on the glucose intake G_{in} and parameter a for a range of the non-linearity indices N_H.

Another scan has been performed on the parameter a_0 and G_{in} (shown in Fig. 3). Again, the oscillatory region widens in G_{in} and shrinks in a_0 if the non-linearity in the Hill functions increases.

4 Conclusions

In this work we studied the stability of the DDE system, which describes a model of glucose-insulin regulation system. We analysed the dependence of the solution types (periodic or asymptotically stable) on the parameters of the glucose-dependent insulin production. We have shown that the stronger the nonlinearity in the Hill functions, which describe the components of the glucose-insulin regulation system, the wider the parameter range (which includes the delay parameter) for which the oscillatory behaviour is observed.

Further study is required for precisely diagnosing the behaviour of the system and connecting it to the physiologically measurable parameters. Also, mathematically, the system of glucose-insulin regulation exhibits both periodic and asymptotically stable solutions. However, normally, only periodic behaviour of the insulin and glucose concentrations is observed in the test environments. It would be interesting to

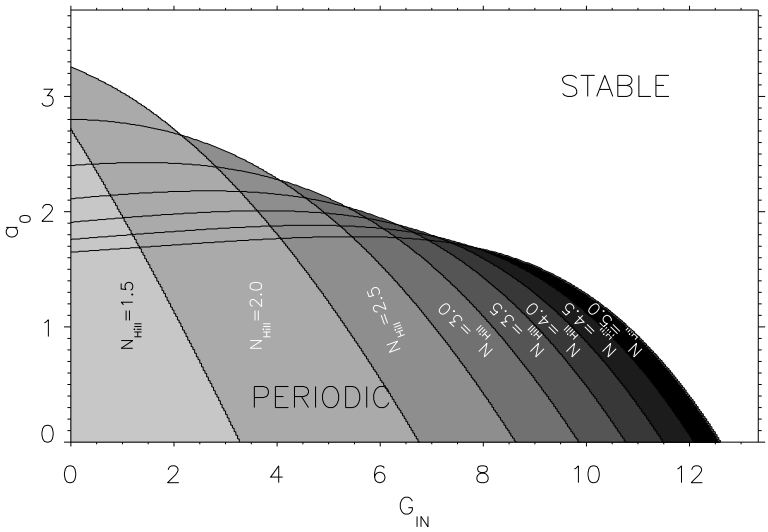


Fig. 7 Solution scan of System (2) on the glucose intake G_{in} and parameter a_0 for a range of the non-linearity indices N_H.

get a better insight into the existence of physiological equivalents of the asymptotically stable solutions of the System (2).

References

1. Kyrychko, Y., and Hogan, S.: On the Use of Delay Equations in Engineering Applications. Journal of Vibration and Control, 16 (2010)
2. Marchetti, L., Reali, F., Dauriz, M., et al.: A Novel Insulin/Glucose Model after a Mixed-Meal Test in Patients with Type 1 Diabetes on Insulin Pump Therapy. Scientific Reports, 6, 36029 (2016)
3. Huard, B., Bridgewater, A., and Angelova, M.: Mathematical investigation of diabetically impared ultradian oscillations in the glucose-insulin regulation. J. Theor. Biology, 418, 66-76 (2017)
4. Bennett, D. L., and Gourley, S. A.: Global stability in a model of the glucose–insulin interaction with time delay. Euro. Jnl of Appl. Math. 15, 203–221 (2004)
5. Li, J., Kuang. Y., and Mason, C.: Modeling the glucose-insuline regulatory system and ultradian insulin secretory oscillations with two time delays. J. Theoret. Biol. 242, 722-735 (2006)
6. Bennett, D. L., and Gourley, S. A.: Periodic oscillations in a model of the glucose–insulin interaction with delay and periodic forcing. Dynamical Systems, 19(2), 109-125 (2004)
7. Bennett, D. L., and Gourley, S. A.: Asymptotic properties of a delay differential equation model for the interaction of glucose with plasma and interstitial insulin. Applied Mathematics and Computation, 151, 189-207 (2004)
8. Huard, B., Easton, J. F., and Angelova, M.: Investigation of stability in a two-delay model of the ultradian oscilaltions in glucose-insulin regulation. Commun. Nonlinear Sci. Numer. Simulat. 26, 211-222 (2015)

9. Angelova, M., Beliakov, G., Ivanov, A., Shelyag, S.: Global Stability and Periodicity in a Glucose-Insulin Regulation Model with a Single Delay. arXiv:2008.11019 (2020)
10. Berrut, J.-P., & Trefethen, L. N.: Barycentric Lagrange Interpolation. SIAM Review, 46, 501 (2004)

Chapter 6

Conservation Laws, Interfaces and Mixing

Effect of adiabatic index on Richtmyer-Meshkov flows induced by strong shocks

Cameron E. Wright and Snezhana I. Abarzhi

Abstract Richtmyer-Meshkov Instability (RMI) is an instability that develops at the interface between fluids of contrasting densities when impacted by a shock wave. Its applications include inertial confinement fusion, supernovae explosions, and the evolution of blast waves. We systematically study the effect of the adiabatic index of the fluids on the dynamics of strong-shock driven flows, particularly the amount of shock energy available for interfacial mixing. Only limited information is currently available about the dynamic properties of matter at these extreme regimes. Smooth Particle Hydrodynamics simulations are employed to ensure accurate shock capturing and interface tracking. A range of adiabatic indexes is considered, approaching limits which, to the best of the author's knowledge, have never been considered before. We analyse the effect of the adiabatic indexes on the interface speed and growth-rate immediately after the shock passage. The simulation results are compared, wherever possible, with rigorous theories and with experiments, achieving good quantitative and qualitative agreement. We find that the more challenging cases for simulations arise where the adiabatic indexes are further apart, and that the initial growth rate is a non-monotone function of the initial perturbation amplitude, which holds across all adiabatic indexes of the fluids considered. The applications of these findings on experiment design are discussed.

1 Introduction

Richtmyer-Meshkov instability (RMI) is a phenomenon in fluid mechanics that describes the evolution of an interface between two fluids of distinct acoustic impedance and distinct densities when a shock wave impacts the interface. The flow evolution is shown in **Figure 1**, where the light (red) fluid travels with velocity to-

Snezhana I. Abarzhi
School of Mathematics and Statistics, The University of Western Australia
e-mail: snezhana.abarzhi@gmail.com

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_18

wards the interface (light green) and heavy fluid (blue) and the Richtmyer-Meshov instability develops as the interface between the two fluids changes shape and size over time. ([Richtmyer, 1960]; [Meshkov, 1969]). If the interface between the fluids is given an initial perturbation a_0, (seen on the right in **Figure 1**) the interface amplitude increases in size with growth-rate v_0 as the wave travels and evolves into a large-scale coherent structure of bubbles and spikes (bottom right of **Figure 1**) ([Abarzhi, 2010]; [Abarzhi, 2008]).

1.1 Motivation

Richtmyer-Meshkov instability (RMI) appears in a variety of processes in high energy density plasmas, controlling fluid transformation under strong impact, governing the formation of hot spots in inertial confinement fusion, determining energy transport in core-collapse supernova, and strongly influences the evolution of blast waves and explosions ([Meshkov, 1969]; [Richtmyer, 1960]). RMI forms in situations characterized by strong impact shocks, sharply and quickly changing flow fields, and by small effects of dissipation and diffusion, often producing small scale structures ([Abarzhi, 2010]). Interaction of a shock wave with a density discontinuity such as in the situation in this work may result in the development of RMI and in extensive interfacial mixing ([Meshkov, 1969]; [Richtmyer, 1960]). Since RMI plays such a large role in these applications, the ability to understand and control this instability is very important.

To the best of the author's knowledge, previous numerical simulations investigating the dynamics of RMI flows have only contained ideal monotonic gases. This study aims to investigate RMI flows for gases with more than one atom per molecule, e.g. diatomic or triatomic gases such as O_2 or H_2O. This is significant for investigation, as it will enable RMI studies to have a wider field of application as it will be able to more accurately model situations with high-speed non-monatomic gases, and aid in control of the instability through the initial parameter set-up. The applications include rocket thrust flow (such as those in scramjets), inertial confinement fusion, and explosion blast waves ([Drake, 2009]; [Bodner et al., 1998]; [Zel'dovich, 1967]).

1.2 Approach

Sometimes, in applications such as inertial confinement fusion, the effects of the Richtmyer-Meshkov instability are undesirable, and it is necessary to control the instability's evolution ([Lindl et al., 2004]). Some methods of achieving this include suppressing RMI completely, or controlling it through adjusting the initial parameters of the system ([Anisimov et al., 2013]; [Abarzhi, 2010]; [Demskoĭ et al., 2006]). In order to do this, information is required about the effect the initial parameters

have on the development of the instability. Of particular interest in controlling the development of the instability is the amount of energy available for interfacial mixing deposited into the interface from the shock wave. This is one of the aims of this study.

The volume of physical experimental data of RMI produced by strong shocks is sparse as the experiments require challenging control of flow implementation and diagnostics. ([Motl et al., 2009]; [Orlicz et al., 2009]; [Jacobs and Krivets, 2005]). Therefore, numerical modeling of RMI is a powerful tool to aid in designing and building systems in which RMI is present. However, the dynamics of RMI are complex and a numerical model should be able to manage numerous competing requirements, such as shock capturing, interface tracking, and accurate accounting for the dissipation processes ([Stanic et al., 2012]; [McFarland et al., 2011]; [Herrmann et al., 2008]; [Dimonte et al., 2004]).

Using a hydrodynamic approximation, we systematically study a broad spectrum of the parameter regime and its influence on the fraction of energy available for interfacial mixing in RM flow. To do this, we will obtain data on three variables of the flow- the interface speed, the interface growth rate, and the initial curvature of the front of the interface. These variables inform us to how much mixing of the interface occurs, and how well the numerical simulations capture small scale structures, which the simulations must do well in order to obtain data on the interfacial mixing. The results of each of these variables are compared with rigorous theoretical theories, finding good quantitative and qualitative agreement.

1.3 RMI Dynamics

RMI develops when a shock impacts an interface between two fluids of differing densities and the energy is distributed throughout the fluids ([Aleshin et al., 1990]). This dissertation will only focus on 2 dimensional RMI case, and with the shock propagating from the light fluid into the heavy fluid. When the shock hits an ideally planar interface (without any perturbation), it splits into a reflected shock travelling back through the light fluid and a transmitted shock travelling through the heavy fluid ([Stanic et al., 2012]; [Herrmann et al., 2008]; [Demskoĭ et al., 2006]). The bulk of the fluid influenced by the shock impact (the bulk between the reflected shock and the transmitted shock, including the fluid interface) all moves with the same velocity v_∞, called the background velocity, seen in **Figure 1**.

The velocity v_∞ quantifies the amount of energy transferred into the fluid bulk by the shock and is a function of the shock strength and fluid properties ([Stanic et al., 2012]; [Richtmyer, 1960]).

If the interface between the two fluids is perturbed (no longer planar), the bulk fluid containing the interface still moves with velocity v_∞, but the interface itself now has growth-rate v_0 due to the impulsive acceleration induced by the shock ([Meshkov, 1969]; [Richtmyer, 1960]). Arrows mark the direction of fluid motion at the tip of the bubble (right) and spike (left). Eventually, the bubble and spike

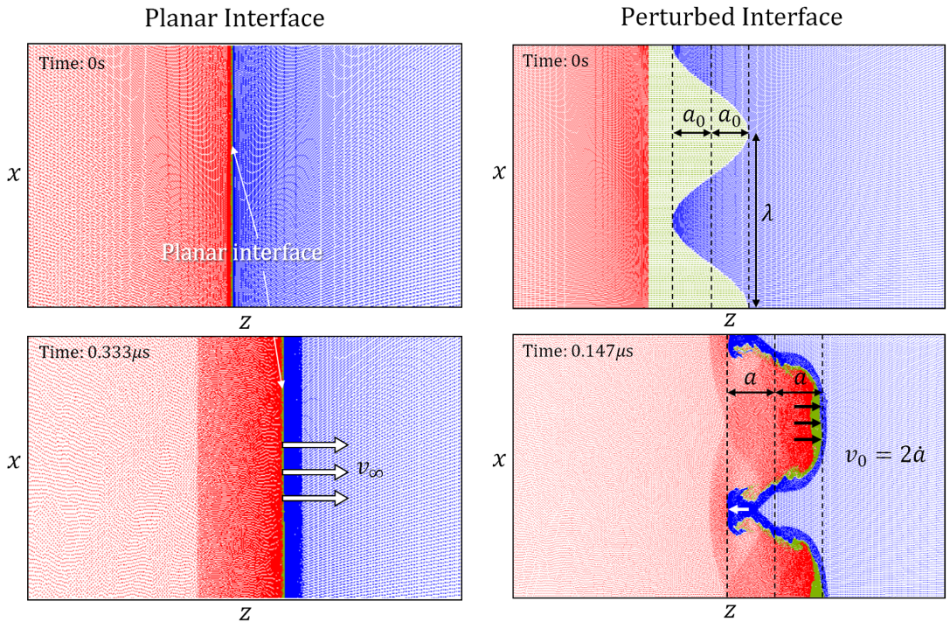


Fig. 1 Evolution of the Richtmyer-Meshkov instability for both the planar interface and perturbed interface cases. In both cases, the interface moves at speed v_∞, and the interface also grows with speed v_0 in the perturbed interface case, whereas it stays flat in the planar interface case.

growth decelerates, and small scale structures appear on the sides of the developing spikes ([Stanic et al., 2012]).

1.4 Parameters of the System

1.4.1 Mach Number

The Mach number, denoted M, is defined as the ratio of the speed of a shock to the speed of sound in the light fluid $c_l = 2.039 \times 10^3$m/s ([Richtmyer, 1960]). For weak shocks, $M \approx 1$, the background motion (the motion of the whole body of fluid) is subsonic relative to the light fluid, $v_\infty/c_l < 1$, where c_l is the speed of sound in the light fluid. For $M \approx 3$ the ratio is $v_\infty/c_l \approx 1$. For shocks with $M \approx 5$ the background motion is supersonic, $v_\infty/c_l > 1$, and for shocks with $M > 7$, the background motion is hypersonic, $v_\infty/c_l \gg 1$ ([Stanic et al., 2013]; [Stanic et al., 2012]; [Herrmann et al., 2008]).

1.4.2 Atwood Number

The Atwood number, denoted A, describes the density difference between two adjacent fluids with a common interface.

$$A = \frac{\rho_h - \rho_l}{\rho_h + \rho_l} \tag{1.4.1}$$

where ρ_h, ρ_l are the densities of the heavy and light fluids respectively.

1.4.3 Adiabatic Index

The adiabatic index of a substance, denoted γ, can be understood from three perspectives. From a thermodynamic point of view, the adiabatic index gives an important relation for an adiabatic process of an ideal gas:

$$PV^{\gamma} = \text{constant} \tag{1.4.2}$$

where P, V, and γ is the pressure, volume, and adiabatic index respectively of the fluid.

It can also be understood as the ratio of the heat capacity at constant pressure C_P to the heat capacity at constant volume C_V,

$$\gamma = \frac{C_P}{C_V}. \tag{1.4.3}$$

From a molecular dynamics point of view, the adiabatic index can also be related to the degrees of freedom f of a molecule as

$$\gamma = 1 + \frac{2}{f}. \tag{1.4.4}$$

To the best of the author's knowledge, previous simulation analyses of the dynamics of RMI flows have been conducted with an adiabatic index of $\gamma = 5/3$, which is the value for ideal monotonic gases. Its value decreases for gases with more than one atom per molecule, e.g. for diatomic gases $\gamma = 7/5$. While γ is known to have a strong influence on the flow dynamics, no systematic study has been undertaken on the effect of γ on the dynamics. The gases analysed in this study are theoretical ones, where we vary the adiabatic index of the heavy and light fluids systematically instead of choosing particular gases.

1.5 Parameter Regime

The parameter regime we investigate is for the Mach and Atwood numbers, $(M, A) = (5, 0.8)$. This pair has been well documented in previous studies ([Stanic et al., 2012];

[Dell et al., 2015]). We vary the adiabatic index of the heavy and light fluid $(\gamma_l, \gamma_h) = (1.2, 1.3, \cdots, 1.6)$. For each pair of γ_l and γ_h, the amplitude of the initial perturbation of the interface between the two bulk fluids was varied from 0% to 100% of the interface wavelength, ie $a_0/\lambda = (0, 0.1, 0.2, \ldots, 1)$ where a_0 is the amplitude of the sinisoidal initial perturbation, and λ is the wavelength of the perturbation. This is a well studied regime, and has been documented well in the past, allowing us to compare our results with previous studies ([Dell et al., 2015]; [Stanic et al., 2012]; [Stanic et al., 2013]).

This regime contains 275 cases, and we ran a numerical simulation for each case. The average simulation takes 36 hours to run, making a total of about 10,000 hours. The simulations were run on three Windows laptop computers with i7 processors and with 8GB, 12GB, and 16GB of RAM.

2 Methods

2.1 Theoretical Approaches

In this work, we compare the results from our numerical simulations to analytical solutions. This analysis takes different forms depending on the progression of the instability. In the initial linear regime of RMI, the interface perturbation grows at a constant rate v_0, which is a function of the amplitude a_0 and wavelength λ of the initial perturbation ([Richtmyer, 1960]). In the following nonlinear regime, the interface perturbation growth-rate decreases and a large coherent structure of spikes and bubbles appears ([Abarzhi, 2010]; [Abarzhi, 2008]). The heavy fluid penetrates the light fluid in spikes as seen on the bottom right of **Figure 1**. As they travel, the spikes decelerate and small scale structures form on the sides of the spikes ([Stanic et al., 2012]).

2.1.1 Zeroth-order theory

An important parameter of RMI dynamics is v_∞, the magnitude of the velocity (hereafter: velocity) of the bulk, or the background motion. This value can be precisely calculated by zeroth-order theory from the conditions of the conservation of mass, momentum, and energy, and the equations of state of the fluids ([Richtmyer, 1960]). For ideal gases, it is a function of the initial shock's Mach number, the adiabatic index of the fluids $\gamma_{h(l)}$, and the Atwood number, $v_\infty = v_\infty(M, A, \gamma_{h(l)})$, and it is useful because it quantifies the amount of energy deposited by the shock into the fluid bulk ([Stanic et al., 2012]). The analysis of the numerical simulations becomes much simpler once v_∞ is obtained and used as the characteristic time scale so that the frame of reference is moving at speed v_∞ ([Dell et al., 2015]).

2.1.2 Linear theory

Another important parameter in RMI dynamics is the initial growth-rate v_0 of the interface. It is a function of M, A, $\gamma_{h(l)}$, and the initial perturbation amplitude and wavelength, $v_0 = v_0(M, A, \gamma_{h(l)}, a_0, \lambda)$ ([Stanic et al., 2012]; [Nishihara et al., 2010]; [Holmes et al., 1999]). For very small amplitude ($a_0/\lambda = 10^{-2}$ or smaller), v_0 is precisely calculated by linear theory, and grows linearly with a_0, $v_0 \approx \frac{a_0}{\lambda} M c_l$ ([Nishihara et al., 2010]; [Wouchuk, 2001]; [Richtmyer, 1960]).

For moderately small values of a_0, the growth rate v_0 becomes non-linear and has been calculated in previous studies by [Velikovich et al., 2014]; [Nishihara et al., 2010], and [Velikovich and Dimonte, 1996]. For larger values of a_0, the rate v_0 may grow with a_0 even slower than the linear and weakly nonlinear theory predict ([Stanic et al., 2012]; [Holmes et al., 1999]).

2.1.3 Highly nonlinear theory

In the late stages of Richtmyer-Meshkov dynamics v_0 has been calculated with group theory consideration ([Abarzhi, 2010]; [Abarzhi et al., 2003]; [Abarzhi, 2002]). At this late stage, the bubbles decelerate and flatten, and there is almost no fluid motion away from the interface, and extensive interfacial mixing occurs ([Stanic et al., 2013]; [Stanic et al., 2012]; [Herrmann et al., 2008]). These late-time dynamics are a complex problem, and many features require better understanding.

2.2 Smoothed Particle Hydrodynamics Simulations

Numerical modelling of RMI in these extreme conditions is a difficult task because the method should be able to accurately handle large speeds, strong shocks, and preserve small scale structures with high precision and accuracy. These small scale structures are embedded in large scale dynamics, moving at high speeds, so the order of precision required is very large ([Anisimov et al., 2013]). To model these complex dynamics we have employed the Smoothed Particle Hydrodynamics code (SPHC), which is an open-source code written in C developed by Dr. Stellingwerf and has free access to a complete set of validation test cases ([Stellingwerf, 1991]). This code has been widely used and tested on a broad variety of shock and flow problems, including RMI dynamics, the Noh problem, and with problems involving plasmas, complex materials, and multiphase flows ([Stanic et al., 2013]; [Stanic et al., 2012]; [Monaghan, 2005]; [Lucy, 1977]; [Stellingwerf, 1990]). Particularly, it has been used by NASA to investigate the Space Shuttle Columbia incident ([Stellingwerf et al., 2004]). SPHC conserves momentum, angular momentum, mass, and energy globally and locally and reflects particles on the boundaries in order to produce the correct boundary solutions, accurate to within 0.001% ([Stellingwerf, 1990]).

2.2.1 SPH Technique

SPH is a grid-free method that represents a continuous fluid with fixed-mass SPH particles, which are each represented by a mathematical basis function (or kernel) ([Stellingwerf, 1990]). In essence, SPHC keeps track of a large array of particles and for each time step, and calculates each interaction between all particles.

2.2.2 SPHC Simulation Setup

In this study, the computational setup is the standard similar to that in [Stanic et al., 2012] in order to extend the results of previous studies.

The amplitude is set at the start of every run, and when the shock hits the interface, the interface amplitude is compressed and then grows as RMI develops. We want to obtain the initial growth-rate of the interface, so we locate the first minimum of the amplitude and take v_0 to be the slope of the linear regression line from the first minimum amplitude over the next few initial data points, as seen in **Figure 2**. Note that v_0 is defined as the time derivative of the difference of the initial positions of the bubble and spike, which is twice the amplitude, as in [Stanic et al., 2012].

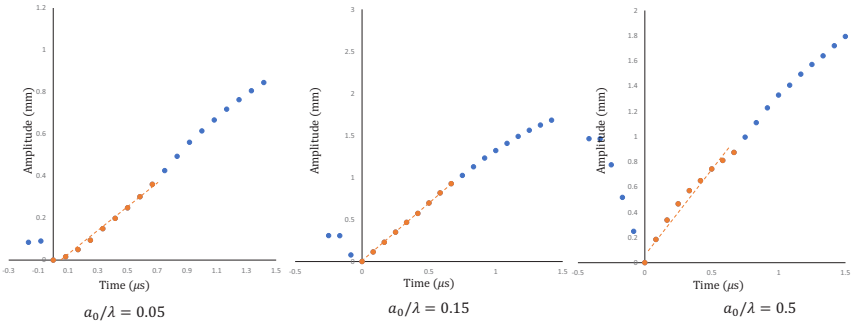


Fig. 2 The method by which v_0 is measured. We take the gradient of the the data points where $0 < t < 0.8\tau$ (marked in orange). The cases displayed are for $\gamma_l = 1.5$, $\gamma_h = 1.5$.

However, an issue arises when choosing what data points to include, because the value of v_0 changes significantly depending on where the cutoff for the initial growth is defined. To solve this problem, we notice that the shape of the amplitude-time curve for early time changes depending on the value of the initial perturbation, marked in orange in **Figure 2**. For $a_0/\lambda = 0.1$ the curve is concave, for $a_0/\lambda = (0.2, 0.3)$ the curve is linear, and for $a_0/\lambda = (0.4, 0.5, \cdots, 1)$ the curve is convex, forming a bubble like shape. Each of these shapes have clearly defined endpoints, marked on the diagram by the first blue points: the first finishes at the inflection point (the transition from concave to convex), the second where the growth starts to become non linear, and the third finishes at the end of the convex "bubble" (and

where the second bubble begins). These endpoints for all three shapes all finish at 0.8τ, where $\tau = v_\infty/\lambda$, giving $0.8\tau = 6.67 \times 10^{-7}$. These patterns are consistent across all different values of $\gamma_{h(l)}$ and a_0/λ considered, and this method produces good results, which indicates that this is a good choice.

To find v_∞, we measure the speed of the interface in the ideally planar interface case, with $a_0/\lambda = 1 \times 10^{-12} \approx 0$ (which is the smallest perturbation the simulation software would allow). The position of the interface is taken to be the leftmost position of the interface, as in [Dell et al., 2015]. The planar interface position motion is almost perfectly linear in time, allowing very accurate calculation of v_∞ by taking the slope of the linear regression line of the interface position over time. When the interface is not planar, the growth of the interface interferes with the measurement of the speed of the bulk, making accurate measurements difficult. However, because the amplitude of the interface doesn't affect the speed of the bulk, the values of v_∞ calculated for the planar case are taken to be the same for cases where $a_0/\lambda \neq 0$ ([Dell et al., 2015]).

3 Results

3.1 Background Motion for Planar Interface Case

Using the SPHC simulations and the method described above, we calculated the value of v_∞ for the cases $(M,A) = (5,0.8)$, with γ_h and γ_l ranging from $1.2, 1.3, \cdots$, 1.6. The results of these calculations are shown in **Figure 3**. The shape of the curves formed for each γ_h value when γ_l ranges from 1.2 to 1.6 appears to be decreasing in a non-linear fashion, which may be approaching a value asymptotically. In order to determine the behaviour accurately, more data points are required and may need further investigation in future studies.

We compare these results to the analytical calculations produced by zero-order theory [Richtmyer, 1960], where v_∞ can be calculated precisely from zero-order theory in the planar interface case (when the initial perturbation is 0) ([Stanic et al., 2012]; [Richtmyer, 1960]). The percentage error between the simulation results and the zero-order theory is shown in **Table 1**.

γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6
1.2	0.641	4.87	9.52	12.9	16.7
1.3	5.75	0.732	3.56	7.54	10.5
1.4	9.75	4.87	0.597	-2.79	5.68
1.5	12.8	8.04	4.00	0.822	-2.15
1.6	15.3	10.5	6.74	3.68	0.836

Table 1 Percentage error for v_∞ calculated from the SPHC simulations for the cases $a_0/\lambda = 0, \gamma_l, \gamma_h = (1.2, 1.3, \cdots, 1.6)$ when compared to the zero-order theory predictions. Values with high error ($> 7\%$) are marked in red.

We found that when gamma heavy and gamma light are close, simulation agreement with the linear series is close- <7%. Values with errors above these thresholds are marked in red. In particular, when $\gamma_l = \gamma_h$, the agreement is excellent, $> 99\%$.

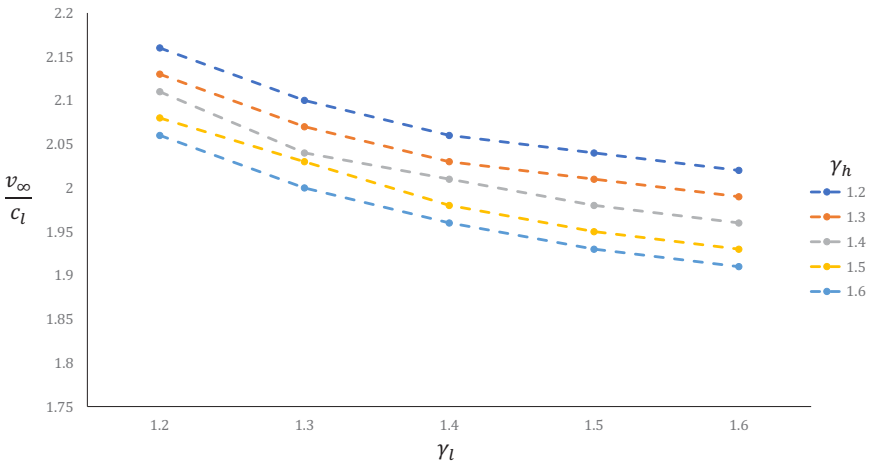


Fig. 3 Plot of v_∞/c_l for values of γ_l (on the x axis) and γ_h (coloured data points)

We found that if γ_l and γ_h are too different with $\gamma_{h(l)} \geq \gamma_{l(h)} + 2$ for $\gamma_{l(h)} = (1.2, 1.3)$ or $\gamma_{h(l)} \geq \gamma_{l(h)} + 3$ for $\gamma_{l(h)} = (1.4, 1.5, 1.6)$, the simulations don't handle those situations well and the results are too inaccurate to make predictions and are excluded from further simulation analysis in this work. To the best of the author's knowledge, this observation has not been made before, and may prove useful in the understanding of scenarios with varying adiabatic indexes.

γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6
1.2	0.432	0.426			
1.3	0.421	0.414	0.409		
1.4		0.406	0.402	0.396	0.391
1.5			0.396	0.391	0.386
1.6			0.393	0.387	0.382

Table 2 Results for $v_\infty/(M \cdot c_l)$ for each value of γ_l and γ_h with $a_0/\lambda = 0$. Values with error >7% are excluded.

Table 2 shows the results for v_∞, scaled by c_l and M, the speed of sound in the light fluid and the Mach number, respectively. We see that the velocity of the background motion, v_∞ is only a fraction of the shock velocity, ranging from ∼20% to ∼40%.

3.2 Initial Amplitude Growth Rate for Perturbed Interface

3.2.1 Simulation Results

After the shock passes through the interface, the interface amplitude grows approximately linearly with time, as we see in **Figure 2**. When this value is calculated from the simulations, we scale it by the velocity of the bulk fluid, v_∞, which is done to compare the interface growth rate to the background motion. This value v_0/v_∞ quantifies the distribution of the energy imparted by the shock wave between the interfacial fluid and the bulk fluid ([Dell et al., 2015]).

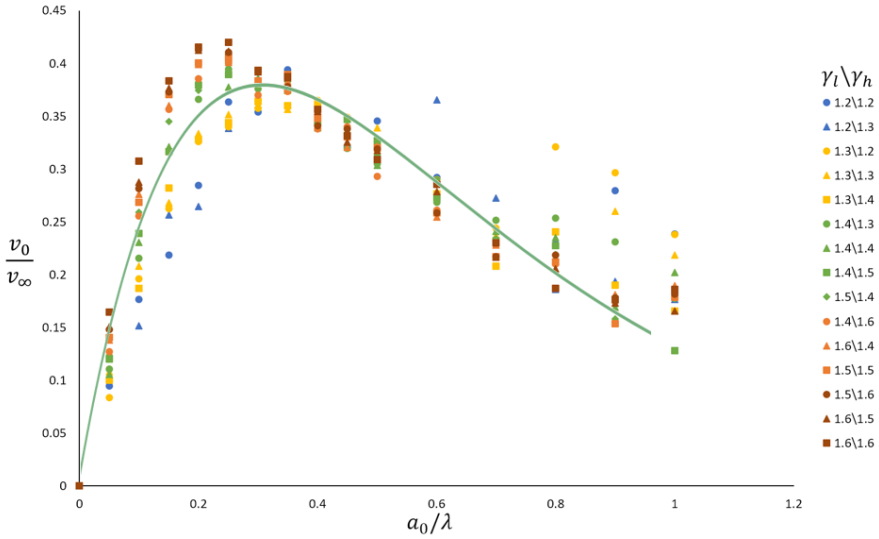


Fig. 4 Plot of v_0/v_∞ against a_0/λ with an approximate curve fitted.

We ran simulations for $(M,A) = (5,0.8), \gamma_l, \gamma_h = (1.2, 1.3, \ldots, 1.6)$, and $a_0/\lambda = (0, 0.1, 0.2 \ldots, 1)$. Determining the interface amplitude growth-rate v_0 and bulk velocity v_∞ from the simulations using the method as described in Section 2.2.2, we plot their ratio v_0/v_∞ for each value of a_0/λ, displayed in **Figure 4**. A well-defined shape is formed, which increases linearly for early time, becomes non linear and eventually peaks and decreases for late time, asymptotically approaching 0. As in [Abarzhi et al., 2019]; [Dell et al., 2017], and [Dell et al., 2015], a function satisfying these criteria is

$$\frac{v_0}{v_\infty} \cdot \frac{1}{A} = c_1 \cdot \frac{a_0}{\lambda} \cdot e^{-c_2 \cdot \frac{a_0}{\lambda}}. \tag{3.2.1}$$

This depends on the Atwood number and two constants, c_1 and c_2. Our objective is to ascertain how well the simulation data fits this curve and to find these constants to compare them with linear theory and to assess the accuracy of our simulations.

Letting $x = \frac{a_0}{\lambda}$ and $y = \frac{v_0}{v_\infty} \cdot \frac{1}{A}$, we have

$$y = c_1 x e^{-c_2 x}. \tag{3.2.2}$$

As we see in **Figure 4**, the curve follows the data closely.

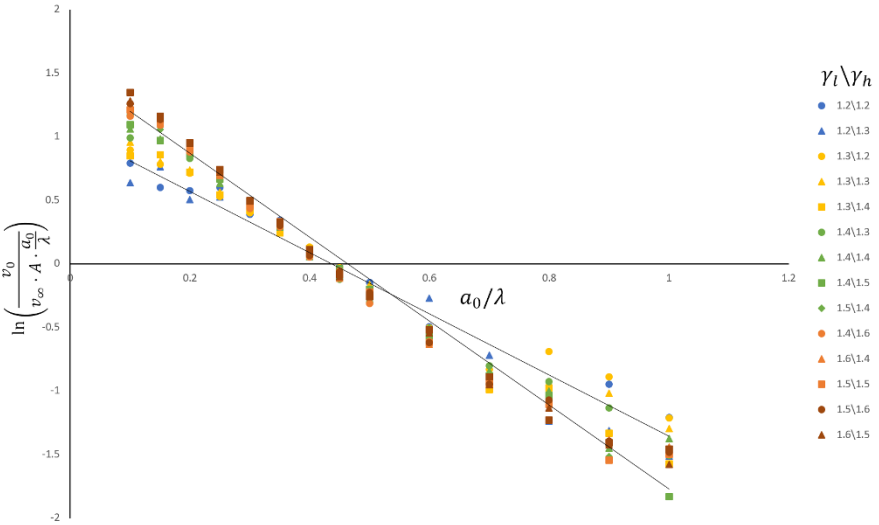


Fig. 5 The combined calculated values of v_0/v_∞ for all pairs of γ_l and γ_h linearised according to the proposed equation that describes their behaviour with a linear regression line fitted.

To determine the constants c_1 and c_2 of the equation of the curve and to quantify the strength of the relationship between the data and the curve, we rearrange Equation 3.2.2 to form a linear relationship so that a linear regression line can be fitted to the data:

$$\begin{aligned} y &= c_1 x e^{-c_2 x} \\ \frac{y}{c_1 x} &= e^{-c_2 x} \\ \ln\left(\frac{y}{c_1 x}\right) &= -c_2 x \\ \ln\left(\frac{y}{x}\right) - \ln(c_1) &= -c_2 x \\ \ln\left(\frac{y}{x}\right) &= (-c_2)x + \ln c_1 \end{aligned} \tag{3.2.3}$$

Letting $\hat{y} = \ln\left(\frac{y}{x}\right), m = -c_2, n = \ln(c_1)$, we have the linear relation

$$\hat{y} = m \cdot x + n. \tag{3.2.4}$$

c_1						c_2					
γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6	γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6
1.2	2.88	2.76				1.2	2.36	2.27			
1.3	3.39	3.53	3.46			1.3	2.78	2.87	2.86		
1.4		4.28	4.45	4.40	5.00	1.4		3.40	3.47	3.39	3.80
1.5			4.86	5.18	5.40	1.5			3.67	3.81	3.90
1.6			5.21	5.51	5.80	1.6			3.82	3.95	4.09

Table 3 Values of c_1 and c_2 for each value of γ_l and γ_h in our parameter regime.

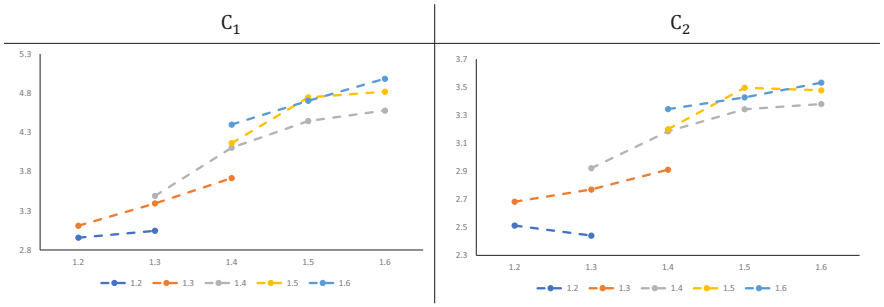


Fig. 6 Plots of c_l and c_2 for values of γ_l (on the x axis) and γ_h (coloured data points)

We plot $\hat{y}$ against x for all chosen values of γ_l, γ_h (excluding those found to produce high errors) and for $a_0/\lambda = (0.1, 0.2, \cdots, 1)$. The plot of all the values of γ_h and γ_l combined is shown in **Figure 5**, where the relationship can be seen to be strongly negative and linear. The value of c_1 is calculated by taking the exponential of the y-intercept of the linear regression line, m, and c_2 is the negative of the slope of the line, n, which follows from the setup of Equation 3.2.2. The values of c_1 and c_2 for each pair of γ_l, γ_h are calculated from the lines of best fit in **Figure 5**, and are shown in **Table 3**.

The plot of the results of the constants c_1 and c_2 is shown in **Figure 6**. These plots reveal some information about the variation in the constants with respect to γ_l and γ_h. Both values of c_1 and c_2 have roughly the same shapes, but with $c_1 > c_2$. Higher values of γ_h produce higher values of c_1 and c_2. We see a linear increase in c_1 and c_2 for small γ_l, which becomes slower than linear for higher values of γ_l.

We found that if we ordered the adiabatic indexes from small γ_h and γ_l to large γ_h and γ_l in a "diagonal" fashion (shown in **Figure 7**), we notice a trend in the values of c_1 and c_2. They appear to increase in a linear fashion, which to the best of

the author's knowledge has not been discussed before, and may benefit from future study.

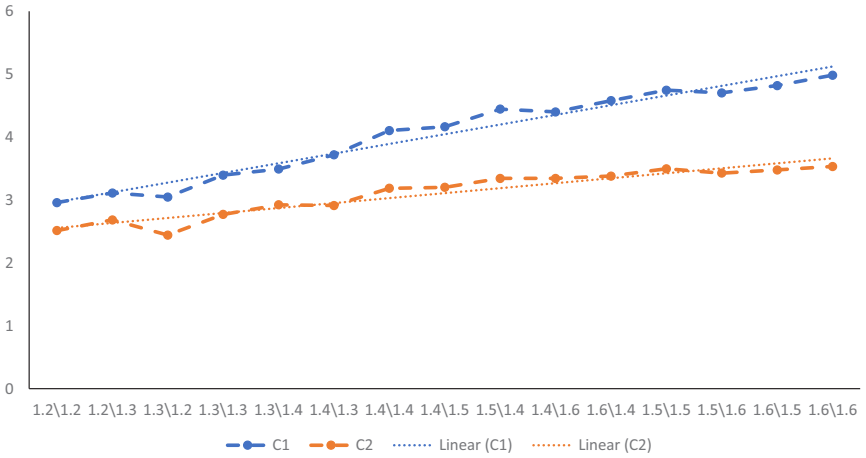


Fig. 7 Plot of the values of c_1 and c_2 for the ordered values of γ_l and γ_h. The equation of the c_1 line is $y = 0.1539x + 2.8127$, and the equation of the c_2 line is $y = 0.0791x + 2.4754$

3.2.2 Comparison of Simulation Results with Linear Theory

For small a_0/λ, $a_0/\lambda \leq 0.1$, the initial growth rate v_0 linearly depends on a_0. Linear theory can predict this growth for these small amplitudes, as seen in **Figure 8** ([Dell et al., 2015]). We compare this linear theory with our simulation results. Restating our equation relating interface growth rate and initial perturbation amplitude, we have

$$\frac{1}{A} \cdot \frac{v_0}{v_\infty} = c_1 \cdot \frac{a_0}{\lambda} \cdot e^{-c_2 \cdot \frac{a_0}{\lambda}}. \tag{3.2.5}$$

The linear theory finds values of $v_0/(v_\infty \cdot \frac{a_0}{\lambda})$ for small a_0/λ, so we choose the smallest initial amplitude, $a_0/\lambda = 0.1 \ll 1$. We rearrange Equation 3.2.5 to get

$$\left[\frac{v_0}{v_\infty \cdot \frac{a_0}{\lambda}}\right]_T = A \cdot c_1 \cdot e^{-0.1c_2}, \tag{3.2.6}$$

where $[\cdot]_T$ is the value obtained from the linear theory. The values of the theoretical $\left[\frac{v_0}{v_\infty \cdot \frac{a_0}{\lambda}}\right]_T$ are compared to the simulation data $A \cdot c_1 \cdot e^{-0.1c_2}$ in **Table 4**, as well as the percentage error. The simulation results are in good agreement with the theoretical values, with an average error of 4.68%. Only one of the fifteen cases

have an error of more than 10%, making this a very accurate prediction. In order to investigate the linear approximation more closely, we ran simulations over a finer increment of a_0/λ, increasing by 0.05 instead of 0.1, as displayed in **Figure 8**. The approximations for linear theory assume the perturbation amplitude to be very small, $ka_0 \ll 1$ or $a_0/\lambda < 0.05$, but we see that the approximation holds for larger amplitudes, up to $a_0/\lambda < 0.1$, which is a result consistent with previous studies ([Dell et al., 2017]; [Dell et al., 2015]).

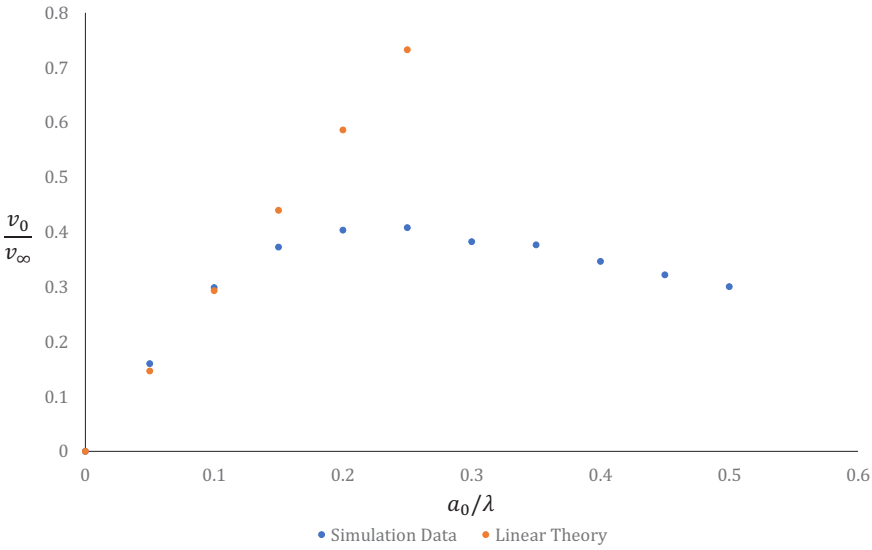


Fig. 8 Plot showing values of v_0/v_∞ from the simulations (blue), and the prediction from linear theory (orange). We see that linear theory only has predicting power for values of $a_0/\lambda \leq 0.1$.

3.2.3 Maximum Interface Growth-Rate

Of interest in experiments is the amount of energy that can be deposited into the interface by the shock, which determines the amount of energy available for interfacial mixing, discussed in **Section 3.3**. The maximum scaled interface growth rate v_0/v_∞ quantifies this amount, so is of interest in this study. We have found in **Section 3.2.1** that the interface growth-rate is a non-monotone function of the initial perturbation amplitude, and is described by the relationship

$$\frac{v_0}{v_\infty \cdot A} = c_1 \frac{a_0}{\lambda} e^{-c_2 \cdot \frac{a_0}{\lambda}}. \tag{3.2.7}$$

Letting $x = \frac{a_0}{\lambda}$ and $y = \frac{v_0}{v_\infty} \cdot \frac{1}{A}$, we have

	γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6
Simulation	1.2	1.82	1.76			
	1.3	2.05	2.12	2.08		
	1.4		2.44	2.51	2.51	2.74
	1.5			2.69	2.83	2.92
	1.6			2.84	2.97	3.08
	γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6
Theoretical	1.2	1.88	2.07			
	1.3	2.01	2.24	2.39		
	1.4		2.35	2.52	2.64	2.72
	1.5			2.62	2.75	2.84
	1.6			2.69	2.83	2.93
	γ_l \ γ_h	1.2	1.3	1.4	1.5	1.6
Error (%)	1.2	3.34	15.2			
	1.3	2.05	5.45	13.0		
	1.4		3.60	0.37	4.83	0.63
	1.5			2.72	3.20	2.88
	1.6			5.56	4.80	5.25

Table 4 Comparison of the values of $\frac{v_0}{v_\infty a_0/\lambda}$ from the linear theory and the simulation results, obtained using Equation 3.2.6.

$$y = c_1 x e^{-c_2 x}. \tag{3.2.8}$$

We find the maximum scaled interface growth rate $[v_0/v_\infty]_{max}$, from the data in **Section 3.2.1**. In order to find the a_0/λ at this maximum, we find where the gradient of Equation 3.2.8 is zero,

$$\begin{aligned} y'(x) &= c_1 e^{-c_2 x} - c_1 c_2 x e^{-c_2 x} = 0 \\ 0 &= c_1 - c_1 c_2 x \\ x &= \frac{1}{c_2}. \end{aligned} \tag{3.2.9}$$

The results for the maximum growth rate and the value of a_0/λ at which this maximum occurs are plotted in **Figure 9**. The plots range the adiabatic indexes from small γ_h and γ_l to large γ_h and γ_l in a "diagonal" fashion because it is easier and simpler to display the data in this way, and because we noticed a trend in the data when displayed in this fashion. This trend has not been discussed before, and may benefit from future study. The maximum growth rate hits a minimum at $(\gamma_l, \gamma_h) \approx (1.3, 1.4)$ before increasing in a linear fashion as (γ_l, γ_h) ranges to $(1.6, 1.6)$. The amplitude a_0/λ at which these occur decreases as γ_l, γ_h increases. The value v_0/v_∞ quantifies the fraction of energy imparted into the interface by the shock, and these results show that on average, $\sim 45\%$ of the bulk velocity is available for interfacial mixing. From **Table 2** we know that at on average $\sim 30\%$ of the shock velocity is imparted into the bulk motion, meaning that on average, only $\sim 15\%$ of the shock velocity is

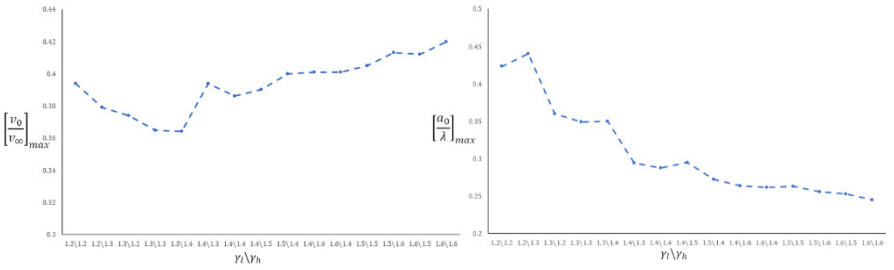


Fig. 9 Plots showing the maximum scaled interface growth-rate (top) and the initial perturbation amplitude at which this occurs (bottom) for all cases of γ_l and γ_h, arranged from lowest $\gamma_{h(l)}$ to highest $\gamma_{h(l)}$.

available for interfacial mixing. Despite this very small scale of mixing, our results are in good agreement with the theory, demonstrating the ability of SPHC to capture small scale dynamics embedded in large scale dynamics very accurately.

3.3 Discussion and Conclusion

In this study, through use of numerical simulations we have systematically studied the effect of a previously unconsidered regime of adiabatic indexes of the fluids on the early time dynamics of RMI, specifically the extent of the interfacial mixing. The key properties of the dynamics we have analysed are the velocity, growth-rate, and curvature of the interface. Our regime included the Mach and Atwood numbers, $(M,A) = (5, 0.8)$, a range of adiabatic indexes $\gamma_l, \gamma_h = (1.2, 1.3, \cdots, 1.6)$ and a range of initial perturbations from 0% to 100% of the wavelength. In this regime, the simulations are repeatable and qualitatively similar, and we found good quantitative and qualitative agreement between the simulation results and the theory, demonstrating the accuracy at which SPHC simulations are able to capture small scale dynamics embedded within large scale dynamics in extreme and challenging situations.

In order to find the appropriate scale for the amount of energy deposited into the interface by the shock, we obtained the velocity of the background motion. In experiments, the background motion makes reliable diagnostics of RMI challenging because flow measurements must be taken from a quickly moving interface. Numerical simulations have the ability to scale the dynamics by the bulk velocity via the use of a Galilean transformation to a moving frame of reference. In order to scale our results, we obtained the bulk velocity v_∞ in **Section 3.1**, and we found that the more challenging cases for the simulations to model occur when the adiabatic indexes of the two fluids are further apart (**Table 1**), which to the best of the author's knowledge is the first time this observation has been made. To ensure our results were reliable, we removed the cases that were challenging for our simulations to model accurately from consideration.

In order to quantify the amount of energy deposited into the interface by the shock, we found the initial growth rate of the interface v_0, and scaled it by the velocity of the background motion v_∞. The speed at which the interface spreads out indicates how much energy the interface received from the initial shock. We found that the initial growth rate is a non-monotone function of the initial perturbation amplitude, and that this relationship is insensitive to the adiabatic indexes of the fluids (**Figure 4**). For each pair of adiabatic indexes, we found the maximum values of the growth-rate scaled by the background motion, v_0/v_∞, and at which values of a_0/λ this occurred. We found that this maximum changes with γ_l and γ_h (**Figure 9**). We found that on average, only $\sim 15\%$ of the shock velocity is transferred to the growth of the interface, indicating that only a fraction of the energy from the shock is deposited into the interface and is hence available for interfacial mixing. We compared our simulation results to linear theory, and found the average error to be less than 5% for the bulk velocity (**Table 1**) and the interface growth-rate (**Table 4**). This indicates that the SPHC simulations can handle small scale dynamics embedded in large scale dynamics with high accuracy.

Our results provide good benchmarks for further studies and experiments and open up further avenues of investigation for non-ideal adiabatic indexes. Our results also have implications for hydrodynamic instabilities and mixing in inertial confinement fusion (ICF). To achieve ICF ignition, the ability to avoid or control the Richtmyer-Meshkov instability that forms during the implosion process is necessary ([Lindl et al., 2004]). One method is to fully suppress the development of RMI, which is based on traditional scenarios of RMI that suggest that the development of RMI may produce uncontrolled growth of small-scale imperfections and lead to disordered mixing that is similar to canonical turbulence ([Dimonte et al., 2004]). However, research conducted in studies like this through simulations and theoretical analysis suggests that the interfacial mixing may keep a significant degree of order, shown by the ability to accurately predict the evolution of the interface through theoretical analysis. These findings suggest that the dynamics can be controlled through the initial perturbation, so that turbulent mixing may be prevented without having to completely suppress RMI, which may be easier to implement ([Dell et al., 2015]; [Anisimov et al., 2013]; [Abarzhi, 2010]; [Demskoĭ et al., 2006]). This study also suggests there is reason to have confidence in the ability of the numerical simulations produced by SPHC in accurately capturing small scale dynamics embedded in large scale structures despite its difficulty.

References

[Abarzhi, 2002] Abarzhi, S. (2002). A new type of the evolution of the bubble front in the Richtmyer–Meshkov instability. *Physics Letters A*, 294(2):95–100.

[Abarzhi, 2008] Abarzhi, S. (2008). Review of nonlinear dynamics of the unstable fluid interface: conservation laws and group theory. *Physica Scripta*, T132:014012.

[Abarzhi et al., 2003] Abarzhi, S., Nishihara, K., and Glimm, J. (2003). Rayleigh–Taylor and Richtmyer-Meshkov instabilities for fluids with a finite density ratio. *Physics Letters A*,

317(5):470–476.

[Abarzhi, 2010] Abarzhi, S. I. (2010). Review of theoretical modeling approaches of Rayleigh-Taylor instabilities and turbulent mixing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1916):1809–1828.

[Abarzhi et al., 2019] Abarzhi, S. I., Bhowmick, A. K., Naveh, A., Pandian, A., Swisher, N. C., Stellingwerf, R. F., and Arnett, W. D. (2019). Supernova, nuclear synthesis, fluid instabilities, and interfacial mixing. *Proceedings of the National Academy of Sciences*, 116(37):18184–18192.

[Aleshin et al., 1990] Aleshin, A. N., Lazareva, E. V., Zaitsev, S. G., Rozanov, V. B., Gamalii, E. G., and Lebo, I. G. (1990). Linear, nonlinear, and transient stages in the development of the Richtmyer-Meshkov instability. *Soviet Physics Doklady*, 35:159.

[Anisimov et al., 2013] Anisimov, S. I., Drake, R. P., Gauthier, S., Meshkov, E. E., and Abarzhi, S. I. (2013). What is certain and what is not so certain in our knowledge of Rayleigh-Taylor mixing? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(2003):20130266.

[Bodner et al., 1998] Bodner, S. E., Colombant, D. G., Gardner, J. H., Lehmberg, R. H., Obenschain, S. P., Phillips, L., Schmitt, A. J., Sethian, J. D., McCrory, R. L., Seka, W., Verdon, C. P., Knauer, J. P., Afeyan, B. B., and Powell, H. T. (1998). Direct-drive laser fusion: Status and prospects. *Physics of Plasmas*, 5(5):1901–1918.

[Dell et al., 2015] Dell, Z., Stellingwerf, R. F., and Abarzhi, S. I. (2015). Effect of initial perturbation amplitude on Richtmyer-Meshkov flows induced by strong shocks. *Physics of Plasmas*, 22(9):092711.

[Dell et al., 2017] Dell, Z. R., Pandian, A., Bhowmick, A. K., Swisher, N. C., Stanic, M., Stellingwerf, R. F., and Abarzhi, S. I. (2017). Maximum initial growth-rate of strong-shock-driven Richtmyer-Meshkov instability. *Physics of Plasmas*, 24(9):090702.

[Demskoĭ et al., 2006] Demskoĭ, D. K., Marikhin, V. G., and Meshkov, A. G. (2006). Lax representations for triplets of two-dimensional scalar fields of chiral type. *Teoret. Mat. Fiz.*, 148(2):189–205.

[Dimonte et al., 2004] Dimonte, G., Youngs, D. L., Dimits, A., Weber, S., Marinak, M., Wunsch, S., Garasi, C., Robinson, A., Andrews, M. J., Ramaprabhu, P., Calder, A. C., Fryxell, B., Biello, J., Dursi, L., MacNeice, P., Olson, K., Ricker, P., Rosner, R., Timmes, F., Tufo, H., Young, Y.-N., and Zingale, M. (2004). A comparative study of the turbulent Rayleigh–Taylor instability using high-resolution three-dimensional numerical simulations: The Alpha-Group collaboration. *Physics of Fluids*, 16(5):1668–1693.

[Drake, 2009] Drake, R. P. (2009). Perspectives on high-energy-density physics. *Physics of Plasmas*, 16(5):055501.

[Herrmann et al., 2008] Herrmann, M., Moin, P., and Abarzhi, S. I. (2008). Nonlinear evolution of the Richtmyer-Meshkov instability. *J. Fluid Mech.*, 612:311–338.

[Holmes et al., 1999] Holmes, R. L., Dimonte, G., Fryxell, B., Gittings, M. L., Grove, J. W., Schneider, M., Sharp, D. H., Velikovich, A. L., Weaver, R. P., and Zhang, Q. (1999). Richtmyer–Meshkov instability growth: experiment, simulation and theory. *Journal of Fluid Mechanics*, 389:55–79.

[Jacobs and Krivets, 2005] Jacobs, J. W. and Krivets, V. V. (2005). Experiments on the late-time development of single-mode Richtmyer-Meshkov instability. *Physics of Fluids*, 17(3):034105.

[Lindl et al., 2004] Lindl, J. D., Amendt, P., Berger, R. L., Glendinning, S. G., Glenzer, S. H., Haan, S. W., Kauffman, R. L., Landen, O. L., and Suter, L. J. (2004). The physics basis for ignition using indirect-drive targets on the National Ignition Facility. *Physics of Plasmas*, 11(2):339–491.

[Lucy, 1977] Lucy, L. B. (1977). A numerical approach to the testing of the fission hypothesis. *Astronomical Journal*, 82:1013–1024.

[McFarland et al., 2011] McFarland, J. A., Greenough, J. A., and Ranjan, D. (2011). Computational parametric study of a Richtmyer-Meshkov instability for an inclined interface. *Phys. Rev. E*, 84:026303.

[Meshkov, 1969] Meshkov, E. E. (1969). Instability of the interface of two gases accelerated by a shock wave. *Fluid Dynamics*, 4(5):101–104.

[Monaghan, 2005] Monaghan, J. J. (2005). Smoothed particle hydrodynamics. *Reports on Progress in Physics*, 68(8):1703–1759.

[Motl et al., 2009] Motl, B., Oakley, J., Ranjan, D., Weber, C., Anderson, M., and Bonazza, R. (2009). Experimental validation of a Richtmyer–Meshkov scaling law over large density ratio and shock strength ranges. *Physics of Fluids*, 21(12):126102.

[Nishihara et al., 2010] Nishihara, K., Wouchuk, J. G., Matsuoka, C., Ishizaki, R., and Zhakhovsky, V. V. (2010). Richtmyer-Meshkov instability: Theory of linear and nonlinear evolution. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1916):1769–1807.

[Orlicz et al., 2009] Orlicz, G. C., Balakumar, B. J., Tomkins, C. D., and Prestridge, K. P. (2009). A Mach number study of the Richtmyer–Meshkov instability in a varicose, heavy-gas curtain. *Physics of Fluids*, 21(6):064102.

[Richtmyer, 1960] Richtmyer, R. D. (1960). Taylor instability in shock acceleration of compressible fluids. *Communications on Pure and Applied Mathematics*, 13(2):297–319.

[Stanic et al., 2013] Stanic, M., McFarland, J., Stellingwerf, R. F., Cassibry, J. T., Ranjan, D., Bonazza, R., Greenough, J. A., and Abarzhi, S. I. (2013). Non-uniform volumetric structures in Richtmyer-Meshkov flows. *Physics of Fluids*, 25(10):106107.

[Stanic et al., 2012] Stanic, M., Stellingwerf, R. F., Cassibry, J. T., and Abarzhi, S. I. (2012). Scale coupling in Richtmyer-Meshkov flows induced by strong shocks. *Physics of Plasmas*, 19(8):082706.

[Stellingwerf, 1991] Stellingwerf, B. (1991). Stellingwerf Consulting. www.stellingwerf.com.

[Stellingwerf et al., 2004] Stellingwerf, R., Robinson, J., Richardson, S., Evans, S., Stallworth, R., and Hovater, M. (2004). Foam on tile impact modeling for the STS-107 investigation. 5.

[Stellingwerf, 1990] Stellingwerf, R. F. (1990). *Smooth particle hydrodynamics*, chapter 25, pages 239–247. Springer Verlag.

[Velikovich and Dimonte, 1996] Velikovich, A. L. and Dimonte, G. (1996). Nonlinear perturbation theory of the incompressible Richtmyer-Meshkov instability. *Phys. Rev. Lett.*, 76:3112–3115.

[Velikovich et al., 2014] Velikovich, A. L., Herrmann, M., and Abarzhi, S. I. (2014). Perturbation theory and numerical modelling of weakly and moderately nonlinear dynamics of the incompressible Richtmyer-Meshkov instability. *Journal of Fluid Mechanics*, 751:432–479.

[Wouchuk, 2001] Wouchuk, J. G. (2001). Growth rate of the linear Richtmyer-Meshkov instability when a shock is reflected. *Phys. Rev. E*, 63:056303.

[Zel'dovich, 1967] Zel'dovich, I. B. (1967). *Physics of shock waves and high-temperature hydrodynamic phenomena*. Academic Press, New York.

Compressibility Effect on Markstein Number for a Flame Front in Long-Wavelength Approximation

Keigo Wada and Yasuhide Fukumoto

Abstract The effect of compressibility on the Markstein number for a planar front of a premixed flame is examined, at small Mach numbers, in the form of M^2-expansions. The method of matched asymptotic expansions is used to analyze the solution in the preheat zone in a power series in two small parameters, the relative thickness of the preheat zone and the Mach number. We employ a specific form of perturbations, valid at long wavelengths, for the thermodynamic variables, which produces the correction term, to the Markstein number, of second order in the Mach number in drastically simple form. Our analysis accounts for the pressure variation as a source term in the heat-conduction equation and calls for the Navier-Stokes equation. The suppression effect of the front curvature on the Darrieus-Landau instability is enhanced by the viscous effect if $Pr > 4/3$, but is weakened if otherwise.

1 Introduction

The pioneering work of the linear stability analysis of a planar front of a premixed flame was made in the low-Mach-number limit by Darrieus [6] and Landau [13, 14] independently. They treated a flame front as a density discontinuity interface accompanied by an essential parameter of the thermal expansion, or the heat release. Their conclusion is that a planar flame front is unstable for small perturbations of any wavelength. This result is now called the Darrieus-Landau instability (DLI),

Keigo Wada
Center of Coevolutionary Research for Sustainable Communities, Kyushu University,
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan,
e-mail: k-wada@imi.kyushu-u.ac.jp

Yasuhide Fukumoto
Institute of Mathematics for Industry, Kyushu University,
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: yasuhide@imi.kyushu-u.ac.jp

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_19

or the hydrodynamic instability. Although the DLI effectively explains the intrinsic instability of a planar flame front, stable flame fronts are observed in the laboratory, some contradiction between the DLI and the observation. The DLI imposes a major assumption on the boundary condition that the flame speed S_f is constant, which is expressed in non-dimensional form as

$$S_f = 1. \tag{1}$$

The flame speed S_f is defined as the incoming normal velocity of a gas relative to a flame front and evaluated at the edge of a flame front on the unburned side. The constancy of S_f was modified phenomenologically by Markstein [16]. It is regarded as the curvature effect [5].

$$S_f = 1 - Mr\Delta F, \tag{2}$$

where Mr is the Markstein number and F is the infinitesimal displacement of a planar flame front. Based on (2), not (1), Markstein showed that the DLI is stabilized at sufficiently large values of wavenumber, or small wavelengths, of perturbations. In order to find the expression of the Markstein number Mr in (2), many subsequent works have investigated the transport process inside a flame front in detail. For example, the effect of diffusion properties of the mixture on the flame speed was clarified by Eckhaus [7, 8]. These results were generalized to a more comprehensive concept of flame stretch [2, 17, 18, 22]. However, most of the previous research stayed at the low-Mach-number limit. In such a treatment, the isobaric condition prevails, without having to consider pressure variation.

In this paper, we highlight the compressibility effect and shall derive S_f corrected by the compressibility effect in a tidy form as

$$S_f = 1 - Mr_M\Delta F, \tag{3}$$

with

$$Mr_M = \delta\left\{1 + Ma^2(\gamma - 1)\left(\frac{4}{3}Pr - 1\right)\right\}, \tag{4}$$

where, as will be defined in (11), δ, Ma, γ and Pr are respectively the scale factor of a preheat zone which is an inner structure of a flame front, the Mach number, the specific heats ratio and the Prandtl number (= kinematic viscosity/thermal diffusivity). We reveal that the compressibility effect is accompanied with the viscous effect as evidenced by Pr in (4) for Mr_M which is the extension of the Markstein number Mr to the compressible case. If the value of Pr is larger than $3/4$, the compressibility acts to weaken the DLI for any values of the Mach number. On the other hand, if the value of Pr lies in the range from 0 to $3/4$, then the compressibility reduces the curvature effect.

The flame speed condition (3) is derived from the study of the preheat zone which is the inner structure of a flame front. In the preheat zone, the transport process of the heat and the mass is dominant rather than the exothermic chemical reaction. The effect of the chemical kinetics is confined in the reaction zone, which is the innermost structure of a flame front and is sandwiched by the preheat and the burned

zones. Although the investigation of the preheat zone requests the jump (boundary) conditions across the reaction zone contained in it [20, 21], we may dispense with such conditions by postulating the long-wavelength approximation [3], or the translational symmetry. Thanks to the collaboration of the matched asymptotic expansions with respect to δ and the long-wavelength approximation, we reach the substantially compact representation of the Markstein number (4) affected by the compressibility effect.

There are several works on the DLI with the compressibility effect incorporated. Some employed the same assumption as (1) with the density perturbation omitted and concluded that the DLI is enhanced at small values of the Mach number [9, 11, 12]. In contrast to them, works based on the long-wavelength approximation [3, 15] successfully included the density perturbation in the flame speed condition, although the viscous effect is ignored. In [15], the second-order effect of the wavenumber is analysed and the suppression of the DLI by the increasing wavenumber is shown numerically. In this paper, we reveal that the effect of viscosity comes into play for the DLI, with the Navier-Stokes equations coupled to the heat-conduction equations via pressure variation.

We explore the compressibility effect in the form of the M^2 expansions for small Mach numbers Ma ($Ma^2 \ll 1$), under the long-wavelength approximation, as will be exposed in Sect. 2. The scheme of the matched asymptotic expansions with respect to δ ($\ll 1$) and Ma, for deriving the condition of the flame speed based on the first principle, is sketched in Sect. 3. This paper sidesteps handling the reaction term in the heat-conduction equation, but instead, resort to the large activation energy asymptotics. The detailed analysis of the translational symmetry is performed to gain (3), the flame speed with a correction from the weak compressibility effect, in Sect. 4. In Sect. 5, the dispersion relation of a planar flame front is calculated, showing that the DLI can be suppressed depending on the Prandtl number Pr and the Mach number Ma. This paper is closed with a summary in Sect. 6.

2 Non-Dimensional Governing Equations

The Cartesian coordinate system (x, y, z) and the velocity field are made dimensionless by use of the hydrodynamic length scale $\tilde{L}$ and the laminar flame speed $\tilde{S}_L$, the speed of a flat flame. We consider the situation where a planar flame front propagates in the negative z-direction. The velocity field is partitioned into the tangential and normal components as $\vec{V} + W\vec{e}_z$. Besides, the differential operator is defined for the x-y plane as $\nabla = (\partial/\partial x, \partial/\partial y)$ and $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$. Then we deal with the following equations governing dimensionless hydrodynamic variables [18, 24]:

$$\frac{\partial R}{\partial t}+\nabla\cdot(R\mathbf{V})+\frac{\partial}{\partial z}(RW)=0, \tag{5}$$

$$R\left(\frac{\partial \mathbf{V}}{\partial t}+\left(\mathbf{V}\cdot\nabla+W\frac{\partial}{\partial z}\right)\mathbf{V}\right) = -\frac{1}{\gamma Ma^2}\nabla P+\delta Pr\left(\left(\Delta+\frac{\partial^2}{\partial z^2}\right)\mathbf{V}+\frac{1}{3}\nabla\left(\nabla\cdot\mathbf{V}+\frac{\partial W}{\partial z}\right)\right), \tag{6}$$

$$R\left(\frac{\partial W}{\partial t}+\left(\mathbf{V}\cdot\nabla+W\frac{\partial}{\partial z}\right)W\right) = -\frac{1}{\gamma Ma^2}\frac{\partial P}{\partial z}+\delta Pr\left(\left(\Delta+\frac{\partial^2}{\partial z^2}\right)W+\frac{1}{3}\frac{\partial}{\partial z}\left(\nabla\cdot\mathbf{V}+\frac{\partial W}{\partial z}\right)\right), \tag{7}$$

$$R\left(\frac{\partial T}{\partial t}+\left(\mathbf{V}\cdot\nabla+W\frac{\partial}{\partial z}\right)T\right) = \delta\left(\Delta+\frac{\partial^2}{\partial z^2}\right)T+\frac{\gamma-1}{\gamma}\left(\frac{\partial P}{\partial t}+\left(\mathbf{V}\cdot\nabla+W\frac{\partial}{\partial z}\right)P\right)+qQ, \tag{8}$$

$$P=RT, \tag{9}$$

where the non-dimensional variables, R, T and P are the density of the mixture, the temperature and the pressure, respectively. All of the variables are made dimensionless with use of those of the fresh mixture at a position far from a flame front, where the flow field is assumed to be uniform with its velocity identified with $\tilde{S}_L$ in the coordinate frame relative to the flame front. Therefore, each quantity is subject to the following boundary condition in the far field on the unburned side ($z<0$).

$$R=W=T=P=1, \quad \vec{V}=\vec{0} \quad \text{as} \quad z\to-\infty. \tag{10}$$

Our aim is to derive the condition of the flame speed (3). For this, we need to investigate the preheat zone whose length scale is represented by $\tilde{l}_d=\tilde{D}_{th}/\tilde{S}_L$, with $\tilde{D}_{th}$ being the thermal diffusivity. We find that (3) is highly influenced by the scale factor of the preheat zone δ, the Mach number Ma, the Prandtl number Pr and the specific heats ratio γ. These non-dimensional parameters are defined as

$$\delta=\frac{\tilde{l}_d}{\tilde{L}}, \quad Ma=\frac{\tilde{S}_L}{\tilde{c}_s}, \quad Pr=\frac{\tilde{\nu}}{\tilde{D}_{th}} \quad \gamma=\frac{\tilde{c}_p}{\tilde{c}_v}, \tag{11}$$

where $\tilde{c}_s$, $\tilde{\nu}$, $\tilde{c}_p$ and $\tilde{c}_v$ are the adiabatic sound speed defined in the fresh mixture, the kinematic viscosity, the specific heats at constant pressure and volume, respectively.

The reaction term Q in (8) is not dealt with in this paper by resorting to the large-activation-energy asymptotics. The coefficient q represents the non-dimensional heat release whose value is positive for an exothermic chemical reaction. A detailed analysis, with the compressibility effect taken into consideration, is relegated to an independent investigation [23].

It is noteworthy that Bychkov *et al.* [3] ignored the viscous terms in (7), though it naturally enters through the coupling of the heat-conduction equations with the Navier-Stokes equation, via pressure variation. However, as seen from (4), the vis-

cous effect is indispensable for the compressible correction to the Markstein number and thence to the DLI. The compressible correction is sensitive to the value of Pr.

2.1 Perturbations in Hydrodynamic Zone

We superimpose an infinitesimal perturbation to a plane flame front, which coincides with the x-y plane parametrized by $\vec{x} = (x, y)$.

$$F(x, y, t) = f \exp(\mathrm{i}\vec{x} \cdot \vec{k} + \Omega t),$$

where $\vec{k} = (k_x, k_y)$ and Ω are the wavenumber, with $k = (k_x^2 + k_y^2)^{1/2}$, and the growth rate of the perturbation, which are made dimensionless as follows.

$$\vec{k} = \tilde{\vec{k}}\tilde{L}, \quad \Omega = \tilde{\Omega}\tilde{L}/\tilde{S}_L. \tag{12}$$

Any hydrodynamic variable Φ is partitioned into a steady planar solution $\bar{\Phi}(z)$ and a small perturbation $\Phi'(x, y, z, t)$ to it as

$$\Phi = \bar{\Phi}(z) + \Phi'(x, y, z, t), \tag{13}$$

with

$$\Phi' = \tilde{\Phi}(z) \exp(\mathrm{i}\vec{x} \cdot \vec{k} + \Omega t)\ . \tag{14}$$

For our purpose of taking compressibility into account, we expand all the functions in powers of a small parameter $Ma^2 (\ll 1)$, up to $O(Ma^2)$. For the basic flow, M^2 expansions tale the form as

$$\begin{aligned} &\bar{R} = \bar{R}_{0M} + Ma^2 \bar{R}_{2M}, \quad \bar{W} = \bar{W}_{0M} + Ma^2 \bar{W}_{2M}, \quad \bar{M} = \bar{M}_{0M} + Ma^2 \bar{M}_{2M}, \\ &\bar{T} = \bar{T}_{0M} + Ma^2 \bar{T}_{2M}, \quad \bar{P} = 1 + \gamma Ma^2 \bar{P}_{2M}, \end{aligned} \tag{15}$$

and, for the perturbations, as

$$\begin{aligned} &R' = R'_{0M} + Ma^2 R'_{2M}, \quad W' = W'_{0M} + Ma^2 W'_{2M}, \quad M' = M'_{0M} + Ma^2 M'_{2M}, \\ &T' = T'_{0M} + Ma^2 T'_{2M}, \quad P' = \gamma Ma^2 P'_{2M} + \ldots, \quad \vec{V}' = \vec{V}'_{0M} + \ldots, \\ &F = F_{0M} + Ma^2 F_{2M}, \end{aligned} \tag{16}$$

where the mass flux perpendicular to a plane flame front is defined by

$$M = RW. \tag{17}$$

We note that the leading term of the pressure is constant under $Ma^2 \ll 1$ because of the first term on the right hand side of (6) and (7) with the boundary condition (10). Furthermore, each quantity is expanded with respect to δ as, for the basic flow,

$$\bar{R}_{0M} = \bar{R}_0 + \delta\bar{R}_1, \qquad \bar{W}_{0M} = \bar{W}_0 + \delta\bar{W}_1,$$
$$\bar{R}_{2M} = \bar{R}_{2M,0} + \delta\bar{R}_{2M,1}, \quad \bar{W}_{2M} = \bar{W}_{2M,0} + \delta\bar{W}_{2M,1}, \tag{18}$$

and, for the perturbations,

$$R'_{0M} = R'_0 + \delta R'_1, \qquad W'_{0M} = W'_0 + \delta W'_1, \qquad F_{0M} = F_0 + \delta F_1,$$
$$R'_{2M} = R'_{2M,0} + \delta R'_{2M,1}, \quad W'_{2M} = W'_{2M,0} + \delta W'_{2M,1}, \quad F_{2M} = F_{2M,0} + \delta F_{2M,1}. \tag{19}$$

The steady planar state of (5), (7) and (8) should satisfy, in the language of the notation (13),

$$\frac{d\bar{M}}{dz} = 0, \tag{20}$$

$$\bar{M}\frac{d\bar{W}}{dz} = -\frac{1}{\gamma M a^2}\frac{d\bar{P}}{dz} + \frac{4}{3}\delta Pr\frac{d^2\bar{W}}{dz^2}, \tag{21}$$

$$\bar{M}\frac{d\bar{T}}{dz} = \delta\frac{d^2\bar{T}}{dz^2} + \frac{\gamma-1}{\gamma}\bar{W}\frac{d\bar{P}}{dz}, \tag{22}$$

where the reaction term Q is omitted from (23) by use of the assumption of the large-activation-energy asymptotics. We easily find from (20) that the steady planar mass flux is constant. Especially, $\bar{M} = 1$ from the boundary condition (10) on the unburned side of a flame front.

The perturbed heat-conduction equation is deduced, from (8) with substitution from (13), as

$$\bar{R}\frac{\partial T'}{\partial t} + \frac{\partial T'}{\partial z} + M'\frac{d\bar{T}}{dz}$$
$$= \delta\left(\frac{\partial^2}{\partial z^2} + \Delta\right)T' + \frac{\gamma-1}{\gamma}\left(\frac{\partial P'}{\partial t} + \bar{W}\frac{\partial P'}{\partial z} + W'\frac{d\bar{P}}{dz}\right). \tag{23}$$

In deriving the condition for a flame speed, we need to explore (23) in the preheat zone scale. We omit the detailed derivation of the jump conditions for the hydrodynamic variables, gained by taking the outer limit of the solution of (23), which is treated in [15] (See also refs [1, 4, 10, 19]).

2.2 Long-Wavelength Approximation

We focus on the specific form of perturbations of the temperature and the density posed by Bychkov *et al.* [3].

$$T' = -F\frac{d\bar{T}}{dz}, \quad R' = -F\frac{d\bar{R}}{dz}, \tag{24}$$

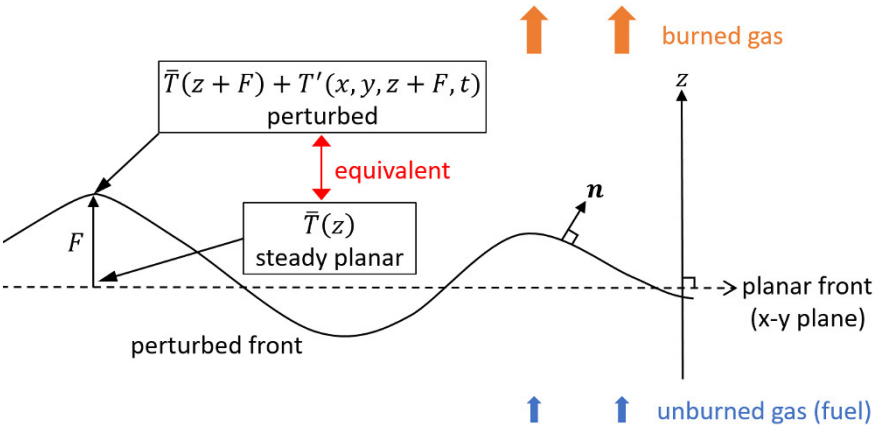


Fig. 1 Schematic illustration of translational symmetry for small perturbations ($F \ll 1$).

where $F(x,y,t)$ is a small-amplitude displacement of the perturbed flame front as shown in Fig. 1. This form is interpreted to come from the translational symmetry possessed by the temperature and the density as illustrated by Fig. 1. It follows from (9), the equation of state, that the pressure inherits the same symmetry.

$$P' = R'\bar{T} + \bar{R}T' = -F\frac{d\bar{R}}{dz}\bar{T} - \bar{R}F\frac{d\bar{T}}{dz} = -F\frac{d\bar{P}}{dz}. \tag{25}$$

Bychkov *et al.* [3] proved that these forms of the perturbations lead to the certain relation for the mass flux valid in the compressible case. This provides the missing boundary condition for the DLI.

By substitution from the asymptotic expansions (15), (16), (18) and (19), we rewrite the perturbation of density in (24) as

$$\begin{aligned} R'_0 &= -F_0\frac{d\bar{R}_0}{dz}, \quad R'_1 = -\left(F_1\frac{d\bar{R}_0}{dz} + F_0\frac{d\bar{R}_1}{dz}\right), \\ R'_{2M,0} &= -\left(F_{2M,0}\frac{d\bar{R}_0}{dz} + F_0\frac{d\bar{R}_{2M,0}}{dz}\right), \\ R'_{2M,1} &= -\left(F_{2M,1}\frac{d\bar{R}_0}{dz} + F_1\frac{d\bar{R}_{2M,0}}{dz} + F_{2M,0}\frac{d\bar{R}_1}{dz} + F_0\frac{d\bar{R}_{2M,1}}{dz}\right). \end{aligned} \tag{26}$$

3 Equations in Preheat Zone

In this section, we write down the equations for a steady planar flow and perturbations to it in the preheat zone which are used to derive the boundary condition of the

mass flux in Sect. 4. Our approach successfully links the heat-conduction equation to the Navier-Stokes equation, via pressure variation, which is absent in the low-Mach-number limit. In accordance, the viscous effect was neglected in the previous work [3]. Under the long-wavelength approximation, the stretching transformation $z - F = \delta\zeta$ is employed to analyse the preheat zone of a planar flame front by the inner variable ζ with the assumption $\delta \ll 1$.

3.1 Steady Planar Flow

From (9) and (20)-(22), the governing equations for a steady planar flow are

$$\bar{M} = 1, \tag{27}$$

$$\frac{d\bar{W}}{d\zeta} = -\frac{1}{\gamma Ma^2}\frac{d\bar{P}}{d\zeta} + \frac{4}{3}Pr\frac{d^2\bar{W}}{d\zeta^2}, \tag{28}$$

$$\frac{d\bar{T}}{d\zeta} = \frac{d^2\bar{T}}{d\zeta^2} + \frac{\gamma-1}{\gamma}\bar{W}\frac{d\bar{P}}{d\zeta}, \tag{29}$$

$$\bar{P} = \bar{R}\bar{T}. \tag{30}$$

It follows from (15), (17), (27) and (30) that

$$\begin{aligned} \bar{R}_{0M}\bar{W}_{0M} = 1, \quad 1 = \bar{R}_{0M}\bar{T}_{0M}, \\ \bar{R}_{2M}\bar{W}_{0M} + \bar{R}_{0M}\bar{W}_{2M} = 0, \quad 0 = \bar{R}_{2M}\bar{T}_{0M} + \bar{R}_{0M}\bar{T}_{2M}. \end{aligned} \tag{31}$$

Remembering that $\bar{P}_{0M} = 1$, (29) is rewritten as

$$\frac{d\bar{T}_{0M}}{d\zeta} = \frac{d^2\bar{T}_{0M}}{d\zeta^2}. \tag{32}$$

Because of $\bar{W}_{0M} = \bar{T}_{0M}$ from (31), (28) yields

$$\frac{d\bar{P}_{2M}}{d\zeta} = \left(\frac{4}{3}Pr - 1\right)\frac{d\bar{T}_{0M}}{d\zeta}. \tag{33}$$

The asymptotic expansions of the steady planar solutions and small amplitude of perturbation with respect to δ in the preheat zone, by use of (15) and (16), are

$$\begin{aligned}
\bar{T}_{0M} &= \bar{\theta}_0 + \delta\bar{\theta}_1 + \delta^2\bar{\theta}_2 + \delta^3\bar{\theta}_3,\ \bar{R}_{0M} = \bar{\rho}_0 + \delta\bar{\rho}_1 + \delta^2\bar{\rho}_2 + \delta^3\bar{\rho}_3,\\
F_{0M} &= F_0 + \delta F_1 + \delta^2 F_2 + \delta^3 F_3,\ \bar{W}_{0M} = \bar{w}_0 + \delta\bar{w}_1 + \delta^2\bar{w}_2 + \delta^3\bar{w}_3,\\
\bar{T}_{2M} &= \bar{\theta}_{2M,0} + \delta\bar{\theta}_{2M,1} + \delta^2\bar{\theta}_{2M,2} + \delta^3\bar{\theta}_{2M,3},\\
\bar{R}_{2M} &= \bar{\rho}_{2M,0} + \delta\bar{\rho}_{2M,1} + \delta^2\bar{\rho}_{2M,2} + \delta^3\bar{\rho}_{2M,3},\\
\bar{W}_{2M} &= \bar{w}_{2M,0} + \delta\bar{w}_{2M,1} + \delta^2\bar{w}_{2M,2} + \delta^3\bar{w}_{2M,3},\\
\bar{P}_{2M} &= \bar{p}_{2M,0} + \delta\bar{p}_{2M,1} + \delta^2\bar{p}_{2M,2} + \delta^3\bar{p}_{2M,3},\\
F_{2M} &= F_{2M,0} + \delta F_{2M,1} + \delta^2 F_{2M,2} + \delta^3 F_{2M,3}.
\end{aligned} \tag{34}$$

By introducing (34), (31) and (33) reduce to

$$\bar{\rho}_0\bar{w}_0 = 1, \quad 1 = \bar{p}_0 = \bar{\rho}_0\bar{\theta}_0, \tag{35}$$

$$\bar{\rho}_{2M,0}\bar{w}_0 + \bar{\rho}_0\bar{w}_{2M,0} = 0, \quad 0 = \bar{\rho}_{2M,0}\bar{\theta}_0 + \bar{\rho}_0\bar{\theta}_{2M,0}, \tag{36}$$

$$\frac{d\bar{p}_{2M,0}}{d\zeta} = \left(\frac{4}{3}Pr - 1\right)\frac{d\bar{\theta}_0}{d\zeta}. \tag{37}$$

3.2 Perturbations with Translational Symmetry

The linearised heat-conduction equation (23), valid in the preheat zone, is

$$\begin{aligned}
&\bar{R}\frac{\partial T'}{\partial t} + \frac{\bar{M}}{\delta}\frac{\partial T'}{\partial \zeta} + \frac{M'}{\delta}\frac{d\bar{T}}{d\zeta}\\
&= \delta\left(\frac{1}{\delta^2}\frac{\partial^2}{\partial\zeta^2} + \Delta\right)T' + \frac{\gamma-1}{\gamma}\left(\frac{\partial P'}{\partial t} + \frac{\bar{W}}{\delta}\frac{\partial P'}{\partial\zeta} + \frac{W'}{\delta}\frac{d\bar{P}}{d\zeta}\right).
\end{aligned} \tag{38}$$

The perturbations (24) and (25) are written in terms of the inner variable as

$$T' = -\frac{F}{\delta}\frac{d\bar{T}}{d\zeta}, \quad R' = -\frac{F}{\delta}\frac{d\bar{R}}{d\zeta}, \quad P' = -\frac{F}{\delta}\frac{d\bar{P}}{d\zeta}. \tag{39}$$

Furthermore, (39) is expanded in powers of Ma^2, with the help of (15) and (16), as

$$\begin{aligned}
T'_{0M} &= -\frac{F_{0M}}{\delta}\frac{d\bar{T}_{0M}}{d\zeta}\ ,\ T'_{2M} = -\frac{F_{2M}}{\delta}\frac{d\bar{T}_{0M}}{d\zeta} - \frac{F_{0M}}{\delta}\frac{d\bar{T}_{2M}}{d\zeta},\\
R'_{0M} &= -\frac{F_{0M}}{\delta}\frac{d\bar{R}_{0M}}{d\zeta}\ ,\ R'_{2M} = -\frac{F_{2M}}{\delta}\frac{d\bar{R}_{0M}}{d\zeta} - \frac{F_{0M}}{\delta}\frac{d\bar{R}_{2M}}{d\zeta},\\
P'_{2M} &= -\frac{F_{0M}}{\delta}\frac{d\bar{P}_{2M}}{d\zeta}.
\end{aligned} \tag{40}$$

By substitution from the asymptotic expansions (34), we rewrite the perturbations (40) as

$$
\begin{aligned}
T'_{0M} &= \frac{\theta_{-1}}{\delta} + \theta_0 + \delta\theta_1 + \delta^2\theta_2, \\
R'_{0M} &= \frac{\rho_{-1}}{\delta} + \rho_0 + \delta\rho_1 + \delta^2\rho_2, \\
T'_{2M} &= \frac{\theta_{2M,-1}}{\delta} + \theta_{2M,0} + \delta\theta_{2M,1} + \delta^2\theta_{2M,2}, \\
R'_{2M} &= \frac{\rho_{2M,-1}}{\delta} + \rho_{2M,0} + \delta\rho_{2M,1} + \delta^2\rho_{2M,2}, \\
P'_{2M} &= \frac{p_{2M,-1}}{\delta} + p_{2M,0} + \delta p_{2M,1} + \delta^2 p_{2M,2},
\end{aligned}
\tag{41}
$$

where each term is written, for instance, as

$$
\begin{aligned}
\theta_{-1} &= -F_0\frac{d\bar{\theta}_0}{d\zeta}, \quad \theta_0 = -\left(F_1\frac{d\bar{\theta}_0}{d\zeta} + F_0\frac{d\bar{\theta}_1}{d\zeta}\right), \\
\theta_1 &= -\left(F_2\frac{d\bar{\theta}_0}{d\zeta} + F_1\frac{d\bar{\theta}_1}{d\zeta} + F_0\frac{d\bar{\theta}_2}{d\zeta}\right), \\
\theta_2 &= -\left(F_3\frac{d\bar{\theta}_0}{d\zeta} + F_2\frac{d\bar{\theta}_1}{d\zeta} + F_1\frac{d\bar{\theta}_2}{d\zeta} + F_0\frac{d\bar{\theta}_3}{d\zeta}\right), \\
\theta_{2M,-1} &= -\left(F_{2M,0}\frac{d\bar{\theta}_0}{d\zeta} + F_0\frac{d\bar{\theta}_{2M,0}}{d\zeta}\right), \\
\theta_{2M,0} &= -\left(F_{2M,1}\frac{d\bar{\theta}_0}{d\zeta} + F_1\frac{d\bar{\theta}_{2M,0}}{d\zeta} + F_{2M,0}\frac{d\bar{\theta}_1}{d\zeta} + F_0\frac{d\bar{\theta}_{2M,1}}{d\zeta}\right), \\
\theta_{2M,1} &= -\left(F_{2M,2}\frac{d\bar{\theta}_0}{d\zeta} + F_2\frac{d\bar{\theta}_{2M,0}}{d\zeta} + F_{2M,1}\frac{d\bar{\theta}_1}{d\zeta}\right. \\
&\qquad \left. + F_1\frac{d\bar{\theta}_{2M,1}}{d\zeta} + F_{2M,0}\frac{d\bar{\theta}_2}{d\zeta} + F_0\frac{d\bar{\theta}_{2M,2}}{d\zeta}\right).
\end{aligned}
\tag{42}
$$

The similar is true for other quantities. Following ref [3], the form associated with he translational symmetry, like (24), is not postulated for the disturbance of the normal component of the velocity field. Formally we pose the following expansions.

$$
\begin{aligned}
W'_{0M} &= w_0 + \delta w_1 + \delta^2 w_2, \\
W'_{2M} &= w_{2M,0} + \delta w_{2M,1} + \delta^2 w_{2M,2}.
\end{aligned}
\tag{43}
$$

For the perturbation of the mass flux which is introduced in (16), we have, upon substitution from (15), (41) and (43),

$$
\begin{aligned}
M'_{0M} &= \frac{m_{-1}}{\delta} + m_0 + \delta m_1 + \delta^2 m_2, \\
M'_{2M} &= \frac{m_{2M,-1}}{\delta} + m_{2M,0} + \delta m_{2M,1} + \delta^2 m_{2M,2},
\end{aligned}
\tag{44}
$$

where each term is expressed as

$$
\begin{aligned}
m_{-1} &= \rho_{-1}\bar{w}_0 = -F_0\frac{d\bar{\rho}_0}{d\zeta}\bar{w}_0,\\
m_0 &= \rho_0\bar{w}_0 + \rho_{-1}\bar{w}_1 + \bar{\rho}_0 w_0\\
&= -\left(F_1\frac{d\bar{\rho}_0}{d\zeta} + F_0\frac{d\bar{\rho}_1}{d\zeta}\right)\bar{w}_0 - F_0\frac{d\bar{\rho}_0}{d\zeta}\bar{w}_1 + \bar{\rho}_0 w_0,
\end{aligned}
\tag{45}
$$

and the same procedure is repeated for the higher-order terms.

4 Boundary Condition of Mass Flux

We seek the compressibility effect on the flame speed condition, whose original form of the DLI is given by (1). For this purpose, by leaving the detailed analysis of the preheat zone to our next paper, we treat, in this investigation, restricted solutions valid for long wavelengths brought by the constraint of translational symmetry dictated by the previous section. We extend the mass-flux condition by [3] to $O(\delta Ma^2)$. At this order, the viscous effect switches on by coupling the Navier-Stokes equation with the heat-conduction equation. The end product, equation (74), reflects, remarkably in a tidy form, the compressible effect on the Markstein number.

4.1 Matching Condition

In order to calculate the mass-flux condition, we use the matching conditions in the overlapping region between the preheat and the hydrodynamic zones. For the density, these conditions read, on the unburned side,

$$
\bar{\rho}_0|_{\zeta\to-\infty} = \bar{R}_0|_{z\to F_-}, \qquad \left.\frac{d\bar{\rho}_1}{d\zeta}\right|_{-\infty} = \left.\frac{d\bar{R}_0}{dz}\right|_{-}, \tag{46}
$$

$$
\bar{\rho}_{1-\infty} = \bar{R}_{1-} + \zeta\left.\frac{d\bar{R}_0}{dz}\right|_{-}, \quad \left.\frac{d\bar{\rho}_2}{d\zeta}\right|_{-\infty} = \left.\frac{d\bar{R}_1}{dz}\right|_{-} + \zeta\left.\frac{d^2\bar{R}_0}{dz^2}\right|_{-}, \tag{47}
$$

$$
\left.\frac{d^2\bar{\rho}_1}{d\zeta^2}\right|_{-\infty} = 0, \qquad \left.\frac{d^2\bar{\rho}_2}{d\zeta^2}\right|_{-\infty} = \left.\frac{d^2\bar{R}_0}{dz^2}\right|_{-}. \tag{48}
$$

The similar conditions apply to the other quantities. Matching conditions (46)-(48) also hold at $O(Ma^2)$, for instance, $\bar{\rho}_{2M,0-\infty} = \bar{R}_{2M,0-}$, $d\bar{\rho}_{2M,1}/d\zeta|_{-\infty} = d\bar{R}_{2M,0}/dz|_{-}$ and so on.

4.2 Low-Mach-Number Limit

We begin with the derivation of the mass-flux condition in the incompressible limit. Collecting the terms of $O(\delta^{-2}Ma^0)$ in (38), we get

$$\frac{\partial \theta_{-1}}{\partial \zeta} + m_{-1}\frac{d\bar{\theta}_0}{d\zeta} = \frac{\partial^2 \theta_{-1}}{\partial \zeta^2}. \tag{49}$$

Substitution from (42) and (45), (49) becomes

$$F_0 \bar{w}_0 \left(\frac{d\bar{\theta}_0}{d\zeta}\right)^2 = -F_0 \frac{d}{d\zeta}\left(\frac{d^2\bar{\theta}_0}{d\zeta^2} - \frac{d\bar{\theta}_0}{d\zeta}\right), \tag{50}$$

by virtue of (35). The right-hand side of (50) is zero, because the last term of (29) vanishes by $\bar{p}_0 = 1$, the second of (35). Requirement of $\bar{w}_0 \neq 0$ enforces

$$\frac{d\bar{\theta}_0}{d\zeta} = 0. \tag{51}$$

In view of (35) and (37), (51) leads to

$$\frac{d\bar{w}_0}{d\zeta} = \frac{d\bar{\rho}_0}{d\zeta} = \frac{d\bar{p}_{2M,0}}{d\zeta} = 0. \tag{52}$$

The terms of $O(\delta^{-1}Ma^0)$ do not bring any new information. We proceed to the next order. Collecting the terms of $O(\delta^0 Ma^0)$ in (38), we have

$$\bar{\rho}_1 \frac{\partial \theta_{-1}}{\partial t} + \bar{\rho}_0 \frac{\partial \theta_0}{\partial t} + \frac{\partial \theta_1}{\partial \zeta} + m_1 \frac{d\bar{\theta}_0}{d\zeta} + m_0 \frac{d\bar{\theta}_1}{d\zeta} + m_{-1}\frac{d\bar{\theta}_2}{d\zeta} = \frac{\partial^2 \theta_1}{\partial \zeta^2} + \Delta\theta_{-1}. \tag{53}$$

By substitution from (32), (42), (45), (51) and (52), we are left with

$$-F_0 \frac{d\bar{\rho}_1}{d\zeta}\bar{w}_0 + \bar{\rho}_0\left(w_0 - \frac{\partial F_0}{\partial t}\right) = 0. \tag{54}$$

Taking the outer limit $\zeta \to -\infty$, (46) gives rise to the matching condition for the hydrodynamic zone on the unburned side, resulting in

$$R'_{0-}\bar{W}_{0-} + \bar{R}_{0-}\left(W'_{0-} - \frac{\partial F_0}{\partial t}\right) = 0, \tag{55}$$

where use has been made of (26) for the density perturbation. This is the desired mass-flux condition on the unburned side of a flame front in the hydrodynamic zone, correcting Landau's assumption [13, 14] with the first term incorporating the compressibility effect. This condition coincides with that of Bychkov *et al*. [3].

We are ready to go on to the first-order solution in δ, in the preheat zone, to deal with the curvature effect, embodying the Markstein effect [16]. Collecting the terms

of $O(\delta Ma^0)$ in (38), we have, using (29), (42), (45), (51), (52) and (54),

$$-\left(F_1\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_2}{d\zeta}\right)\bar{w}_0-F_0\frac{d\bar{\rho}_1}{d\zeta}\bar{w}_1$$
$$+\bar{\rho}_1\left(w_0-\frac{\partial F_0}{\partial t}\right)+\bar{\rho}_0\left(w_1-\frac{\partial F_1}{\partial t}\right)+\Delta F_0=0. \tag{56}$$

Matching with the hydrodynamic zone, by taking the limit $\zeta\to-\infty$ of (56), with use of (46) and (47), leads to

$$-\left(F_1\frac{d\bar{R}_0}{dz}\bigg|_-+F_0\frac{d\bar{R}_1}{dz}\bigg|_-\right)\bar{W}_{0-}-F_0\frac{d\bar{R}_0}{dz}\bigg|_-\bar{W}_{1-}$$
$$+\bar{R}_{1-}\left(W'_{0-}-\frac{\partial F_0}{\partial t}\right)+\bar{R}_{0-}\left(W'_{1-}-\frac{\partial F_1}{\partial t}\right)+\Delta F_0$$
$$=\zeta\left\{F_0\frac{d^2\bar{R}_0}{dz^2}\bigg|_-\bar{W}_{0-}+F_0\frac{d\bar{R}_0}{dz}\bigg|_-\frac{d\bar{W}_0}{dz}\bigg|_-\right.$$
$$\left.-\frac{d\bar{R}_0}{dz}\bigg|_-\left(W'_{0-}-\frac{\partial F_0}{\partial t}\right)-\bar{R}_{0-}\frac{dW'_0}{dz}\bigg|_-\right\}. \tag{57}$$

The right-hand side of (57) diverges in the limit of $\zeta\to-\infty$. But this difficulty is rescued by the equation obtained from the derivative of (56) with respect to ζ,

$$-\left(F_1\frac{d^2\bar{\rho}_1}{d\zeta^2}+F_0\frac{d^2\bar{\rho}_2}{d\zeta^2}\right)\bar{w}_0-\left(F_1\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_2}{d\zeta}\right)\frac{d\bar{w}_0}{d\zeta}-F_0\frac{d^2\bar{\rho}_1}{d\zeta^2}\bar{w}_1-F_0\frac{d\bar{\rho}_1}{d\zeta}\frac{d\bar{w}_1}{d\zeta}$$
$$+\frac{d\bar{\rho}_1}{d\zeta}\left(w_0-\frac{\partial F_0}{\partial t}\right)+\bar{\rho}_1\frac{d\bar{w}_0}{d\zeta}+\frac{d\bar{\rho}_0}{d\zeta}\left(w_1-\frac{\partial F_1}{\partial t}\right)+\bar{\rho}_0\frac{d\bar{w}_1}{d\zeta}=0. \tag{58}$$

The outer limit ($\zeta\to-\infty$) of (58), with application of (46), (47) and (48), results in

$$-F_0\frac{d^2\bar{R}_0}{dz^2}\bigg|_-\bar{W}_{0-}-F_0\frac{d\bar{R}_0}{dz}\bigg|_-\frac{d\bar{W}_0}{dz}\bigg|_-+\frac{d\bar{R}_0}{dz}\bigg|_-\left(W'_{0-}-\frac{\partial F_0}{\partial t}\right)+\bar{R}_{0-}\frac{dW'_0}{dz}\bigg|_-=0.$$

Thus, the diverging terms in (57) cancel each other and we eventually reach the mass flux of $O(\delta)$ by using (26) for the density perturbation.

$$R'_{1-}\bar{W}_{0-}+R'_{0-}\bar{W}_{1-}+\bar{R}_{1-}\left(W'_{0-}-\frac{\partial F_0}{\partial t}\right)+\bar{R}_{0-}\left(W'_{1-}-\frac{\partial F_1}{\partial t}\right)=-\Delta F_0. \tag{59}$$

This is the desired mass-flux condition reflecting the curvature of a flame front, a feature of the Markstein condition [16]. In the context of the long-wave approximation, the previous investigation [3] did not enter into $O(\delta)$, and the condition (59) is new.

4.3 Compressibility Effect

We make headway to deduce the mass flux on the unburned side of a flame front, with the compressibility effect taken into account, based only on the heat-conduction equation. Collecting the terms of $O(\delta^{-2}Ma^2)$ in (38), we have

$$\frac{\partial\theta_{2M,-1}}{\partial\zeta}+m_{2M,-1}\frac{d\bar{\theta}_0}{d\zeta}+m_{-1}\frac{d\bar{\theta}_{2M,0}}{d\zeta}=\frac{\partial^2\theta_{2M,-1}}{\partial\zeta^2}+(\gamma-1)\bar{w}_0\frac{\partial p_{2M,-1}}{\partial\zeta}. \quad (60)$$

The last term vanishes because the analogue of the first of (42) reads $p_{2M,-1}=-F_0 d\bar{p}_{2M,0}/d\zeta=0$, the latter equality coming from (52). By use of (45), (51), (52) and the variant of (42), (60) becomes

$$\frac{\partial}{\partial\zeta}\left(-F_0\frac{d\bar{\theta}_{2M,0}}{d\zeta}\right)=\frac{\partial^2}{\partial\zeta^2}\left(-F_0\frac{d\bar{\theta}_{2M,0}}{d\zeta}\right),$$

which is satisfied by (29) because of (52).

Next, collecting the terms of $O(\delta^{-1}Ma^2)$ in (38), we have

$$\begin{aligned}&\bar{\rho}_{2M,0}\frac{\partial\theta_{-1}}{\partial t}+\bar{\rho}_0\frac{\partial\theta_{2M,-1}}{\partial t}+\frac{\partial\theta_{2M,0}}{\partial\zeta}+m_{2M,0}\frac{d\bar{\theta}_{0M}}{d\zeta}\\&+m_{2M,-1}\frac{d\bar{\theta}_1}{d\zeta}+m_0\frac{d\bar{\theta}_{2M,0}}{d\zeta}+m_{-1}\frac{d\bar{\theta}_{2M,1}}{d\zeta}\\&=\frac{\partial^2\theta_{2M,0}}{\partial\zeta^2}+(\gamma-1)\left(\frac{\partial p_{2M,-1}}{\partial t}+\bar{w}_1\frac{\partial p_{2M,-1}}{\partial\zeta}+\bar{w}_0\frac{\partial p_{2M,0}}{\partial\zeta}+w_0\frac{d\bar{p}_{2M,0}}{d\zeta}\right).\end{aligned}$$

By use of (29), (42), (45), (51), (52) and (54), we are left only with

$$-F_0\frac{d\bar{\rho}_{2M,0}}{d\zeta}\bar{w}_0\frac{d\bar{\theta}_1}{d\zeta}=0.$$

The temperature gradient $d\bar{\theta}_1/d\zeta$ should not be zero because the θ_0 term in (42) should not be zero due to the matching condition $\theta_0|_{\zeta\to-\infty}\to T_{0-}\neq 0$ and $d\bar{\theta}_0/d\zeta=0$ by (51). Consequently, we have no choice but to put the density gradient zero.

$$\frac{d\bar{\rho}_{2M,0}}{d\zeta}=0. \quad (61)$$

It follows from (36) that

$$\frac{d\bar{w}_{2M,0}}{d\zeta}=0,\quad \frac{d\bar{\theta}_{2M,0}}{d\zeta}=0, \quad (62)$$

with the help of (51) and (52).

Collecting the terms of $O(\delta^0 Ma^2)$ in (38), taking account of (29), (42), (45), (51), (52), (61) and (62), we have

$$\left\{-\bar{\rho}_{2M,0}\frac{\partial F_0}{\partial t}-\bar{\rho}_0\frac{\partial F_{2M,0}}{\partial t}-\left(F_{2M,0}\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_{2M,1}}{d\zeta}\right)\bar{w}_0\right. \tag{63}$$

$$\left.-F_0\frac{d\bar{\rho}_1}{d\zeta}\bar{w}_{2M,0}+\bar{\rho}_{2M,0}w_0+\bar{\rho}_0 w_{2M,0}\right\}\frac{d\bar{\theta}_1}{d\zeta} \tag{64}$$

$$=\frac{\gamma-1}{\bar{\rho}_0}\left(\bar{\rho}_0 w_0-\bar{\rho}_0\frac{\partial F_0}{\partial t}-\bar{w}_0 F_0\frac{d\bar{\rho}_1}{d\zeta}\right)\frac{d\bar{p}_{2M,1}}{d\zeta}, \tag{65}$$

where, in the same manner as $O(\delta^0 Ma^0)$, the first of (42), valid in the long-wavelength approximation, has dictated vanishing of $\theta_{2M,-1}$ and therefore of $\Delta\theta_{2M,-1}$ because of (62). The right-hand side is eliminated owing to (54), and (65) further simplifies to

$$-\left(F_{2M,0}\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_{2M,1}}{d\zeta}\right)\bar{w}_0-F_0\frac{d\bar{\rho}_1}{d\zeta}\bar{w}_{2M,0}$$
$$+\bar{\rho}_{2M,0}\left(w_0-\frac{\partial F_0}{\partial t}\right)+\bar{\rho}_0\left(w_{2M,0}-\frac{\partial F_{2M,0}}{\partial t}\right)=0. \tag{66}$$

Taking the outer limit $\zeta\to-\infty$, with use of (46), gives the boundary condition on the hydrodynamic zone.

$$R'_{2M,0-}\bar{W}_{0-}+R'_{0-}\bar{W}_{2M,0-}$$
$$+\bar{R}_{2M,0-}\left(W'_{0-}-\frac{\partial F_0}{\partial t}\right)+\bar{R}_{0-}\left(W'_{2M,0-}-\frac{\partial F_{2M,0}}{\partial t}\right)=0, \tag{67}$$

where we notice from the second of (26) and (46) that the first two terms of (66) become the first term of (67) as $\zeta\to-\infty$. This implies that the perturbed mass flux on the unburned side of the flame front is absent, an extension of Landau's assumption to the compressible case. The condition (67) coincides with that of Bychkov *et al.* [3].

In the long-wave approximation admitting the translational symmetry, to the leading order in δ, the perturbed mass flux is zero at the flame front even when the compressibility effect is included. The compressibility has a non-trivial influence on the mass flux at the next order in δ, at which the contributions from the curvature of the flame front and the viscosity play an vital role. Collecting the terms of $O(\delta Ma^2)$ in (38), using the conditions (29), (42), (45), (51), (52), (61) and (62), we have

$$
\begin{aligned}
&\frac{d\bar{\theta}_1}{d\zeta}\left\{-\bar{\rho}_{2M,1}\frac{\partial F_0}{\partial t}-\bar{\rho}_{2M,0}\frac{\partial F_1}{\partial t}-\bar{\rho}_1\frac{\partial F_{2M,0}}{\partial t}-\bar{\rho}_0\frac{\partial F_{2M,1}}{\partial t}\right.\\
&-\left(F_{2M,1}\frac{d\bar{\rho}_1}{d\zeta}+F_{2M,0}\frac{d\bar{\rho}_2}{d\zeta}+F_1\frac{d\bar{\rho}_{2M,1}}{d\zeta}+F_0\frac{d\bar{\rho}_{2M,2}}{d\zeta}\right)\bar{w}_0\\
&-\left(F_{2M,0}\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_{2M,1}}{d\zeta}\right)\bar{w}_1-\left(F_1\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_2}{d\zeta}\right)\bar{w}_{2M,0}\\
&\left.-F_0\frac{d\bar{\rho}_1}{d\zeta}\bar{w}_{2M,1}+\bar{\rho}_{2M,1}w_0+\bar{\rho}_{2M,0}w_1+\bar{\rho}_1w_{2M,0}+\bar{\rho}_0w_{2M,1}\right\}\\
&=-\Delta F_{2M,0}\frac{d\bar{\theta}_1}{d\zeta}+(\gamma-1)\left\{\left(-\frac{\partial F_1}{\partial t}+F_0\frac{d\bar{w}_2}{d\zeta}+F_1\frac{d\bar{w}_1}{d\zeta}+w_1\right)\frac{d\bar{p}_{2M,1}}{d\zeta}\right.\\
&\qquad\left.+\left(-\frac{\partial F_0}{\partial t}+F_0\frac{d\bar{w}_1}{d\zeta}+w_0\right)\frac{d\bar{p}_{2M,2}}{d\zeta}\right\}. \qquad (68)
\end{aligned}
$$

In order to reduce the right-hand side of (68), we invoke (31) and (33). We see from (31) that

$$
\begin{aligned}
\bar{\rho}_1\bar{w}_0+\bar{\rho}_0\bar{w}_1&=0,\\
\bar{\rho}_2\bar{w}_0+\bar{\rho}_1\bar{w}_1+\bar{\rho}_0\bar{w}_2&=0.
\end{aligned}
$$

Then, because of (52), $d\bar{w}_1/d\zeta$ and $d\bar{w}_2/d\zeta$ are rewritten as

$$
\begin{aligned}
\frac{d\bar{w}_1}{d\zeta}&=-\frac{\bar{w}_0}{\bar{\rho}_0}\frac{d\bar{\rho}_1}{d\zeta},\\
\frac{d\bar{w}_2}{d\zeta}&=-\frac{\bar{w}_0}{\bar{\rho}_0}\frac{d\bar{\rho}_2}{d\zeta}-\frac{d\bar{\rho}_1}{d\zeta}\frac{\bar{w}_1}{\bar{\rho}_0}-\frac{\bar{\rho}_1}{\bar{\rho}_0}\frac{d\bar{w}_1}{d\zeta}. \qquad (69)
\end{aligned}
$$

Upon substitution of (34) into (33), we get

$$
\frac{d\bar{p}_{2M,1}}{d\zeta}=\left(\frac{4}{3}Pr-1\right)\frac{d\bar{\theta}_1}{d\zeta}. \qquad (70)
$$

By taking advantage of (56), (69) and (70), we reduce (68) to

$$
\begin{aligned}
&-\bar{\rho}_{2M,1}\frac{\partial F_0}{\partial t}-\bar{\rho}_{2M,0}\frac{\partial F_1}{\partial t}-\bar{\rho}_1\frac{\partial F_{2M,0}}{\partial t}-\bar{\rho}_0\frac{\partial F_{2M,1}}{\partial t}\\
&-\left(F_{2M,1}\frac{d\bar{\rho}_1}{d\zeta}+F_{2M,0}\frac{d\bar{\rho}_2}{d\zeta}+F_1\frac{d\bar{\rho}_{2M,1}}{d\zeta}+F_0\frac{d\bar{\rho}_{2M,2}}{d\zeta}\right)\bar{w}_0\\
&-\left(F_{2M,0}\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_{2M,1}}{d\zeta}\right)\bar{w}_1-\left(F_1\frac{d\bar{\rho}_1}{d\zeta}+F_0\frac{d\bar{\rho}_2}{d\zeta}\right)\bar{w}_{2M,0}\\
&-F_0\frac{d\bar{\rho}_1}{d\zeta}\bar{w}_{2M,1}+\bar{\rho}_{2M,1}w_0+\bar{\rho}_{2M,0}w_1+\bar{\rho}_1w_{2M,0}+\bar{\rho}_0w_{2M,1}\\
&=-\Delta F_{2M,0}-\frac{\gamma-1}{\bar{\rho}_0}\left(\frac{4}{3}Pr-1\right)\Delta F_0. \qquad (71)
\end{aligned}
$$

The outer limit $\zeta \to -\infty$ of (71) produces terms proportional to ζ, and we are requested to show cancellation of these, otherwise diverging, terms. To confirm this, it suffices to take the derivative of (71) with respect to ζ, to take the outer limit, and to impose the matching conditions (46), (47) and (48). The resulting equation is

$$
\begin{aligned}
&-\frac{d\bar{R}_{2M,0}}{dz}\bigg|_{-}\frac{\partial F_0}{\partial t}-\frac{d\bar{R}_0}{dz}\bigg|_{-}\frac{\partial F_{2M,0}}{\partial t}-\left(F_{2M,0}\frac{d^2\bar{R}_0}{dz^2}\bigg|_{-}+F_0\frac{d^2\bar{R}_{2M,0}}{dz^2}\bigg|_{-}\right)\bar{W}_{0-}\\
&-\left(F_{2M,0}\frac{d\bar{R}_0}{dz}\bigg|_{-}+F_0\frac{d\bar{R}_{2M,0}}{dz}\bigg|_{-}\right)\frac{d\bar{W}_0}{dz}\bigg|_{-}-F_0\frac{d^2\bar{R}_0}{dz^2}\bigg|_{-}\bar{W}_{2M,0-}-F_0\frac{d\bar{R}_0}{dz}\bigg|_{-}\frac{d\bar{W}_{2M,0}}{dz}\bigg|_{-}\\
&+\frac{d\bar{R}_{2M,0}}{dz}\bigg|_{-}W'_{0-}+\bar{R}_{2M,0-}\frac{\partial W'_0}{\partial z}\bigg|_{-}+\frac{d\bar{R}_0}{dz}\bigg|_{-}W'_{2M,0-}+\bar{R}_{0-}\frac{\partial W'_{2M,0}}{\partial z}\bigg|_{-}=0. \qquad (72)
\end{aligned}
$$

The remaining task is to take the outer limit of (71), by imposing (46) and (47), to get the mass flux of $O(\delta Ma^2)$, on the unburned side, leaving, with the help of (26) and (72),

$$
\begin{aligned}
&R'_{2M,1-}\bar{W}_{0-}+R'_{2M,0-}\bar{W}_{1-}+R'_{1-}\bar{W}_{2M,0-}+R'_{0-}\bar{W}_{2M,1-}+\bar{R}_{2M,1-}\left(W'_{0-}-\frac{\partial F_0}{\partial t}\right)\\
&+\bar{R}_{2M,0-}\left(W'_{1-}-\frac{\partial F_1}{\partial t}\right)+\bar{R}_{1-}\left(W'_{2M,0-}-\frac{\partial F_{2M,0}}{\partial t}\right)+\bar{R}_{0-}\left(W'_{2M,1-}-\frac{\partial F_{2M,1}}{\partial t}\right)\\
&=-\Delta F_{2M,0}-\frac{\gamma-1}{\bar{R}_{0-}}\left(\frac{4}{3}Pr-1\right)\Delta F_0. \qquad (73)
\end{aligned}
$$

This result attains an extension of the Markstein effect to the compressible case. The curvature effect, in combination with the Prandtl number, appears for the mass flux. This implies that the viscosity should be retained when we consider the compressible flow field.

The above results (55), (59), (67) and (73) are summarized as

$$
R_-\left(W_- -\frac{\partial F}{\partial t}\right)=1-\delta\left(1+Ma^2(\gamma-1)\left(\frac{4}{3}Pr-1\right)\right)\Delta F. \qquad (74)
$$

The flame speed (3) is obtained from (74) by assuming that the flow field is incompressible in the hydrodynamic scale as

$$
R=\begin{cases}1 & (z<F)\\ 1/(1+q) & (z>F)\end{cases}. \qquad (75)
$$

5 Effect of Compressible Markstein Number on DLI

We are now in a position to look into how the compressibility modifies the DLI. As indicated by (3), the compressibility effect is incorporated into the condition of a flame speed, though the flow field is assumed to be incompressible in the hydro-

dynamic regions. The flame speed S_f is defined as the normal velocity of the fluid relative to that of a flame front, which is evaluated at the edge of the front on the unburned side.

$$S_f = (\vec{V} - \vec{V}_f)|_{z=F_-} \cdot \vec{n} \approx \bar{W}|_{z=F_-} + W'|_{z=F_-} - \frac{\partial F}{\partial t}, \tag{76}$$

where the normal velocity of a flame front and the unit normal vector are given by

$$\vec{V}_f \cdot \vec{n} \approx \frac{\partial F}{\partial t}, \quad \vec{n} \approx \left(-\frac{\partial F}{\partial x}, -\frac{\partial F}{\partial y}, 1 \right). \tag{77}$$

At $O(\delta^0 Ma^0)$, we solve the following linearised equations of (5)-(7) for the perturbation form of (13) on the unburned and burned sides of a flame front.

$$\frac{\partial W_0'}{\partial z} + \nabla \cdot \mathbf{V}_0' = 0, \tag{78}$$

$$R\frac{\partial W_0'}{\partial t} + \frac{\partial W_0'}{\partial z} = -\frac{\partial P_{2M,0}'}{\partial z}, \tag{79}$$

$$R\frac{\partial \mathbf{V}_0'}{\partial t} + \frac{\partial \mathbf{V}_0'}{\partial z} = -\nabla P_{2M,0}', \tag{80}$$

where the density R is assumed to be constant as given by (75). The following jump conditions are imposed at a flame front, $z = F_\pm$, at $O(\delta^0 Ma^0)$.

$$[\![R(\vec{V}_0 - \vec{V}_f) \cdot \vec{n}]\!] = 0, \tag{81}$$

$$[\![\vec{V}_0 \times \vec{n}]\!] = \vec{0}, \tag{82}$$

$$[\![P_{2M,0} + R((\vec{V}_0 - \vec{V}_f) \cdot \vec{n})^2]\!] = 0, \tag{83}$$

where, for any quantity ϕ, $[\![\phi]\!] = \phi(z = F_+) - \phi(z = F_-)$ denotes the jump across a flame front in the hydrodynamic zone. In addition to (81)-(83), the condition of a flame speed for perturbations is given by (3), with the help of (76), as

$$W_0'|_{z=F_-} - \frac{\partial F}{\partial t} = -Mr_M \Delta F. \tag{84}$$

Enforcing (81)-(84) on the solutions of (78)-(80), we gain

$$(\sigma+1)\Omega^2 + 2(1 + Mr_M k)\sigma k\Omega - (\sigma - 1 - 2\sigma Mr_M k)\sigma k^2 = 0, \tag{85}$$

where the non-dimensional growth rate Ω is defined by (12) and $\sigma = 1 + q$ is the thermal expansion ratio, with q (> 0) the non-dimensional heat release as defined in Sect. 2. By imposing the condition of $\Omega = 0$, we find the critical wavenumber as

$$k_c = \frac{\sigma - 1}{2\sigma Mr_M}. \tag{86}$$

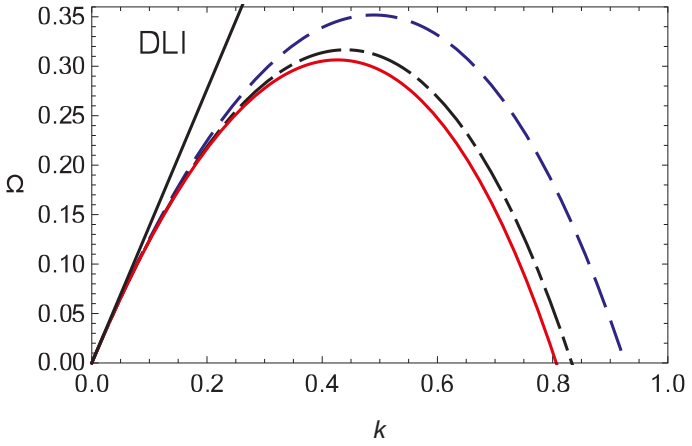


Fig. 2 Growth rate Ω v.s. wavenumber k with $\gamma = 1.4$, $\sigma = 6$, $Ma = 0.5$ and $\delta = 0.5$ for several values of Prandtl number: $Pr = 0$ (dashed), $Pr = 3/4$ (dot-dashed) and $Pr = 1$ (solid). The DLI is also plotted for comparison.

Because positivity of the Markstein number, $Mr_M > 0$, brings the decrease of the growth rate, we need the following requirement for the suppression of the DLI for $k > k_c$.

$$1 + Ma^2(\gamma - 1)\left(\frac{4}{3}Pr - 1\right) > 0. \tag{87}$$

This condition means that if $Pr > 3/4$, the increase of the value of Ma absolutely reinforce the suppression of the DLI. On the other hand, if $0 < Pr < 3/4$, there is a possibility of the enhancement of the DLI, or $Mr_M < 0$, by the compressibility effect. However, under the condition of $Ma^2 \ll 1$, this is unlikely to occur.

Finally, we plot the solution of (85) in Figs. 2 and 3, which is given by

$$\Omega = -\frac{\sigma}{1+\sigma}(1 + Mr_M k)k + \frac{1}{1+\sigma}\left\{\sigma^2(1 + Mr_M k)^2 + \sigma(\sigma^2 - 1)\left(1 - 2\frac{\sigma}{\sigma - 1}Mr_M k\right)\right\}^{1/2} k. \tag{88}$$

We observe from Fig. 2 that the DLI is suppressed at the critical wavenumber given by (86). In the range of $Pr > 3/4$, the reduction of the growth rate is achieved by the increase of the Mach number as shown in Figure 3. Conversely, the rise of the growth rate is caused by the compressibility in the range of $0 < Pr < 3/4$, though such a growth rate is still less than the DLI.

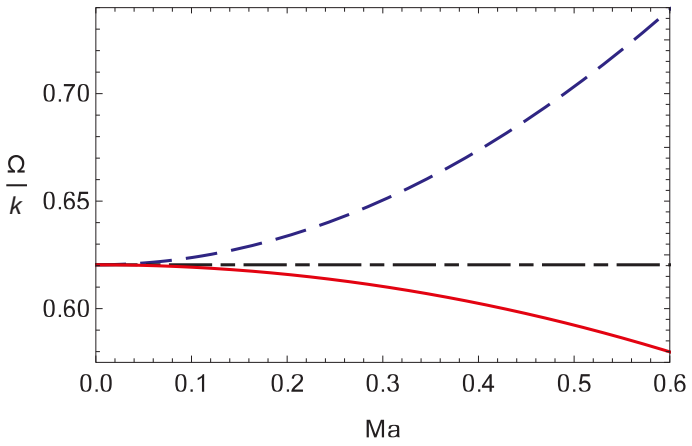


Fig. 3 Growth rate Ω/k v.s. Mach number Ma with the same parameters as Fig. 2 but for $k = 0.5$.

6 Conclusions

We have investigated the effect of the compressibility on the Markstein number by use of the M^2 expansions. Our analysis has been performed on the scale of the preheat zone, represented by δ, by employing the matched asymptotic expansions. The compressibility brings the pressure variation term as a heat source in the heat-conduction equation. The pressure term connects the heat-conduction equation with the Navier-Stokes equation. As a consequence, the viscous effect tales part in the compressibility correction to the Markstein number.

In this investigation, we have appealed to the long-wavelength approximation for the perturbations. The resulting condition of the mass flux (74) implies no perturbation of the mass flux to $O(\delta^0)$. However, at $O(\delta)$, the perturbation of the mass flux is generated due to the curvature effect which is virtually equivalent to the Markstein effect [16]. The term of $O(\delta Ma^2)$ is new, of compressibility origin, which modifies the magnitude of the Markstein number.

The influence of the compressibility on the Darrieus-Landau instability (DLI) is discussed in Sect. 5. Enhancement or reduction of the Markstein effect is sensitive to the value of the Prandtl number Pr. In the range of $0 < Pr < 3/4$, the compressibility leads to the increase of the growth rate of infinitesimal perturbations, though its value is still less than that of the DLI. On the other hand, if $Pr > 3/4$, then the growth rate decreases as the Mach number increases.

The ansatz (24) for the form of infinitesimal perturbation drastically facilitates the integration of the coupled system of the heat-conduction and the Navier-Stokes equations. In a companion paper [23], we tackle with the burning-rate eigenvalue problem in the reaction zone, with allowance for compressibility, and thereby manipulate the laminar flame speed. The present investigation establishes a concise formula (4) for the Markstein number with the compressibility taken into account,

though assuming the laminar flame speed to be unity, the first term of (3). These two pieces of papers complements each other. There are a lot to be examined concerning the compressibility effect on combustions, for instance, the inertia and acceleration effects [1, 10] and the deflagration to detonation transition (DDT) [24]. These effects are left for future study.

Acknowledgements We are grateful to Snezhana Abarzhi, Moshe Matalon, Kaname Matsue and Michael Tribelsky for helpful discussions and invaluable comments. Y.F. was supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (grant no.19K03672).

References

1. Abarzhi, S.I., Fukumoto, Y., Kadanoff, L.P.: Stability of a hydrodynamic discontinuity. Phys. Scr. **90**, 018,002 (2015)
2. Buckmaster, J.: The quenching of two-dimensional premixed flames. Acta Astronautica **6**, 741–769 (1979)
3. Bychkov, V.V., Modestov, M., Marklund, M.: The darrieus-landau instability in fast deflagration and laser ablation. Phys. Plasmas **15**, 032,702 (2008)
4. Class, A.G., Matkowsky, B.J., Klimenko, A.Y.: Stability of planar flames as gasdynamic discontinuities. J. Fluid Mech. **491**, 51–63 (2003)
5. Class, A.G., Matkowsky, B.J., Klimenko, A.Y.: A unified model of flames as gasdynamic discontinuities. J. Fluid Mech. **491**, 11–49 (2003)
6. Darrieus, G.: unpublished works presented at la technique moderne (1938)
7. Eckhaus, W.: On the stability of laminar flame-fronts. M.I.T Fluid Dynamics Research Group Report (59-4) (1959)
8. Eckhaus, W.: Theory of flame-front stability. J. Fluid Mech. **10**, 80–100 (1961)
9. He, L.: Analysis of compressibility effects on darrieus-landau instability of deflagration wave. Europhys. Lett. **49**, 576–582 (2000)
10. Ilyin, D.V., Fukumoto, Y., Goddard III, W.A., Abarzhi, S.I.: Analysis of dynamics, stability, and flow fields' structure of an accelerated hydrodynamic discontinuity with interfacial mass flux by a general matrix method. Phys. Plasmas **25**, 112,105 (2018)
11. Kadowaki, S.: Instability of a deflagration wave propagating with finite mach number. Phys. Fluids **7**, 220–222 (1995)
12. Kadowaki, S., Mashiko, T., Kobayashi, H.: Unstable behavior of premixed flames generated by hydrodynamic and diffusive-thermal effects. J. Combust. Soc. Japan **45**, 177–183 (2003)
13. Landau, L.D.: On the theory of slow combustion. Acta Phys. (USSR) **19** (1944)
14. Landau, L.D., Lifshitz, E.M.: Fluid Mechanics : Course of Theoretical Physics Vol. 6, 2nd edn. Butterworth-Heinemann (1987)
15. Liberman, M.A., Bychkov, V.V., Golberg, S.M., Book, D.L.: stability of a planar flame front in the slow-combustion regime. Phys. Rev. **49**, 445–453 (1994)
16. Markstein, G.H.: Experimental and theoretical studies of flame-front stability. J. Aero. Sci. **18**, 199–209 (1951)
17. Matalon, M.: On flame stretch. Combust. Sci. Tech. **31**, 169–181 (1983)
18. Matalon, M., Matkowsky, B.J.: Flames as gasdynamic discontinuities. J. Fluid Mech. **124**, 239–259 (1982)
19. Matkowsky, B.J.: On flames as discontinuity surfaces in gasdynamic flows. A Celebration of Mathematical Modeling **The Joseph B. Keller Anniversary Volume**, 137–160 (2004)

20. Matkowsky, B.J., Sivashinsky, G.I.: An asymptotic derivation of two models in flame theory associated with the constant density approximation. SIAM J. Appl. Math. **37**, 686–699 (1979)
21. Sivashinsky, G.I.: Structure of bunsen flames. J. Chem. Phys. **62**, 638–643 (1975)
22. Sivashinsky, G.I.: On a distorted flame front as a hydrodynamic discontinuity. Acta Astronautica **3**, 889–918 (1976)
23. Wada, K., Fukumoto, Y.: Mallard-le-chatelier formula for laminar flame speed with volumetric heat loss caused by compressibility effect. preprint
24. Williams, F.A.: Combustion Theory: The Fundamental Theory of Chemically Reacting Flow Systems, 2nd edn. Addison-Wesley (1985)

Computational fluid dynamics modelling of a transient solids concentration in a lagoon

Ashfaq A. Khan, Yan Ding

Abstract Investigation of slurry flows is important for the mineral industry, biomass processing and waste processing. In the design of slurry handling systems such as channel flows, separators where solids concentrates are separated from clear liquid streams, knowledge of physics underlying slurry flows is required. In this study, slurry flows in tanks have been investigated. The transient profiles of the solids concentration along the length have been modelled using computational fluid dynamics(CFD). This investigation examines multiphase flows with settling solids in a non-Newtonian flow. The dynamical model gives guidance in determining formation accumulation of solids as a sludge blanket. In addition the clear liquid solids interface position has been determined this is needed for the recycle of the clear water for water conservation.

1 Introduction

Biological treatment processes are widely used in wastewater treatment plants. One of the key factors that control the efficiencies of the plant is the separation processes that remove the solids concentrated streams from the clear liquids as discussed by Li([2]). The resulting sludge is caused by sedimentation that consists of settling solids which have to be of sufficient size to get an efficient settling velocity. In order to accomplish this the fluid dynamics underlying the multiphase non-Newtonian flow needs to be understood. For wastewater systems this is done in large tanks or circular basins in order to achieve enough hydraulic time to settle the solids flocs into a sludge layer. In a previous work, Zhou and McCorquodale([1]) modelled the flow in a rectangular tank in a simplified model which did not consider the rheological effects of the solids. Lakehal et al([5]) further included non-Newtonian effects in

Ashfaq A. Khan and Yan Ding
Mathematical and Geospatial Sciences, School of Science, RMIT University, Melbourne, Australia
e-mail: ash.khan@rmit.edu.au, e-mail: yan.ding@rmit.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_20

modelling sludge flows in a wastewater basin. In this investigation, we adopted the approach used by Lakehal et al([5]). We combine the rheology of wastewater with large scale settling in turbulent flows and investigated the sludge flow in a lagoon. Our model flows is represented by a two dimensional tank that is 8 meters high and 50 meters in length. In this tank we assume a feed of slurry at one end in the top three meters. There are two outlets in the next end; a top outlet flow for the clear liquid withdrawal, and the bottom outlet for the sludge withdrawal; both these outlets are one meter in length. To our knowledge this approach has not been used on such scale for wastewater systems.

2 Governing equations

The system of equations include the continuity equation and the general equations of motion. They have been modified to include the complex physics of the clarification process. The overall multiphase flow process is turbulent and we have used the $k-\varepsilon$ model for this. Detailed explanation of the symbols need to be referred to Lakehal et al([5]). We have incorporated these modifications as User Defined Functions(UDFs) into Fluent([6]).

Continuity and X - Momentum Equations

$$\frac{\partial V_x}{\partial x}+\frac{\partial V_y}{\partial y}=0 \tag{1}$$

$$\rho\frac{\partial V_x}{\partial t}+\rho\frac{\partial V_x^2}{\partial x}+\rho\frac{\partial V_xV_y}{\partial y}=-\frac{\partial p}{\partial x}+\frac{\partial}{\partial x}(2\mu_t\frac{\partial V_x}{\partial x})+\frac{\partial}{\partial y}[\mu_t(\frac{\partial V_x}{\partial y}+\frac{\partial V_y}{\partial x})]+\frac{gC(\rho_p-\rho_w)}{\rho_w} \tag{2}$$

The last term in this equation is a source term for momentum in the X direction. The density difference provides a buoyancy effect.

Y - Momentum Equations

$$\rho\frac{\partial V_y}{\partial t}+\rho\frac{\partial V_y^2}{\partial y}+\rho\frac{\partial V_xV_y}{\partial x}=-\frac{\partial p}{\partial y}+\frac{\partial}{\partial x}(\mu_t(\frac{\partial V_x}{\partial y}+\frac{\partial V_y}{\partial x})+\frac{\partial}{\partial y}[2\mu_t(\frac{\partial V_x}{\partial y})] \tag{3}$$

The turbulence is described by k and ε by

$$\rho\frac{\partial k}{\partial t}+\rho\frac{\partial V_xk}{\partial x}+\rho\frac{\partial V_yk}{\partial y}=\frac{\partial}{\partial x}[(\mu+\frac{\mu_t}{\sigma_k})\frac{\partial k}{\partial x}]+\frac{\partial}{\partial y}[(\mu+\frac{\mu_t}{\sigma_k})(\frac{\partial k}{\partial y})]+G_k+G_b-\rho\varepsilon \tag{4}$$

$$\rho\frac{\partial\varepsilon}{\partial t}+\rho\frac{\partial V_x\varepsilon}{\partial x}+\rho\frac{\partial V_y\varepsilon}{\partial y}=\frac{\partial}{\partial x}[(\mu+\frac{\mu_t}{\sigma_\varepsilon})\frac{\partial\varepsilon}{\partial x}]+\frac{\partial}{\partial y}[(\mu+\frac{\mu_t}{\sigma_\varepsilon})(\frac{\partial\varepsilon}{\partial y})]+$$
$$C_1\varepsilon\frac{\varepsilon}{k}(G_k-C_3\varepsilon G_b)-\rho C_2\varepsilon\frac{\varepsilon^2}{k} \quad (5)$$

The convection-diffusion equation is used to compute the field of suspended solids concentration C

$$\rho\frac{\partial C}{\partial t}+\rho\frac{\partial(V_x+V_s)C}{\partial x}+\rho\frac{\partial(V_yC)}{\partial y}=\frac{\partial}{\partial x}\left[\frac{\mu_t}{\sigma_c}\frac{\partial C}{\partial x}\right]+\frac{\partial}{\partial y}\left[\frac{\mu_t}{\sigma_c}\frac{\partial C}{\partial y}\right] \quad (6)$$

The value for the turbulent Schmidt number is 0.7 which is a typical value for free flow and near wall flow as applied to this situation. The solids settling velocity V_s is modelled using a settling function of Takacs[3]).

$$V_s=V_{s0}exp[-r_h(C-C_{ns})]-V_{s0}exp[-r_p(C-C_{ns})] \quad (7)$$

This approach adequately describes the hindered settling of activated sludge.
To physically characterize the rheology of the sludge, we have used the Bingham turbulent constitutive equation used by Dahl[4]) to characterize the slurry. The yield stress τ_b is function of the solids concentration. The shear stress is given as

$$\tau_{xy}=-(\frac{\tau_b}{2\gamma}+\mu_p+\mu_t)(\frac{\partial V_x}{\partial y}+\frac{\partial V_y}{\partial x}), \quad (8)$$

where the turbulent viscosity μ_t is dependent on k and ε,

$$\mu_t=\rho C_\mu\frac{k^2}{\varepsilon},$$

and the yield stress τ_b is given by

$$\tau_b=\beta_1 exp(\beta_2 C).$$

Table 1 lists the values given to the parameters that are used in this simulation

3 Results and Discussion

We are interested in flows in lagoons with long lengths; in which in this case is 50 meters long and 8 meters high. The initial condition has water in the tank with no solids and zero velocity. The solids are introduced in the inlet stream at a steady

Table 1 Parameters used for the simulation

Parameters	Description	Value
U_in	Inflow Velocity	0.019 $\frac{m}{s}$
C_in	Inflow Particle Concentration	3.2 $\frac{kg}{m^3}$
ρ_p	Dry Particle Density	1450 $\frac{kg}{m^3}$
ρ_w	Clear Water Density	1000 $\frac{kg}{m^3}$
σ_c	Schmidt Number	0.7
US0	Reference Settling Velocity	0.005 $\frac{m}{s}$
RH	Floc Settling Parameter	0.7 $\frac{m^3}{kg}$
RP	Colloidal Settling Parameter	5 $\frac{m^3}{kg}$
CMIN	Nonsettleable Concentration	0.01 $\frac{kg}{m^3}$
USMAX	Maximum Settling Velocity	0.002 $\frac{m}{s}$

flow of 0.019 m/s, in which the concentration of solids is 3.2 kg/m3 as shown in Table 1. This work investigates the accumulations of solids as a sludge blankets by a transient two dimensional analysis.

The systems Equations (2) - (8) were solved using finite volume method in ANSYS Fluent 14.2. The source momentum terms and the rheological properties are implemented by User Defined Functions (UDFs) of Fluent ([6]). Using a similar approach to Lakehal et al ([5]) we have solved slurry model to depict the dynamics of solids concentration in a lagoon. The suspended solids concentration is determined by C in Equation (6). The buoyancy effects that result from the solids settling due to gravity with the density differences cause temporary circulation effects with the resultant non-uniform sludge layering effects.

The 2D model transient solution provides the profiles of solid concentration, designated by Scalar 0 at different times. Figure 1 and Figure 2 we present the contours on the solids concentration(Scalar 0) at times t = 1000s and t = 6000s respectively. In the case of 1000s, as shown in Figure 1, high concentration flow of solids from the inlet are pulled by gravity along the wall until the bottom is reached. After the solids reach the bottom they move along the length towards the outlets due to the momentum of the flow. However because of gravity there is a stratification with higher concentration of solids at about $1.5kg/m^3$ along the bottom. At about 12 meters from the end there is a solids accumulation spot in the bottom of about $5kg/m^3$. The reason for this is the recirculation in flow. This causes stagnation in which the gravity effects dominate and cause a pile up of solids this point. Note that as the flow get approaches the end, at about 5 meters there is another smaller solids build up on the bottom. This is again caused by the circulatory currents. In this case since the overall solids concentration is higher because if dispersion mixing, the overall settling effect is lower since the negative buoyancy effect is decreased. However, at the top exit point we see a slight increase in solids concentration to about $0.5kg/m^3$, which results in turbid liquid to be extracted from the top end.

Figure2 contours shows a solids concentration at t = 6000s which approaches a quasi-steady-state in which a distinctive separation layer between the solids and

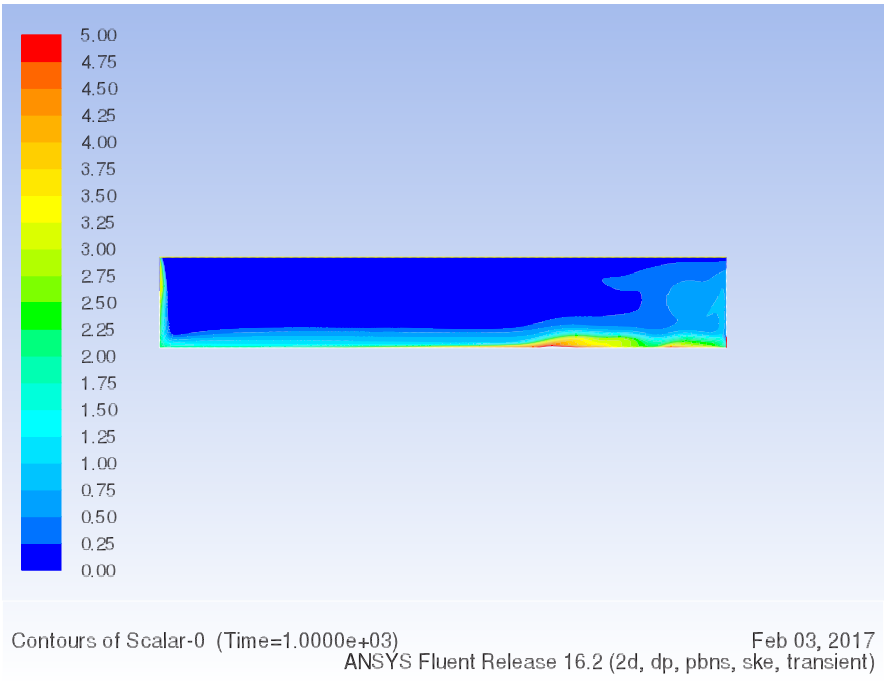


Fig. 1 Variation of solids concentration at T=1000s.

water has formed. This top layer is free of solids and can be recycle for use as water. Also, it can be seen from Figure2 that the solids concentration increases sharply from approximately $0.5kg/m^3$ to about $2kg/m^3$. Furthermore we also note that a thin layer of high concentration of solids is greater than $4.5kg/m^3$ along the bottom. The thickness of this layer is approximately constant after about 4 meters from the inlet wall.

We have quantified the solids concentration profiles at these times by extracting numerical results of the biomass(solids) concentration along three positions in the flow direction (y = 4, 20 and 40 meters). These are shown in Figure 3 (t = 1000s), Figure 4(t = 6000s). In Figure 3, we see that overall solids concentration increases along the y axis. This is consistent with the fact that near the end the recirculation causes stirring of the solids and hence increases the concentration. Also it should be noted that the inlet is constantly feeding solids which initially flow along the bottom along the y axis and then circulate near the end causing an increase in concentration with height. We confirm that a steady-state is approached at 6000s as shown in Figure 4 which shows that the Scalar 0 value jumps from approximately $0kg/m^3$ to greater than $9kg/m^3$ at a depth of 3 meters from the surface. From Figure 4 we can also confirm a formation of a highly concentrated layer of about $3kg/m^3$ to $9kg/m^3$ near the bottom.

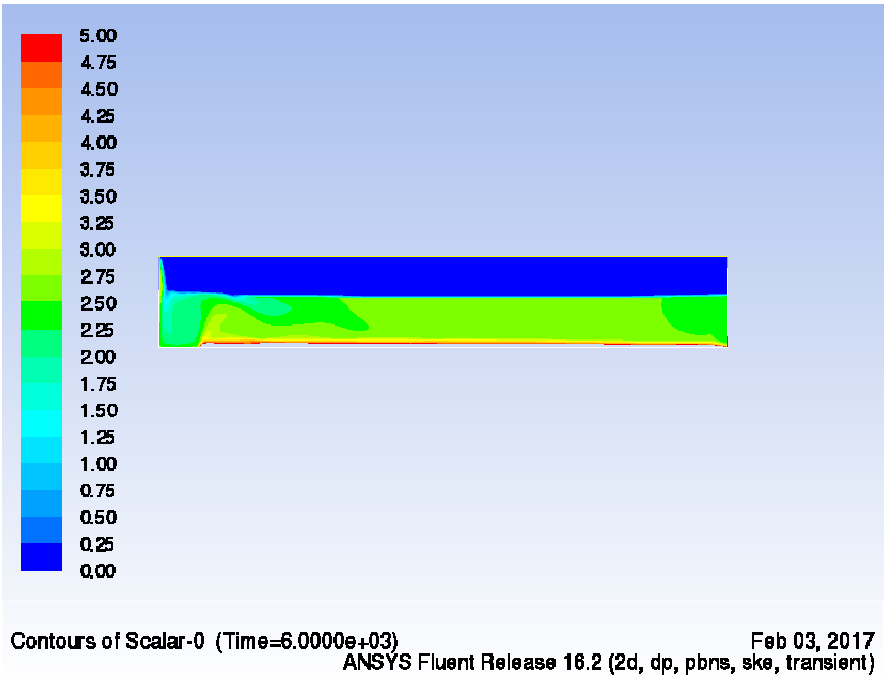


Fig. 2 Variation of solids concentration at T=6000s

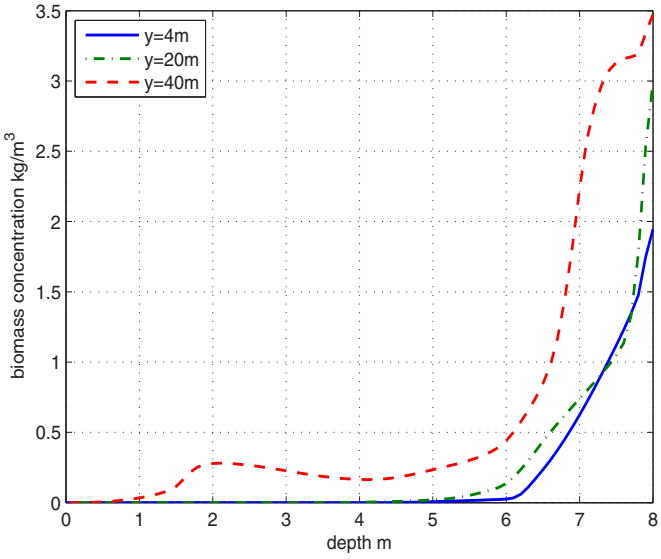


Fig. 3 Variation of solids concentration profiles at T=1000s

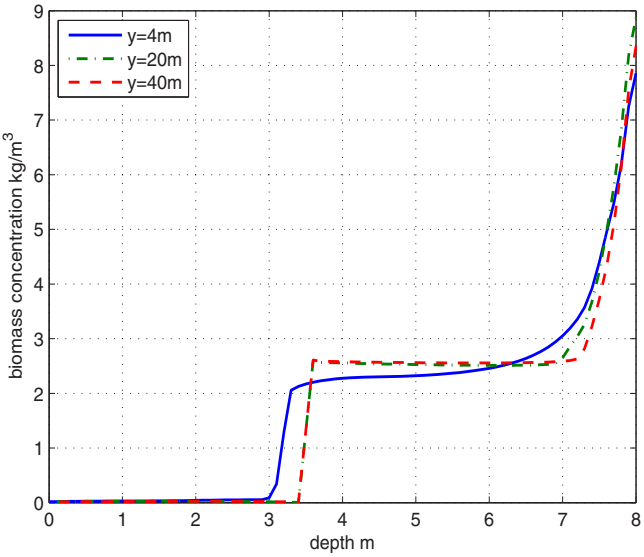


Fig. 4 Variation of solids concentration profiles at T=6000s

4 Conclusion

The results for solids concentration at t = 1000s show two pockets of solids build up on the bottom near the outlet. This is due to flow recirculation due to a combined effect of turbulence in the slurry and the effect of gravity. However it is noticed that for a longer time of 6000 s these pockets of solids disappear and a quasi-steady state is reached with a thin layer of concentrated solids at the bottom and a distinctive separation of clear water and solids starts at about 3 meters from the surface. Thus it can be concluded that providing we supply an adequate outlet near the bottom of the lagoon the solids buildup should be stabilized. Also, overall we have shown that this method can be used to determine the transient concentration profile which is an important issue for lagoon solids management for providing guidelines for water treatment strategy and sludge removal maintenance schedule. Future work will be focussed on parameter modelling for optimal design of lagoons.

References

1. S.Zhou and J.A. McCorquodale, Modeling of rectangular settling tanks, *Journal of Hydraulic Engineering*, **118(10)**:1391-1405, 1992.
2. B.Li and M. K. Stenstrom, Dynamic one-dimensional modeling of secondary settling tanks and design impacts of sizing decisions, *Water Research*, **50**:160-170, 2014.

3. I. Takacs, G. G. Patry and D. Nolasco, A dynamic model of the clarification-thickening process, *Water Research*, **25**:1263-1271,1991.
4. C. Dahl, T. Larsen and O. Petersen, Numerical modelling and measurement in a test secondary settling tank, *Water Science technology*,**30**:219-229,1994.
5. D. Lakehal, P. Krebs, J. Krijgsman and W. Rodi, Computing shear flow and sludge blanket in secondary clarifiers, *Journal of Hydraulic Engineering*, **125(3)**:253-262,1999.
6. ANSYS Inc.,*Ansys Fluent UDF Manual Release 14.2*,Canonsburg, PA, USA, 2012.

Regular and Singular Behaviours and New Morphologies in the Rayleigh Taylor Instability

Kurt Williams, Desmond L. Hill, Snezhana I. Abarzhi

Abstract The Rayleigh Taylor Instability is a fluid instability that develops when fluids of different densities are accelerated against their density gradient. Its applications include inertial confinement fusion, supernovae explosion, fossil fuel extraction and nano fabrication. We study Rayleigh Taylor instability developing at an interface with a spatially periodic perturbation under a time varying acceleration using group theoretic methods. For the first time, to our knowledge, both regular and singular nonlinear solutions are found, which correspond to the structure of bubbles and spikes emerging at the interface. We find that the dynamics of bubbles is regular, and the dynamics of spikes is singular. The parameters affecting the behaviour of both bubble and spikes are discussed, including the inter-facial shear, which is shown to have a profound effect. The results set key theoretical benchmarks for future analysis.

1 Introduction

1.1 Rayleigh Taylor Instability

The problem of Rayleigh Taylor instability was first systematically studied in 1883[12] by Lord Rayleigh, who proposed an experiment in which a dense fluid (eg: water) is balanced on top of a less dense fluid (eg: oil). The system, if perfectly balanced, would remain at rest - with the dense fluid on top being unable to penetrate the lighter fluid. However, any perturbation or deviation away from this equilibrium state causes the system to rapidly accelerate away from the equilibrium state. Later experiments by Taylor [7] would confirm the unstable nature of such a system and

Snezhana I. Abarzhi
School of Mathematics and Statistics, The University of Western Australia
e-mail: snezhana.abarzhi@gmail.com

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_21

provide geometric insight about the problem.

In the most general of terms, the Rayleigh Taylor Instability can be defined to be a system of two fluids of different densities undergoing a prolonged acceleration normal to the interface between the fluids. In a system with this configuration, the less dense fluid "bubbles" up and penetrates the denser fluid, which itself penetrates the lighter fluid as "spikes". Both "bubble" and "spike" structures are observed to have a finger-like structure that is paraboloidal in nature for early time, but may evolve into more irregular structures in the late-time "mixing" regime.

In a complete description of the system, shearing forces that emerge at the interface are responsible for deformations of these bubble and spike structures. Whilst at very small scales, this vortical behaviour can be described independently, as has been done in research of Kelvin Helmholtz instabilities, for the Rayleigh Taylor instability, these structures are tiny, and shear is best described as a global property of the system that can affect its growth and other behaviour.

Rayleigh Taylor behaviours are observed in a broad range of circumstances and scales. Examples in nature include supernovae[4], galactic evolution [4], and ocean dynamics [1]. Industrial examples include laser micromachining [13] (including laser ablation [5]), inertial confinement fusion [9], optical telecommunications [11] and aeronautics [6]. With such a large range of fundamental processes being driven by Rayleigh Taylor dynamics, it is vital that effective theoretical benchmarks are set.

2 Theoretical Approach

2.1 Governing Equations

2.1.1 Euler Equations

The analytic description of the system begins with the Euler equations for incompressible fluids of uniform density:

$$\frac{D\mathbf{u}}{Dt} = -\nabla\omega + \mathbf{g} \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0. \tag{2}$$

By considering an infinitesimal volume dV of fluid, the conservation of mass, momentum and energy lead to the following conservation equations:

$$\frac{\partial}{\partial x_i}(\rho v_i) = \frac{\partial \rho}{\partial t} \tag{3}$$

$$\frac{\partial \rho v_i}{\partial t} + \sum_{j=1}^{3} \frac{\partial \rho v_i v_j}{\partial x_i} + \frac{\partial P}{\partial x_i} = 0 \tag{4}$$

$$\frac{\partial E}{\partial t} + \frac{\partial (E+P) v_i}{\partial x_i} = 0 \tag{5}$$

where in these equations, ρ is the density, p is the momentum, v_i are the components of the velocity field, x_i are the spatial coordinates of the system and E and P are the energy and pressure respectively. The energy can also be expressed as $E = \rho(e + \mathbf{v}^2/2)$ for specific internal energy e.

2.2 Interface Conditions

In order to separate the dynamics of both the bulk and the interface, we introduce a scalar function $\theta(\rho, \mathbf{v}, P, E)$ which has derivatives of at least first order (i.e. $\nabla\theta$ and $\dot{\theta}$ exist), with θ being 0 at the interface of the two fluids. Then the denser fluid is located in the region $\theta > 0$ and the less dense fluid fills the region $\theta < 0$.

Since these two fluids are perfectly separated by this boundary of $\theta = 0$, we may express our total domain as $(\rho, \mathbf{v}, P, E) = (\rho, \mathbf{v}, P, E)_h H(\theta) + (\rho, \mathbf{v}, P, E)_l H(-\theta)$. Substituting into the conservation equations, we obtain the following conditions at the interface:

$$\begin{aligned} [\mathbf{j}\cdot\mathbf{n}] = 0 \qquad & [(P + \frac{(\mathbf{j}\cdot n)^2}{\rho})\mathbf{n}] = 0 \\ [(\mathbf{j}\cdot\mathbf{n})(\frac{(\mathbf{j}\cdot\tau)}{\rho})\tau] = 0 \qquad & [(\mathbf{j}\cdot\mathbf{n})(W + \frac{(\mathbf{j})^2}{2\rho^2})] = 0 \\ \mathbf{n} = \frac{\nabla\theta}{|\nabla\theta|} \qquad & \mathbf{n}\cdot\tau = 0, \end{aligned} \tag{6}$$

where the square brackets [...] denote the "jump" of the function across the interface - essentially the limit of the derivative with respect to θ. The mass flux is expressed as $\mathbf{j}$.

In the case in which there is no mass flux across the interface ($\mathbf{j}\cdot\mathbf{n}|_{\theta=0^{\pm}} = 0$), these boundary conditions at the interface become:

$$[\mathbf{v}\cdot\mathbf{n}] = 0, [P] = 0, [\mathbf{v}\cdot\tau] = arbitrary, [W] = arbitrary, \tag{7}$$

and at infinity:

$$\lim_{z\to\infty} v_h = 0, \lim_{z\to-\infty} v_l = 0. \tag{8}$$

Whilst there exist two natural time scales for Rayleigh Taylor systems with time-varying acceleration [10], we will focus on the the timescale of acceleration-driven dynamics. The two timescales are $\tau_g = (kG)^{-1/(a+2)}$ and $\tau_0 = 1/(k|v_0|)$ with $|v_0|$ some initial growth rate for the system. Furthermore, there is a unique length scale $1/k$ imposed by the wave vector.

2.3 Large Scale Dynamics

Any vector field can be expressed as the sum of the gradient a scalar potential plus the curl of a vector potential field. In this way, we may express our vector field as:

$$\mathbf{v} = \nabla\Phi + \nabla\times\phi. \tag{9}$$

The large scale dynamics are assumed to be irrotational, since no discontinuities or circulations occur. The small-scale dynamics are rotational, but by Kelvin's Circulation Theorem the large scale dynamics are irrotational in the bulk [14]. We hence set $\nabla\times\phi = 0$. This means that the velocity field $\mathbf{v}$ can be expressed as $\nabla\Phi$.

By substituting this expression for $\mathbf{v}$ into the conservation equations, we obtain the following:

$$\Delta\Phi = 0 \tag{10}$$

$$\rho(\frac{\partial\Phi}{\partial t} + \frac{\nabla\Phi^2}{2}) + P = 0. \tag{11}$$

Now substituting the expression for $\mathbf{v}$ into 7, we obtain a system of equations to solve:

$$\rho_h(\nabla\Phi_h\cdot\mathbf{n} + \frac{\dot\theta}{|\nabla\theta|}) = \rho_l(\nabla\Phi_l\cdot\mathbf{n} + \frac{\dot\theta}{|\nabla\theta|}) = 0 \tag{12}$$

$$\nabla\Phi_h\cdot\tau - \nabla\Phi_l\cdot\tau = arbitrary \tag{13}$$

$$\rho_h(\frac{\partial\Phi_h}{\partial t} + \frac{|\nabla\Phi_h|^2}{2} + (g(t) + \frac{\partial v}{\partial z})z) = \rho_l(\frac{\partial\Phi_l}{\partial t} + \frac{|\nabla\Phi_l|^2}{2} + (g(t) + \frac{\partial v}{\partial z})z), \tag{14}$$

where $g(t) = Gt^a$, a power-law function of time. In the frame of reference that moves with the bubble tip, the boundary conditions are instead expressed:

$$\nabla\Phi_h|_{z\to\infty} = (0,0,-v(t)), \nabla\Phi_l|_{z\to-\infty} = (0,0,-v(t)). \tag{15}$$

2.4 Group Theory

In order to capture the highly-symmetric nature of our solution, we appeal to group theory. The interface between the two fluids is initially flat, or rather very close to flat and so is essentially $\mathbb{R}^2$, a group of Lie Type, which is to say that $\mathbb{R}^2$ under some set of transformations, can be considered to be both a group and a manifold. The elements of the irreducible representations of this group will inform the structure of Fourier series over the group. Since we are seeking a solution which has symmetries over the entire space, we infer that the group operations in question must be symmetry transforms on $\mathbb{R}^2$. There are seventeen groups of invariants under these transformations, but by imposing the condition that our structures must have inversions along the interfacial plane, and must be repeating, we need only consider the two-dimensional groups p2mm, p4mm, p6mm, p2 and cmm for three dimensional flows; and the one-dimensional group p1m for two-dimensional flows. These groups are referred to using the international notation [2]. In this notation, m's denote the number of reflective or "mirroring" symmetries a cell has, p's indicate primitive cells - which have natural translational symmetries, c's denote face-centred cells and free numbers indicate the rotational symmetry of each cell.

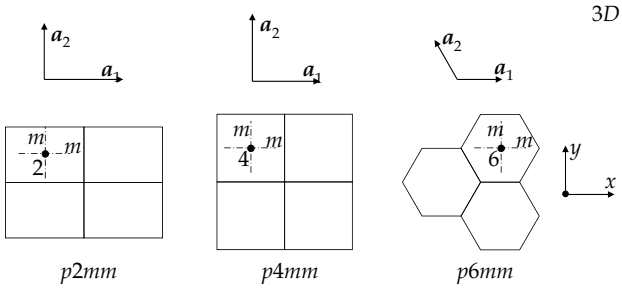


Fig. 1 A selection of the seventeen unique wallpaper groups.

We will be examining the symmetry group p6mm, which has six rotational symmetries, 2 reflective symmetries and three directions of equal magnitude translational symmetry, the third of these being the direct sum of the first two. The naive treatment of such a structure would be to construct vectors $\mathbf{a}_i$ along each edge and express our solution in terms of these. However, to actually express our solution, we need to map these "lattice vectors" into reciprocal space - a non-euclidean manifold in which the metric for distances between two points is:

$$d(x,y) = 1/d_E(x,y), \tag{16}$$

where d_E is the Euclidean metric. Each of the vectors $\mathbf{k}_j$ in inverse space obeys the relationship $\mathbf{a}_i \cdot \mathbf{k}_j = 2\pi$, for $i, j \in 1,2$ and $\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2$. Since the lattice vec-

tors are $(2\pi)\mathbf{a}_i = \{(1,0),(-0.5,\sqrt{3}/2),(-0.5,-\sqrt{3}/2)\}$, our reciprocal vectors are $(\sqrt{3}\lambda/4\pi)\mathbf{k}_j = \{(\sqrt{3}/2,1/2),(0,1),$
$(\sqrt{3}/2,-1/2)\}$. There is an interesting geometric relation between the lattice vectors and the reciprocal lattice vectors in euclidean space - the reciprocal vectors form a basis for the centre of each cell (figure 2). As such, these vectors are the basis for a Fourier series of structures along the interface.

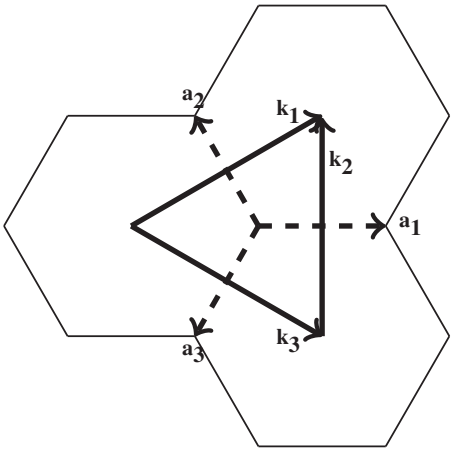


Fig. 2 The geometric relationship between the lattice vectors $\mathbf{a}_i$ (dashed) and the reciprocal vectors $\mathbf{k}_j$ (solid)

So then, since we require the maxima of our Fourier series along the interface to be at the centre of the hexagonal cells, we expect our Fourier series for the velocity potential to be of the form:

$$\Phi \sim \sum_{j=1}^{3} \cos(\alpha \mathbf{k}_j \cdot \mathbf{r}).$$

Then summing over all the modes of harmonics and the boundary conditions for the heavy and light fluids, we obtain in a single step:

$$\Phi_h(\mathbf{r},z,t) = \sum_{m=0}^{\infty} \Phi_m(t)(z + e^{\frac{-mkz}{3mk}} \sum_{j=1}^{3} \cos(m\mathbf{k}_j \cdot \mathbf{r})) + f_h \tag{17}$$

$$\Phi_l(\mathbf{r},z,t) = \sum_{m=0}^{\infty} \hat{\Phi}_m(t)(-z + e^{\frac{mkz}{3mk}} \sum_{j=1}^{3} \cos(m\mathbf{k}_j \cdot \mathbf{r})) + f_l, \tag{18}$$

where $\mathbf{r} = (x,y)$ is the position along the interface, Φ_m and $\hat{\Phi}_m$ are the Fourier amplitudes, with $m \in \mathbb{Z}$. It is worth noting that $|\mathbf{a}_i| = \lambda$, $k = |\mathbf{k}_i| = 4\pi/(\lambda\sqrt{3})$.

We are interested in the motion at the tips of the bubbles and spikes, so we are able to Taylor expand the scalar function $\theta(z) = z - z^*(x,y,t)$. Knowing that the structures are symmetric about the centre of each "symmetry cell", the expansion is:

$$z^*(x,y,t) = \sum_{N=1}^{\infty} \zeta_N(t)\mathbf{r}^{2N}. \tag{19}$$

To the first order (N=1), the tip expansion is $z^* = \zeta(x^2+y^2)$.

2.5 The Moments Expansion

Since the equation is expanded in terms of harmonics of standing waves (17), it would be natural to truncate the series to a few terms and analyse those. However, much of the behaviour of the system is governed by the interplay between these harmonics. In order to preserve these harmonics in our equations we introduce weighted sums over all the harmonics known as "moments":

$$M_n = \sum_{m=0}^{\infty} \Phi_m(t)k^n m^n \tag{20}$$

$$\hat{M}_n = \sum_{m=0}^{\infty} \hat{\Phi}_m(t)k^n m^n. \tag{21}$$

2.6 The Dynamical System

Finally, having attained local expressions for the potential field and the interface, we substitute into 12-2.14 and obtain the following system of equations:

$$(1+A)(\dot{\zeta} - 2\zeta M_1 - \frac{M_2}{4}) = (1-A)(\dot{\zeta} - 2\zeta\hat{M}_1 + \frac{\hat{M}_2}{4}) = 0 \tag{22}$$

$$(1+A)(\frac{\dot{M}_1}{4} + \zeta\dot{M}_0 - \frac{M_1^2}{8} + \zeta g) = (1-A)(\frac{\dot{\hat{M}}_1}{4} + \zeta\dot{\hat{M}}_0 - \frac{\hat{M}_1^2}{8} + \zeta g) \tag{23}$$

$$M_1 - \hat{M}_1 = arbitrary, \qquad M_0 = -\hat{M}_0 = -v. \tag{24}$$

Where $A = (\rho_h - \rho_l)/(\rho_h + \rho_l)$, the Atwood number and $g = g(t) = Gt^a$.

2.7 Early Time Solutions

For the early time solutions, only the first harmonics are retained in the expressions, yielding:

$$\begin{aligned} M_i &= -\hat{M}_i = -k^i v, \qquad v = \frac{4}{k^2}\dot{\zeta} \\ (1+A)(\ddot{\zeta} - \zeta G t^a) &= (1-A)(-\ddot{\zeta} - \zeta G t^a). \end{aligned} \tag{25}$$

This system of equations has general solution:

$$\zeta(t) = c_1\sqrt{\frac{t}{\tau}} I_{\frac{1}{2s}}\left(\sqrt{AG}\frac{(\frac{t}{\tau})^s}{s}\right) + c_2\sqrt{\frac{t}{\tau}} I_{-\frac{1}{2s}}\left(\sqrt{AG}\frac{(\frac{t}{\tau})^s}{s}\right). \tag{26}$$

With $\tau = \tau_g$ being the characteristic timescale of the time-dependent acceleration force. The case $a = 0$ yields the classic result:

$$\zeta(t) = c_1 \exp\left(\sqrt{AG}\frac{t}{\tau}\right) + c_2 \exp\left(\sqrt{AG}\frac{t}{\tau}\right). \tag{27}$$

2.8 Nonlinear Dynamics

In general the dynamical system is not solvable. We can, however, generate an asymptotic solution in the regime of $t \to \infty$. In such a regime, we assume that each of the modes of oscillation grow at the same rate, since otherwise our problem is dominated by a single mode which could be analysed in exactly the same way as our linear, early time dynamics. We also assume that the rate of growth goes as some power law expansion of time, which is to say that the growth of these modes is governed by the external acceleration, which has power-law dependence on time. We presume that asymptotically:

$$\zeta \sim t^a, \qquad (M, \hat{M}, \Phi, \hat{\Phi}) \sim (m, \hat{m}, \phi, \hat{\phi}) t^\beta$$

In order to ensure our solution resolves issues of closure, and captures the interaction between harmonic modes, we expand our moments to the second mode of oscillation:

$$M_n(t) = (\Phi_1(t) + 2^n \Phi_2(t))k^n, \qquad \hat{M}_n(t) = (\hat{\Phi}_1(t) + 2^n \hat{\Phi}_2(t))k^n \tag{28}$$

Which yields the following:

$$\zeta_1 = -\frac{m_2}{8m_1} \qquad \zeta_1 = -\frac{\hat{m}_2}{8\hat{m}_1} \tag{29}$$

$$m_1 = \frac{2m_0 k}{3-8p} \qquad \hat{m}_1 = \frac{2\hat{m}_0}{3+8p} \tag{30}$$

$$m_2 = 3km_1 - 2k^2 m_0 \qquad \hat{m}_2 = 3k\hat{m}_1 - 2k^2\hat{m}_0 \tag{31}$$

$$p = -\frac{\zeta}{k} \tag{32}$$

Substituting these into equation 23, we obtain:

$$\begin{aligned}(1+A)(\frac{bm_1}{4}+\zeta_1 bm_0-\frac{m_1^2}{8}t^{1+b}-\zeta_1 Gt^{1+a-b})\\ =(1-A)(\frac{b\hat{m}_1}{4}-\zeta_1 b\hat{m}_0-\frac{\hat{m}_1^2}{8}t^{1+b}-\zeta_1 Gt^{1+a-b}).\end{aligned} \tag{33}$$

In determining solutions to this equation, we find three potential balances:

$$a<-2, b=-1 \qquad a=-2, b=-1 \qquad -2<a, b=\frac{a}{2}$$

If a is sufficiently small ($a<-2$), then the external acceleration will have negligible effect, and the motion will essentially be of Richtmyer Meshkov type. The second case will be a threshold point at which the dynamics will be both of Rayleigh Taylor and Richtmyer Meshkov type. The Rayleigh Taylor dynamics are given in the third case. In letting $b=a/2$ and $-2<a<0$, we solve for the velocity:

$$v(t)=-\frac{\sqrt{G(t/\tau)^a}}{k}(64p^2-9)\sqrt{\frac{2Ap}{48p+A(64p^2+9)}} \tag{34}$$

$$=-\frac{1}{\tau k}(\frac{t}{\tau})^{a/2}(64p^2-9)\sqrt{\frac{2Ap}{48p+A(64p^2+9)}}. \tag{35}$$

The structure with the fastest velocity for a given Atwood number is known as the Atwood structure. Setting the time derivative of $v(t)$ to be zero yields the following condition for the curvature ($p=p^*$) and velocity ($v(t)=v^*(t)$) of an Atwood structure:

$$\begin{aligned}p^{*4}+\frac{1}{A}p^{*3}+\frac{9}{32}p^{*2}-(\frac{3}{16})^3=0\\ \Longrightarrow\ v^*(t)=-\frac{1}{\tau k}(\frac{t}{\tau})^{a/2}(8p^*)^{\frac{3}{2}}.\end{aligned} \tag{36}$$

We also seek to account for the vortical structures that emerge strictly at the interface. Although they do not cause global circulation, they do provide a means by which the two fluids can move or "shear" past each other. We thus introduce a global parameter to quantitatively measure this interfacial shearing:

$$\Gamma(\zeta,t)=M_1(t)-\hat{M}_1(t)=\frac{12k}{64p^2-9}v(t). \tag{37}$$

In all of our equations, the curvature (ζ) and wavelength (k) are natural parameters of the system. Since the wavelength is fixed by the initial configuration of our system it is natural to assume, as Garabedian[8] did, that solutions to the problem of Rayleigh Taylor instability form a one-parameter family of solutions, and that the dynamics are single-scale in nature. So then, our full description of the system is:

$$v(t) = -\frac{1}{\tau k}\left(\frac{t}{\tau}\right)^{a/2}(64p^2-9)\sqrt{\frac{2Ap}{48p+A(64p^2+9)}}$$
$$\Gamma(\zeta,t) = \frac{12k}{64p^2-9}v(t) \tag{38}$$

3 Bubble Dynamics

Bubbles are formed when the lighter fluid penetrates into the heavy fluid. As such, they are concave downwards in z and have negative curvature ($\zeta < 0$, $p > 0$). The velocity function is shown in figure 3.

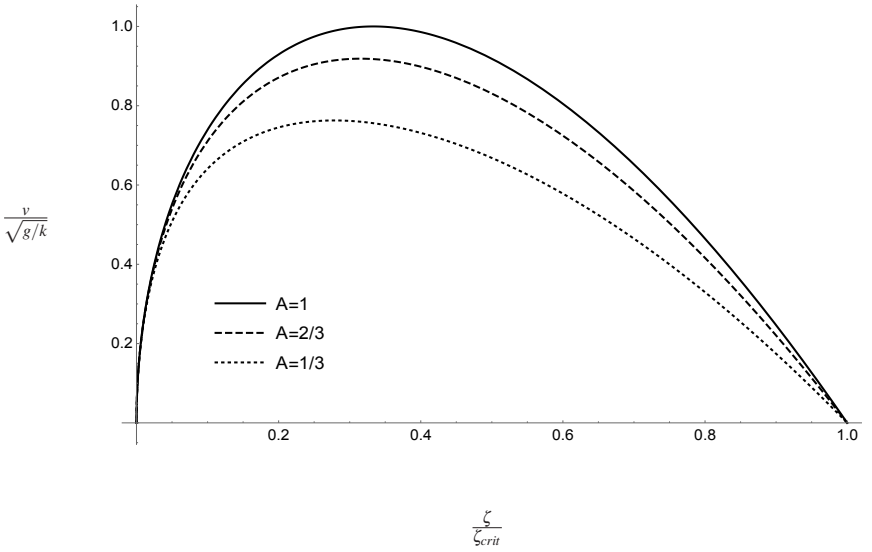


Fig. 3 The velocity (v) of a rising bubble scaled by growth rate ($\sqrt{g/k}$, $g = Gt^a$) as a function of its curvature (ζ).

We clearly observe that there is a one-parameter family of solutions. For any given Atwood number, there is a broad range of possible curvatures, each with its own velocity. The curvature of each solution is uniquely determined by the initial interface perturbation.

The behaviour seen in figure 3 makes physical sense. If the interface is perfectly flat ($\zeta = 0$), then there is no dynamic motion and the velocity is zero. However, any curvature in the interface will allow the heavy fluid to sink and the resultant bubble to rise up. Thinner bubbles grow faster, and it appears that there is a positive cor-

relation between curvature and velocity. However, at a sufficient curvature (which depends on the Atwood number), there is a maximally fast bubble, and at curvatures higher than this, the velocity becomes decreases with curvature. The velocity eventually approaches zero at the critical curvature ($\zeta = -3/8k$). This unique curvature is a stagnation point for the system.

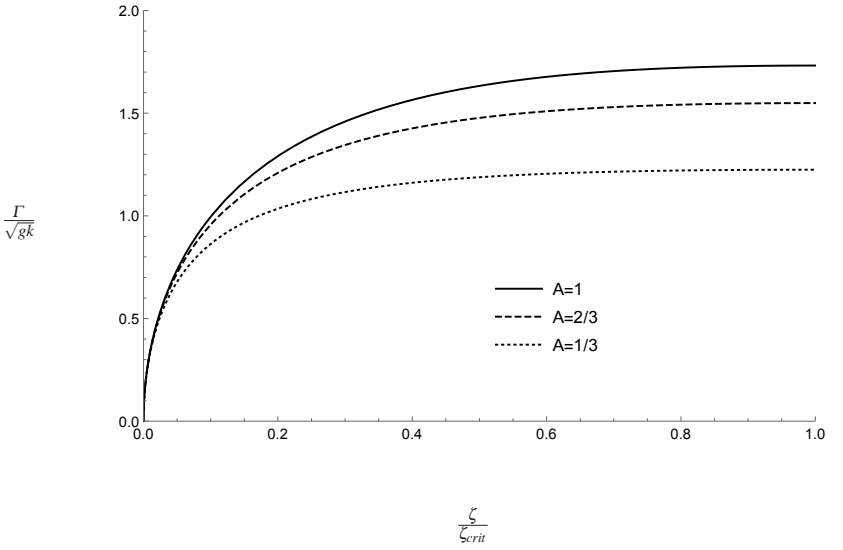


Fig. 4 The intefacial shear (Γ) scaled by growth rate ($\sqrt{gk}$, $g = Gt^a$) as a function of its curvature (ζ), note that larger Atwood numbers give a larger interfacial shear.

The interfacial shearing continuously grows with curvature (figure 4), and is maximal at $\zeta = -3/8k$. We conclude that whilst the interfacial shearing does grow with velocity, it eventually dominates the dynamics and those solutions with maximal shear do have a lower velocity. Thus, whilst the interfacial shearing is dependent on the velocity, it is a competing mechanism in the dynamics and resists rising bubbles reaching their maximal velocity. The velocity is highly sensitive to this interfacial shearing and nonlinear bubbles have a multiscale dependence on both the curvature (a parameter governing its effective acceleration) and the interfacial shear.

In any case, we expect the dynamics to be dominated by the bubble exhibiting the highest velocity, which we herein call the "Atwood" bubble. Numerical simulations involving competing bubbles of various velocities demonstrate that asymptotic dynamics are dominated by bubbles with the highest velocity [3]. We thus expect the Atwood solution to be the physically relevant one.

We can therefore conclude that there is a one-parameter family of solutions that

arises due to the multiscale character of the dynamics. The dynamics is multiscale and governed by the interaction of acceleration and the interfacial shearing. However, any system will be dominated by the fastest or "Atwood" type solution in the asymptotic regime.

4 Spike Dynamics

Spikes are the complementary structure of bubbles. They are concave in the positive θ direction and flow from the heavy fluid into the light ($v < 0$), they have positive curvature ($\zeta > 0$, $p < 0$). The velocity has the form seen in figure 5.

There are a number of unique features in the asymptotic velocities of the spikes.

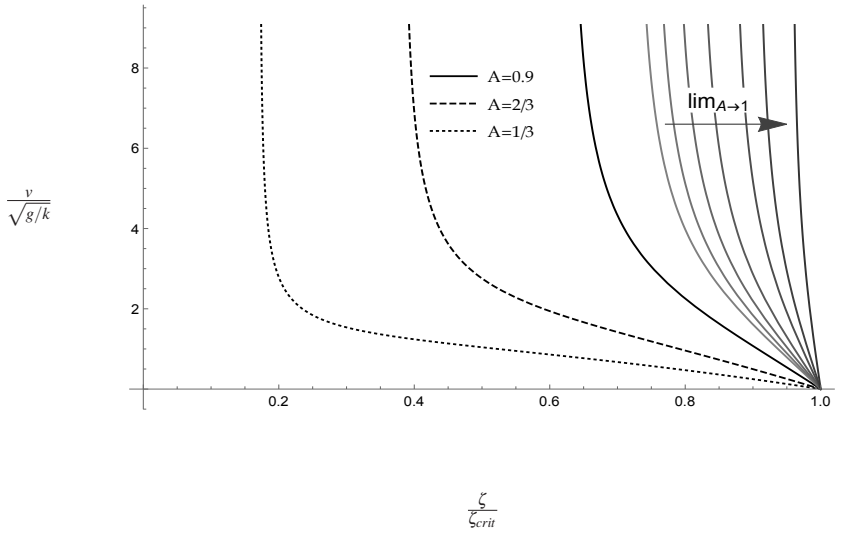


Fig. 5 The dependence of spike velocity (v) scaled by growth rate ($\sqrt{g/k}$, $g = Gt^a$) on the curvature (ζ). The dynamics as $A \to 1$ is unbounded for all curvatures less than the critical curvature.

Much like bubbles, there is a critical curvature ($\zeta = 3/8k$) which forms a stagnation point for spikes. Unlike bubbles, however, spikes with very small curvatures ($\zeta \to 0$) do not tend towards stagnation, but reach unbounded growth. It may be tempting to suggest that these singularities are nonphysical, but our analysis has restricted itself to finding dynamics on the order $\sim k$. A singular velocity suggests that the dynamics of the system outgrows this scale. This growth is likely the mechanism for the transition between the nonlinear dynamics and the mixing regime. Thus, this analysis could open the door to understanding the hitherto unexplored mixing regime.

The exact scaled curvature at which the spike velocity becomes asymptotic is $\kappa(A)k$, where:

$$\kappa(A) = \frac{3}{8}\frac{1-\sqrt{1-A^2}}{A} \tag{39}$$

And it is interesting to note that at this curvature, the interfacial shearing is also singular (figure 6), suggesting that the dynamics grows beyond $\sim k$ in both of the associated scales (wavelength and amplitude). As such, the effect of interfacial shear is not dominated by this unbounded growth in velocity and the dynamics of the spikes is also to be understood as a multiscale phenomenon. It should also be noted that in the limit of the density of the lighter fluid tending towards zero ($A \to 1$), the spike velocity becomes unbounded for all curvatures less than the stagnant critical curvature ($\zeta = 3/8k$).

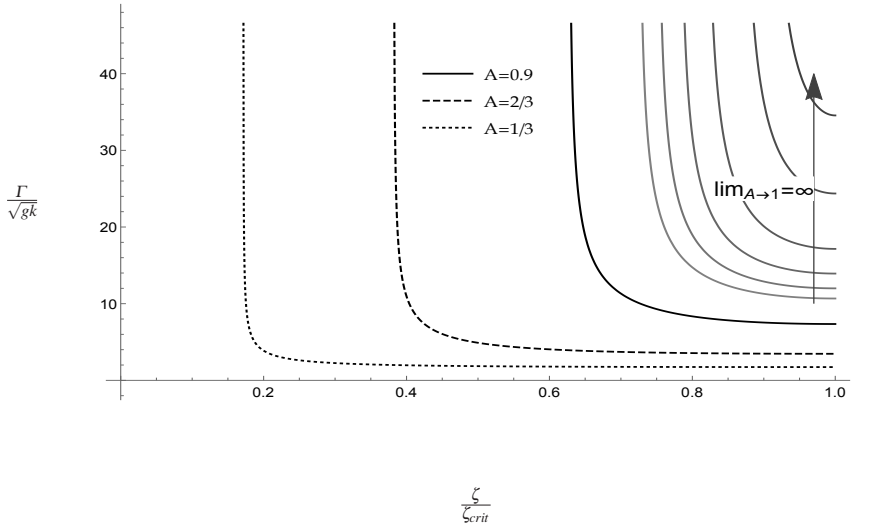


Fig. 6 The dependence of shear (Γ) scaled by growth rate ($\sqrt{gk}$, $g = Gt^a$) on the curvature (ζ). The dynamics as $A \to 1$ is unbounded for all curvatures. In the rescaling, $g = Gt^a$

5 Conclusion

By using group theoretic methods, we have explored the linear and non-linear dynamics of the large-scale structures in Rayleigh Taylor instabilities under variable acceleration. We have considered an interface with two translational symmetries under a time-varying acceleration with power-law dependence - in particular, power-

law with exponents larger than -2. By invoking the theory of group representations. we have expanded the flow fields, derived a dynamical system from the governing equations and then found solutions for both the bubbles and spikes that emerged (Equations 12).

For the early time regime, we found that the behaviour of bubbles and spikes can be described using a linear combination of Bessel functions (Equation 26). For non-linear bubbles and spikes, however we found asymptotic solutions with power-law time dependence. For non-linear bubbles, we have observed that for small enough curvatures, the velocity is small (Figure 4). For spikes, we have observed that the velocity does fall away for sufficiently large curvatures, but is also singular at a curvature determined by the Atwood number (Figure 5). We linked this unexpected and unusual behaviour to the interfacial shearing. For non-linear bubbles, the interfacial shear mediates the decrease in velocity that occurs at large curvatures. For the non-linear spikes, the interfacial shear induces the velocity bounding at large curvatures, but it also grows with the singular velocity that appears at sufficiently small curvatures.

We found that the shear dominates the acceleration induced dynamics in bubbles and spikes of sufficient curvature, meaning that the velocity is dependent on the interfacial shearing. The problem of Rayleigh Taylor instability therefore exhibits multi-scale dynamics and has a one-parameter family of solutions.

To conclude, we have studied the problem of Rayleigh-Taylor instability in time-varying acceleration using group theoretic methods. We have found the interface dynamics to directly depend on the interfacial shearing and revealed the multi-scale dynamics of late-time Rayleigh-Taylor nature. Our analysis has achieved excellent agreement with available observations, and gives new theoretical benchmarks for future analysis, experiments and simulations.

References

1. SI Abarzhi, Katsunobu Nishihara, and R Rosner. Multiscale character of the nonlinear coherent dynamics in the rayleigh-taylor instability. *Physical Review E*, 73(3):036310, 2006.
2. Snezhana I. Abarzhi. Review of nonlinear dynamics of the unstable fluid interface: conservation laws and group theory. *Physica Scripta*, T132:014012, dec 2008.
3. U. Alon, J. Hecht, D. Ofer, and D. Shvarts. Power laws and similarity of rayleigh-taylor and richtmyer-meshkov mixing fronts at all density ratios. *Phys. Rev. Lett.*, 74:534–537, Jan 1995.
4. David Arnett and Roger A. Chevalier. Supernovae and Nucleosynthesis: An Investigation of the History of Matter, from the Big Bang to the Present. *Physics Today*, 49(10):68,70, 1996.
5. Stephen E. Bodner, Denis G. Colombant, John H. Gardner, Robert H. Lehmberg, Stephen P. Obenschain, Lee. Phillips, Andrew J. Schmitt, John D. Sethian, Robert L. McCrory, Wolf. Seka, et al. Direct-drive laser fusion: Status and prospects. *Physics of Plasmas*, 5(5):1901–1918, 1998.
6. JP. Choi and VWS. Chan. Predicting and adapting satellite channels with weather-induced impairments. *IEEE Transactions on Aerospace and Electronic Systems*, 38(3):779–790, 2002.

7. R. M. Davies and Geoffrey Ingram Taylor. The mechanics of large bubbles rising through extended liquids and through liquids in tubes. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 200(1062):375–390, 1950.
8. PR Garabedian. On steady-state bubbles generated by taylor instability. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 241(1226):423–431, 1957.
9. Abbas Ghasemizad, Hanif Zarringhalam, and Leila Gholamzadeh. The investigation of Rayleigh-Taylor instability growth rate in inertial confinement fusion. *J. Plasma Fusion Res*, 8:1234–1238, 2009.
10. Desmond L. Hill, Aklant K. Bhowmick, Dan V. Ilyin, and Snezhana I. Abarzhi. Group theory analysis of early-time scale-dependent dynamics of the rayleigh-taylor instability with time varying acceleration. *Phys. Rev. Fluids*, 4:063905, Jun 2019.
11. Ivan P. Kaminow, Chris R. Doerr, Corrado. Dragone, Tom. Koch, Uzi. Koren, Adel AM. Saleh, Alan J. Kirby, CM. Ozveren, B. Schofield, and Robert E. Thomas. A wideband all-optical WDM network. *IEEE Journal on Selected Areas in Communications*, 14(5):780–799, 1996.
12. Rayleigh Lord. Investigation of the character of the equilibrium of an incompressible heavy fluid of variable density. *Scientific papers*, pages 200–207, 1900.
13. B. Luk'yanchuk, N. Bityurin, S. Anisimov, and D. Bäuerle. The role of excited species in UV-laser materials ablation. *Applied Physics A*, 57(4):367–374, Oct 1993.
14. Grétar Tryggvason. Numerical simulations of the rayleigh-taylor instability. *Journal of Computational Physics*, 75(2):253 – 282, 1988.

The extended Prandtl closure model applied to the two-dimensional turbulent classical far wake

Ashleigh J. Hutchinson

Abstract Prandtl's mixing length closure model has been used extensively in turbulent wake flows. Although the simplicity of this model is advantageous, it contains mathematical and physical limitations. In particular, this model results in a poor estimation of the flow on the center-line and near the wake boundary. Prandtl constructed an improved model, which will be referred to as the extended mixing length model, in an attempt to address many of the limitations of the original model. In this work, the extended Prandtl model is considered. A similarity solution that leaves both the governing equation for the stream-wise mean velocity deficit and the conserved quantity invariant is obtained. The governing partial differential equation is reduced to an ordinary differential equation. The ordinary differential equation, which must be solved subject to appropriate boundary conditions and the conserved quantity, cannot be solved analytically and thus a double-shooting method is developed to obtain the stream-wise mean velocity deficit. A plot of the mean velocity deficit is then produced.

Key words: Extended Prandtl's mixing length, turbulent classical wake, conserved quantity, mean velocity deficit

1 Introduction

In turbulent flows, the time averaged Navier-Stokes equation is used to solve for the mean flow variables. Unknown Reynolds stress terms arise, resulting in an incomplete system of equations. In order to complete the system of equations, a closure model is needed. Many closure models have been proposed. Algebraic closure mod-

Ashleigh J. Hutchinson
School of Computer Science and Applied Mathematics, Johannesburg
DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences, South Africa
e-mail: Ashleigh.Hutchinson@wits.ac.za

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_22

els are the simplest type. In algebraic closure models, the Reynolds stresses which are co-variances, are related to a single mean velocity gradient by a turbulent viscosity function [1]. The effective viscosity is expressed as the sum of the kinematic viscosity, which is an intrinsic property of the fluid, and the turbulent or eddy viscosity which is not a characteristic of the fluid [2, 3]. In most algebraic closure models, the kinematic viscosity is neglected as it is negligible when compared to the turbulent viscosity.

Prandtl's mixing length closure model [4] falls under the class of algebraic closure models. In Prandtl's original model [4], the concept of a mixing length is introduced. The Reynolds stresses are written in terms of the square of this mixing length, and in terms of the square of the mean velocity gradient perpendicular to the axis of the wake. This model has been successfully applied to the turbulent classical far wake and other free shear flows [5]. Prandtl's mixing length model is convenient in that it is fairly easy to implement mathematically.

Prandtl's mixing length model has various limitations. When applied to turbulent wake flows, the predicted width of the wake is underestimated [4, 6]. Another failing of Prandtl's mixing length model is that the mixing length cannot be derived from the model and its form must be independently imposed. Prandtl assumed that the mixing length is proportional to the width of the wake. These limitations have been addressed [7] by modifying Prandtl's model by including the kinematic viscosity. Prandtl neglected the kinematic viscosity in his analysis and it is shown that by including the kinematic viscosity, the mixing length can be derived using a systematic method [7]. It is also shown that when the kinematic viscosity is included, the predicted width of the wake lies outside of the predicted width when the kinematic viscosity is neglected.

Prandtl realised the limitations of his closure model and put forth a new extended version. In this model, the kinematic viscosity is still neglected. Instead, two mixing lengths are introduced and the turbulent viscosity is considered as a function of both the first and second derivatives of the mean velocity deficit perpendicular to the axis of the wake [8]. This new form increases the mathematical complexity of the model. However, this model suffers from the same issue as the original model in that the two mixing lengths have to be independently specified. This issue can again be addressed by including the kinematic viscosity. However, including the kinematic viscosity further complicates the model and the numerical method and so is excluded in the current paper. Prandtl's hypothesis that each mixing length is proportional to the width of the wake is used to specify the form of each mixing length.

The aim of this work is to obtain an expression for the mean flow variables when the extended Prandtl model is applied to the two-dimensional turbulent classical far wake. A similarity solution, that leaves both the conserved quantity and the governing equation for the stream-wise mean velocity deficit invariant, is obtained. The partial differential equation is reduced to an ordinary differential equation which cannot be solved analytically. As an initial study, the kinematic viscosity is not included which simplifies the numerical method significantly.

This paper is presented as follows. In Section 2, the derivation of the governing equations, boundary conditions, and conserved quantity for the two-dimensional turbulent classical far wake is provided. The extended Prandtl closure model is used to complete the system of equations. In Section 3, similarity solutions are considered. Each mixing length is assumed to be proportional to the width of the wake. In Section 3, a numerical method is developed to solve the reduced ordinary differential equation. The similarity velocity profile is then plotted. Conclusions and further work are given in Section 4.

2 Mathematical model for the two-dimensional turbulent classical far wake

In this section a brief review of the derivation of the momentum equation for the two-dimensional turbulent classical wake far downstream of a stationary slender object is provided. A more in-depth derivation can be obtained from [9, 7]. The mean velocity profile for the two-dimensional turbulent classical wake is illustrated in Figure 1. A Cartesian coordinate system is used with the origin positioned at the trailing edge of the slender object. A laminar incompressible fluid with constant velocity $(U,0)$ flows past the stationary slender object. Downstream of the object, a wake is formed. For large Reynolds number flows, the wake that forms is turbulent. The turbulent wake region merges smoothly with the laminar mainstream flow.

In this work, the components of the mean velocity deficit, $\overline{v}_x$ and $\overline{v}_y$, in the wake region are considered. The work conducted in [7] is expanded upon by considering an effective viscosity of the form $E = E\left(x, y, \dfrac{\partial \overline{v}_x}{\partial y}, \dfrac{\partial^2 \overline{v}_x}{\partial y^2}\right)$ so that the extended Prandtl model can be investigated.

In algebraic closure models, the effective viscosity is expressed as the sum of the kinematic viscosity, ν, and the turbulent or eddy viscosity, ν_T [1]:

$$E = \frac{\mu + \mu_T}{\rho} = \nu + \nu_T. \tag{2.1}$$

For Prandtl's extended model, the turbulent viscosity is of the form

$$\nu_T = \nu_T\left(x, y, \frac{\partial \overline{v}_x}{\partial y}, \frac{\partial^2 \overline{v}_x}{\partial y^2}\right). \tag{2.2}$$

The Reynolds number for the mean flow is defined as

$$Re = \frac{UL}{E_C}. \tag{2.3}$$

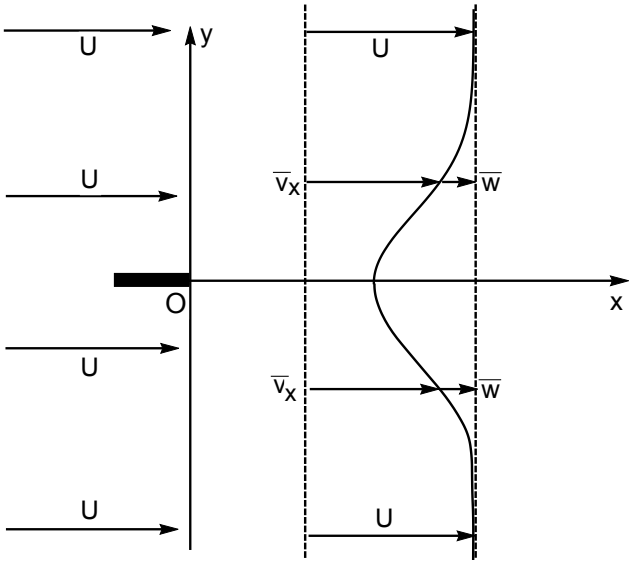


Fig. 1 The two-dimensional turbulent classical far wake behind a thin symmetric planar body aligned with a uniform flow. The mean velocity in the x–direction, $\overline{v}_x$, and the mean velocity deficit, $\overline{w}$, are shown.

Here, L is the length downstream over which the reduction in velocity is not negligible, and $E_C = \nu + \nu_{TC}$ is the characteristic effective viscosity where ν is the viscosity, and ν_{TC} is the characteristic turbulent viscosity.

Boundary layer theory is used to describe wake flows. Dimensionless variables for the x and y coordinates, the mean velocity components $\overline{v}_x$ and $\overline{v}_y$, and the mean fluid pressure $\overline{p}$ are now introduced [10]:

$$x^* = \frac{x}{L}, \quad y^* = \frac{y}{\delta} = y\frac{\sqrt{Re}}{L},$$

$$\overline{v}_x{}^* = \frac{\overline{v}_x}{U}, \quad \overline{v}_y{}^* = \overline{v}_y\frac{\sqrt{Re}}{U}, \quad \overline{p}^* = \frac{\overline{p}}{\rho U^2}, \quad E^* = \frac{E}{E_C}. \tag{2.4}$$

The dimensionless effective viscosity is

$$E^* = \frac{\nu}{\nu + \nu_{TC}} + \frac{\nu_{TC}}{\nu + \nu_{TC}}\nu_T^*, \tag{2.5}$$

where $\nu_T^* = \nu_T/\nu_{TC}$ is the dimensionless turbulent viscosity. The dimensionless mean velocities are given by

$$\overline{v}_x{}^*(x^*, y^*) = 1 - \overline{w}^*(x^*, y^*), \quad \overline{v}_y{}^*(x^*, y^*) = 0 + \overline{v}_y{}^*(x^*, y^*), \tag{2.6}$$

where $\overline{w}^*(x^*, y^*)$ is the dimensionless mean velocity deficit in the x–direction. In order to derive the appropriate approximation of the momentum equation for the far wake flow, these variables are substituted into the momentum equation, the boundary layer approximation is implemented, and terms which are products and powers of the velocity deficits or their derivatives are neglected. The y– component of the momentum equation simply gives $\dfrac{d\overline{p}^*}{dy^*} = 0$, and from mainstream matching, $\dfrac{d\overline{p}^*}{dx^*} = 0$. Thus, there is no external pressure gradient. The momentum equation in the x– direction reduces to

$$\frac{\partial \overline{w}}{\partial x} = \frac{\partial}{\partial y}\left[E\left(x, y, \frac{\partial \overline{w}}{\partial y}, \frac{\partial^2 \overline{w}}{\partial y^2}\right)\frac{\partial \overline{w}}{\partial y}\right], \tag{2.7}$$

where the star notation has been suppressed for convenience. The momentum equation must be solved subject to

$$\overline{w}(x, \pm y_b) = 0, \qquad \frac{\partial \overline{w}}{\partial y}(x, \pm y_b) = 0, \tag{2.8}$$

$$\frac{\partial \overline{w}}{\partial y}(x, 0) = 0, \tag{2.9}$$

where the boundary $y = \pm y_b(x)$ is left unspecified. For a wake that extends to infinity in the y–direction, $y_b(x) = \infty$. For the purpose of obtaining numerical solutions, $y_b(x)$ can be considered to be the effective half width of the wake. The first conditions, (2.8), state that the turbulent wake flow merges smoothly with the mainstream flow. The second condition, (2.9), expresses the condition that the mean velocity deficit is a maximum on the center-line.

In order to derive the conserved quantity, Equation (2.7) is integrated with respect to y over the width of the wake. The boundary conditions, (2.8), are imposed. This results in the condition

$$2\int_0^{y_b(x)} \overline{w} dy = D, \tag{2.10}$$

where D is the drag.

Prandtl's extended mixing length model states that the effective viscosity is of the form

$$E(x, \overline{w}_y, \overline{w}_{yy}) = \nu + l_1^2(x)\left[(\overline{w}_y)^2 + l_2^2(x)(\overline{w}_{yy})^2\right]^{1/2}, \tag{2.11}$$

where l_1 and l_2 are known as the mixing lengths. In order to express this in the form

$$E^* = \frac{\nu}{\nu + \nu_{TC}} + \frac{\nu_{TC}}{\nu + \nu_{TC}}\nu_T^*, \tag{2.12}$$

dimensionless mixing lengths corresponding to l_1 and l_2 need to be defined. The mixing lengths are chosen to scale with the boundary layer thickness, δ. In other words,

$$l_1^* = \frac{l_1}{\delta}, \quad l_2^* = \frac{l_2}{\delta}. \tag{2.13}$$

In terms of the dimensionless variables, $\nu_{TC}\nu_T^*$ is given by

$$\nu_{TC}\nu_T^* = U\delta(l_1^*)^2\left[\left(\overline{w}_{y^*}^*\right)^2 + (l_2^*)^2\left(\overline{w}_{y^*y^*}^*\right)^2\right]^{1/2}, \tag{2.14}$$

which shows that

$$\nu_{TC} = U\delta. \tag{2.15}$$

Suppressing the star notation for convenience,

$$E\left(x, \frac{\partial\overline{w}}{\partial y}, \frac{\partial^2\overline{w}}{\partial y^2}\right) = \frac{\nu}{\nu+\nu_{TC}} + \frac{\nu_{TC}}{\nu+\nu_{TC}} l_1^2(x)\left[\left(\frac{\partial\overline{w}}{\partial y}\right)^2 + l_2^2(x)\left(\frac{\partial^2\overline{w}}{\partial y^2}\right)^2\right]^{1/2}. \tag{2.16}$$

Substituting (2.16) into (2.7) gives

$$\frac{\partial\overline{w}}{\partial x} = \frac{\partial}{\partial y}\left[\frac{\nu}{\nu+\nu_{TC}} + \frac{\nu_{TC}}{\nu+\nu_{TC}} l_1^2(x)\left[\left(\frac{\partial\overline{w}}{\partial y}\right)^2 + l_2^2(x)\left(\frac{\partial^2\overline{w}}{\partial y^2}\right)^2\right]^{1/2}\frac{\partial\overline{w}}{\partial y}\right]. \tag{2.17}$$

As an initial investigation, the kinematic viscosity is neglected. The first term on the right hand side of Equation (2.17) can be neglected, and $\nu + \nu_{TC} \approx \nu_{TC}$. Neglecting the kinematic viscosity leads to

$$\frac{\partial\overline{w}}{\partial x} = \frac{\partial}{\partial y}\left[l_1^2(x)\left[\left(\frac{\partial\overline{w}}{\partial y}\right)^2 + l_2^2(x)\left(\frac{\partial^2\overline{w}}{\partial y^2}\right)^2\right]^{1/2}\frac{\partial\overline{w}}{\partial y}\right]. \tag{2.18}$$

3 Similarity solutions

In this section, similarity solutions admitted by (2.18) are considered. The partial differential equation is then reduced to an ordinary differential equation. Expressions for l_1 and l_2 cannot be obtained when the kinematic viscosity is neglected. Instead, Prandtl's hypothesis that these mixing lengths are proportional to the width of the wake, which behaves as $\sqrt{2x}$, is imposed. This gives

$$l_1(x) = l_{01}\sqrt{2x}, \tag{3.1}$$

$$l_2(x) = l_{02}\sqrt{2x}, \tag{3.2}$$

where l_{01} and l_{02} are constants that can be obtained either numerically or from experimental results. Because there is no extrinsic length scale for this problem, it is reasonable to seek for similarity solutions. The width of the wake behaves like $\sqrt{2x}$, so the similarity variable

$$\xi(x,y) = \frac{y}{\sqrt{2x}}, \tag{3.3}$$

is defined. Let

$$\overline{w}(x,y) = \frac{F(\xi)}{\sqrt{2x}}, \tag{3.4}$$

where F is a function to be determined. Substituting (3.3) and (3.4) into (2.18) results in the ordinary differential equation

$$\frac{d}{d\xi}\left[l_{01}^2\left[\left(F'\right)^2 + l_{02}^2\left(F''\right)^2\right]^{1/2}F'\right] + \frac{d}{d\xi}\left[\xi F\right] = 0. \tag{3.5}$$

In terms of the similarity variables, the conserved quantity, (2.10), becomes

$$\int_0^{y_b(x)/\sqrt{2x}} F d\xi = \frac{D}{2}, \tag{3.6}$$

and because this is independent of x,

$$y_b(x) = \xi_b\sqrt{2x}, \tag{3.7}$$

where ξ_b is a constant that remains to be determined. The conserved quantity becomes

$$\int_0^{\xi_b} F d\xi = \frac{D}{2}. \tag{3.8}$$

The boundary conditions from Equations (2.8) and (2.9) are, in terms of F

$$F(\pm\xi_b) = 0, \quad F'(\pm\xi_b) = 0, \tag{3.9}$$

$$F'(0) = 0. \tag{3.10}$$

Equation (3.5) can be integrated once. Applying the boundary conditions, (3.9), results in a zero constant of integration. Thus,

$$l_{01}^2\left[\left(F'\right)^2 + l_{02}^2\left(F''\right)^2\right]^{1/2}F' + \xi F = 0. \tag{3.11}$$

4 Numerical results

In this section the numerical method used to solve the ordinary differential equation, (3.11), subject to the boundary conditions and the conserved quantity is presented. Because the wake is symmetric about the x–axis, it is convenient to consider only the upper half of the wake. First let

$$\overline{\xi} = \frac{\xi}{\xi_b}. \tag{4.1}$$

Substituting into (3.8)–(3.11) and omitting the bars for convenience gives

$$\frac{l_{01}^2}{\xi_b^3}\left[\left(F'\right)^2+\frac{l_{02}^2}{\xi_b^2}\left(F''\right)^2\right]^{1/2}F'+\xi F=0, \tag{4.2}$$

$$F(1)=0,\quad F'(1)=0, \tag{4.3}$$

$$F'(0)=0, \tag{4.4}$$

$$\int_0^1 F(\xi)\,d\xi=\frac{D}{2\xi_b}. \tag{4.5}$$

Now, in the upper half of the wake, $F'\leq 0$ and so (4.2) can be written in the form

$$l_{01}^2\left[\left(F'\right)^2+\frac{l_{02}^2}{\xi_b^2}\left(F''\right)^2\right]^{1/2}\left|F'\right|=\xi_b^3\xi F. \tag{4.6}$$

Squaring both sides and solving for F'' leads to

$$F''=\pm\frac{\xi_b^4}{l_{02}l_{01}^2}\left[\frac{\xi^2F^2}{\left(F'\right)^2}-\frac{l_{01}^4}{\xi_b^6}\left(F'\right)^2\right]^{1/2}. \tag{4.7}$$

Let

$$G=F'. \tag{4.8}$$

Then Equation (4.7) can be written as two first order differential equations:

$$F'=G, \tag{4.9}$$

$$G'=\pm\frac{\xi_b^4}{l_{02}l_{01}^2}\left[\frac{\xi^2F^2}{G^2}-\frac{l_{01}^4}{\xi_b^6}G^2\right]^{1/2}. \tag{4.10}$$

In terms of F and G, the boundary conditions (4.3) and (4.4) become

$$F(1)=0,\quad G(1)=0, \tag{4.11}$$

$$G(0)=0. \tag{4.12}$$

Although both boundary conditions on G are not required since the differential equation for G is of first order, using the condition $G(0)=0$ is convenient as solving for G results in solving an initial value problem. However, the only condition on F is at a boundary, and so a shooting method is required to solve for F.

In the upper half of the wake, $F'=G\leq 0$. From $G(0)=0$ and the fact that $G\leq 0$, it is seen that the negative root in Equation (4.10) must be taken. Using a forward difference scheme,

$$F_{n+1}=\Delta\xi G_n+F_n, \tag{4.13}$$

$$G_{n+1} = -\Delta\xi \frac{\xi_b^4}{l_{02}l_{01}^2}\left[\frac{\xi_n^2 F_n^2}{G_n^2} - \frac{l_{01}^4}{\xi_b^6}G_n^2\right]^{1/2} + G_n, \tag{4.14}$$

where $\Delta\xi$ is the chosen step-size. The initial value for G is $G_1 = 0$ and at the end boundary, $F_N = 0$. As mentioned previously, a shooting method is required to solve for F. An initial guess for F_1 is obtained from the conserved quantity and the most basic approximation to it:

$$\int_0^1 F(\xi)d\xi \approx \frac{1}{2}(F_1 + F_N) = \frac{1}{2}F_1 = \frac{D}{2\xi_b}. \tag{4.15}$$

So,

$$F_1 = \frac{D}{\xi_b}. \tag{4.16}$$

The value of ξ_b is also not known and must be determined from the conserved quantity. Thus a double shooting method is required to determine the boundary value problem for F, and the value of ξ_b.

The process is as follows: An initial value for ξ_b is chosen. Initially, choose $\xi_b = 1$. The value for F_1 is obtained from (4.16). For the chosen ξ_b value, the boundary value problem for F is solved. Once the solutions for F and G are obtained, the conserved quantity is evaluated and the value of ξ_b is updated. This process is continued until convergence is achieved.

For illustration purposes, let

$$\frac{\xi_b^4}{l_{02}l_{01}^2} = 1, \quad \frac{l_{01}^4}{\xi_b^6} = 0.05, \quad D = 0.1. \tag{4.17}$$

A step size value of $\Delta\xi = 0.001$ is used. A plot of the similarity profile is shown in Figure 2.

5 Further work and conclusions

In this work, the extended Prandtl closure model was applied to the two-dimensional turbulent classical far wake. A similarity solution that left both the governing equation for the stream-wise mean velocity deficit in the x–direction and the conserved quantity invariant, was obtained. The governing partial differential equation was reduced to a second order ordinary differential equation. This second order differential equation was then expressed as two first order ordinary differential equations. Numerical methods were required to solve these two equations. The numerical method of choice involved using a double shooting method to solve a boundary value problem and the unknown value of the boundary which was determined from the conserved quantity. A plot of the similarity velocity profile was provided for illustrative purposes.

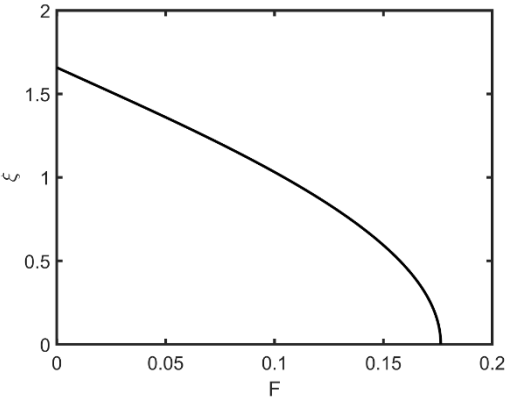


Fig. 2 Similarity profile of the mean velocity deficit.

The numerical scheme presented in Section 4 takes time to converge and is very sensitive to the initial choices for ξ_b and F_1. It also appears that convergence is only achieved for a limited range of values of l_{01} and l_{02}. A much more in-depth analysis of this numerical scheme is required. Alternative schemes need to be developed that allow for faster convergence without compromising accuracy.

The values for l_{01} and l_{02} were chosen arbitrarily. In order to obtain the correct order of magnitude for these values, the numerical result must be compared to experimental data. To date, the skills required for fitting parameters in a model to data are being investigated.

Once an improved numerical scheme is developed, and parameter fitting methods are well-understood, the aim is then to compare the different closure models.

Acknowledgements A J Hutchinson thanks R Gusinow and K Born, University of the Witwatersrand, for proof reading this paper.

References

1. Pope S B. Turbulent Flows, Cambridge University Press, Cambridge; 2000.
2. Cebeci T. Analysis of turbulent flows, Elsevier, London 2004.
3. Boussinesq J. Théorie de l'écoulement tourbillant. Mém. prés. Acad. Sci. 1877;23:46.
4. Prandtl L. Proceedings of the Second International Congress for Applied Mechanics, Zurich 1926, p 62.
5. Rudy D H, Bushnell D M. A rational approach to the use of Prandtl's mixing length model in free turbulent shear flow calculations. NASA Langley Research Center, 1973.
6. Swain L M. On the turbulent wake behind a body of revolution. Proceedings of the Royal Society of London. Series A, 1929; 125: 647-659.
7. Hutchinson A J, Mason D P. Revised Prandtl mixing length model applied to the two-dimensional turbulent classical wake. Int. J. Non-Linear Mech. 2015; 77: 162-171.

8. Prandtl L. Report on investigation of developed turbulence. United States, Washington, D.C: National Advisory Committee for Aeronautics 1949.
9. Hutchinson A J, Mason D P, Mahomed F M. Solutions for the turbulent classical wake using Lie symmetry methods. Commun Nonlinear Sci Numer Simulat 2015; 23: 51-70.
10. Schlichting H. Boundary-Layer Theory, M^{c}Graw-Hill, New York, Sixth Edition 1968, Chs 8, 18 and 19.

Mixing, tunnelling and the direction of time in the context of Reichenbach's principles

Alexander Y. Klimenko

Abstract This work reviews the understanding of the direction of time introduced by Hans Reichenbach, including the fundamental relation of the perceived flow of time to the second law of thermodynamics (i.e. the Boltzmann time hypothesis), and the principle of parallelism of entropy increase. An example of a mixing process with quantum effects, which is advanced here in conjunction with Reichenbach's ideas, indicates the existence of a physical mechanism that reflects global conditions prevailing in the universe and enacts the direction of time locally (i.e. the "time primer"). Generally, this mechanism, whose effects are often enacted by presuming antecedent causality, remains unknown at present. The possibility of experimental detection of the time primer is also discussed: if the time primer is CPT-invariant, its detection may be possible in high-energy experiments under the current level of technology.

Key words: the direction of time, the second law of thermodynamics, mixing, decoherence, quantum tunnelling, the time primer

> It appears that mixing processes, in the most general sense of the term, are the instruments which indicate a direction of time
>
> Hans Reichenbach

Alexander Y. Klimenko
SoMME, The University of Queensland, St.Lucia, QLD 4072, Australia
e-mail: klimenko@mech.uq.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_23

1 Introduction

Discussing time is always difficult since the notion of time is deeply embedded into both our language and our intuition. Many key words in English (e.g. "then","follows", "since") and most other languages and cultures imply both a logical link and a temporal arrangement. The perceived flow of time and conceptual inferences are almost indistinguishable, or at least they are not properly distinguished by most languages we use. Immanuel Kant [1] wrote in 1781:

> Time is a necessary representation that grounds all intuitions. In regard to appearances in general one cannot remove time, though one can very well take the appearances away from time. Time is therefore given a priori.

One one hand this intuition assists us in everyday life and in the formulation of scientific theories not related to the nature of time. One the other hand, this intuition needs to be subordinated to rational thought when the nature of time is discussed, and this can be difficult. It is worthwhile to note that the conventional intuitive interpretation of the flow of time is the most common interpretation, but certainly not the only one possible: there are indigenous tribes living in the North-Western part of Queensland, who intuitively perceive time as being directed from East to West.

The perceived flow of time is thought to reflect causality — the fundamental directional connection between events unfolding in time, as well as the possibility of explaining observed phenomena in terms of more basic principles. The two sides of causality, related to 1) atemporal logical statements of a generic nature (e.g. objects fall because of the action of gravity) and 2) directional dependence between specific consecutive events (the vase is shattered because it was pushed from the table), may be interpreted synergistically [2] or be clearly distinguished [3]. It is the second interpretation, which is often referred to as *antecedent causality*, that we are interested most in this work. In the 1st half of the 20th century, there was a common belief that the directional properties of the perceived flow of time can be explained in terms of more objective casual relations that are postulated a priori as one of the fundamental intrinsic properties of nature. This belief had to face mounting difficulties in defining causality, and largely evaporated toward the end of the 20 century. As early as in 1914, Bertrand Russell [4] noted that

> The view that the law of causality itself is a priori cannot, I think, be maintained by anyone who realises what a complicated principle it is.

The conceptual understanding of causality has grown to accommodate randomness, counterfactual logic, etc. but, overall, our interpretation of causality remains largely intuitive and rather short of being the basis of rational thought. Antecedent causality is now explained in terms of physical laws rather than placed at the foundation of these laws. Dowe [5] defines the direction of casual action in terms of physical laws that possess temporal asymmetry: either the second law of thermodynamics or CP violations in the quantum world. Tying causality to the second law of thermodynamics in one form or another has become the central element of conventional thinking about the problem ([6–9]). The strongest form of the link between

the direction of time and the second law of thermodynamics is given by the Boltzmann time hypothesis, which proclaims that the arrow of time and the second law are two sides of the same physical effect [10–13]. Hawking [12] explains this: "the second law of thermodynamics is really a tautology", since the direction of our perceived flow of time is, in fact, determined by the second law. The physical side of the direction of time is covered in a number of principal publications [14–17].

The second half of the 20th century is marked by two seminal, yet very different, books that endeavour to bridge philosophical and physical arguments about the direction of time [11, 18]. The book by Huw Price is well written and delivers its message

> I have been trying to correct a variety of common mistakes and misconceptions about time in contemporary physics — mistakes and misconceptions whose origins lie in the distorting influence of our own ordinary temporal perspective, and especially of the time asymmetry of that perspective

in a clear and articulate form. The other book is the last book written by Hans Reichenbach. He was not able to complete his work and the book was published by Mrs. Reichenbach in 1956, after her husband's death in 1953. The book tends to mix philosophical and physical arguments in a way that might be confusing for both philosophers and physicists, yet Reichenbach's book is probably the greatest book about time ever written. According to his wife, Reichenbach considered his last book to be the culmination of his contribution to philosophy. The *Boltzmann time hypothesis*, the *principle of parallelism of entropy increase* and *the principle of the common cause* are, perhaps, the most important contributions presented in the book. While the Boltzmann time hypothesis gradually became accepted by many philosophers and physicists, the principle of parallelism of entropy increase is still a subject of debates [7, 8, 19–22].

The present work is, of course, not intended to review all issues related to the arrow of time and causality within a short article. Conceptual issues are discussed only in the context of selected examples that can illustrate physical statements in a concise and transparent manner. Without attempting to overview or replace the comprehensive publications cited above, this work focuses on select few problems. Section 2 briefly overviews the understanding of the directionality of time suggested by Reichenbach. Section 3 analyses an example of a mixing process and demonstrates the significance of time priming pointing to existence of unknown physical mechanisms of very small magnitude associated with the direction of time. Section 4 discusses a wider scope of issues focusing on the possibility of experimental evaluation of these mechanisms. The Appendix considers the example of Section 3 and involves evaluation of a quantum system in thermodynamic conditions when decoherence or recoherence are present.

2 The direction of time and the second law

Our experience of time is very directional — we remember the past but cannot possibly remember the future and our photographs always show us younger than we are now. If we see dents on bumpers of two cars that are standing next to each other, we conclude that these cars have just collided and, certainly, not that they are going to collide in the future. At an intuitive level, we characterise these directional properties of time as "time flow" but, according to the fundamental Boltzmann time hypotheses, these properties of time reflect the objective reality and directional nature of the second law of thermodynamics. Unlike most physical theories (e.g. classical and quantum mechanics, relativity and electromagnetism) which are time-symmetric, this law is time-directional, stating that, in an isolated system, entropy must increase (or stay the same) forward in time. Following Reichenbach, the Boltzmann time hypotheses is explained below by using a gedanken experiment called "footsteps on a beach".

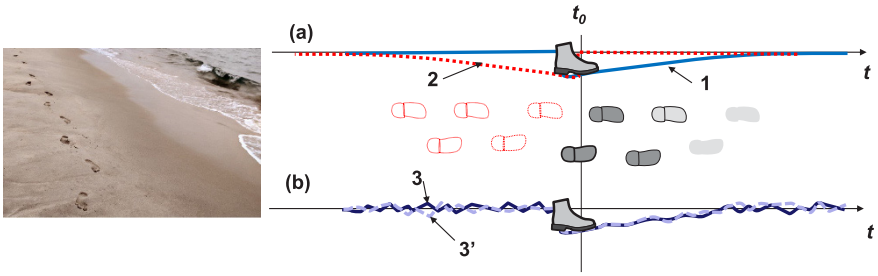


Fig. 1 Footprints on a beach: a) effect of the second law of thermodynamics and b) effect of random disturbances. Curves: 1-realistic; 2-violating the second law; 3,3'- realistic disturbed by wind.

2.1 Why don't we remember the future?

The sand on a beach is always levelled by wind and water – this is the state of maximal entropy where all specific information is destroyed. We might try to change this by stepping on the sand and leaving footprints. These footprints, however, cannot stay forever and will soon disappear. This process, shown by line 1 in Figure 1, is perfectly consistent with the second law of thermodynamics. Another possibility is shown by line 2 — footsteps gradually appear and then are removed by a walking man. The second scenario is not realistic as it contradicts the second law of thermodynamics: footsteps cannot appear forward in time under the influence of random factors such as wind and waves. To be more precise, this can, in principle, happen,

but the probability of such event is so small so that it can be safely neglected. The second law of thermodynamics is a probabilistic law — it predicts the behaviour of entropy not with absolute certainty but with overwhelming probability.

If we see footsteps on the beach, do they mean that someone walked on the beach in the past (line 1) or that someone will walk on the beach in the future (line 2)? According to the second law, footsteps cannot possibly appear without a reason (i.e. a man walking) in the past but do not need a reason to disappear. In the same way marks, photos, memories, scratches of car paint, etc. reflect past events but tell us nothing about future events. This conclusion is obvious but its link to the second law of thermodynamics is not trivial.

The Boltzmann time hypothesis has not been accepted universally. Karl Popper, one of the most distinguished philosophers of the 20th century, argued that the Boltzmann time hypothesis cannot be true due to thermodynamic fluctuations and that Boltzmann would not suggest his hypothesis if he knew more about these fluctuations [23]. Popper's remarks are usually accurate, sharp and impressively prescient, but this statement seems rather controversial. First, Boltzmann was well aware of thermodynamic fluctuations and even interpreted (for the sake of illustration) his imaginary world of reversed time as a gigantic galactic fluctuation [10]. Second, exactly the same fluctuation argument was later used not against but in support of connection between the arrow of time and the second law of thermodynamics [5]. The flow of time is a powerful illusion; it is very useful in real life and even in scientific applications, but, as noted by Price [18], it can easily produce a distorted view when issues related to the direction of time are discussed. Although details of specific opinions may vary, most philosophers and physicists tend to accept the existence of deep underlying link between the perceived direction of time and the action of the second law of thermodynamics [5, 10–12, 18, 24].

2.2 Parallelism of entropy increase

The importance of this principle was stressed by Reichenbach, who considered the main system to be divided into branch systems (i.e semi-independent subsystems branching from the main system) and suggested that "*in the vast majority of branch systems, the directions toward higher entropy are parallel to one another and to that of the main system*". Since "the main system" can be deemed to encompass the whole universe, its direction toward higher entropy is the temporal direction of overall entropy increase in the universe. This principle does not preclude occasional fluctuations that might slightly decrease local entropy and, therefore, it is not clear to what extent this principle represents an independent statement. For example, Boltzmann believed that local entropy trends simply reflect global increase of entropy in the observable part of the universe, while Reichenbach insisted that parallelism of entropy increase is an independent principle, which, generally, cannot be derived from the global temporal conditions imposed on the universe: despite the presence of fluctuations, entropy increases in branch systems are more consistent

than it can be inferred from the global entropy increase. Since a microstate of each branch system can be characterised by a point in a phase space of very large dimensions, the state of maximal entropy corresponds to the uniform distribution of such points over all possible microstates. Reichenbach interprets increase of entropy as a *generalised mixing* process, which is associated with diffusion of particles or points towards being distributed over larger volumes in the physical and/or phase spaces. This interpretation of the entropic directionality as a trend to expand distributions in phase spaces of large dimensions is often used by physicists [15]. The principle of parallelism of entropy increase is presented and discussed in a few publications, most notably in books by Davies [19] and Sklar [20].

While association of causality with the second law is now widely acknowledged, the physical origins of the second law remain unclear. The second law is fundamental but largely empirical: it declares that entropy increases forward in time but does not explain why. Since all major physical laws and theories are time symmetric, the most common explanation is that the temporal asymmetry of the second law is due to asymmetric temporal boundary conditions imposed on the universe (these conditions can be referred to as the past hypothesis or low-entropy Big Bang). Albert [7, 8] believes that this explanation is perfectly sufficient but, according to Reichenbach, the principle of parallelism of entropy increase is needed (in addition to the commonly presumed low-entropy conditions in the early universe) to explain the observed consistency of the second law [11]. Winsberg [21] agrees with Reichenbach, while North [22] supports Albert. As discussed further in Section 3, there are reasonable arguments on both sides of this debate but, overall, it seems rather unlikely that the second law can be replaced by a combination time-symmetric physical laws and time-assymetric temporal boundary (i.e. initial and final) conditions.

The principle of parallelism of entropy increase allows us to apply entropic considerations to relatively small thermodynamic systems or even to non-thermodynamic macroscopic objects. We often imply this principle when we commingle macroscopic and microscopic considerations. For example, one can associate an entropy change to random reshuffling of playing cards, although this change is insignificant compared to changes in thermodynamic entropy — the latter is larger by a factor of $\sim 1/k_B$, where k_B is the Boltzmann constant. While applying entropic considerations to macroscopic objects mostly produces reasonable outcomes and good intuitive illustrations of thermodynamic principles, such applications are less rigorous compared to the very high level of statistical certainty associated with the laws involving the thermodynamic entropy. Macroscopic interpretations of entropy are subject to conditions that are difficult to stipulate in a rigorous and universal manner and, therefore, may produce incorrect inferences if taken out of context. Reichenbach notes that we can put cards back into their original order if we need to, but we cannot possibly reorder molecules exactly into their original positions. The grains of sand from the example shown in Figure 1 may be very small, but they are macroscopic objects.

2.3 The principle of the common cause

Reichenbach states this principle as "if an improbable coincidence has occurred, there must exist a common cause"; this cause should be in the past as common effects in the future cannot cause improbable coincidences. The term "improbable coincidence" for two events A and B, refers to the simultaneous occurrence of A and B in excess of $P(\mathrm{A})P(\mathrm{B})$ — the probability if they were independent events. Price refers to this property as the principle of the independence of incoming influences (PI^3) — indeed incoming influences (i.e. those that do not have a common cause) must be statistically independent. This principle is intuitively obvious but, again, the essence of the Boltzmann time hypothesis is that this effect is, in fact, a consequence of the second law. Figure 1b illustrates this point. Consider a model when wind and waves naturally impose some degree of roughness on the sand level. The lines 3 and 3' shown in this figure correspond to the effect of wind and waves causing the surface at two selected points to fluctuate at random. These points level out only if only someone steps on them. Levelling, however, does not last for long, since wind and waves gradually introduce new disturbances, which erase the footprints. The usual state of the surface is rough and influences of events cannot propagate back in time (since this propagation specified by curve 2 contradicts the second law) — these conditions require that dependences are induced by past events.

It is probably true that Reichenbach's treatment of mutual causes and mutual effects in his last book presents a combination of physical and philosophical arguments, intermixing them to extent that may become puzzling for both physicists [12] and philosophers [25]. Perhaps applying these ideas to conventional elements of statistical physics can provide a more transparent illustration. In the next subsection, we give an example of chemical kinetics that illustrates Reichenbach's key point — the link between the principle of the common cause and the second law of thermodynamics.

2.4 Chemical kinetics and causality

Consider the following reactions

$$1)\ A + B \longrightarrow AB, \quad 2)\ AB \longrightarrow A + B \tag{1}$$

which are assumed not to have any heat effect. As illustrated in Figure 2, these reactions can be interpreted as open (left) and closed (right) casual forks analysed by Reichenbach, who denoted AB by C (cause) or E (effect). Events A, B, AB respectively denote appearance of molecules A, B, AB in a volume V, which is much smaller than V_t — the total volume under consideration. In the first reaction, A and B are causes that have a common effect AB, while in the second reaction, A and B are effects that have a common cause AB. Hence, according to the principle of the common cause $P(\mathrm{A{+}B}) = P(\mathrm{A})P(\mathrm{B})$ for the first reaction but not for the second.

Here, $P(\text{A+B})$ is the probability of simultaneous presence of A and B in the volume V. If $P(\text{A+B})$ is significantly larger than $P(\text{A})P(\text{B})$, then, in accordance with the third principle of Reichenbach, there must be a common cause — the second reaction in (1).

Considering that A and B are independent causes of the first reaction and AB is the cause of the second reaction, the overall reaction rates of the first and second reactions can be expressed by

$$W_1 = V_t K \frac{N_A}{V_t}\frac{N_B}{V_t},\ W_2 = V_t K \frac{N_{AB}}{V_t} \tag{2}$$

where $P(X) = N_X V / V_t$ for any $X = A, B, AB$, N_X is the total number of molecules X in the volume V_t and K is the reaction rate constant. Note that kinetic equation

$$\frac{dN_A}{dt} = \frac{dN_B}{dt} = -\frac{dN_{AB}}{dt} = W_2 - W_1 \tag{3}$$

implies that the entropy defined as

$$S = N_A \ln\left(e\frac{V_t}{N_A}\right) + N_B \ln\left(e\frac{V_t}{N_B}\right) + N_{AB} \ln\left(e\frac{V_t}{N_{AB}}\right) \tag{4}$$

cannot decrease; i.e.

$$\frac{dS}{dt} = \frac{dN_A}{dt} \ln\left(\frac{V_t N_{AB}}{N_A N_B}\right) = K\left(N_{AB} - \frac{N_A N_B}{V_t}\right) \ln\left(\frac{V_t N_{AB}}{N_A N_B}\right) \geq 0 \tag{5}$$

in accordance with the second law of thermodynamics.

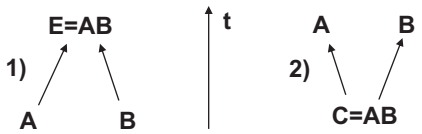


Fig. 2 Chemical reations shown in the form of casual forks.

We may try alternative anticasual arrangements when causes are located in the future and effects are in the past. According to the anticasual assumptions, the first reaction is caused by AB while the second reaction is caused by two independent events A and B. This means that the overall reaction rates are now

$$W_1 = V_t K \frac{N_{AB}}{V_t},\ W_2 = V_t K \frac{N_A}{V_t}\frac{N_B}{V_t} \tag{6}$$

so that the entropy change rate is given by

$$\frac{dS}{dt} = \frac{dN_A}{dt} \ln\left(\frac{V_t N_{AB}}{N_A N_B}\right) = K\left(\frac{N_A N_B}{V_t} - N_{AB}\right) \ln\left(\frac{V_t N_{AB}}{N_A N_B}\right) \leq 0 \qquad (7)$$

This illustrates that casual or anticasual assumptions imply the following trends for the entropy: increasing in time for the former and decreasing in time for the latter. Of course, only the casual case corresponds to the real world.

If quantum effects are to be considered (it is arguable that interactions of atoms are determined by quantum effects), then the casual case (2)-(5) corresponds to persistent decoherence of the molecules before and after the reaction, while the anticasual case (6)-(7) corresponds to persistent recoherence [26]. There is a physical connection between causality and the temporal direction of decoherence [26, 27]. The second law of thermodynamics is a macroscopic law, but it is enacted by microscopic irreversible processes — quantum decoherences and collapses [15–18]. (We tend to use these the terms "decoherences" and "collapse" interchangeably, as there is a significant overlap between implications of these terms — see Appendix of Ref. [13] for details.)

3 Why mixing is time-directional?

Despite temporal symmetry of the overwhelming majority of the physical laws, entropy tends to increase or stay the same with a high degree of certainty for any thermodynamic system, small or large. The temporal boundary conditions imposed on the universe (e.g. a low-entropy Big Bang) must play a key role in this trend — these conditions are often sufficient to explain many effects associated with directionality of time even if physical laws are deemed to be completely time-symmetric. Indeed, if the universe has a very strong overall trend to increase the entropy and the universe is divided into semi-autonomous subsystems (branches according to Reichenbach), then increase of entropy must be more likely than decrease of entropy in these subsystems. While the low-entropy initial conditions imposed on the universe are important and instrumental in explaining entropy increase for many physical phenomena, this does not mean that all observed physical effects can be directly explained by imposing these conditions while assuming that all physical laws are strictly time-symmetric. Therefore, the principle of parallelism of entropy increase is indicative of some fundamental properties of the universe that need to be understood and examined further.

These points are illustrated here by analysing time-directional properties of mixing. We consider diffusion of N_t molecules (called particles) of a substance in a gas. The number N_t is relatively small so that molecules do not interact with each other; the admixture remains passive and does not affect major thermodynamic quantities such as pressure and density, although N_t is large enough in absolute terms to produce reliable statistical quantities that can be observed macroscopically as concentrations.

3.1 Importance of the initial conditions

The particles (molecules) $j = 1, ..., N_t$ are released at the same location $x_j = x_0$ at $t = t_1$ and diffuse forward in time $t > t_1$. The particle trajectories $x_j(t)$ represent Brownian motion, while the average concentration of particles $f(x,t)$ satisfies the diffusion equation

$$\frac{\partial f}{\partial t} = D\frac{\partial^2 f}{\partial x^2} \tag{8}$$

Note that particle trajectories are time-symmetric — that is we cannot distinguish trajectories running forward in time from those running backward in time. The direction of diffusion is determined by the initial conditions $x_j = x_0$ at $t = t_1$. If we impose final conditions at $t = t_2 > t_1$ (for example, this can be done by considering the following process $x'_j(t) = x_j(t) - x_j(t_2) + x_0$, which satisfies $x'_j = x_0$ at $t = t_2$), then the concentration of trajectories $x'_j(t)$ would be characterised by diffusion equation (8) but with a negative diffusion coefficient $D' = -D$; i.e. this is diffusion occurring backward in time[1].

This seem to favour temporal boundary conditions as a driving force behind irreversibility. The processes described by the diffusion equation with positive and negative diffusion coefficients are radically different. The direction of the diffusion is determined not by the random variations of particle positions, which do not have a time arrow, but by imposing the initial or the final conditions. The influence of initial or final conditions, however, disappears in the equilibrium state $f = \text{const}$ of fully mixed components (assuming that the diffusion takes place in a finite volume). Indeed, once the steady-state is achieved, say within the interval $t_1^\circ < t < t_2^\circ$ where $t_1 < t_1^\circ < t_2^\circ < t_2$, it is impossible to tell the direction of the diffusion process, even if the most detailed current characteristics of trajectories are monitored — information about initial or final conditions has been lost. Setting initial conditions at $t = t_1$ cannot be distinguished from setting the final conditions at $t = t_2$ by observing equilibrium solution at $t_1^\circ < t < t_2^\circ$. Equilibria achieve maximal entropy and destroy information.

The example of this subsection reflects the *lattice of mixture model* examined by Reichenbach [11]. We see that, within limitations of this model, the overall initial conditions imposed on the whole system are sufficient to ensure directionality of mixing processes in every macroscopic subsystem. The evolution of the universe can be interpreted as a generalised mixing process where particles diffuse to occupy a larger and larger number of microstates. Since the universe was presumably formed with low-entropy initial conditions and has not achieved its equilibrium state, this consideration provides a justification for generally preferring initial conditions to final conditions in today's environment. It might seem that Reichenbach's principle of parallelism of entropy increase is excessive — the low-entropy initial condition

[1] Note that this reversal is different from the reversal of the Kolmogorov backward equation and time reversal of Markov diffusion processes preserving $f(x,t)$ — see ref. [28]. It is also possible to use both conditions at $t = t_1$ and $t = t_2$, leading to so called Brownian bridge, but this case is not considered here.

imposed on the universe ensures both overall entropy increase and, as long as overall equilibrium is not reached, proper directionality of various local thermodynamical processes. While under some idealised conditions, global entropy increase induces entropy increases in local processes, there are important details that are missing in this inference. The lattice of mixture model reveals some useful properties but, nevertheless, is a significant oversimplification of the physical reality.

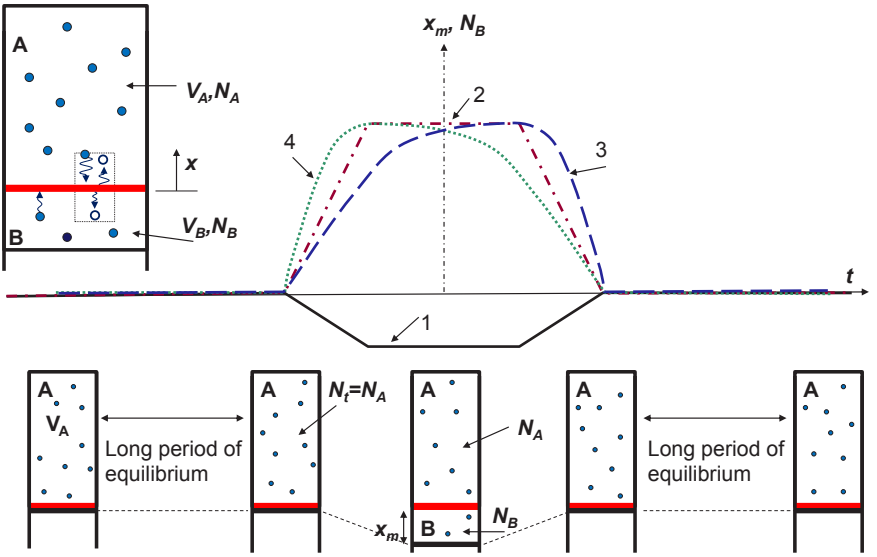


Fig. 3 Experiment with admixture passing through semi-permeable membrane. Curves: 1 – position of the piston $x_m(t)$; 2 – equilibrium $N_B(t)$; 3 – $N_B(t)$ for $C = +1$; 4 – $N_B(t)$ for $C = -1$.

3.2 Why is the principle of parallelism of entropy increase essential?

At this point we consider a modified experiment, which is illustrated in Figure 3. A cylinder having a finite volume V_t contains N_t particles of the passive admixture (as considered previously) and is kept in a state of thermodynamic equilibrium for a long time. The cylinder is located in a remote part of the universe away from any possible influences of the matter that populates the universe. The piston remains at $x = 0$ for a long time so that $V_\mathrm{A} = V_t$ and $V_\mathrm{B} = 0$, then moves down and up in a time-symmetric manner so that $V_\mathrm{B} > 0$ as shown in Figure 3 and then, again, remains at $x = 0$ for a very long time so that $V_\mathrm{B} = 0$. In addition to admixture molecules, the cylinder may also be filled by a gas to ensure that the system under

consideration is thermodynamic. The volumes A and B are divided by a very thin semi-permeable membrane that is fully permeable for the gas (if gas is present) and only partially permeable for the molecules of admixture, so that these molecules can occasionally tunnel through the membrane. When considered from a quantum-mechanical perspective, the membrane is interpreted as a potential barrier that can be tunneled through, while the rest of the walls are formed by impervious barriers of a high potential. We note that such experiments are not only conceptually possible but, due to recent technological advances [29], also practically feasible. Obviously, $N_A + N_B = N_t = \text{const}$ and $V_A = \text{const}$. The number of particles N_t is sufficiently large to ensure that N_A and N_B are macroscopic parameters, which can be measured by classical instruments.

For simplicity of evaluation, the probability of successful tunnelling of admixture molecules through the membrane is assumed to be small so that the concentrations of particles remain uniform in volumes A and B (although not necessarily the same on both sides of the membrane — see Figure 3). Since particles do not interact, they can be considered autonomously from one another. The concentrations of particles on both sides of the membrane are determined by quantum tunnelling through the membrane. Classical statistics is assumed so that most of the quantum states are vacant: all states have the same probability of occupation determined by the concentrations of the particles: N_A/V_A on one side and N_B/V_B on the other.

As the particles tunnel through the membrane, they must decohere since, otherwise they would be simultaneously present in volumes A and B, be governed by unitary evolutions and not subject to the laws of statistical physics (see Ref. [30]). We, however, do not have any experimental evidence that this can happen when an object is progressively screened from the direct influence of the initial and final conditions imposed on the universe. If decoherence is terminated, we would effectively obtain a less cruel version of Schrödinger's cat — a substance whose particles are not located in volumes A or B but are in superposition states between these volumes (strictly speaking, N_A and N_B are not classically defined in this case). After decoherence and collapse of the wave function, particles appear either on one side of the membrane or the other with some classical probability. As we do not wish to discriminate the direction of time a priori, we must admit that the particles can decohere or recohere (i.e. decohere backward in time), as discussed in the Appendix. The concentration of particles is governed by the equation (see Appendix and Refs. [26, 27])

$$\frac{dN_B}{dt} = -\frac{dN_A}{dt} = CK\left(\frac{N_A}{V_A} - \frac{N_B}{V_B}\right) \tag{9}$$

where K is the rate constant for transition through the membrane, which, as shown in the Appendix, must be the same for transitions from A to B and from B to A. The constant $C = +1$ corresponds to predominant decoherence and $C = -1$ to predominant recoherence (i.e. decoherence back in time). In principle, we also need to consider the case of $C = 0$ (assuming that intensities of decoherence and recoherence exactly match each other) but this case is not realistic. Indeed, if the piston moves very slowly, then the densities of particles must approach the same values on both sides of the membrane $N_A/V_A = N_B/V_B = N_t/\ V_t$ and, obviously, $N_A(t) = V_A N_t/$

$V_t(t)$. On the one hand, $N_{\mathrm{A}}(t)$ changes but, on the other hand, equation (9) with $C = 0$ enforces that $dN_{\mathrm{A}}/dt = 0$. Therefore, particles must either predominately decohere or predominantly recohere. This can be easily determined by moving the piston a bit faster so that the solution of equation (9) deviates from the equilibrium given by $N_{\mathrm{A}}(t) = V_{\mathrm{A}}N_t/\ V_t(t)$, as illustrated in Figure 3. We can observe either the behaviour indicated by line 3, which corresponds to $C = +1$, or the behaviour indicated by line 4, which corresponds to $C = -1$. The difference between the two cases is in the definition of the direction of time. As we use the conventional definition of the direction of time, where entropy increases toward $t = +\infty$, then $C = +1$ and particles predominantly decohere.

From the perspective of quantum mechanics, the state of equilibrium corresponds to the maximally mixed quantum state, where the density matrix is proportional to the unit matrix and the entropy is maximal. This state of maximal entropy cannot be altered without external interference; neither by unitary evolution (which cannot change entropy), nor by decoherence (which cannot reduce entropy). The effect of decoherence, therefore, is not observable in equilibrium conditions (as it should be — equilibrium states do not evolve). It would be rather unphysical to assume that decoherence, which exists at its full strength under smallest deviations from equilibrium, physically disappears once full equilibrium state is reached. It is the statistical effect of decoherence that disappears, not decoherence itself: it still affects individual particles at microscopic level. This can be illustrated by the Ehrenfest urn model: balls are located in two urns are picked up at random and are placed into another urn (possibly with a fixed probability reflecting the transmission rate between the urns). Each act of redistribution of balls increases uncertainty of ball locations, and therefore, increases the corresponding entropy. Once equilibrium is reached and the two urns have the same number of balls, the process (which still continues physically) does not change the distribution and does not change the entropy.

We observe a very interesting situation: the system stays in complete equilibrium for a very long time and should not be affected by any initial conditions that were imposed on the system or on the whole universe a long time before the experiment. According to the conditions of the experiment, all external influences must be macroscopic. These influences are limited to the movements of the piston, which are conducted in a time-symmetric manner and cannot possibly create any directionality of time. The known laws of classical, quantum and relativistic physics are also time-symmetric. Why do the particles behave in a time-directional manner (decohere and not recohere)? Reichenbach's principle of parallelism of entropy increase clearly requires that $C = +1$ in (9) and, at least under conditions shown in Figure 3, this cannot be directly explained by the low-entropy initial conditions imposed on the universe. Something must be missing.

3.3 The time primer

We, of course, do not suggest that predominance of decoherence ($C = +1$ in (9)) is not related to the low-entropy initial conditions imposed on the universe, but rather observe that there must be a physical mechanism that connects decohering properties of matter to the fundamental state of the universe. There is, however, no obvious or known mechanism that translates a low-entropy Big Bang into the fact that matter predominantly decoheres under conditions when matter is screened from the Big Bang by an equilibrium state (presumably destroying all information about the previous states of the universe). We can call this mechanism the "time primer"[13]. The time primer is related to the most fundamental properties of matter and its primary effect should be predominance of quantum decoherence, resulting in the second law of thermodynamics, causality and in the perceived "flow of time". The time primer must exist and, at least in principle, should be represented by a mechanism that can be detected in experiments but, as discussed in the rest of this paper, this is likely to be a very difficult task. The time primer may, of course, reflect environmental interferences but these interferences should be measurable and enacting the arrow of time without presuming antecedent causality.

The conventional quantum theories [31–33] explain the physical mechanism of decoherence quite well, but only under conditions, in which the direction of time is discriminated by implied causality: setting initial (and not final) conditions is essential for these theories. Therefore, we are trapped in a logical loop: we explain causality by the second law, the second law by decoherence, and decoherence by causality (Figure 4). The time primer points to an unknown physical effect that is needed to break this loop. For the case illustrated in Figure 3 there is no obvious justification for discriminating the directions of time by preferring the initial conditions to the final conditions. Price [18] noted that physical theories often discriminate the directions of time by intuitively implying time-directional causality — these may be valuable theories in many respects but they cannot serve, as physical explanations of the directional properties of time as these properties are presumed and not deduced.

4 Discussion

The current state of arguments about direction of time (illustrated in a simplified form by Figure 4) reflects persisting confusion: philosophers seek the assistance of the physical laws (and especially that of the second law of thermodynamics) in defining antecedent causality, while physicists base their justifications of physical laws and theories on implications of causality (often tacitly or implicitly). This state forms an unsatisfactory explanatory loop, in which antecedent causality is associated with the action of the second law and the second law is explained by the effects of antecedent causality. While the action of the second law can be related at an elementary level to implications of quantum decoherence and collapse, the quantum theory, as was remarked by Einstein half-a-century ago, still cannot provide a uni-

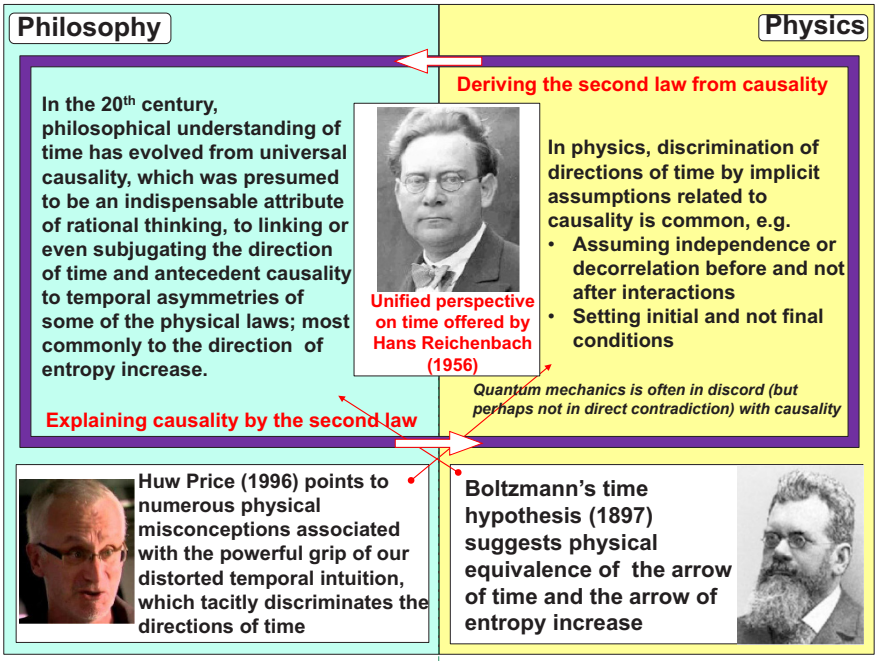


Fig. 4 If considered from a transdisciplinary perspective, arguments commonly used in physics and philosophy in explaining antecedent causality and the second law of thermodynamics form a logical circle

fied picture of physical reality. The time primer is not a physical theory but rather a placeholder for such a theory, recognising that something important is missing in our understanding of thermodynamic time.

Over the last few decades, the direction of time has experienced a gradual drift from the domain of philosophy to the domain of physics. While the influence of physical ideas and theories gradually increases, this transition has not been completed yet since the possibility of experimental validation is a necessary attribute of any physical theory. The possibility of experimental testing of the time priming is discussed in this section.

4.1 Environmental time priming

Decoherence may be induced by relatively weak interactions with the rest of the universe, since our universe is far away from equilibrium and (at least in principle) can induce time-directional effects in the system. Numerous quantum theories point to environmental interferences as the mechanism responsible for decoherence

and thermodynamic behaviour in quantum systems [33–39]. These theories, however, are not specific with respect to the physical mechanism of interactions, which make experimental validation of these interactions rather difficult and uncertain. Environmental interference of CP-violating and CPT-invariant quantum systems is expected to produce apparent CPT violations [13, 40] and detected CPT discrepancies [41] may be related to interference from the environment. The problem with experimental validation of environmental interference is that, even if this interference is detected, there is no guarantee that it is this interference and not something else that represents the principal mechanism controlling the time priming. To prove this point we need to reduce this interference and expect a corresponding reduction in consistency of time priming.

Radiation is likely to be the first suspect for thermodynamic interactions. Since radiation itself must be decoherence-neutral [26], its role should be in connecting the equilibrated system (in Figure 3) to matter that populates the universe and remains far from equilibrium. If the experiment is located in a remote area of the universe, incoming radiation can be interpreted as a random signal. This signal can stimulate decoherence, but it seems that presuming causality is unavoidable under these conditions [32, 33, 35].

If a system is placed far away from all other matter, a reduction in effectiveness of interactions can be expected. At present, however, we do not have any evidence that thermodynamic time slows down when a system is screened from the influence of (or placed far from) other thermodynamic systems. Would radioactive decays become any slower if an radioactive object is placed in a very remote part of the universe? There is no direct evidence that this would be the case. Reichenbach believed that complete insulation of a subsystem would not affect the rate of its entropy increase. This does not rule out environmental mechanisms of time priming, but it does illustrate that obtaining experimental proof of environmental time priming would be very difficult. In principle, there might be a "time field" that is present everywhere, and the direction of time is determined by very weak, yet very important, interactions with this field. This case, however, is practically indistinguishable from intrinsic mechanisms of decoherence.

4.2 Intrinsic mechanisms of time priming

Various theories modifying equations of quantum mechanics to incorporate quantum collapses and decoherences have been suggested [34, 42–45]. These theories, however, assume causality rather than attempt to explain causality (and some are merely empirical). The physical mechanism of entropy-increasing processes at microscopic level remains uncertain. Penrose [42, 46] suggested a physical mechanism that can "prime" the direction of time. This theory (due to Diosi and Penrose) points to gravitational effects as a culprit of irreversibilities observed in the quantum world. Gravity induces quantum violations causing collapses of otherwise reversible uni-

tary evolutions. This provides a very good illustration of how small these violations might be and how difficult it would be to directly detect them in experiments [46].

Considering that radiation is expected to remain decoherence-neutral we might extend this inference to all bosons and expect that the intrinsic source of decoherence must be hidden in the properties of matter, most likely in quark - containing particles (e.g. neutrons and protons) since quarks are known of being capable to violate time symmetry in weak interactions (e.g. known CP violations in mesons, which, in conjunction with CPT invariance, imply T violation). One may prefer to imagine that baryons are capable of accumulating and confining very large quantities of information (i.e. baryons have effectively infinite numbers of internal degrees of freedom that are not externally accessible under normal conditions). In this case, there remain two possibilities: baryons and antibaryons can violate unitarity of external quantum evolutions in a symmetric or antisymmetric manner, which result in either symmetric or antisymmetric extension of thermodynamics from matter into antimatter [27, 47]. Symmetric and antisymmetric versions of thermodynamics respectively correspond to CP- and CPT-invariant time priming and can not be valid simultaneously — only one of them can be (and is) real. The antisymmetric version may or may not correspond to the real world but, conceptually, it is quite attractive due to a number of reasons. One of these reasons is that, if antisymmetric thermodynamics is valid, it kinetically favours conversion of antimatter into matter and, at the same time, explains the present arrow of time by the relative abundance of matter over antimatter [27, 47]. If detected in experiments, antisymmetric thermodynamics can pinpoint at the intrinsic mechanisms of time priming. If it is the symmetric version that is real, then experimental examination of the intrinsic mechanisms of time priming becomes a more difficult task.

4.3 Testing the symmetry of time priming.

From a theoretical perspective, testing whether thermodynamics possesses symmetric or antisymmetric properties may seem straightforward — we just need to create thermodynamically significant quantities of antimatter and see which thermodynamic properties it has. Practically, producing significant quantities of antimatter can be extremely difficult. It might be possible, however, to test the symmetric/antisymmetric properties of thermodynamics at the present level of technology.

It seems that a system with some thermodynamic properties (i.e. quark-gluon plasma [48]) can be created at very small scales as a result of collision of high-energy protons and nuclei. For example, two protons may collide elastically producing two protons with different momenta or inelastically producing jets of multiple particles. While the former collisions are unitary, we are tempted to assume that the latter collisions have some thermodynamic features. If this thermodynamic interpretation of inelastic collisions is correct, collisions of two antiprotons should be the same as collisions of protons according to symmetric thermodynamics, and can be expected to be different from collisions of protons according to antisym-

metric thermodynamics. While the overall energy, momentum and other conserved properties must always be preserved, antisymmetric thermodynamics involves opposite entropy trends for matter and antimatter. Therefore, assuming that thermodynamic effects can play a role within very short times associated with collisions (which is a big assumption, of course), antisymmetric thermodynamics predicts that antiproton-antiproton collisions should tend to have smaller inelastic collision cross-sections than the inelastic cross-sections of the proton-proton collisions under the same conditions. In simple terms, collisions of antiprotons should be biased towards elastic collisions compared to collisions of protons under the same conditions. This attributes the action of the time primer to complex interactions of partons inside baryons, which are clearly revealed only when collision energies are sufficiently high. The extent of the differences between baryons and antibaryons is determined by persistency of the time primer (i.e. it might be difficult to collide two antiprotons inelastically). Symmetric thermodynamics does not predict any differences between inelastic cross-sections of protons and antiprotons.

Note that the implications of antisymmetric thermodynamics may produce an impression of CPT violations: protons and antiprotons can have different overall inelastic collision cross-sections [49]. According to interpretation given above, this conclusion would be incorrect — antisymmetric thermodynamics is based on complete CPT symmetry exhibited both at small and large scales. This effect is similar to apparent CPT violations that can be observed due to the presence of environmental mechanisms of time priming — see Ref [40] for details. It seems that microscopic action of time priming can be detected (due to its interference with unitarity) as apparently present CPT violations in systems that in fact strictly preserve the CPT symmetry.

Another possibility for testing the extension of thermodynamics from matter to antimatter is investigation of photon absorption and radiation by atoms and antiatoms under the same conditions. The antiatoms need to be trapped and cooled down, which is not easy but still possible [50]. The kinetics of light absorption and radiation is the same for atoms and antiatoms in symmetric thermodynamics and different in antisymmetric thermodynamics[26]. In simple terms, if antiatoms are somewhat more reluctant to adsorb photons than the corresponding atoms under the same conditions, then this would indicate validity of antisymmetric thermodynamics. Again, if such effects are detected, they must not be confused with CPT violations — antisymmetric thermodynamics is very much consistent with the CPT invariance.

5 Conclusions

This work briefly reviews and explains the principal ideas about time that were brought by the late Hans Reichenbach in his last book. The Boltzmann time hypothesis and the Reichenbach principle of parallelism of entropy increase seem to be most important among these ideas. While the Boltzmann time hypothesis tends

to be accepted by modern philosophers and physicists (at least by those who have thought about or investigated these issues), the principle of parallelism of entropy increase is still subject to debate. In the present work, we consider a mixing process involving quantum effects and demonstrate that, although the low-entropy initial conditions that characterised early universe are most important, there should be an unknown mechanism that delivers the influence of these initial conditions to thermodynamic subsystems observed in the real world. We call this mechanism the "time primer". The time primer is responsible for prevailing forward-time decoherence in quantum systems, which increases entropy and, according to the Boltzmann time hypothesis, introduces antecedent causality and other components of the perceived flow of time.

The possibility of experimental detection of the time primer is discussed in the last section — in general, this task is quite difficult. If, however, the time primer is CPT-invariant (rather than CP-invariant) and objects with some thermodynamic properties emerge at small scales in inelastic high-energy collisions, the direct effects of the time primer may be detected under the current level of technology.

Acknowledgements The author thanks the Mathematical Research Institute MATRIX and the Department of Mathematics and Statistics at The University of Western Australia for productive discussions and financial support. The author also appreciates fruitful discussion at the Centre for Time (The University of Sydney).

References

[1] Immanuel Kant. *Critique of Pure Reason*. Palgrave Macmillan UK, London, 2007 (1781-1788).
[2] Mario Bunge. *Causality and Modern Science*. Taylor and Francis, 2017 (1959).
[3] Max Born. *Natural philosophy of cause and chance : together with a new essay, Symbol and reality.* Waynflete lectures ; 1948. Dover Pub., New York, 1964.
[4] B. Russell. *Our Knowledge of the External World*. Taylor and Francis, Florence, 2009 (1914).
[5] P. Dowe. Process causality and asymmetry. *Erkenntnis*, 37(2):179–196, 1992.
[6] Jan Faye. *Causation, Reversibility and the Direction of Time*, pages 237–266. Springer Netherlands, Dordrecht, 1997.
[7] David Z Albert. *Time and chance*. Harvard University Press, Cambridge, Mass., 2000.
[8] David Z Albert. *After Physics*. Harvard University Press, 2015.
[9] Barry Loewer. Counterfactuals and the second law. In Huw Price and Richard Corry, editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press, 2007.
[10] L. Boltzmann. *Lecures on gas thoery*. English translation by S.G. Brush. University of California Press, Berkeley and L.A., 1964 (1895,1897).
[11] H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956, (reprinted 1971).
[12] S.W. Hawking. The no-boundary proposal and the arrow of time. *Vistas in Astronomy*, 37: 559 – 568, 1993.
[13] A Y Klimenko. The direction of time and Boltzmann's time hypothesis. *Physica Scripta*, 94: 034002, 2019.

[14] I. Prigogine. *From being to becoming : time and complexity in the physical sciences*. W. H. Freeman, San Francisco, 1980.
[15] R. Penrose. *Road to Reality: A Complete Guide to the Laws of the Universe*. A. Knopf Inc., 2005.
[16] H. D. Zeh. *The physical basis of the direction of time*. Springer, New York;Berlin;, 5th edition, 2007.
[17] S. W. Hawking. *A brief history of time : from the big bang to black holes*. Bantam, London, 2011.
[18] H. Price. *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. Oxford Univ. Press, Oxford, UK, 1996.
[19] P. C. W. Davies. *The Physics of Time Asymmetry.* University of California Press, Berkeley, 1977.
[20] Lawrence Sklar. *Physics and chance : Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press, Cambridge, 1993.
[21] Eric Winsberg. Laws and statistical mechanics. *Philosophy of Science*, 71(5):707–718, 2004.
[22] J. North. Time in thermodynamics. In *The Oxford Handbook of Philosophy of Time*. Oxford University Press, 2011.
[23] Karl R. Popper. Irreversible processes in physical theory. *Nature*, 181(4606):402–403, 1958.
[24] J. Faye, U. Scheffler, and M. Urchs. *Perspectives on time*. Boston studies in the philosophy of science ; v. 189. Kluwer Academic Publishers, Dordrecht ; Boston, 1997.
[25] Clark Glymour and Frederick Eberhardt. Hans reichenbach. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016.
[26] A. Y. Klimenko. Kinetics of interactions of matter, antimatter and radiation consistent with antisymmetric (CPT-invariant) thermodynamics. *Entropy*, 19:202, 2017.
[27] A. Y. Klimenko. Symmetric and antisymmetric forms of the Pauli master equation. *Scientific Reports (nature.com)*, 6:29942, 2016.
[28] A. Y. Klimenko. On the inverse parabolicity of pdf equations. *QJMAM*, 57:79–93, 2004.
[29] Igor Poltavsky, Limin Zheng, Majid Mortazavi, and Alexandre Tkatchenko. Quantum tunneling of thermal protons through pristine graphene. *The Journal of Chemical Physics*, 148 (20), 2018.
[30] A. Y. Klimenko. On quantum tunnelling with and without decoherence and the direction of time. *In preparation*, 2019.
[31] W. H. Zurek. Decoherence, einselection, and the quantum origins of the classical. *Reviews of Modern Physics*, 75(3), 2003.
[32] N. Linden, S. Popescu, A. J. Short, and A. Winter. Quantum mechanical evolution towards thermal equilibrium. *Phys. Rev. E*, 79:061103, 2009.
[33] V.I. Yukalov. Equilibration and thermalization in finite quantum systems. *Laser Phys. Lett.*, 8(7):485—-507, 2011.
[34] P. C. E. Stamp. Environmental decoherence versus intrinsic decoherence. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 370(1975):4429, 2012.
[35] W. H. Zurek. Environment-induced superselection rules. *Phys. Rev. Lett.*, 26(8):1862–1888, 1982.
[36] E. Joos, C. Kiefer, and H. D. Zeh. *Decoherence and the Appearance of a Classical World in Quantum Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2 edition, 2003.
[37] M. Schlosshauer. Decoherence, the measurement problem, and interpretations of quantum mechanics. *Rev. Mod. Phys.*, 76:1267–1305, 2005.
[38] S. Goldstein, J. L. Lebowitz, R. Tumulka, and N. Zanghi. Canonical typicality. *Phys. Rev. Lett.*, 96:050403, 2006.
[39] S. Popescu, A. J. Short, and A. Winter. Entanglement and the foundations of statistical mechanics. *Nature Physics*, 2(11):754–758, 2006.
[40] A.Y. Klimenko. Note on invariant properties of a quantum system placed into thermodynamic environment. *Physica A: Statistical Mechanics and its Applications*, 398:65 – 75, 2014.

[41] Lees, J. P. et. al. (BABAR Collaboration). Tests of cpt symmetry in $B^0-\overline{b}^0$ mixing and in $B^0 \to c\bar{c}K^0$ decays. *Phys. Rev. D*, 94:011101, 2016.
[42] R. Penrose. On gravity's role in quantum state reduction. *General Relativity and Gravitation*, 28(5):581–600, 1996.
[43] W. H. Zurek. Decoherence and the transition from quantum to classical – revisited. *Los Alamos Science*, (27):1–26, 2002.
[44] A. Bassia and G. Ghirardi. Dynamical reduction models. *Physics Reports.*, 379:257–426, 2003.
[45] G. P. Beretta. On the general equation of motion of quantum thermodynamics and the distinction between quantal and nonquantal uncertainties (MIT, 1981). arXiv: quant-ph/0509116, 2005.
[46] R. Penrose. On the gravitization of quantum mechanics 1: Quantum state reduction. *Foundations of Physics*, 44(5):557–575, 2014.
[47] A. Y. Klimenko and U. Maas. One antimatter- two possible thermodynamics. *Entropy*, 16 (3):1191–1210, 2014.
[48] P. Braun-Munzinger and J. Stachel. The quest for the quark-gluon plasma. *Nature*, 448 (7151):302–309, 2007.
[49] McL. Emmerson J. *Symmetry principles in particle physics*. Clarendon Press, Oxford, 1972.
[50] Niels Madsen. Cold antihydrogen: a new frontier in fundamental physics. *Phil.Trans. Roy. Soc. A*, 368(1924):3671–3682, 2010.
[51] L. D. Landau and E. M. Lifshits. *Course of Theoretical Physics vol.3: Qunatum mechanics*. Butterworth-Heinemann, Oxford, 1980.
[52] W. Pauli. Uber das h-theorem vom anwachsen der entropie vom standpunkt der neuen quantenmechanik. In *Probleme der Modernen Physik. Arnold Sommerfeld zum 60 Geburtstage*, pages 30–45. Hirzel, Leipzig, 1928.

Appendix. Quantum tunnelling and decoherence

The quantum outcomes of tunnelling can be expressed by the scattering matrix $\mathbb{S}$, which is a unitary matrix (i.e. $\mathbb{S}\mathbb{S}^\dagger = \mathbb{I}$) that connects the amplitudes A^- and B^- of incoming waves with the amplitudes of the outgoing waves A^+ and B^+ (see Figure 5) so that:

$$\underbrace{\begin{bmatrix}\tilde{A}^+\\ \tilde{B}^+\end{bmatrix}}_{\psi(t_+)} = \underbrace{\begin{bmatrix}r & q\\ q & r\end{bmatrix}}_{\mathbb{S}} \underbrace{\begin{bmatrix}\tilde{A}^-\\ \tilde{B}^-\end{bmatrix}}_{\psi(t_-)}, \quad \underbrace{\begin{bmatrix}\tilde{A}^-\\ \tilde{B}^-\end{bmatrix}}_{\psi(t_-)} = \underbrace{\begin{bmatrix}r^* & q^*\\ q^* & r^*\end{bmatrix}}_{\mathbb{S}^\dagger} \underbrace{\begin{bmatrix}\tilde{A}^+\\ \tilde{B}^+\end{bmatrix}}_{\psi(t_+)} \tag{10}$$

where $\tilde{A}^+ = A^+ e^{ik\Delta/2}$, $\tilde{A}^- = A^- e^{-ik\Delta/2}$, $\tilde{B}^+ = B^+ e^{ik\Delta/2}$, $\tilde{B}^- = B^- e^{-ik\Delta/2}$ are the corresponding wave amplitudes evaluated at the boundaries of the barrier at $x = \pm\Delta/2$, the asterisk denotes complex conjugates and the values of q and r are specified below. The quantum barrier is assumed to be symmetric, which corresponds to a symmetric matrix $\mathbb{S}$. The matrix $\mathbb{S}$ should not be confused with the commonly used transfer matrix that links the wave amplitudes on one side of the barrier to the wave amplitudes on the other side. Note that $|q|^2 + |r|^2 = 1$ and $\left|r^2 - q^2\right| = 1$ due to the unitary of $\mathbb{S}$. The tunnelling parameters q and r can be determined for specific shape of the potential barrier $U(x)$, which is assumed to have a rectangular shape as shown in Figure 5. The solution of this problem can be found in standard textbooks [51]:

$$\begin{aligned} r &= (k^2+\kappa^2)\frac{(1-Q^2)}{W},\ q = 4ik\kappa\frac{Q}{W},\ Q = \exp(-\kappa\Delta)\\ W &= (k+i\kappa)^2 - (k-i\kappa)^2 Q^2,\ k = \frac{\sqrt{2mE}}{},\ \kappa = \frac{\sqrt{2m(U_0-E)}}{} \end{aligned}$$

$$|q|^{-2} = 1 + \frac{1}{4}\frac{(k^2+\kappa^2)^2}{k^2\kappa^2}\sinh^2(\kappa\Delta) \underset{U_0 \gg E}{\approx} \frac{1}{4}\frac{U_0}{E}\sinh^2\left(\Delta\frac{\sqrt{2mU_0}}{}\right) \tag{11}$$

where E is the energy of the particle, is the Planck constant and $|q|^2$ is the transmission coefficient. The barrier is assumed to be thin: i.e. its thickness Δ is small but its magnitude U_0 is large. We can assume that $U_0 \gg E$ and therefore $q \ll 1$, $r \sim 1$.

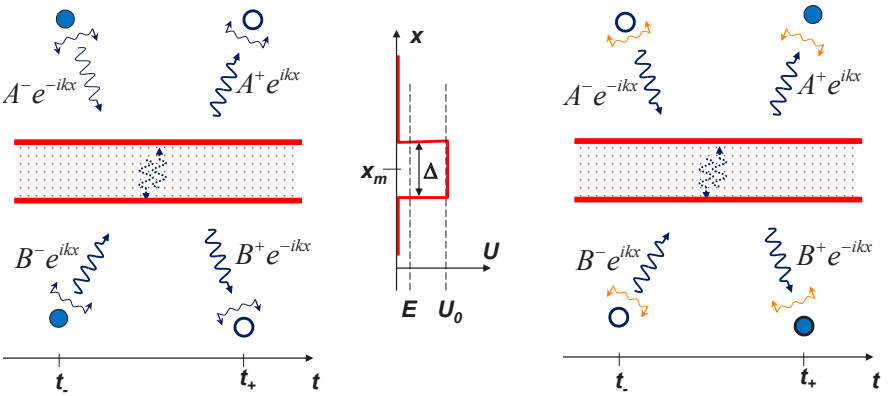


Fig. 5 Tunnelling of a particle through the membrane: left – with decoherence, right – with recoherence, middle – the membrane potential $U = U(x)$

The quantum description of tunnelling specified by (10) is time-symmetric, while its effect on the thermodynamic system considered here (Figure 3) is determined by the decoherence of quantum waves as shown in Figure 5. The decoherence transforms the time-reversible Schrodinger equation into the Pauli master equation, which is the principal equation that combines quantum description with directionality of time [52]. The Pauli master equations are general equations that incorporate decoherence, which determines the direction of the entropy increase, into the quantum world; i.e. different forms of the Pauli master equation are obtained for the same quantum system depending on properties of decoherence and recoherence [26, 27].

Since particles do not interact and classical statistics is implied (i.e. most quantum states are not occupied), one can consider the wave function ψ_j of a single particle. The Pauli master equation for the probabilities $p_j = \psi_j \psi_j^*$ (no summation over j) is given by [27]

$$\frac{dp_j}{dt} = \sum_k C w_j^k p_k - \sum_k C w_k^j p_j \tag{12}$$

where $C = +1$ corresponds to dominant decoherence and $C = -1$ corresponds to dominant recoherence, and $w_j^k = w_k^j$ are transitional probabilities. Note that, unlike in Ref. [26, 27], the predominant direction of the time priming is assumed to be the same for all quantum states. Consider states $a = a_1, a_2, \ldots$ on side A of the membrane and states $b = b_1, b_2, \ldots$ on side B of the membrane so that $j = a_1, a_2, \ldots b_1, b_2, \ldots$ and the states a_i and b_i, $i = 1, 2, 3, \ldots$ correspond to interacting waves with the same energy E_i. Evaluation of the two sums over $j = a_1, a_2, \ldots$ and over $j = b_1, b_2, \ldots$ in equation (12) while taking into account

$$\sum_a p_a = \frac{N_\mathrm{A}}{N_t}, \quad \sum_b p_b = \frac{N_\mathrm{B}}{N_t}, \quad N_t = N_\mathrm{A} + N_\mathrm{B} \tag{13}$$

yields

$$\frac{1}{N_t}\frac{dN_\mathrm{B}}{dt} = -\frac{1}{N_t}\frac{dN_\mathrm{A}}{dt} = C \sum_b \sum_a \left(w_b^a p_a - w_a^b p_b \right) \tag{14}$$

Substituting the equilibrium distribution g_j° (which are assumed to be classical Gibbs distributions due to $g_j^\circ \ll 1$) and the density of quantum states ρ_j

$$p_a = \frac{N_\mathrm{A}}{N_t V_\mathrm{A}} \rho_a g_a^\circ, \quad p_b = \frac{N_\mathrm{B}}{N_t V_\mathrm{B}} \rho_b g_b^\circ, \quad g_{a_i}^\circ = g_{b_i}^\circ = \exp\left(\frac{\mu - E_i}{k_B T}\right) \tag{15}$$

where μ is the chemical potential, we obtain

$$\frac{dN_\mathrm{B}}{dt} = -\frac{dN_\mathrm{A}}{dt} = C\left(K_1 \frac{N_\mathrm{A}}{V_\mathrm{A}} - K_2 \frac{N_\mathrm{B}}{V_\mathrm{B}}\right) \tag{16}$$

and

$$K_1 = \sum_b \sum_a w_b^a \rho_a g_a^\circ = \sum_i w_{b_i}^{a_i} \rho_{a_i} g_{a_i}^\circ, \quad K_2 = \sum_b \sum_a w_a^b \rho_b g_b^\circ = \sum_i w_{a_i}^{b_i} \rho_{b_i} g_{b_i}^\circ \tag{17}$$

since only the corresponding states a_i and b_i interact, i.e. $w_{a_j}^{b_i} = w_{b_j}^{a_i} = 0$ for $j \neq i$. The symmetry of the coefficients $w_a^b = w_b^a$ and equilibrium distributions $g_{a_i}^\circ = g_{b_i}^\circ = g_i^\circ$ and the same conditions on both sides of the membrane $\rho_{a_i} = \rho_{b_i} = \rho_i$ yield equation (9) with $K = K_1 = K_2$. The direction of thermodynamic time in this equation is determined by the temporal direction of decoherence (i.e. by $C = +1$ or $C = -1$). As expected [26], *the transmission rate is proportional to the concentration of decohered particles and does not depend on the concentration of recohered particles, irrespective of the temporal direction of decoherence or recoherence*.

Considering that the scattering matrix is close to unity, one can write $\mathbb{S} = \mathbb{I} + i\mathbb{T}$ where $\mathbb{T}$ is small (since $|q|^2 \ll 1$) and Hermitian $\mathbb{T}^\dagger = \mathbb{T}$ at the leading order. The operator $\mathbb{T}$ can be conventionally be expressed in terms of the interaction Hamiltonian by using perturbation methods, but this is not needed here as we already have the exact solution for the tunnelling problem. Substituting $w_{a_i}^{b_i} = w_{b_i}^{a_i} = \mathfrak{A} u_i |q|_i^2 / 2$, where $\mathfrak{A}$ is the area of the membrane, $|q|_i^2 \approx 4E_i e^{-2\kappa\Delta}/U_0$ is the transmission coefficient, $u_i^2 = 2E_i/m$ and $\kappa = (2mU_0)^{1/2}/$, into (17) results in

$$K = 2^{2\frac{1}{2}} \frac{\mathfrak{A}}{U_0 m^{\frac{1}{2}}} e^{-2\kappa\Delta} \sum_i E_i^{1\frac{1}{2}} \rho_i g_i^\circ \tag{18}$$

A more detailed analysis of tunnelling without decoherence under these conditions can be found in Ref. [30].

Controlling stability of longwave oscillatory Marangoni patterns

Anna Samoilova and Alexander Nepomnyashchy

Abstract We apply nonlinear feedback control to govern the stability of long-wave oscillatory Marangoni patterns. We focus on the patterns caused by instability in thin liquid film heated from below with a deformable free surface. This instability emerges in the case of substrate of low thermal conductivity, when two monotonic long-wave instabilities, Pearson's and deformational, are coupled. We provide weakly nonlinear analysis within the amplitude equations, which govern the evolution of the layer thickness and the temperature deviation. The action of the nonlinear feedback control on the nonlinear interaction of two standing waves is investigated. It is shown that quadratic feedback control can produce additional stable structures (standing rolls and standing squares), which are subject to instability leading to traveling wave in the uncontrolled case.

1 Introduction

The onset and development of oscillatory Marangoni convection in a thin film heated from below without control was recently investigated by Shklyaev *et al.* [5]. Among all the variety of possible patterns, only a few were stable: one-dimensional traveling waves, traveling rectangles and alternating rolls. In this paper we aim at revealing more complex and exotic stable patterns, such as alternating rolls and standing squares.

We have recently considered the influence of the feedback control on the oscillatory Marangoni instability in a thin film heated from below. We have shown that

Anna Samoilova
Department of Theoretical Physics, Perm State University, Perm 614990, Russia, e-mail: annsomeoil@gmail.com

Alexander Nepomnayshchy
Department of Mathematics, Technion–Israel Institute of Technology, Haifa 32000, Israel, e-mail: nepom@technion.ac.il

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_24

a linear control gain can delay the onset of instability [3] and a quadratic control gain can eliminate the subcritical excitation of instability [4]. The analysis of pattern formation was done for an infinite region, nonlinear interaction of the traveling waves was considered. In the case of traveling waves we showed that quadratic feedback control can produce additional stable structures, besides conventional traveling rolls. However, in a realistic system the reflection of waves on the lateral boundaries results in emergence of standing waves, which can interact to each other. Extending our previous investigation, we examine here the effect of nonlinear feedback control on development of Marangoni instability in a system of standing waves propagating with a definite angle between the wave vectors.

The paper is organized as follows. We start with the mathematical formulation of the long-wave Marangoni convection problem in Sec. 2. There we present a set of coupled amplitude equations which governs the evolution of the layer thickness and the temperature deviation under nonlinear feedback control [4]. In Sec. 3 we perform the weakly nonlinear stability analysis of wave patterns within these amplitude equations. Nonlinear interaction of standing waves is investigated by means of the analysis of a system of four complex Landau equations. The paper concludes with summary in Sec. 4.

2 Amplitude Equations

We consider a horizontal liquid layer confined between a deformable free upper surface and a solid bottom wall. The layer is heated from below; the thermal conductivity of the liquid λ is assumed to be large in comparison with that of the substrate, so that the vertical component of the heat flux λA is fixed. The unperturbed layer thickness H is assumed sufficiently small, so that the influence of buoyancy is negligible and the free surface deformation is important. The surface tension decreases linearly with the temperature: $\sigma = \sigma_0 - \sigma_T T$, where T is the deviation of the temperature from a reference one, which is the temperature of the gas above the liquid layer. The heat flux from the free surface is governed by Newton's law of cooling, which describes the rate of heat transfer from the liquid to the ambient gas phase with the heat transfer coefficient q. The Cartesian reference frame is chosen in such a way that the x- and y-axes are in the substrate plane and the z-axis is normal to the substrate.

The problem of convective instability in the given system is characterized by the following dimensionless parameters,

$$Ca = \frac{\sigma_0 H}{\rho \nu \chi},\ Bi = \frac{qH}{\lambda},\ Ga = \frac{gH^3}{\nu \chi},\ Ma = \frac{\sigma_T A H^2}{\rho \nu \chi},$$

which are the capillary, Biot, Galileo and Marangoni numbers, respectively. Here g is the gravitational acceleration, χ is the thermal diffusivity, ρ is the density, and ν is the kinematic viscosity.

In the uncontrolled case, the oscillatory long-wave Marangoni instability was revealed in [5]. To govern this instability we apply the feedback control based on the measurement of the temperature deviation on the free surface from its value in the conductive state. This feedback control strategy was recently demonstrated as the most effective one to delay the onset of instability under consideration [3].The heat flux applied on the solid substrate is changed as

$$\left.\frac{\partial T}{\partial z}\right|_{z=0} = -1 - K(f) f,\ f = T|_{z=h} - \left. T^{(0)}\right|_{z=1}\ , \tag{1}$$

where $T^{(0)}$ is the temperature of no-motion state, h is the local layer thickness, K is the non-dimensional scalar control gain.

Within the lubrication approximation we employ a standard long-wave scaling

$$x = \varepsilon^{-1} X,\ y = \varepsilon^{-1} Y,\ t = \varepsilon^{-2}\tau \tag{2}$$

and restrict ourselves to following assumptions

$$Ca = \varepsilon^{-2} C,\ Bi = \varepsilon^2 \beta,\ K = \varepsilon^2 \varkappa\,, \tag{3}$$

where $\varepsilon \ll 1$ can be thought of as the ratio of H to a typical horizontal lengthscale.

The long-wave Marangoni convection in this layer is governed by the following system of dimensionless amplitude equations [4]

$$\frac{\partial h}{\partial \tau} = \nabla \cdot \left(\frac{h^3}{3} \nabla P + Ma \frac{h^2}{2} \nabla f \right) \equiv \nabla \cdot \vec{j}\,, \tag{4}$$

$$h \frac{\partial \Theta}{\partial \tau} = \nabla \cdot (h \nabla \Theta) - \frac{1}{2} (\nabla h)^2 - (\beta - \varkappa(f))\, f + \vec{j} \cdot \nabla f + \nabla \cdot \left(\frac{h^4}{8} \nabla P + \frac{h^3}{6} Ma \nabla f \right)\ , \tag{5}$$

where $\Theta(X, Y, \tau)$ is the temperature deviation from its conductive value

$$T = -z + \frac{1}{Bi} + \Theta\ . \tag{6}$$

Here $P = Gah - C\nabla^2 h$, $f = \Theta - h$ has a meaning of perturbation of the free surface temperature; $\nabla = (\partial/\partial X, \partial/\partial Y, 0)$. The vector $-\vec{j}$ has a meaning of the longitudinal flux of a liquid integrated across the layer.

Hereinafter we assume that the term corresponding to the feedback control in (5) is a quadratic polynomial of the free surface temperature perturbation:

$$\varkappa(f) f = \varkappa_l f + \varkappa_q f^2\ , \tag{7}$$

where $\varkappa_l$ and $\varkappa_q$ are constant.

The influence of the linear part of control gain $\varkappa_l$ can be expressed as replacement $\beta \to \beta - \varkappa_l$ in formulas describing the instability threshold [3]. The quadratic part of control gain $\varkappa_q$ affects the nonlinear development of instability. In the following sections we investigate the influence of a nonlinear feedback control on the pattern formation (the linear part $\varkappa_l$ will be omitted). Specifically, we are interested in the elimination of subcritical instability.

3 Weakly Nonlinear Analysis

Below we study the nonlinear dynamics of small perturbations close to the threshold of the oscillatory instability Ma_0

$$Ma - Ma_0 = \delta^2 Ma_2, \delta \ll 1 \,, \tag{8}$$

where $Ma_0 = 3 + Ga + Ck^2 + 3\beta/k^2$ is obtained from the linear analysis [3].

3.1 Basic Expansions

We present h, Θ, Ma and the time derivative as a series in power of the small parameter δ:

$$h = 1 + \delta\xi_1 + \delta^2\xi_2 + \dots, \Theta = 1 + \delta\theta_1 + \delta^2\theta_2 + \dots, \frac{\partial}{\partial\tau} = \frac{\partial}{\partial\tau_0} + \delta^2\frac{\partial}{\partial\tau_2} + \dots, \tag{9}$$

where two time variables, τ_0 and τ_2, are introduced according to the multiscale approach [2] as the dynamics of wave patterns is characterized by two different time scales. The frequency of oscillations is of order of 1, while the growth rate of disturbances is of the order of $Ma - Ma_0$, i.e. $O(\delta^2)$

Substituting the ansatz (9) into equations (4)-(5), and collecting the terms of equal powers in δ, we obtain at the first order the linear stability problem. Its solution can be presented as

$$\xi_1 = \sum_{j=1}^{n} A_j(\tau_2)\, exp\left(\mathrm{i}\vec{k_j}\cdot\vec{r} - \mathrm{i}\omega\tau_0\right) + c.c. \,, \tag{10}$$

$$\theta_1 = (\alpha+1)\sum_{j=1}^{n} A_j(\tau_2)\, exp\left(\mathrm{i}\vec{k_j}\cdot\vec{r} - \mathrm{i}\omega\tau_0\right) + c.c. \,, \tag{11}$$

where c.c. denotes complex conjugate terms, $|\vec{k_j}| = k$ is the wavenumber, $\alpha = -2\left(Ga + Ck^2\right)/3Ma_0 + 2\mathrm{i}\omega/Ma_0k^2$. Frequency of neutral perturbations is determined by formula

$$\omega = \frac{k^2}{12}\sqrt{(72+Ga+Ck^2)\,(Ma_{mon}-Ma_0)},$$

where

$$Ma_{mon} = \frac{48\left(\beta+k^2\right)\left(Ga+Ck^2\right)}{k^2\,(72+Ga+Ck^2)}$$

is the threshold of a monotonic instability [3].

The analysis can be done for any k, but the case of the critical wavenumber k_c, corresponding to the minimum of the neutral curve, is especially important, because one can expect that patterns with the wavenumber k_c will appear in the natural way by the growth of Ma. Below we consider the nonlinear interaction of disturbances and the wave patterns supported by that interaction. The computations will be done for $k = k_c$, $k_c^2 = \sqrt{3\beta}$.

3.2 Interaction of Waves

In order to investigate the nonlinear interaction of waves, consider the class of solutions corresponding to two pairs of waves with the wave vectors $\pm\vec{k}_1$, $\pm\vec{k}_2$, where $\vec{k}_1 = (k,0)$ and $\vec{k}_2 = (k\cos\phi, k\sin\phi)$, that propagate with a phase velocity ω/k and complex amplitudes $A_{1,2}$ and $B_{1,2}$

$$\xi_1 = \left[A_1(\tau_2)\mathrm{e}^{\mathrm{i}kX} + A_2(\tau_2)\mathrm{e}^{-\mathrm{i}kX} + B_1(\tau_2)\mathrm{e}^{\mathrm{i}\vec{k}_2\cdot\vec{r}} + B_2(\tau_2)\mathrm{e}^{-\mathrm{i}\vec{k}_2\cdot\vec{r}}\right]\mathrm{e}^{\mathrm{i}\omega\tau_0} + c.c. \quad (12)$$

Here ϕ is an arbitrary angle different from 0 and π. That class of solutions includes travelling and standing waves as particular cases.

At the second order we obtain

$$\frac{\partial\xi_2}{\partial\tau_0} - \Delta\left(\frac{1}{3}P_2 + \frac{Ma_0}{2}f_2\right) = \nabla\cdot\left(\xi_1\nabla P_1 + Ma_0\xi_1\nabla f_1\right), \quad (13)$$

$$\frac{\partial\theta_2}{\partial\tau_0} - \Delta\left(\theta_2 + \frac{1}{8}P_2 + \frac{Ma_0}{6}f_2\right) + \beta f_2 = -\xi_1\frac{\partial\theta_1}{\partial\tau_0} + \nabla\cdot(\xi_1\nabla\theta_1) - \frac{1}{2}(\nabla\xi_1)^2 + \varkappa_q f_1^2 + \left(\frac{1}{3}P_1 + \frac{Ma_0}{2}f_1\right)\cdot\nabla f_1 + \nabla\cdot\left(\frac{\xi_1}{2}\nabla P_1 + \frac{Ma_0}{2}\xi_1\nabla f_1\right), \quad (14)$$

where $P_{1,2} = Ga\xi_{1,2} - C\Delta\xi_{1,2}$, $f_{1,2} = \theta_{1,2} - \xi_{1,2}$. The solution can be chosen in the form

$$\xi_2 = a_{10}\left(A_1B_2^* + A_2B_1^*\right)\mathrm{e}^{\mathrm{i}\psi_+} + a_{1-0}\left(A_1B_1^* + A_2B_1^*\right)\mathrm{e}^{\mathrm{i}\psi_-}$$
$$+\left[a_{11}\left(A_1B_1\mathrm{e}^{\mathrm{i}\psi_+} + A_2B_2\mathrm{e}^{-\mathrm{i}\psi_+}\right) + a_{1-1}\left(A_1B_2\mathrm{e}^{\mathrm{i}\psi_-} + A_2B_1\mathrm{e}^{-\mathrm{i}\psi_-}\right)\right.$$
$$\left. + a_{22}\left(A_1^2\mathrm{e}^{2\mathrm{i}kX} + A_2^2\mathrm{e}^{-2\mathrm{i}kX} + B_1^2\mathrm{e}^{2\mathrm{i}\vec{k}_2\cdot\vec{r}} + B_2^2\mathrm{e}^{-2\mathrm{i}\vec{k}_2\cdot\vec{r}}\right)\right]\mathrm{e}^{2\mathrm{i}\omega\tau_0}$$
$$+a_{20}\left(A_1A_2^*\mathrm{e}^{2\mathrm{i}kX} + B_1B_2^*\mathrm{e}^{2\mathrm{i}\vec{k}_2\cdot\vec{r}}\right) + c.c. \qquad (15)$$

$$\theta_2 = b_{20}\left(A_1A_2^*\mathrm{e}^{2\mathrm{i}kX} + B_1B_2^*\mathrm{e}^{2\mathrm{i}\vec{k}_2\cdot\vec{r}}\right) + b_{02}\left(A_1A_2 + B_1B_2\right)\mathrm{e}^{2\mathrm{i}\omega\tau_0}$$
$$+b_{10}\left(A_1B_2^* + A_2B_1^*\right)\mathrm{e}^{\mathrm{i}\psi_+} + B_{1-0}\left(A_1B_1^* + A_2B_1^*\right)\mathrm{e}^{\mathrm{i}\psi_-}$$
$$+\left[b_{11}\left(A_1B_1\mathrm{e}^{\mathrm{i}\psi_+} + A_2B_2\mathrm{e}^{-\mathrm{i}\psi_+}\right) + B_{1-1}\left(A_1B_2\mathrm{e}^{\mathrm{i}\psi_-} + A_2B_1\mathrm{e}^{-\mathrm{i}\psi_-}\right)\right.$$
$$\left. + b_{22}\left(A_1^2\mathrm{e}^{2\mathrm{i}kX} + A_2^2\mathrm{e}^{-2\mathrm{i}kX} + B_1^2\mathrm{e}^{2\mathrm{i}\vec{k}_2\cdot\vec{r}} + B_2^2\mathrm{e}^{-2\mathrm{i}\vec{k}_2\cdot\vec{r}}\right)\right]\mathrm{e}^{2\mathrm{i}\omega\tau_0}$$
$$+b_{00}\left(|A_1|^2 + |A_2|^2 + |B_1|^2 + |B_2|^2\right) + c.c., \qquad (16)$$

where $\psi_+ = kX + \vec{k}_2\cdot\vec{r}$, $\psi_- = kX - \vec{k}_2\cdot\vec{r}$. Hereafter the asterisk denotes the complex-conjugate term; b_{00}, b_{02}, a_{10}, b_{10},..., b_{1-1} are constants, which are very cumbersome and therefore they are not given here.

At the third order in δ, we obtain

$$\frac{\partial\xi_3}{\partial\tau_0} - \Delta\left(\frac{1}{3}P_3 + \frac{Ma_0}{2}f_3\right) = F^{(1)}, \qquad (17)$$

$$\frac{\partial\theta_3}{\partial\tau_0} - \Delta\left(\theta_3 + \frac{1}{8}P_3 + \frac{Ma_0}{6}f_3\right) + \beta f_3 = F^{(2)}, \qquad (18)$$

where $P_3 = Ga\xi_3 - C\Delta\xi_3$, $f_3 = \theta_3 - \xi_3$; inhomogeneities $F^{(1,2)}$ are defined as

$$F^{(1)} = -\frac{\partial\xi_1}{\partial\tau_2} + \frac{1}{2}Ma_2\Delta f_1 + \nabla\cdot\left(Ma_0\xi_1\nabla f_2 + \xi_1\nabla P_2\right)$$
$$+\nabla\cdot\left[\xi_1^2\left(\nabla P_1 + \frac{Ma_0}{2}\nabla f_1\right) + \xi_2\left(\nabla P_2 + Ma_0\nabla f_1\right)\right], \qquad (19)$$

$$F^{(2)} = -\frac{\partial\theta_1}{\partial\tau_2} - \xi_2\frac{\partial\theta_1}{\partial\tau_0} - \xi_1\frac{\partial\theta_2}{\partial\tau_0} + 2\varkappa_q f_1 f_2 + \frac{1}{6}Ma_2\Delta f_1$$
$$-\nabla\xi_1\cdot\nabla\xi_2 + \nabla\cdot\left(\xi_1\nabla\theta_2 + \xi_2\nabla\theta_1\right) + \frac{1}{3}\nabla P_2\cdot\nabla f_1$$
$$+\nabla P_1\cdot\left(\xi_1\nabla f_1 + \frac{1}{3}\nabla f_2\right) + \frac{3}{4}\nabla\cdot\left(\xi_1^2\nabla P_1\right) + \frac{1}{2}\nabla\cdot\left(\xi_1\nabla P_2 + \xi_2\nabla P_1\right)$$
$$+Ma_0\left[\xi_1\nabla f_1{}^2 + \nabla f_1\cdot\nabla f_2 + \frac{1}{2}\nabla\cdot\left[\left(\xi_1^2 + \xi_2\right)\nabla f_1\right] + \frac{1}{2}\nabla\cdot\left(\xi_1\nabla f_2\right)\right]. \qquad (20)$$

The solvability condition at the third order can be formulated as

$$\left(\mathrm{i}\omega+\frac{Ma_0k^2}{6}+k^2+\beta\right)F_{sec}^{(1)}=\frac{Ma_0k^2}{2}F_{sec}^{(2)}\,, \tag{21}$$

where $F_{sec}^{(1,2)}$ are secular parts of inhomogeneities. It yields a set of four complex differential equations that govern the evolution of wave amplitudes $A_{1,2}$ and $B_{1,2}$

$$\begin{aligned}
\frac{dA_1}{d\tau_2}&=\left(\gamma-K_0|A_1|^2-K_1|A_2|^2-K_2(\phi)|B_1|^2-K_2(\pi-\phi)|B_2|^2\right)A_1-K_3(\phi)A_2{}^*B_1B_2\\
\frac{dA_2}{d\tau_2}&=\left(\gamma-K_0|A_2|^2-K_1|A_1|^2-K_2(\phi)|B_2|^2-K_2(\pi-\phi)|B_1|^2\right)A_2-K_3(\phi)A_1{}^*B_1B_2\\
\frac{dB_1}{d\tau_2}&=\left(\gamma-K_0|B_1|^2-K_1|B_2|^2-K_2(\phi)|A_1|^2-K_2(\pi-\phi)|A_2|^2\right)B_1-K_3(\phi)B_2{}^*A_1A_2\\
\frac{dB_2}{d\tau_2}&=\left(\gamma-K_0|B_2|^2-K_1|B_1|^2-K_2(\phi)|A_2|^2-K_2(\pi-\phi)|A_1|^2\right)B_2-K_3(\phi)B_1{}^*A_1A_2
\end{aligned} \tag{22}$$

Here

$$\gamma=\frac{k^2Ma_2}{2}\left(1-\mathrm{i}\frac{3k^2\left(Ga+Ck^2+72\right)}{2\omega_0}\right)\,,$$

expressions for Landau coefficients K_0, K_1, $K_2(\phi)$ and $K_3(\phi)$ are very cumbersome and therefore they are not given here.

Equations (22) were studied in detail by [6] in the case of square symmetry, i.e. for $\phi=\pi/2$. They found six types of solutions.

(i)	Traveling rolls (TR)	$\lvert A_1\rvert^2=\gamma_r/K_{0r},\ A_2=B_1=B_2=0.$
(ii)	Standing rolls (SR)	$A_1=A_2,\ \lvert A_1\rvert^2=\gamma_r/\left(K_{0r}+K_{1r}\right),\ B_1=B_2=0.$
(iii)	Traveling squares (TS)	$A_1=B_1,\ \lvert A_1\rvert^2=\gamma_r/\left(K_{0r}+K_{2r}\right),\ A_2=B_2=0.$
(iv)	Standing squares (SSq)	$A_1=A_2=B_1=B_2,$ $\lvert A_1\rvert^2=\gamma_r/\left(K_{0r}+K_{1r}+2K_{2r}+K_{3r}\right).$
(v)	Alternating rolls (AR)	$A_1=A_2=\mathrm{i}B_1=\mathrm{i}B_2,$ $\lvert A_1\rvert^2=\gamma_r/\left(K_{0r}+K_{1r}+2K_{2r}-K_{3r}\right).$
(vi)	Standing cross-rolls (SCR)	$A_1=A_2,\ B_1=B_2,\ \lvert A_1\rvert\neq\lvert B_1\rvert.$

For any parameters, we use notation $K_r=\mathrm{Re}K$, $K_i=\mathrm{Im}K$.

A stability analysis for the patterns on the square lattice shows that they are selected if they emerge through the direct Hopf bifurcation ($\gamma_r>0$). The remaining stability conditions also obtained by [6], are as follows

(TR): $K_{0r} < K_{1r},\ K_{0r} < K_{2r}$

(SR): $K_{0r} > K_{1r},\ K_{0r} + K_{1r} - 2K_{2r} < 0,\ |K_0 + K_1 - 2K_2|^2 > |K_3|^2$.

(TS): $K_{0r} > K_{2r},\ K_{0r} - K_{1r} - K_{3r} < 0,\ K_{0r} - K_{1r} + K_{3r} < 0$.

(SSq): $K_{0r} + K_{1r} - 2K_{2r} - 3K_{3r} > 0,\ K_{0r} - K_{1r} - K_{3r} > 0,$
$[K_3^*(K_0 + K_1 - 2K_2)]_r < |K_3|^2$.

(AR): $K_{0r} + K_{1r} - 2K_{2r} + 3K_{3r} > 0,\ K_{0r} - K_{1r} + K_{3r} > 0,$
$-[K_3^*(K_0 + K_1 - 2K_2)]_r < |K_3|^2$.

(SCR) is always unstable.

These conditions provide boundaries of selection between two stable patterns. Obviously, equation $K_0 = K_{2r}$ defines boundary between stable TR and TS; equation $K_{0r} = K_{1r}$ – between stable TR and SR. Equations $K_{0r} - K_{1r} = K_{3r}$ and $K_{0r} - K_{1r} = -K_{3r}$ define selection between stable TS and SSq or AR, respectively.

Below we apply the general results described above to the particular problem, which is the subject of the present paper. Our goal is the computation of coefficients K_{0r}, K_{1r}, $K_{2r}(\pi/2)$ and $K_{3r}(\pi/2)$ as functions of the problem parameters, which are β, Ga and $\varkappa_q$.

For uncontrolled convection, pattern selection was investigated previously in the case $\phi = \pi/2$ [5]. It was shown that a small area of stable alternating rolls was discovered (see Fig.1 (a)). However, this area intersects with the domain of subcritical traveling rolls, so here depending on the initial condition the system either approaches AR or demonstrates the infinite growth of one of the amplitudes. Note, that the boundary of stability for alternating rolls here is defined by condition $K_{0r} - K_{1r} + K_{3r} < 0$ corresponding to the boundary between AR and TS. Thus, alternating rolls first become unstable against traveling squares, that in turn become unstable against traveling rolls.

3.3 Nonlinear Feedback Control

Quadratic control gain varies Landau coefficients, resulting in a change of stability boundaries for the patterns.

Influence of the quadratic feedback control on pattern selection for $\phi = \pi/2$ is presented in Fig.1 (b). Recall that the oscillatory instability is critical only inside the domain bounded by the dashed line in Fig.1.

Positive control gain reduces the domain of stability for traveling wave, whereas the domain of subcriticality for traveling wave is enlarged. Additional domain of subcriticality arises due to the standing squares. Stable standing squares emerge for $\varkappa_q = 0.1$ instead of alternating rolls in the uncontrolled case. However, the domain of stable standing squares intersects with the domain of subcritical traveling rolls, so here depending on the initial condition the system either approaches SSq or

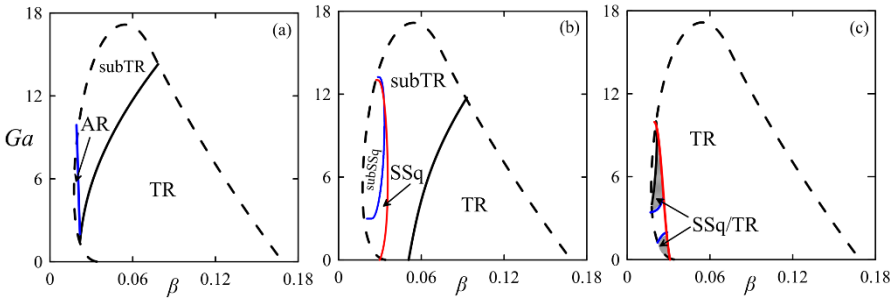


Fig. 1 Pattern selection for $\phi = \pi/2$ in the case of uncontrolled convection **(a)**,for $\varkappa_q = 0.1$ **(b)** and for $\varkappa_q = -0.1$. Domains of stability for traveling rolls, standing squares and alternating rolls are marked by "TR", "SSq" and "AR", respectively. The domains of subcriticality for traveling rolls and standing squares are marked by "subTR" and "subSSq", respectively. Domains of bistability of traveling rolls and standing squares is shaded and marked by "SSq/TR".

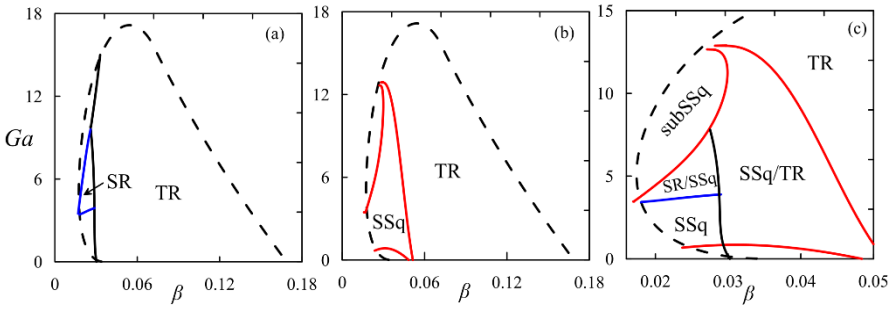


Fig. 2 Pattern selection for $\phi = \pi/2$, $\varkappa_q = -0.2$. Domains of stability for traveling rolls, standing rolls and standing squares are marked "TR", "SR" and "SSq", respectively. Panel **(c)** shows zoomed-in domains of bistability (marked "SR/SSq" and SSq/TR). "subSSq" marks domain of subcriticality for standing squares.

demonstrates the infinite growth of one of the amplitudes. Note, that the boundary of stability for standing squares here is defined by condition $K_{0r} - K_{1r} - K_{3r} < 0$ corresponding to the boundary between SSq and TS. Thus, standing squares first become unstable against traveling squares, that in turn become unstable against traveling rolls.

For negative control gain traveling rolls are stable within the whole domain, where the oscillatory mode is critical, see Fig.1 (c). However, there is a domain of subcriticality for standing squares. Moreover, there are two small areas of stable standing squares, which intersect the domain of stable traveling rolls, resulting in the bistability.

Pattern selection for $\phi = \pi/2$ under the control gain $\varkappa_q = -0.2$ is presented in Fig. 2. Small areas of subcritical traveling rolls, standing rolls and standing squares exist for a small values of β.

Traveling rolls are stable in most of the domain, where the oscillatory mode is critical. But there are also domains of stability for standing rolls and standing squares, see Figs.2(a) and (b), respectively. Note that domains of stability for TR and SSq, SR and SSq intersect partially, resulting in bistability.

4 Conclusions and Discussion

We have studied pattern formation of oscillatory Marangoni instability in a thin film under nonlinear feedback control.

We have performed a weakly nonlinear analysis within the amplitude equations, which describe coupled evolution of the thickness and temperature of thin film in the presence of the nonlinear control. Our analysis is based on the consideration of the nonlinear interaction of a pair of standing waves propagating at the angle ϕ between the wave vectors. That consideration leads to a set of four complex Landau equations that govern the evolution of wave amplitudes. The coefficients of Landau equations, which define pattern formation, have been calculated in the case $\phi = \pi/2$ for different values of the control gain, Galileo and Biot numbers. We have demonstrated, that besides conventional traveling rolls an additional stable patterns (such as standing rolls and standing squares) emerges under nonlinear feedback control. In the case of negative control gain, we have shown that a quadratic control can eliminate the subcritical excitation of instability within entire domain, where oscillatory mode is critical.

Acknowledgements This research was supported by the Israel Science Foundation (grant No. 843/18).
The authors are grateful to the organisers of the programme "Conservation laws, interfaces and mixing" at MATRIX (Creswick, Victoria, November 4-8 2019) for the invitation to present the results of their research.

References

1. Nayfeh, A.H.: Introduction in Perturbation Techniques. Wiley-VCH, New York (1993)
2. Samoilova, A.E., Nepomnyashchy, A.: Feedback control of Marangoni convection in a thin film heated from below. J. Fluid Mech. (2019) doi:10.1017/jfm.2019.578
3. Samoilova, A.E., Nepomnyashchy, A.: Nonlinear feedback control of Marangoni wave patterns in a thin film heated from below. Phys. Rev. E (2020) doi:10.1016/j.physd.2020.132627
4. Shklyaev, S., Khenner, M., Alabuzhev, A. A.: Long-wave Marangoni convection in a thin film heated from below. Phys. Rev. E (2012) doi:10.1103/PhysRevE.85.016328

5. Silber, M., Knobloch, E.: Hopf bifurcation on a square lattice. Nonlinearity (1991) doi:10.1088/0951-7715/4/4/003

Rigorous modelling of nonlocal interactions determines a macroscale advection-diffusion PDE

Prof A.J. Roberts

Abstract A slowly-varying or thin-layer multiscale assumption empowers macroscale understanding of many physical scenarios from dispersion in pipes and rivers, including beams, shells, and the modulation of nonlinear waves, to homogenisation of micro-structures. Here we begin a new exploration of the scenario where the given physics has non-local microscale interactions. We rigorously analyse the dynamics of a basic example of shear dispersion. Near each cross-section, the dynamics is expressed in the local moments of the microscale non-local effects. Centre manifold theory then supports the local modelling of the system's dynamics with coupling to neighbouring cross-sections as a non-autonomous forcing. The union over all cross-sections then provides powerful new support for the existence and emergence of a macroscale model advection-diffusion PDE global in the large, finite-sized, domain. The approach quantifies the accuracy of macroscale advection-diffusion approximations, and has the potential to open previously intractable multiscale issues to new insights.

1 Introduction

This paper introduces a new rigorous approach to the multiscale challenge of systematically modelling by macroscale PDEs the dynamics of microscale, *spatially nonlocal*, systems. This approach provides a novel quantified error formula. Previous research using this type of approach rigorously modelled systems that were expressed as PDEs on the microscale. This previous research encompassed both cylindrical multiscale domains (Roberts 2015a) and more general multiscale domains (Roberts and Bunder 2017; Bunder and Roberts 2018). But recall that PDEs are themselves mathematical idealisations of physical processes that typically take place on mi-

Prof A.J. Roberts
School of Mathematical Sciences, University of Adelaide, http://orcid.org/0000-0001-8930-1552, e-mail: anthony.roberts@adelaide.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_25

croscale length scales. Hence, here we begin to address the challenges arising when the given mathematical model of a system encodes microscale physical interactions over finite microscale lengths.

Physical systems with nonlocal, microscale, spatial interactions arise in many applications. In neuroscience, a spatial convolution expresses the excitatory/inhibitory effects of a neurone on a nearby neurone, giving rise to nonlocal neural field equations, and "have been quite successful in explaining various experimental findings" (Ermentrout 2015, e.g.). Models of free crack propagation in brittle materials invoke microscale *nonlocal* stress-strain laws, called peridynamics (Silling 2000, e.g.): one challenge is to derive the effective mesoscale PDEs from the nonlocal laws (Silling and Lehoucq 2008; Lipton 2014, e.g.). Nonlocal dispersal and competition models arise in biology (Omelyan and Kozitsky 2018; Duncan et al. 2017, e.g.). Other examples are non-local cell adhesion models (Buttenschön and Hillen 2020, e.g.). In this introduction we begin by exploring the specific example of a so-called 'Zappa' dispersion in a channel (Section 2) in which material is transported by finite jumps along the channel, and also is intermittently thoroughly mixed across the channel.

General scenario

Zappa dispersion is a particular case of the following general scenario—a scenario that is the subject of ongoing research. In generality we consider a field $u(x,y,t)$, on a 'cylindrical' spatial domain $\mathbb{X}\times\mathbb{Y}$ (where $\mathbb{X}\subseteq\mathbb{R}$ and where $\mathbb{Y}$ denotes the cross-section). We suppose the field u is governed by a given autonomous system in the form

$$\frac{\partial u}{\partial t} = \int_{\mathbb{Y}}\int_{\mathbb{X}} k(x,\xi,y,\eta)u(\xi,\eta,t)\,d\xi\,d\eta\,, \tag{1}$$

where the given kernel $k(x,\xi,y,\eta)$ expresses both nonlocal and local physical effects at position (x,y) from the field at position (ξ,η), both within the cylindrical domain $\mathbb{X}\times\mathbb{Y}$. We allow the kernel to be a generalised function so that local derivatives may be represented by derivatives of the Dirac delta function δ: for example, a component $\delta'(x-\xi)\delta(y-\eta)$ in the kernel k encodes the differential term $-\partial u/\partial x$ in the right-hand side of (1). In general the physical effects encoded in the kernel k may be heterogeneous in space. But, as is common and apart from boundaries, Zappa dispersion is homogeneous in space (translationally invariant) in which case some significant simplifications ensue.

The nonlocal system (1) is linear for simplicity, but we invoke the framework of centre manifold theory so the approach should, with future development, apply to nonlinear generalisations as in previous work on such modelling where the system is expressed as PDEs on the microscale (Roberts 2015a).

Our aim is to rigorously establish that the emergent dynamics of the nonlocal system (1) are captured over the 1D spatial domain $\mathbb{X}$ by a mean/averaged/coarse/macroscale variable $U(x,t)$ that satisfies a macroscale, second-order, advection-diffusion PDE of the form

$$\frac{\partial U}{\partial t} \approx A_1 \frac{\partial U}{\partial x} + A_2 \frac{\partial^2 U}{\partial x^2}, \quad x \in \mathbb{X}, \tag{2}$$

for some derived coefficients A_1 and A_2.[1] This macroscale PDE (2) is to model the dynamics of the microscale nonlocal (1) after transients have decayed exponentially quickly in time, and to the novel quantified error (6d).

2 Zappa shear dispersion

This section introduces a basic example system (non-dimensional) of nonlocal microscale jumps by a particle (inspired by W. R. Young, private communication). Section 3 systematically derives an advection-diffusion PDE (2) for the particle that is valid over macroscale space-time. Consider a particle in a channel $-1 < y < 1$, $\mathbb{Y} = (-1, 1)$, and of notionally infinite extent in x, $\mathbb{X} = \mathbb{R}$. Let $u(x, y, t)$ be the probability density function (PDF) for the particle's location: equivalently, view $u(x, y, t)$ as the concentration of some continuum material.

The 'Zappa' dynamics of the particle's PDF is encoded by

$$\frac{\partial u}{\partial t} = \Big[\tfrac{1}{v(y)} \underbrace{\int_{-\infty}^{x} e^{-(x-\xi)/v(y)} u(\xi, y, t)\, d\xi}_{= e^{-x/v(y)} \star u, \text{ the convolution (5)}} - u\Big] + \Big[\tfrac{1}{2}\int_{-1}^{1} u\, dy - u\Big] \tag{3}$$

for some jump profile $v(y) > 0$—$v(y)$ is an effective velocity along the channel. That is, the kernel of the Zappa system is the generalised function

$$\begin{aligned} k(x, \xi, y, \eta) = &\Big[\tfrac{1}{v(y)} e^{-(x-\xi)/v(y)} H(x-\xi) - \delta(x-\xi)\Big]\, \delta(y-\eta) \\ &+ \big[\tfrac{1}{2} - \delta(y-\eta)\big]\, \delta(x-\xi), \end{aligned} \tag{4}$$

where $H(x)$ is the unit step function. The nonlocal equation (3) governs the PDF of the particle in Zappa dispersion through the following two physical mechanisms.

- We suppose that, at exponentially distributed time intervals with mean one, the particle gets 'zapped' across the channel (by a burst of intermittent turbulence for example) and lands at any cross channel position y with uniform distribution. Consequently the Fokker–Planck PDE (3) for the PDF contains the terms $u_t = \big[\frac{1}{2}\int_{-1}^{1} u\, dy - u\big] + \cdots$.
- Further, suppose that, at exponentially distributed time intervals with mean one, the particle jumps in x a distance to the right, a distance which is exponentially distributed with some given mean $v(y)$. Consequently the Fokker–Planck PDE (3) for the PDF contains the terms $u_t = \big[\frac{1}{v(y)} e^{-x/v(y)} \star u - u\big] + \cdots$, in terms of the upstream convolution

[1] Ongoing research aims to generalise the approach here to certify the accuracy of PDEs truncated to Nth-order for every N.

$$e^{-x/v(y)} \star u = \int_{-\infty}^{x} e^{-(x-\xi)/v(y)} u(\xi,y,t)\,d\xi\,. \tag{5}$$

We derive the macroscale model that the cross-sectional mean field $U(x,t)$ evolves according to an advection-diffusion PDE: $U_t \approx A_1 U_x + A_2 U_{xx}$. The field $U(x,t)$ may be viewed as the marginal probability density of the particle being at x, averaged over the cross-section y. Innovatively, we put the macroscale modelling on a rigorous basis that additionally quantifies the error.

In particular, say we choose $v(y) := 1-y^2$ then computer algebra (Section 6) readily derives that over large space-time scales, and after transients decay roughly like e^{-t}, from every initial condition the Zappa system (3) has the quasistationary distribution (Pollett and Roberts 1990, e.g.)

$$u(x,y,t) \approx U + (y^2 - \tfrac{1}{3})\frac{\partial U}{\partial x} + (2y^4 - \tfrac{8}{3}y^2 + \tfrac{22}{45})\frac{\partial^2 U}{\partial x^2}\,, \tag{6a}$$

$$\text{such that} \qquad \frac{\partial U}{\partial t} = -\frac{2}{3}\frac{\partial U}{\partial x} + \frac{28}{45}\frac{\partial^2 U}{\partial x^2} + \rho\,, \tag{6b}$$

in terms of a macroscale variable here chosen to be the cross-sectional mean,

$$U(x,t) := \frac{1}{2}\int_{-1}^{1} u(x,y,t)\,dy\,. \tag{6c}$$

The macroscale PDE (6b) is a precise equality because we include the error terms in our analysis to find a precise, albeit complicated, expression for the final error ρ. The remainder error ρ in (6b) has the form

$$\begin{aligned}\rho :=\ & r_0 + \langle Z_0, \mathscr{W}_{0:}\mathscr{B}e^{\mathscr{B}t}\star\mathbf{r}'\rangle + \langle Z_0, \mathscr{W}_{0:}\mathbf{r}'\rangle \\ & - A_1\langle Z_0, \mathscr{W}_{1:}e^{\mathscr{B}t}\star\mathbf{r}'\rangle - A_2\langle Z_0, \mathscr{W}_{2:}e^{\mathscr{B}t}\star\mathbf{r}'\rangle\end{aligned} \tag{6d}$$

where here the convolutions are over time, $f(t)\star g(t) = \int_0^t f(t-s)g(s)\,ds$, and other symbols are introduced in the next Section 3. We anticipate this error ρ is

- 'small' in regions of slow variations in space, small gradients, and
- 'large' in regions of relatively large gradients such as spatial boundary layers.

Then, simply, the macroscale PDE model (6b) is valid whenever and wherever the error ρ is small enough for the application purposes at hand. The next section includes deriving this error term and clarifies the notation.

3 Many kernels generate local models

Inspired by earlier research (Roberts 2015a, Proposition 1), this section's aim is to rigorously derive and justify the model (6) that governs the emergent macroscale evolution of Zappa dispersion. The algebra starts to 'explode'—Section 4 discusses

how to compactly do the algebra in physically meaningful forms, and connect to other mathematical methodologies.

To derive the advection-diffusion model (6b) we truncate the analysis to second order quadratic terms. Higher-orders appear to be similar in nature, but much more involved algebraically, and are left for later development.

3.1 Rewrite the equations for local dynamics

Let's analyse the dynamics in the spatial locale about a generic longitudinal cross-section $X \in \mathbb{X}$. Then invoke Lagrange's Remainder Theorem—which empowers us to track errors—to expand the PDF as

$$u(x,y,t) = u_0(X,y,t) + u_1(X,y,t)(x-X) + u_2(X,x,y,t)\frac{(x-X)^2}{2!}, \tag{7}$$

where $u_0 := u$ and $u_1 := \partial u/\partial x$ both evaluated at the cross-section $x = X$, and where $u_2 := \partial^2 u/\partial x^2$ evaluated at some point $x = \hat{x}(X,x,y,t)$ which is some definite (but usually unknown) function of cross-section X, longitudinal position x, cross-section position y, and time t. By the Lagrange Remainder Theorem, the location $\hat{x}$ satisfies $X \lesseqgtr \hat{x} \lesseqgtr x$. The function $\hat{x}$ is implicit in our analysis because it is hidden in the dependency upon x of the second derivative $u_2(X,x,y,t)$.

Substitute (7) into the Zappa nonlocal equation (3) to obtain

$$\begin{aligned}
&\frac{\partial u_0}{\partial t} + \frac{\partial u_1}{\partial t}(x-X) + \frac{\partial u_2}{\partial t}\frac{(x-X)^2}{2!} \\
&= \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} k(x,\xi,y,\eta)\, d\xi \right] u_0(X,\eta,t)\, d\eta \\
&\quad + \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} k(x,\xi,y,\eta)(\xi - X)\, d\xi \right] u_1(X,\eta,t)\, d\eta \\
&\quad + \int_{\mathbb{Y}} \int_{\mathbb{X}} k(x,\xi,y,\eta)\frac{(\xi-X)^2}{2!} u_2(X,\hat{\xi},\eta,t)\, d\xi\, d\eta\,.
\end{aligned} \tag{8}$$

The effect at cross-section x of the nth moment of the kernel at cross-section X is summarised in the integrals $\int_{\mathbb{X}} k(x,\xi,y,\eta)\frac{(\xi-X)^n}{n!}\, d\xi$. So define the local nth moment of the kernel to be, for every $n \geq 0$,

$$\begin{aligned}
k_n(X,y,\eta) &:= \int_{\mathbb{X}} k(X,\xi,y,\eta)\frac{(\xi-X)^n}{n!}\, d\xi \\
&= \left[(-v)^n - \delta_{n0}\right]\delta(y-\eta) + \left[\tfrac{1}{2} - \delta(y-\eta)\right]\delta_{0n}
\end{aligned} \tag{9}$$

upon substituting the Zappa kernel (4). This Zappa problem is homogeneous in x, as are many problems, and so the kernel moments k_n are independent of the cross-section X (except near the boundary inlet and outlet).

The last integral term in the local expansion (8) requires special consideration: apply Lagrange's Remainder Theorem to write $u_2(X,\xi,\eta,t) = u_2(X,X,\eta,t) + (\xi - X)u_{2x}(X,\hat{\xi},\eta,t)$ for some uncertain function $\hat{\xi}(X,\xi,\eta,t)$ that satisfies $X \leqslant \hat{\xi} \leqslant \xi$ for every η,t, and where $u_{2x} := \partial/\partial x[u_2(X,x,\eta,t)]$. Then the last term distributes into two:

$$\begin{aligned}
&\int_{\mathbb{Y}}\int_{\mathbb{X}} k(x,\xi,y,\eta)\frac{(\xi-X)^2}{2!}u_2(X,\hat{\xi},\eta,t)\,d\xi\,d\eta \\
&= \int_{\mathbb{Y}} \underbrace{\int_{\mathbb{X}} k(x,\xi,y,\eta)\frac{(\xi-X)^2}{2!}\,d\xi}_{k_2(X,y,\eta)} u_2(X,X,\eta,t)\,d\eta \\
&\quad + \underbrace{\int_{\mathbb{Y}}\int_{\mathbb{X}} k(x,\xi,y,\eta)3\frac{(\xi-X)^3}{3!}u_{2x}(X,\hat{\xi},\eta,t)\,d\xi\,d\eta}_{\text{a remainder, with a third } x \text{ derivative in } u_{2x}} .
\end{aligned}$$

Define $u_2(X,y,\eta) := u_2(X,X,y,\eta)$ for notational consistency with lower moments—see the definition (9).

The local equation (8) is exact everywhere, but is most useful in the vicinity of the cross-section X, that is, for small $(x-X)$. Notionally we want to 'equate coefficients' of powers of $(x-X)$ in (8), but to be precise we must carefully evaluate $\lim_{x\to X}$ of various x-derivatives of (8). For example, let $x \to X$ in (8), then

$$\begin{aligned}
\frac{\partial u_0}{\partial t} = &\int_{\mathbb{Y}} k_0(X,y,\eta)u_0(X,\eta,t)\,d\eta + \int_{\mathbb{Y}} k_1(X,y,\eta)u_1(X,\eta,t)\,d\eta \\
&+ \int_{\mathbb{Y}} k_2(X,y,\eta)u_2(X,\eta,t)\,d\eta \\
&+ 3\int_{\mathbb{Y}}\int_{\mathbb{X}} k(X,\xi,y,\eta)\frac{(\xi-X)^3}{3!}u_{2x}(X,\hat{\xi},\eta,t)\,d\xi\,d\eta .
\end{aligned}$$

Rewrite this conveniently and compactly as the integro-differential equation (IDE)

$$\frac{\partial u_0}{\partial t} = \mathfrak{L}_0 u_0 + \mathfrak{L}_1 u_1 + \mathfrak{L}_2 u_2 + r_0 , \tag{10}$$

for y-operators defined to be, from the moments (9),

$$\mathfrak{L}_n u := \int_{\mathbb{Y}} k_n(X,y,\eta)u|_{y=\eta}\,d\eta = \begin{cases} \frac{1}{2}\int_{-1}^{1} u\,dy - u , & n=0 , \\ [-v(y)]^n u , & n=1,2,\ldots . \end{cases} \tag{11}$$

The IDE (10) also has the remainder r_0 which couples the local dynamics to neighbouring locales via u_{2x} and is the $n=0$ case of

$$r_n(X,y,t) := 3\int_{\mathbb{Y}}\int_{\mathbb{X}} \frac{\partial^n k}{\partial x^n}\Big|_{x=X}\frac{(\xi-X)^3}{3!}u_{2x}(X,\hat{\xi},\eta,t)\,d\xi\,d\eta . \tag{12}$$

Now we can see how this approach to modelling the spatial dynamics works: given that the y-operators (11) are evaluated at X, the spatially local power series with remainder, in IDEs like (10), 'pushes' the coupling with neighbouring locales to a higher-order derivative term in r_0, here third-order via the u_{2x} factor. Hence the local dynamics in u_0, u_1, u_2 are essentially isolated from all other cross-sections whenever and wherever the coupling r_0 is small enough for the purposes at hand—here when third derivatives are small—that is, when the solutions are, in space, slowly varying enough.

The previous paragraph obtains the IDE for u_0 by simply taking the limit of (8) as $x \to X$. We straightforwardly and similarly obtain IDEs for u_1 and u_2 by finding the limits of spatial derivatives of (8):

$$\lim_{x\to X} \frac{\partial(8)}{\partial x} \implies \frac{\partial u_1}{\partial t} = \mathfrak{L}_0 u_1 + \mathfrak{L}_1 u_2 + r_1\,; \tag{13a}$$

$$\lim_{x\to X} \frac{\partial^2(8)}{\partial x^2} \implies \frac{\partial u_2}{\partial t} = \mathfrak{L}_0 u_2 + r_2\,; \tag{13b}$$

for local coupling remainders r_1 and r_2 defined by (12).

3.2 Local-to-global system modelling theory

This section considers the collection of 'local' systems as one 'global' (in space X) system. Then theory establishes that the advection-diffusion PDE (6b) arises as a globally valid, macroscale, model PDE.

Denote the vector of coefficients $\mathbf{u}(X,y,t) := (u_0, u_1, u_2)$, and similarly for the local coupling remainder $\mathbf{r}(X,y,t) := (r_0, r_1, r_2)$. Then write the IDEs (10) and (13), in the form of the 'forced' linear system

$$\frac{d\mathbf{u}}{dt} = \underbrace{\begin{bmatrix} \mathfrak{L}_0 & \mathfrak{L}_1 & \mathfrak{L}_2 \\ 0 & \mathfrak{L}_0 & \mathfrak{L}_1 \\ 0 & 0 & \mathfrak{L}_0 \end{bmatrix}}_{\mathscr{L}} \mathbf{u} + \mathbf{r}(X,t). \tag{14}$$

for upper triangular matrix/operator $\mathscr{L}$. The system (14) might appear closed, but it is coupled via the derivative u_{2x}, through the 'forcing' remainders $\mathbf{r}$, to the dynamics of cross-sections that neighbour X.

At each locale $X \in \mathbb{X}$, treat the remainder coupling $\mathbf{r}$ (third-order) as a perturbation (and if this was a nonlinear problem, then the nonlinearity would also be part of the perturbation). Thus to a useful approximation the global system satisfies the local linear ODEs $d\mathbf{u}/dt \approx \mathscr{L}\mathbf{u}$ for each $X \in \mathbb{X}$. Hence, the linear operator $\mathscr{L}$ is crucial to modelling the dynamics: all solutions are characterised by the eigenvalues of $\mathscr{L}$. Since $\mathscr{L}$ is block triangular, a structure exploited previously (Roberts 2015a,

§2), its spectrum is thrice that of $\mathfrak{L}_0 = \frac{1}{2}\int_{-1}^{1} u\,dy - u$ (definition (11)). Here it is straightforward to verify that the y-operator $\mathfrak{L}_0$ has:

- one 0 eigenvalue corresponding to eigenfunctions constant across the channel; and
- an 'infinity' of eigenvalue -1 corresponding to all functions with zero average across the channel.

Then globally in space, with $d\mathbf{u}/dt = \mathscr{L}\mathbf{u} + (\text{perturbation})$ at every $X \in \mathbb{X}$, and because of the 'infinity' of the continuum $\mathbb{X}$, the linearised system has a 'thrice-infinity' of the 0 eigenvalue, and a 'double-infinity' of eigenvalue -1. Consequently, the theory of Aulbach and Wanner (2000) asserts:

1. there exists a '(3∞)'-D slow manifold—the quasistationary (6a);
2. which is exponentially quickly attractive to all initial conditions, with transients roughly e^{-t}—it is emergent; and
3. which we approximate by approximately solving the governing differential equations (14)—done in encoded form by Section 6.

We obtain a useful approximation to the global slow manifold by neglecting the 'perturbing' remainder $\mathbf{r}$. Because the remainder $\mathbf{r}$ is the only coupling between different locales X this approximation may be constructed independently at each and every cross-section X. Further, because the Zappa system is homogeneous in space, the construction is identical at each and every $X \in \mathbb{X}$. These two properties vastly simplify the construction of the attractive slow manifold.

Neglecting the coupling remainder $\mathbf{r}$ gives the linear problem $d\mathbf{u}/dt = \mathscr{L}\mathbf{u}$. The approximate slow manifold is thus the zero eigenspace of $\mathscr{L}$. We find the zero eigenspace via (generalised) eigenvectors. With the (generalised) eigenvectors in the three columns of block-matrix $\mathscr{V}$, in essence we seek $\mathbf{u}(t) = \mathscr{V}\mathbf{U}(t)$ such that $d\mathbf{U}/dt = \mathscr{A}\mathbf{U}$ for 3×3 matrix $\mathscr{A}$ having all the zero eigenvalues. To be an eigenspace we need to solve $\mathscr{L}\mathscr{V} = \mathscr{V}\mathscr{A}$. Now let's invoke previously established results (Roberts 2015a, §2). The linear operator $\mathscr{L}$, defined in (14), has the same block Toeplitz structure as previously (Roberts 2015a, (7) on p.1496). Consequently (Roberts 2015a, Lemma 4), a basis for the zero eigenspace of $\mathscr{L}$ is the collective columns of

$$\mathscr{V} = \begin{bmatrix} V_0 & V_1 & V_2 \\ 0 & V_0 & V_1 \\ 0 & 0 & V_0 \end{bmatrix}, \qquad \text{and further, } \mathscr{A} = \begin{bmatrix} 0 & A_1 & A_2 \\ 0 & 0 & A_1 \\ 0 & 0 & 0 \end{bmatrix}.$$

The hierarchy of equations to solve for the components of these has been previously established (Roberts 2015a, Lemma 3): the hierarchy is essentially equivalent to the hierarchy one would solve if using the method of multiple scales, but the theoretical framework here is more powerful. The upshot is that for Zappa dispersion, in which overlines denote cross-channel averages,

$$\begin{aligned} V_0 &= 1, & V_1 &= \overline{v} - v, & V_2 &= 2(\overline{v}^2 - \overline{v^2} - \overline{v}v + v^2), \\ & & A_1 &= -\overline{v}, & A_2 &= \overline{(v-\overline{v})^2} + \overline{v^2}. \end{aligned} \tag{15}$$

In the specific case of $v(y) = 1 - y^2$, these expressions reduce to the coefficients and polynomials of the slowly varying, slow manifold, model (6).

So now we know that the evolution on the zero eigenspace, the approximate slow manifold, is $d\mathbf{U}/dt = \mathscr{A}\mathbf{U}$: let's see how this translates into the macroscale PDE (6b). Now, the first line of $d\mathbf{U}/dt = \mathscr{A}\mathbf{U}$ is the ODE $dU_0/dt = A_1U_1 + A_2U_2$. Defining $U_0 = U(X,t) := \overline{u(X,y,t)}$, Proposition 6 of Roberts (2015a) applies, and so generally $U(x,t)$ satisfies the macroscale effective advection-diffusion PDE (2)—a PDE that reduces to the specific (6b) in the case $v(y) = 1 - y^2$.

3.3 Account for the coupling remainder

Now we treat the exact 'local' system $d\mathbf{u}/dt = \mathscr{L}\mathbf{u} + \mathbf{r}$ as non-autonomously 'forced' by coupling to all cross-sections in $\mathbb{X}$ through the remainder (aka Mori–Zwanzig transformation, e.g., Venturi, Cho, and Karniadakis 2015). There are two justifications, both a simple and a rigorous, for being able to project such 'forcing' onto the local model. First, simply, the rational projection of initial conditions for low-dimensional dynamical models leads to a cognate projection of any forcing (Roberts 1989, §7). Second, alternatively and more rigorously, Aulbach and Wanner (2000) developed a general forward theory, that applies here, of centre manifolds for non-autonomous systems in suitable 'infinite-D' state spaces: the theory establishes the existence and emergence of an 'infinity-D' global centre manifold—a centre manifold whose construction (Potzsche and Rasmussen 2006, Prop. 3.6) happens to be symbolically identical at each $X \in \mathbb{X}$. Keep clear the contrasting points of view that contribute: on the one hand we consider the relatively low-dimensional system at each locale X in space, a system that is weakly coupled to its neighbours; on the other-hand we consider the relatively high-dimensional system of all locales $\mathbb{X}$ coupled together and then theory establishes global properties.

The upshot is that here we need to project the coupling remainder $\mathbf{r}(t)$ onto each local slow manifold. Fortunately, the structure of the linear local dynamics (14) is identical to that discussed by Roberts (2015a). Hence, many of the results reported there apply here. Linear algebra involving adjoint eigenvectors Z_0 and $\mathscr{W}_n$ ($\Omega_0^\dagger Z_0 = 0$ and $\mathscr{L}^\dagger\mathscr{W} = \mathscr{W}\mathscr{A}$, Roberts 2015a, §2.3), together with the history of the coupling remainder $e^{-t} \star \mathbf{r}$, leads to the error formula (6d) (equation (23) from Roberts 2015a). Then the general macroscale advection-diffusion model (2) becomes exact with the error term ρ included (here the error (6d) is third-order in spatial derivatives)

$$\frac{\partial U}{\partial t} = A_1\frac{\partial U}{\partial x} + A_2\frac{\partial^2 U}{\partial x^2} + \rho\,.$$

Then, simply, *the macroscale effective advection-diffusion model* PDE (2) *is valid simply whenever and wherever the error term* ρ *is acceptably small. There is: no* ε*; no limit; no required scaling; no 'balancing'; no ad hoc hierarchy of space-time variables.*

4 Compact analysis, and connect to well-known methodology

It is very tedious to perform all the algebraic machinations of Section 3 on the Taylor series coefficients. Instead, we may compactify the analysis by defining the quadratic *generating polynomial* (Roberts 2015a, §3.1)

$$\tilde{u}(X,\zeta,y,t) := u_0(X,y,t) + \zeta u_1(X,Y,t) + \tfrac{1}{2}\zeta^2 u_2(X,X,y,t) \tag{16}$$

(or a higher-order polynomial if the analysis is to higher-order). This generating polynomial then satisfies the exact differential equation (17). Consider $\partial\tilde{u}/\partial t$, at (X,ζ,y,t), and substitute the equations (14) for the Taylor coefficients at (X,y,t):

$$\begin{aligned}
\frac{\partial\tilde{u}}{\partial t} &= \frac{\partial u_0}{\partial t} + \zeta\frac{\partial u_1}{\partial t} + \tfrac{1}{2}\zeta^2\frac{\partial u_2}{\partial t}\\
&= \quad \mathfrak{L}_0 u_0 + \mathfrak{L}_1 u_1 + \mathfrak{L}_2 u_2 + r_0\\
&\quad + \mathfrak{L}_0\zeta u_1 + \mathfrak{L}_1\zeta u_2 \qquad + \zeta r_1\\
&\quad + \mathfrak{L}_0\tfrac{1}{2}\zeta^2 u_2 \qquad\qquad + \tfrac{1}{2}\zeta^2 r_2\\
&= \mathfrak{L}_0\tilde{u} + \mathfrak{L}_1\frac{\partial\tilde{u}}{\partial\zeta} + \mathfrak{L}_2\frac{\partial^2\tilde{u}}{\partial\zeta^2} + \tilde{r}\\
\implies \frac{\partial\tilde{u}}{\partial t} &= \left[\mathfrak{L}_0 + \mathfrak{L}_1\frac{\partial}{\partial\zeta} + \mathfrak{L}_2\frac{\partial^2}{\partial\zeta^2}\right]\tilde{u} + \tilde{r}
\end{aligned} \tag{17}$$

for the generating polynomial of the coupling remainder, $\tilde{r} := r_0 + \zeta r_1 + \frac{1}{2}\zeta^2 r_2$.

Appropriate analysis of the IDE (17) then reproduces the previous Section 3. But the algebra is done much more compactly as the separate components u_0, u_1, u_2 are all encompassed in the one generating polynomial $\tilde{u}$. One important property of the analysis is that although we normally regard the derivative $\partial/\partial\zeta$ as unbounded, in the analysis of IDE (17) the space of functions is just that of quadratic polynomials in ζ, and so here $\partial/\partial\zeta$ is bounded, as well as possessing other nice properties.

Indeed, since we are only interested in the space of quadratic polynomials in ζ, the analysis neglects any term $\mathcal{O}(\zeta^3)$. Equivalently, we would work to 'errors' $\mathcal{O}(\partial^3/\partial\zeta^3)$. This view empowers us to organise the necessary algebra in a framework where we imagine $\partial/\partial\zeta$ is 'small'. Note: in the methodology here $\partial/\partial\zeta$ is *not assumed* small, as we track errors exactly in the remainder $\tilde{r}$, it is just that we may organise the algebra as if $\partial/\partial\zeta$ was small. Such organisation then leads to the same hierarchy of problems as in Section 3.2, just more compactly.

Connect to extant methodology

Since the notionally small $\partial/\partial\zeta$ is effectively a small spatial derivative, we now connect to extant multiscale methods that a priori assume slow variations in space. That is, we now show that the non-remainder part of IDE (17) appears in a conventional multiscale approximation of the governing microscale system (1).

In conventional asymptotics we invoke restrictive scaling assumptions at the start. Here one would assume that the solution field $u(x,y,t)$ is slowly-varying in space x. Then the argument goes that the field may be usefully written near any $X \in \mathbb{X}$ as the local Taylor quadratic approximation[2]

$$u(\xi,y,t) \approx u|_{\xi=X} + (\xi - X)u_\xi|_{\xi=X} + \frac{(\xi-X)^2}{2!}u_{\xi\xi}|_{\xi=X}\,.$$

Substituting into the nonlocal microscale (1) gives, at (X,y,t) and letting dashes/primes denote derivatives with respect to the first argument,

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \int_{\mathbb{Y}}\int_{\mathbb{X}} k(X,\xi,y,\eta)u(\xi,\eta,t)\,d\xi\,d\eta \\
&\approx \int_{\mathbb{Y}}\int_{\mathbb{X}} k(X,\xi,y,\eta)\left[u|_{\xi=X} + (\xi-X)u'|_{\xi=X} + \frac{(\xi-X)^2}{2!}u''|_{\xi=X}\right]d\xi\,d\eta \\
&= \int_{\mathbb{Y}}\int_{\mathbb{X}} k(X,\xi,y,\eta)\,d\xi\,u(X,\eta,t) + \int_{\mathbb{X}} k(X,\xi,y,\eta)(\xi-X)\,d\xi\,u'(X,\eta,t) \\
&\qquad + \int_{\mathbb{X}} k(X,\xi,y,\eta)\frac{(\xi-X)^2}{2!}\,d\xi\,u''(X,\eta,t)\,d\eta \\
&= \int_{\mathbb{Y}} k_0(X,y,\eta)u(X,\eta,t) + k_1(X,y,\eta)u'(X,\eta,t) \\
&\qquad + k_2(X,y,\eta)u''(X,\eta,t)\,d\eta \\
&= \mathfrak{L}_0 u + \mathfrak{L}_1 u' + \mathfrak{L}_2 u''. \qquad (18)
\end{aligned}$$

Now the generating polynomial $\tilde{u}$, defined by (16), is such that $u(X+\zeta,y,t) = \tilde{u}(X,\zeta,y,t) + \mathscr{O}(\zeta^3)$. Hence, rewriting the *approximate* PDE (18) for $u(X+\zeta,y,t)$ at fixed X gives precisely the IDE (17) except that the remainder coupling $\tilde{r}$ is omitted. Consequently, extant multiscale methodologies continuing on from PDE (18) generate equivalent results to that of Section 3, but in a different framework—a framework without the error term (6d).

Most extant multiscale analysis invokes, at the outset, balancing of scaling parameters, requires a small parameter, is only rigorous in the limit of infinite scale separation, and often invents heuristic multiple space-time variables. The approach developed herein connects with such analysis, but is considerably more flexible and, furthermore, justifies a more formal approach developed 30 years ago (Roberts 1988), and implemented in Section 6. Further this approach derives the rigorous error expression (6d) at finite scale separation.

[2] I continue to conjecture that truncations to orders other than quadratic give corresponding analysis and results. Ongoing research will elucidate.

5 Conclusion

This article initiates a new multiscale modelling approach applied to a specific basic problem. This article considers the scenario where the given physical problem (1) has non-local microscale interactions, such as inter-particle forces or dynamics on a lattice. Many extant mathematical methodologies derive, for such physical systems, an approximate macroscale PDE, such as the advection-diffusion (2). The novelty of our approach is that it derives a precise expression for the error of the macroscale approximate PDE, here (6d). Then, simply, and after microscale transients decay, the macroscale advection-diffusion PDE (2) is valid wherever and whenever the quantified error (6d) is acceptable.

Of course, in all such applications, we need the third moment of the microscale interaction kernel $k(x,\xi,y,\eta)$ to exist (see definition (12)) for the error analysis of Section 3.1 to proceed and provide the error term. All moments exist for the Zappa problem, see (9). If, in some application, the third moment does not exist, but the second moment does, then the advection-diffusion PDE (2) may be an appropriate macroscale model, but this work would not provide a quantifiable error.

Another important characteristic of our new approach is that the validity of the macroscale PDE is not confined by a limit '$\varepsilon \to 0$'—the approach holds for finite scale separation in the multiscale problem, in the large but finite domain $\mathbb{X}$. Further, and in contrast to most extant methodologies, the approach here should generalise in further research to arbitrary order models just as it does when the microscale is expressed as PDEs (Roberts 2015a).

The developed scenario here is that of linear nonlocal systems (1). However, key parts of the argument are justified with centre manifold theory (Aulbach and Wanner 2000; Potzsche and Rasmussen 2006; Haragus and Iooss 2011; Roberts 2015b, e.g.). Consequently, further research should be able to show that cognate results hold for nonlinear microscale systems.

With further research, correct boundary conditions for the macroscale PDEs should be derivable by adapting earlier arguments to derive rigorous boundary conditions for approximate PDEs (Roberts 1992; Chen, Roberts, and Bunder 2018).

Interesting applications of this novel approach would arise whenever there are microscale nonlocal interactions in the geometry of problems such as (e.g., Roberts 2015b) dispersion in channels and pipes, the lubrication flow of thin viscous fluids, shallow water approximations whether viscous or turbulent, quasi-elastic beam theory, long waves on heterogeneous lattices, and pattern evolution.

Acknowledgements This research was partly supported by the Australian Research Council with grant DP180100050. I thank Carlo Laing for prompting this direction for research.

6 Appendix: Computer algebra derives macroscale PDE

The following computer algebra derives the effective advection-diffusion PDE (6b), or any higher-order generalisation, for the microscale nonlocal Zappa system (3). This code uses the free computer algebra package Reduce.[3] Analogous code will work for other computer algebra packages, and/or for cognate problems (Roberts 2015b, e.g.).

```
% advection-diffusion PDE of Zappa transport in a channel
% AJR, 20 Jan 2017 -- 20 Jan 2020
on div; off allfac; on revpri; factor d,uu;

let d^5=>0; % truncate to this order of error
operator uu; depend uu,x,t; % uu(n):=df(uu,x,n)
let { df(uu(~n),x)=>uu(n+1), df(uu(~n),t)=>df(g,x,n)  };
operator mean; linear mean; % average across channel
let { mean(1,y)=>1, mean(y^~~p,y)=>(1+(-1)^p)/2/(p+1) };

% Preprocess nonlocal x-jumping: in essence finds the
% kernel integrals are (-v)^n
depend w,x; % dummy function for u(x)
% Taylor expand w(xi)=w(x+z) where z=xi-x
jmp:=for n:=0:deg((1+d)^99,d) sum d^n*df(w,x,n)*z^n/factorial(n)$
jmp:=int(exp(z/v)*jmp,z)$ % integrate exp((xi-x)/v)w(x)
% eval from z=-inf to 0 for the convolution
jmp:=sub(z=0,jmp/v)-w$

% iterate from quasi-equilibrium start
u:=uu(0)$ g:=0$
for it:=1:99 do begin
   res:=-df(u,t)+sub({w=u,v=1-y^2},jmp)+(-u+mean(u,y));
   write lengthres:=length(res);
   g:=g+(gd:=mean(res,y));
   u:=u+res-gd;
   if res=0 then write "Success: ",it:=it+10000;
end;
write "The resulting slow manifold and evolution is";
u:=u; duudt:=g;
end;
```

References

Aulbach, Bernd and Thomas Wanner (2000). "The Hartman–Grobman theorem for Caratheodory-type differential equations in Banach spaces". In: *Nonlinear Analysis* 40, pp. 91–104. DOI: 10.1016/S0362-546X(00)85006-3 (cit. on pp. 430, 431, 434).

[3] http://www.reduce-algebra.com/

Bunder, J. E. and A. J. Roberts (June 2018). *Nonlinear emergent macroscale PDEs, with error bound, for nonlinear microscale systems*. Tech. rep. [https://arxiv.org/abs/1806.10297] (cit. on p. 423).

Buttenschön, Andreas and Thomas Hillen (Jan. 2020). *Non-Local Cell Adhesion Models: Steady States and Bifurcations*. Tech. rep. http://arxiv.org/abs/2001.00286 (cit. on p. 424).

Chen, Chen, A. J. Roberts, and J. E. Bunder (2018). "Boundary conditions for macroscale waves in an elastic system with microscale heterogeneity". In: *IMA Journal of Applied Mathematics* 83.3, pp. 1–33. DOI: 10.1093/imamat/hxy004. (Cit. on p. 434).

Duncan, Jacob P. et al. (Feb. 2017). "Multi-scale methods predict invasion speeds in variable landscapes". In: *Theoretical Ecology*, pp. 1–17. DOI: 10.1007/s12080-017-0329-0 (cit. on p. 424).

Ermentrout, Bard (2015). "Mathematical Neouroscience". In: *Princeton Companion to Applied Mathematics*. Ed. by Nicholas J. Higham et al. Princeton. Chap. VII.21, pp. 873–879 (cit. on p. 424).

Haragus, Mariana and Gerard Iooss (2011). *Local Bifurcations, Center Manifolds, and Normal Forms in Infinite-Dimensional Dynamical Systems*. Springer. DOI: 10.1007/978-0-85729-112-7 (cit. on p. 434).

Lipton, Robert (Oct. 2014). "Dynamic Brittle Fracture as a Small Horizon Limit of Peridynamics". In: *Journal of Elasticity* 117.1, pp. 21–50. DOI: 10.1007/s10659-013-9463-0. (Cit. on p. 424).

Omelyan, Igor and Yuri Kozitsky (2018). *Spatially Inhomogeneous Population Dynamics: Beyond Mean Field Approximation*. Tech. rep. http://arxiv.org/abs/1805.06795 (cit. on p. 424).

Pollett, P. K. and A. J. Roberts (1990). "A description of the long-term behaviour of absorbing continuous time Markov chains using a centre manifold". In: *Advances Applied Probability* 22, pp. 111–128. DOI: 10.2307/1427600 (cit. on p. 426).

Potzsche, Christian and Martin Rasmussen (2006). "Taylor Approximation of Integral Manifolds". In: *Journal of Dynamics and Differential Equations* 18, pp. 427–460. DOI: 10.1007/s10884-006-9011-8 (cit. on pp. 431, 434).

Roberts, A. J. (1988). "The application of centre manifold theory to the evolution of systems which vary slowly in space". In: *J. Austral. Math. Soc. B* 29, pp. 480–500. DOI: 10.1017/S0334270000005968 (cit. on p. 433).

– (1989). "Appropriate initial conditions for asymptotic descriptions of the long term evolution of dynamical systems". In: *J. Austral. Math. Soc. B* 31, pp. 48–75. DOI: 10.1017/S0334270000006470 (cit. on p. 431).

– (1992). "Boundary conditions for approximate differential equations". In: *J. Austral. Math. Soc. B* 34, pp. 54–80. DOI: 10.1017/S0334270000007384 (cit. on p. 434).

– (2015a). "Macroscale, slowly varying, models emerge from the microscale dynamics in long thin domains". In: *IMA Journal of Applied Mathematics* 80.5, pp. 1492–1518. DOI: 10.1093/imamat/hxv004 (cit. on pp. 423, 424, 426, 429–432, 434).

– (2015b). *Model emergent dynamics in complex systems*. SIAM, Philadelphia. ISBN: 9781611973556. http://bookstore.siam.org/mm20/ (cit. on pp. 434, 435).

Roberts, A. J. and J. E. Bunder (2017). "Slowly varying, macroscale models emerge from microscale dynamics over multiscale domains". In: *IMA Journal of Applied Mathematics* 82, pp. 971–1012. DOI: 10.1093/imamat/hxx021. (Cit. on p. 423).

Silling, S. A. and R. B. Lehoucq (Oct. 2008). "Convergence of Peridynamics to Classical Elasticity Theory". In: *Journal of Elasticity* 93.1, p. 13. DOI: 10.1007/s10659-008-9163-3. (Cit. on p. 424).

Silling, Stewart A (2000). "Reformulation of elasticity theory for discontinuities and long-range forces". In: *Journal of the Mechanics and Physics of Solids* 48.1, pp. 175–209. DOI: 10.1016/S0022-5096(99)00029-0 (cit. on p. 424).

Venturi, Daniele, Heyrim Cho, and George Em Karniadakis (2015). "Mori–Zwanzig Approach to Uncertainty Quantification". In: *Handbook of Uncertainty Quantification*. Ed. by Roger Ghanem, David Higdon, and Houman Owhadi. Springer International Publishing, pp. 1–36. DOI: 10.1007/978-3-319-11259-6_28-1 (cit. on p. 431).

Influence of an oblique magnetic field on planar flame front instability

Mako Sato and Yasuhide Fukumoto

Abstract We investigate the effect of external magnetic field on the Darrieus-Landau instability (DLI), the linear instability of a planar premixed flame front, in an electrically conducting fluid. This setting has applicability to combustion phenomena of the astrophysical scale. Without magnetic field, the planar flame front is necessarily unstable. Previous investigation treated independently the normal and tangential magnetic fields. Here we focus on the case of their simultaneous application, namely, oblique magnetic field. Rederiving the jump conditions, across the flame front, of the physical variables based on the ideal magnetohydrodyamics equations, we correct the previous treatment of the Markstein effect and extend it to incorporate the disparity of the magnetic permeability. A genuinely oblique magnetic field has an unusual characteristics that discontinuity in tangential velocity across the flame is induced. It is found that the Kelvin-Helmholtz instability takes over the stabilizing effect on the DLI in a limited parameter regime when the normal Alfvén speed exceeds the normal fluid velocity in both the unburned and burned regions.

1 introduction

Combustion is a multiscale phenomenon of a reacting fluid from the molecular scale, on which a succession of complicated chemical reactions occur, to the hydrodynamic scale, flows of a fuel (= an unburned gas) and a burned gas on the

Mako Sato
Graduate School of Science, Osaka City University,
3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558-8585, Japan
e-mail: m20sa020@tx.osaka-cu.ac.jp

Yasuhide Fukumoto
Institute of Mathematics for Industry, Kyushu University,
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: yasuhide@imi.kyushu-u.ac.jp

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_26

macroscopic scale, as exemplified by the Bunsen burner and rocket engines [5]. It is a phenomenon whose precise description needs consideration of a number of chemical processes and the mixture and diffusion of a number of reactant species and the heat as well. If we consider combustion phenomena in the universe, such as explosion of supernovae, we have to include the magnetic field [9] which is macroscopically described by the Maxwell equations.

It may well be thought that we have to deal with an innumerable number of equations of governing an innumerable number of chemical species, hydrodynamic, thermodynamic variables and electro-magnetic fields. Extensive studies since the middle of the 20th centuries have clarified that the macroscopic scale behavior of a phenomena is somewhat independent from the microscopic processes, now being known as the concept of the scale separation. In this paper, we consider the combustion subjected to imposed magnetic field, on the hydrodynamic scale, of an electrically conducting fluid. We make the linear stability analysis of a planar front, of infinitesimal thickness, of a premixed flame, based on the equations of the ideal magnetohydrodynamics (MHD), consisting of the continuity and the Euler equations augmented by the induction equation for the magnetic field [10].

The Darrieus-Landau instability (DLI) is an exponential amplification, in time, of wavy deformations of a planar flame front of incompressible fluids (=burned and unburned gases), put forward independently by Darrieus (1938) and Landau (1944) [6, 13, 14]. For this treatment of the hydrodynamic instability, the flame front is reckoned as an infinitely thin interface with density discontinuity across it. In accordance, we skip chemical reactions occurring in it. Darrieus and Landau assumed that the flame front advances to the unburned gas (fuel) at a constant speed S_L, and dealt with the Euler equation and the continuity equation for an incompressible fluid. The growth rate n of the DLI is given by

$$\frac{n}{kS_L} = \frac{\alpha}{1+\alpha}\left(\sqrt{1+\alpha-\frac{1}{\alpha}}-1\right),$$

where $\alpha = \rho_u/\rho_b(>1)$, with $\rho_u(\rho_b)$ being the density of unburned (burned) gas, represents the thermal expansion and k is the wavenumber. The result of $n>0$ for all $\alpha(>1)$ means that a planer flame front is necessarily linearly unstable, with the growth rate being proportional to the wavenumber. However stable flame fronts are observed in experiments, which motivated successors to improve the original DLI. Markstein [15] considered effect of the flame front curvature on the flame speed S_f [5]. Matalon and Matkowsky [16] elaborated the Markstein effect based on the heat-conduction equation for the temperature and the diffusion equation for the reactant, and formulated the Markstein effect as the jump conditions, on the hydrodynamic scale, of the hydrodynamic and thermodynamic variables across the flame front. Class *et al.* [3, 4] devised the jump conditions to take account of chemical reactions through the heat release at the interface concomitant with the gas expansion.

For combustion phenomena in the universe, we have to add the induction equation *etc.* to the governing equations for plasmas, because the space around large-scale objects is filled with magnetic field. The magnetic field is expected to sup-

press the instability of an interface, and its influence has extensively studied for the Kelvin-Helmholtz (KHI), the Rayleigh-Taylor and the Richtmeyer-Meshkov instabilities [10, 17], toward the goal of controlling them. However relatively little is known about the its influence on the DLI [9]. Dursi [7] addressed the magnetic DLI, with treating normal and tangential external magnetic field separately, but has gone untouched the coexistent case of the both fields. The latter half of this paper is devoted to handling simultaneous application of the both fields, that is, the oblique external magnetic field. This case exhibits a marked contrast with the cases of a single component, in the sense that only the genuinely oblique magnetic field is able to admit a discontinuity of tangential velocity. For a neutral fluid, the non-zero mass flux is in no way compatible with the tangential-velocity discontinuity, and hence, an oblique magnetic field may drastically alter the DLI. The coexistence of the DLI and the KHI is of our primary concern, and we shall shoe that this is indeed the case.

The detail of derivation of the jump conditions across the flame front is not often described in the literature [7]. We begin, in section 2, with it, following the method of refs [1, 11]. As a reward, a generalization is achieved to incorporate the disparity of the magnetic permeability and to correct the Markstein effect. With these jump conditions as a basis, we revisit, in section 3, the effect of the tangential magnetic field on the DLI. We include the gravity, the surface tension and the Markstein effects, and look into the effect of the disparity of the magnetic permeability, which are missing in [7]. Section 4 focuses on the effect of the obliques magnetic field. As anticipated above, we find a possible emergence of the KHI to compete with the stabilizing action of the magnetic field. We close, in section 5, with a summary and a list of remaining problems.

2 Basic equations and jump conditions

The governing equations of the magnetohydrodynamics of an inviscid, incompressible and perfectly conducting fluid are

$$\frac{\partial \vec{U}}{\partial t} + (\vec{U}\cdot\nabla)\vec{U} - \frac{1}{\rho\mu}(\vec{B}\cdot\nabla)\vec{B} + \nabla\left(\frac{\vec{B}^2}{2\rho\mu}\right) = -\frac{1}{\rho}\nabla P + \vec{g}\frac{\delta\rho}{\rho}, \tag{1}$$

$$\frac{\partial \vec{B}}{\partial t} - \nabla\times(\vec{U}\times\vec{B}) = 0, \tag{2}$$

$$\nabla\cdot\vec{U} = 0, \tag{3}$$

$$\nabla\cdot\vec{B} = 0, \tag{4}$$

where $\vec{U}$ is the velocity, $\vec{B}$ is the magnetic field, ρ is the density, p is the pressure, μ is the magnetic permeability and $\vec{g} = (0,0,-g)$ is gravity acceleration directed in the negative z-axis. For sake of simplicity, we employ the Boussinesq approximation to treat the buoyancy effect.

We put the xy-plane parallel to the unperturbed planer flame front and the z-axis is perpendicular to it with the front lying on $z = 0$. In the sequel, we derive the boundary conditions, or the jump conditions, across the flame front, from the unburned to the burned sides, by utilizing the Heaviside step function $H(\theta)$ defined by

$$H(\theta) = \begin{cases} 1 & (\theta > 0) \\ 0 & (\theta < 0) \end{cases}, \tag{5}$$

[1, 10, 11]. This function is used to express the discontinuity of the basic flow outside a flame front. The Dirac delta function $\delta(\theta)$ appears as the derivative of the Heaviside step function.

$$\delta(\theta) = \frac{d}{d\theta} H(\theta).$$

The position of the perturbed flame front is set to be $z = \zeta(x,y,t)$. In deriving the jump conditions across the flame front, it is useful to introduce $\theta = \zeta(x,y,t) - z$ with the flame located at $\theta = 0$. The unit normal vector to the flame front is expressed as

$$\vec{n} = \nabla\theta/|\nabla\theta| \approx (\nabla\zeta, -1). \tag{6}$$

The normal component of the velocity of a flame front is

$$\vec{u}_f \cdot \vec{n} = -\frac{\partial\zeta}{\partial t}, \tag{7}$$

and $\dot{\theta} = \partial\theta/\partial t = \partial\zeta/\partial t$. We should keep in mind that $\theta < 0$ and $\theta > 0$ represent the burned and unburned regions respectively.

With this setting, the density field $\rho = \rho(\vec{x},t)$, for example, is expressed as

$$\rho(\vec{x},t) = \rho_u(\vec{x},t)H(\theta) + \rho_b(\vec{x},t)H(-\theta). \tag{8}$$

Here the subscripts b and u denote the quantities on the burned and the unburned sides, respectively. Substituting (8) and the corresponding representations of the other hydrodynamic and thermodynamic variables into the MHD equations (2)-(4), each of equations is divided into three independent parts, identified respectively by $H(\theta)$, $H(-\theta)$ and $\delta(\theta)$. The part including $H(\theta)$ corresponding to the equation on the unburned side, and the one including $H(-\theta)$ corresponding to the equation on the burned side.

The jump conditions across the flame front are gained by picking up the terms including $\delta(\theta)$. We introduce the notation for the jump of any function f across a flame front,

$$[f] = f_b|_{\theta=0_-} - f_u|_{\theta=0_+}.$$

The unit normal vector $\vec{n}$ and the tangential vectors $\vec{t}_1$ and $\vec{t}_2$ on the flame front are given by $\vec{n} \approx (\partial\zeta/\partial x, \partial\zeta/\partial y, -1)$, $\vec{t}_1 \approx (1, 0, \partial\zeta/\partial x)$ and $\vec{t}_2 \approx (0, 1, \partial\zeta/\partial y)$. We denote the mass flux across the flame front by $\vec{q} = \rho(\vec{u} + \dot{\theta}\vec{n}/|\nabla\theta|)$, with $|\nabla\theta| \approx 1$ to be understood. We take the surface tension, with its coefficient σ, into consideration in the jump condition. The surface tension term appears only in the jump condition. With this setting, we write down the jump conditions. To first order in perturbation amplitude, they read

$$[\vec{q}\cdot\vec{n}] = 0, \tag{9}$$

$$[\vec{B}\cdot\vec{n}] = 0, \tag{10}$$

$$[\vec{U}\cdot\vec{t}](\vec{q}\cdot\vec{n}) - (\vec{B}\cdot\vec{n})\left[\frac{\vec{B}\cdot\vec{t}}{\mu}\right] = 0, \tag{11}$$

$$[\vec{U}\cdot\vec{n}](\vec{q}\cdot\vec{n}) = -\left[p + \frac{\vec{B}^2}{2\mu}\right] + \left[\frac{1}{\mu}\right](\vec{B}\cdot\vec{n})^2 + \sigma\left(\frac{\partial^2\zeta}{\partial x^2} + \frac{\partial^2\zeta}{\partial y^2}\right), \tag{12}$$

$$(\vec{q}\cdot\vec{n})\left[\frac{\vec{B}\cdot\vec{t}}{\rho}\right] = (\vec{B}\cdot\vec{n})[\vec{u}\cdot\vec{t}]. \tag{13}$$

These jump conditions (11) and (12) accomplish an extension of the previous ones [18, 7, 10] to include the effect of the difference of the magnetic permeability.

It is worthwhile to recollect Landau's assumption. The flame speed S_f is the speed of the gas incoming to the flame front. Noting that vector $\vec{n}$ is directed to the unburned side, it is given by

$$S_f = \vec{U}_u|_{\theta=0_+}\cdot(-\vec{n}) - \vec{u}_f\cdot(-\vec{n}), \tag{14}$$

where $\vec{u}_f\cdot\vec{n}$ is normal perturbation speed of the flame front. Landau's assumption is interpreted as $S_f = S_L$ [6, 13, 14], though generically the flame front is not flat. This assumption is too restrictive. One of the major efforts to improve the original DLI was to incorporate effect of the flame curvature. It is now established as the Markstein effect [15, 5]. In section 3.4, we look into the Markstein effect in the context of the magnetic DLI in the presence of parallel magnetic field.

3 Magnetic DLI subject to tangential magnetic field

In this section, we derive the growth rate when magnetic field parallel to the front is imposed. We follow closely the approach and notation of Dursi [7].

3.1 Hydromagnetic waves

We consider velocity, magnetic fields and pressure of the form

$$\vec{U} = (u, v, w + W) \qquad (u, v, w \ll W), \tag{15}$$

$$\vec{B} = (b_x + B, b_y, b_z) \qquad (b_x, b_y, b_z \ll B), \tag{16}$$

$$P = P_0 + p \qquad (p \ll P_0), \tag{17}$$

where W, B and P_0 are the values of the basic state, which are taken to be constants within each region. The jump conditions (9) and (13) require jump of the values of W and B across the flame front by

$$[\rho W] = 0, \quad [WB] = 0. \tag{18}$$

These physical quantities may vary rapidly but smoothly inside the flame front [16, 3, 4, 19], but we do not pursue it in this paper.

We take the perturbation of normal form $e^{i\vec{k}\cdot\vec{x}+nt}$ with infinitesimal amplitude. Here the wavevector is defined as $\vec{k} = (k_x, k_y)$ with its magnitude being $k = \sqrt{k_x^2 + k_y^2}$ and n is the growth rate of the wave on the flame front. Here, we exclusively deal with two-dimensional deformation $\vec{k} = (k_x, 0)$

The linearized equations are obtained by substituting (15)-(17) into (1)-(4). We find that a perturbed quantity is expressed by a linear combination of the following modes [7]:

$$\left(C_1 e^{kz} + C_2 e^{-kz} + C_3 e^{-\frac{n+iak_x}{W}z} + C_4 e^{-\frac{n-iak_x}{W}z}\right) e^{i\vec{k}\cdot\vec{x}+nt}, \tag{19}$$

where $a = B/\sqrt{\mu\rho}$ is the the Alfvén speed. It should be born in mind that the Alfvén modes, with amplitude C_3 and C_4, have a distinguishing feature of possessing the vorticity. There is another mode in each region, but this mode turns out to vanish and is irrelevant. We note that the perturbation must be finite. Provided that the real part $\mathrm{Re}[n] > 0$ for instability, the unburned side accepts only the mode with amplitude C_1 and the burned side accepts modes with C_2, C_3 and C_4. The vorticity field introduced on the burned side for a neutral fluid [13, 14] is realized by the Alfvén waves, having amplitude C_3 and C_4, for the MHD.

3.2 Linear perturbations in unburned and burned regions

At the outset, we specify a possible combination of waves for b_{ux} and b_{bx}. By integrating the linearized MHD equations, we obtain the expressions for perturbed quantities as follows [7].

On the unburned side, we have

$$b_{ux} = B_u C_1 e^{kz}, \tag{20}$$

$$b_{yu} = B_u \frac{k_y}{k_x} C_1 e^{kz}, \tag{21}$$

$$b_{uz} = -iB_u \frac{k}{k_x} C_1 e^{kz}, \tag{22}$$

$$u_u = -i\frac{(n+W_u k)}{k_x} C_1 e^{kz}, \tag{23}$$

$$v_u = -i\frac{(n+W_u k)k_y}{k_x^2} C_1 e^{kz}, \tag{24}$$

$$w_u = -\frac{(n+W_u k)}{k}(1+\frac{k_y^2}{k_x^2}) C_1 e^{kz}, \tag{25}$$

$$p_u = \frac{\rho_u (n+W_u k)^2}{k_x^2} C_1 e^{kz}. \tag{26}$$

On the burned side, we have

$$b_{bx} = B_b C_2 e^{-kz} + B_b C_3 e^{-\frac{n+ia_b k_x}{W_b} z} + B_b C_4 e^{-\frac{n-ia_b k_x}{W_b} z}, \tag{27}$$

$$b_{yb} = \frac{k_y}{k_x} B_b C_2 e^{-kz} + B_b C_5 e^{-\frac{n+ia_b k_x}{W_b} z} + B_b C_6 e^{-\frac{n-ia_b k_x}{W_b} z}, \tag{28}$$

$$\begin{aligned} b_{bz} = {} & iB_b \frac{k}{k_x} C_2 e^{-kz} + iW_b B_b \frac{k_x C_3 + k_y C_5}{n+ia_b k_x} e^{-\frac{n+ia_b k_x}{W_b} z} \\ & + iW_b B_b \frac{k_x C_4 + k_y C_6}{n-ia_b k_x} e^{-\frac{n-ia_b k_x}{W_b} z}, \end{aligned} \tag{29}$$

$$u_b = -\frac{i}{k_x}(n-W_b k) C_2 e^{-kz} - a_b C_3 e^{-\frac{n+iak_x}{W_b} z} + a_b C_4 e^{-\frac{n-iak_x}{W_b} z}, \tag{30}$$

$$v_b = -ik_y \frac{n-W_b k}{k_x^2} C_2 e^{-kz} - a_b C_5 e^{-\frac{n+iak_x}{W_b} z} + a_b C_5 e^{-\frac{n-iak_x}{W_b} z}, \tag{31}$$

$$\begin{aligned} w_b = {} & \frac{(n-W_b k)k}{k_x^2} C_2 e^{-kz} - i\frac{W_b (k_x C_3 + k_y C_5)}{n+ia_b k_x} a_b e^{-\frac{n+iak_x}{W_b} z} \\ & + i\frac{W_b (k_x C_4 + k_y C_6)}{n-ia_b k_x} a_b e^{-\frac{n-iak_x}{W_b} z}, \end{aligned} \tag{32}$$

$$p_b = \rho_b \frac{(n-W_b k)^2}{k_x^2} C_2 e^{-kz} - (\rho_b a_b^2) C_3 e^{-\frac{n+iak_x}{W_b} z} - (\rho_b a_b^2) C_4 e^{-\frac{n-iak_x}{W_b} z}. \tag{33}$$

3.3 Jump conditions and dispersion relation

We specialize the jump conditions (9)-(13) to the case in which only magnetic field parallel to the flame front is externally imposed in the unperturbed state. These jump conditions reduce to

$$[b_z - \frac{ik_x}{n} wB] = 0, \tag{34}$$

$$\rho W[u + \frac{ik_x}{n} wW] - (b_z - \frac{ik_x}{n} wB)[\frac{B}{\mu}] = 0, \tag{35}$$

$$\rho W[v + \frac{ik_y}{n} wW] = 0, \tag{36}$$

$$[p + \frac{Bb_x}{\mu}] = \frac{gw}{n}[\rho] + \sigma(\frac{\partial^2 \zeta}{\partial x^2} + \frac{\partial^2 \zeta}{\partial y^2}), \tag{37}$$

$$[Wb_x] = 0, \tag{38}$$

$$[Wb_y] = 0. \tag{39}$$

In (37), we see the the gravity effect entering into in jump condition for the pressure.

In view of (7), Landau's assumption dictates that

$$-w_u + \frac{\partial \zeta}{\partial t} = 0. \tag{40}$$

The jump condition (9) then tells us that

$$-w_b + \frac{\partial \zeta}{\partial t} = 0. \tag{41}$$

We take the advantage of (40), or equivalently (41), to eliminate ζ.

According to the zeroth-order relations in jump conditions

$$\frac{\rho_u}{\rho_b} = \frac{W_b}{W_u} = \frac{B_u}{B_b} = \alpha,$$

we make the following replacement $\rho_b = \rho_u/\alpha, W_b = \alpha W_u$ and $B_b = B_u/\alpha$. We introduce dimensionless variables $\bar{n} = n/kW_u$, the growth rate, $\bar{a}_u = a_u/W_u$, $\bar{\sigma} = \sigma k/\rho_u W_u^2$, $\bar{g} = g/kW_u^2$, $\bar{k}_x = k_x/k$, $\bar{k}_y = k_y/k$ and $\nu = \mu_u/\mu_b$. After some manipulation, we cast the jump conditions into a system of linear equations $\vec{M} \cdot (C_1, C_2, C_3, C_4, C_5, C_6)^T = 0$ with a square matrix $\vec{M}$. For the existence of nontrivial perturbation $(C_1, C_2, C_3, C_4, C_5, C_6) \neq \vec{0}$, the determinant of $\vec{M}$ should be zero, and we arrive at the eigenvalue equation for $\bar{n}$.

$$\begin{aligned} &\left\{\bar{a}_u^2 \bar{k}_x^2 + (\bar{n} - \alpha)^2 \alpha \nu\right\} \left\{\bar{a}_u^2 \bar{k}_x^2 (1 + \bar{n} - 2\alpha + \alpha^2 \nu + \bar{n}\alpha^2 \nu)\right. \\ &\left. + (1 + \bar{n})\alpha\nu \left(\bar{g}(-1 + \alpha) + 2\bar{n}\alpha + \bar{n}^2(1 + \alpha) + \alpha(1 - \alpha + \bar{\sigma})\right)\right\} = 0. \end{aligned} \tag{42}$$

This result extends Dursi's dispersion relation [7] to include the difference of the magnetic permeability as indicated by ν. The magnetic permeability depends on the material and the temperature. As a typical astrophysical combustion, the supernova is considered to be a diamagnetic object ($\nu < 1$).

3.4 Markstein effect

This subsection focuses on the Markstein effect, putting aside the effect of the surface tension ($\sigma = 0$). Markstein [15] amended Landau's assumption phenomenologically by including the effect of the flame-front curvature on the flame speed. As a consequence, Landau's condition is augmented with a term proportional to the front curvature.

$$S_f = -(\vec{U}_u \cdot \vec{n} - \nu \cdot \vec{n}) = S_L(1 - \mathscr{L}\Delta\zeta), \tag{43}$$

where the coefficient $\mathscr{L}$ (> 0) is referred to as the Markstein length.

At the zeroth order, the flame velocity coincides with the laminar flame speed S_L and balances with the basic flow on the unburned side.

$$W_u = S_L. \tag{44}$$

The variation of linear order in (43) incorporates the curvature effect.

$$w_u - \frac{\partial\zeta}{\partial t} = -S_L\mathscr{L}\Delta\zeta. \tag{45}$$

As before, we substitute superposition of the basic flow and perturbations to (34)-(39) and (43). Landau's condition (40) is taken over by (45). For a normal mode $\zeta \sim e^{ik\vec{x}+nt}$, (45) reads

$$w_u = (W_u\mathscr{L}k^2 + n)\zeta, \tag{46}$$

by virtue of $S_L = W_u$,

We restrict our attention to the 2D problem on the xz-plane, and take $k_y = 0$ and $C_5 = C_6 = 0$. With the help of (46), we replace ζ by w_u. Repeating the same procedure, we transform the jump conditions to a system of 4 linear equations $\vec{M} \cdot (C_1, C_2, C_3, C_4)^T = 0$. The condition of vanishing the determinant of matrix $\vec{M}$ yields

$$\begin{aligned}&\left\{\bar{a}_u^2\bar{k}_x^2 + (\bar{n} - \alpha)^2\alpha\nu\right\}\left\{\bar{a}_u^2\bar{k}_x^2(1 + \bar{n} + 2(-1 + \overline{\mathscr{L}})\alpha + \alpha^2\nu + \bar{n}\alpha^2\nu)\right.\\ &+(1 + \bar{n})\alpha\nu\left(\bar{g}(-1 + \alpha) + 2(1 + \overline{\mathscr{L}})\bar{n}\alpha + \bar{n}^2(1 + \alpha) + \alpha(1 - \alpha + 2\overline{\mathscr{L}}\alpha + \bar{\sigma})\right)\Big\}\\ &= 0,\end{aligned} \tag{47}$$

where $\overline{\mathscr{L}} = \mathscr{L}k$ is the dimensionless Markstein length, or the Markstein number.

Given typical values of the Markstein number $\overline{\mathscr{L}} = 0.1$ and the dimensionless gravity acceleration $\bar{g} = 2$, we draw in Fig. 1 the stability boundary ($\mathrm{Re}[n] = 0$) in the parameter space of $\bar{a}_u = B_u/\sqrt{\rho_u\mu_u}$ and α, for various values of the ratio $\nu = \mu_u/\mu_b$ of the magnetic permeability. To this aim, we resort to the Routh-Hurwitz or the Lienard and Chipart criterion for roots of a polynomial equation [2, 8, 12]. The boundary curves correspond to $\nu = 4, 1, 0.5, 0.3$ from left to right, except that the stability region is splitted into two parts for $\nu = 0.3$ and 0.5. The dark region, the right hand side of the curve, except for $\nu = 0.3$ and 0.5, corresponds to the stable

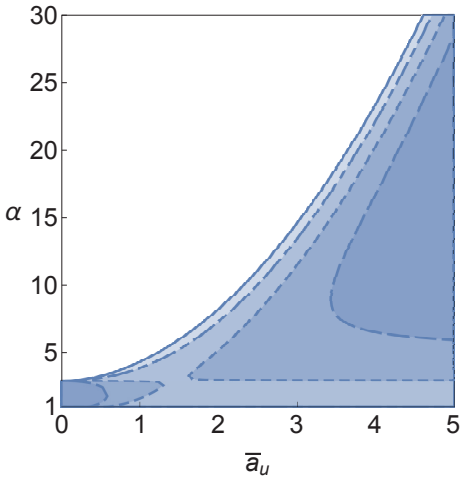


Fig. 1 Dependence on the magnetic-permeability ratio $\nu = \mu_u/\mu_b$ of stability boundary of the magnetic DLI in the plane of the tangential magnetic field $\bar{a}_u = B_u/\sqrt{\rho_u \mu_u}$ and the thermal expansion α ($1 < \alpha < 30$), with the Markstein ($\overline{\mathscr{L}} = 0.1$) and the gravity ($\bar{g} = 2$) effects taken into account. The boundary curves correspond to $\nu = 4, 1, 0.5, 0.3$ from left to right. Colored region (the right hand side of each curve) corresponds to stability region. Notice that stability region is splitted into two regions for $\nu = 0.5$ and 0.3.

parameters, meaning that stronger tangential magnetic field is able to suppress the DLI. For smaller α, the critical value of $\bar{a}_u$ for stability is smaller. The stability region depends sensitively on the magnetic-permeability ratio. The stability region shrinks as ν decreases, implying that the diamagnetic fuel ($\nu < 1$), as is the case of a supernova, enhances the DLI.

We point out that Dursi [7] made an attempt at incorporating the Masrkstein effect into the magnetic DLI. He applied the Markstein condition (45) not only to the unburned side, but also burned side. However, the condition (45) is applicable only to w_u, the burned side.

4 Magnetic DLI subject to oblique magnetic field

We turn to the general case of presence of both the parallel and normal components of external magnetic field, which was not considered in the previous investigation [7]. We reveal that the simultaneous existence of parallel and normal components drastically alters the situation of a single component. Only when the both components are present, discontinuity of the tangential velocity along the flame front is induced, which may cause the Kelvin-Helmholtz instability (KHI). For the sake of simplicity, we disregard the jump of the magnetic permeability across the flame front and take the common value μ_0 of the vacuum for the both sides.

4.1 Jump of basic state

Consider a superposition of the basic state with the unperturbed flame front lying on the plane of $z = 0$, imposed by the external magnetic field (B_x, B_z) imposed, and linear perturbation to it. We restrict our attention to the two-dimensional flow in the xz-plane, and express the flow field as

$$\vec{U} = (u+U, w+W) \qquad (|u|, |w| \ll |U|, |W|), \tag{48}$$

$$\vec{B} = (b_x + B_x, b_z + B_z) \qquad (|b_x|, |b_z| \ll |B_x|, |B_z|), \tag{49}$$

$$\tilde{p} = P + p \qquad (|p| \ll |P|). \tag{50}$$

If we set $U = 0$ and $B_z = 0$, the situation reduces to the case considered in section 3.

We examine the jump conditions (9)-(13) in this general context. Substituting the basic flow field among (48)-(50), these jump conditions become

$$[\rho W] = 0, \tag{51}$$

$$[B_z] = 0, \tag{52}$$

$$[U]\rho W - \frac{B_z}{\mu_0}[B_x] = 0, \tag{53}$$

$$[W B_x] = B_z[U]. \tag{54}$$

The first two conditions give $\alpha W_u = W_b$ and $B_z := B_{uz} = B_{bz}$, by use of the thermal expansion rate $\alpha = \rho_u/\rho_b$. Simultaneous solution of (53) and (54) gives rise to

$$[U] = W_u \frac{a_{ux}}{a_{uz}} \frac{(1-\alpha)a_{uz}^2}{\alpha W_u^2 - a_{uz}^2}, \tag{55}$$

$$B_{bx} = B_{ux} \frac{W_u^2 - a_{uz}^2}{\alpha W_u^2 - a_{uz}^2}, \tag{56}$$

where $a_{ui} = B_{ui}/\sqrt{\rho_u \mu_0}$ $(i = x, z)$ is the Alfvén speed on the unburned side and $a_{bi} = B_{bi}/\sqrt{\rho_b \mu_0}$ the Alfvén speed on the burned side.

It is remarkable that the discontinuity of the tangential velocity U is induced, as opposed to the case of section 3. As is well known, in the absence of the magnetic field, the presence of mass flux penetrating through a discontinuous interface requires continuity of the tangential velocity [14], and hence rules out the possibility of the KHI. The distinguishing feature of the case of oblique magnetic field is emergence of tangential discontinuity $[U](\neq 0)$, which is made possible only by the simultaneous application of B_x and B_z as is seen from (55). It is also noteworthy that the relation between B_{ux} and B_{bx} is different from the case of the parallel magnetic field, in section 3, due to the imposition of $B_z(\neq 0)$.

4.2 Hydromagnetic waves

We send perturbations of the form e^{ikx+nt} and expand the MHD equations (1)-(4) to linear order in the perturbation amplitude.

$$(n+Uik+WD)u-\frac{B_z}{\rho\mu_0}Db_x+\frac{B_z}{\rho\mu_0}ikb_z=-\frac{ik}{\rho}p, \tag{57}$$

$$(n+Uik+WD)w-\frac{B_x}{\rho\mu_0}ikb_z+\frac{B_x}{\rho\mu_0}Db_x=-\frac{1}{\rho}Dp, \tag{58}$$

$$(n+Uik+WD)b_x=(B_xik+B_zD)u, \tag{59}$$

$$(n+Uik+WD)b_z=(B_xik+B_zD)w, \tag{60}$$

$$iku+Dw=0, \tag{61}$$

$$ikb_x+Db_z=0, \tag{62}$$

where we have introduced the differential operator $D=d/dz$.

To gain the z-dependence of perturbed variables, we combine (57)-(62) into a single equation. Applying D to (57) and adding it to $-ik$ times (58), we have

$$(n+Uik+WD)(Du-ikw)-\frac{1}{\rho\mu_0}(B_xik+B_zD)(Db_x+ikb_z)=0. \tag{63}$$

Substituting u from (61) and b_x from (62), we further reduce ik times (63) to

$$(n+Uik+WD)(D^2-k^2)w-\frac{1}{\rho\mu_0}(B_xik+B_zD)(D^2-k^2)b_z=0. \tag{64}$$

Applying (B_xik+B_zD) to (64) to eliminate w, with use of (60), we are eventually left with

$$(D^2-k^2)\left\{(n+Uik+WD)^2-\frac{1}{\rho\mu_0}(B_xik+B_zD)^2\right\}b_z=0. \tag{65}$$

The first factor corresponds to the incompressible limit of the sound wave, and the second factor represents the Alfvén waves for the oblique magnetic field, modified by the Doppler effect. The MHD offers the Alfvén waves as agents of carrying the vorticity with them.

4.3 Linear perturbations in unburned and burned regions

In each region, the perturbation should not diverge in the far region from the flame front, as $z\to-\infty$ on the unburned region and as $z\to\infty$ on the burned side. We seek the instability and suppose that $\mathrm{Re}[n]>0$. We assume $k>0$ without loss of generality.

A possible combination of linear waves, with their z-dependence specified by (65), in each region is given as follows. In the unburned region ($\theta > 0$), the permissible perturbation is, on the condition that $\mathrm{Re}[-n/(W_u - a_{uz})] > 0$,

$$C_1 e^{kz} + C_2 e^{\frac{-n-U_u ik+a_{ux}ik}{W_u-a_{uz}}z}, \tag{66}$$

where C_1 and C_2 are constants. In the burned region ($\theta < 0$), the permissible perturbation is, on the conditions that $\mathrm{Re}[-n/(W_b + a_{bz})] < 0$ and $\mathrm{Re}[-n/(W_b - a_{bz})] < 0$,

$$C_3 e^{-kz} + C_4 e^{\frac{-n-U_b ik-a_{bx}ik}{W_b+a_{bz}}z} + C_5 e^{\frac{-n-U_b ik+a_{bx}ik}{W_b-a_{bz}}z}, \tag{67}$$

where C_3, C_4 and C_5 are constants. The first condition is necessarily fulfilled for $n > 0$. The wave with C_2 is allowable only in the case of $W_u < a_{uz}$ and that of C_5 only in the case $W_b > a_{bz}$. In the marginal case of $W = a_z$ in either the unburned or burned side, the factor $\{(n + Uik + WD)^2 - (a_x ik + a_z D)^2\}$ in (65) degenerates to $(n + Uik - a_x ik)\{n + Uik + a_x ik + (a_z + W)D\}$, and the corresponding wave, with C_2 or C_5, is lost. In such a degenerate case, separate treatment is required [7].

A concrete representation of the perturbation velocity and magnetic field in each region is determined by (57)-(62). For the sake of simplicity, we set $U = 0$ in the unburned side. Given b_{uz}, the other perturbation variables in the unburned side are built so as to satisfy (57)-(62), resulting in

$$b_{uz} = B_z C_1 e^{kz} + B_z C_2 e^{A_2 z}, \tag{68}$$

$$b_{ux} = iB_z C_1 e^{kz} + \frac{i}{k} A_2 B_z C_2 e^{A_2 z}, \tag{69}$$

$$w_u = B_z \frac{n + W_u k}{B_{ux} ik + B_z k} C_1 e^{kz} + a_{uz} C_2 e^{A_2 z}, \tag{70}$$

$$u_u = B_z i \frac{n + W_u k}{B_{ux} ik + B_z k} C_1 e^{kz} + a_{uz} \frac{i}{k} \frac{-n + a_{ux} ik}{W_u - a_{uz}} C_2 e^{A_2 z}, \tag{71}$$

$$p_u = -\frac{\rho_u}{k} \frac{(n + W_u k)^2}{B_{ux} ik + B_z k} B_z C_1 e^{kz} - \rho_u \left\{ a_{uz}^2 + a_{ux} a_{uz} \frac{i}{k} \frac{-n + a_{ux} ik}{W_u - a_{uz}} \right\} C_2 e^{A_2 z}, \tag{72}$$

where

$$A_2 = \frac{-n + a_{ux} ik}{W_u - a_{uz}}. \tag{73}$$

In case $\mathrm{Re}[n] > 0$, $W_u < a_{uz}$ is required for $\mathrm{Re}[A_2] > 0$. Likewise, given b_{bz}, the perturbation variables on the burned side are found to be

$$b_{bz} = B_z C_3 e^{-kz} + B_z C_4 e^{A_4 z} + B_z C_5 e^{A_5 z}, \tag{74}$$

$$b_{bx} = -iB_z C_3 e^{-kz} + \frac{i}{k} A_4 B_z C_4 e^{A_4 z} + \frac{i}{k} A_5 B_z C_5 e^{A_5 z}, \tag{75}$$

$$w_b = \frac{n + U_b ik - kW_b}{B_{bx} ik - B_z k} B_z C_3 e^{-kz} - a_{bz} C_4 e^{A_4 z} + a_{bz} C_5 e^{A_5 z}, \tag{76}$$

$$u_b = -i \frac{n + U_b ik - W_b k}{B_{bx} ik - B_z k} B_z C_3 e^{-kz} - a_{bz} \frac{i}{k} A_4 C_4 e^{A_4 z} + a_{bz} \frac{i}{k} A_5 C_5 e^{A_5 z}, \tag{77}$$

$$\begin{aligned} p_b = &\frac{\rho}{k} \frac{(n + U_b ik - kW_b)^2}{B_x ik - B_z k} B_z C_3 e^{-kz} \\ &- \rho_b \left\{ a_{bz}^2 + a_{bx} a_{bz} \frac{i}{k} A_4 \right\} C_4 e^{A_4 z} - \rho_b \left\{ a_{bz}^2 + a_{bx} a_{bz} \frac{i}{k} A_5 \right\} C_5 e^{A_5 z}, \end{aligned} \tag{78}$$

where

$$A_4 = \frac{-n - U_b ik - a_{bx} ik}{W_b + a_{bz}}, \tag{79}$$

$$A_5 = \frac{-n - U_b ik + a_{bx} ik}{W_b - a_{bz}}. \tag{80}$$

The condition $\mathrm{Re}[A_4] < 0$ is always satisfied. The condition $\mathrm{Re}[A_5] < 0$ requires $W_b > a_{bz}$. Substituting from (55) and (56), the wave numbers A_2, A_4 and A_5 are expressed in terms of dimensionless variables as

$$\frac{A_2}{k} = \frac{-\bar{n} + \bar{a}_{ux} i}{1 - \bar{a}_{uz}}, \tag{81}$$

$$\frac{A_4}{k} = \left(-\bar{n} - \gamma i \frac{(1-\alpha)\bar{a}_{uz}^2}{\alpha - \bar{a}_{uz}^2} - \sqrt{\alpha} \bar{a}_{ux} i \frac{1 - \bar{a}_{uz}^2}{\alpha - \bar{a}_{uz}^2} \right) \Big/ (\alpha + \sqrt{\alpha} \bar{a}_{uz}), \tag{82}$$

$$\frac{A_5}{k} = \left(-\bar{n} - \gamma i \frac{(1-\alpha)\bar{a}_{uz}^2}{\alpha - \bar{a}_{uz}^2} + \sqrt{\alpha} \bar{a}_{ux} i \frac{1 - \bar{a}_{uz}^2}{\alpha - \bar{a}_{uz}^2} \right) \Big/ (\alpha - \sqrt{\alpha} \bar{a}_{uz}), \tag{83}$$

where $\bar{n} = n/kW_u$ is the dimensionless growth rate, $\bar{a}_{iu} = a_{iu}/W_u$ the dimensionless Alfvén speed and we have introduced $\gamma = B_{ux}/B_z$, the measure for the angle of the magnetic field from the normal on the unburned side.

4.4 Jump of perturbation fields

We are in a stage to substitute the solution of each region, written out in section 4.3, into the jump conditions (9)-(13) to connect them at the flame front. These jump conditions are no other than the conservation laws of the mass, the momentum and the magnetic flux and the induction equation in a region, of infinitesimal thickness, centered on the flame front [1, 11]. To spotlight the influence of the oblique external magnetic field, we employ Landau's assumption (40).

$$w_u - \frac{\partial \zeta}{\partial t} = 0, \tag{84}$$

and we ignore the gravity force and the surface tension. Recall that the basic state is constructed so as to comply with the jump conditions (51)-(54) to leading order. The remaining task is to satisfy the conditions to first order in perturbation amplitude.

The jump conditions (9)-(13) linearized for perturbed quantities become

$$U_b \frac{\partial \zeta}{\partial x} - w_b + \frac{\partial \zeta}{\partial t} = 0, \tag{85}$$

$$\left[B_x \frac{\partial \zeta}{\partial x} - b_z \right] = 0, \tag{86}$$

$$\rho_u W_u \left[u + W \frac{\partial \zeta}{\partial x} \right] - \frac{B_z}{\mu_0} [b_x] + \left(B_{ux} \frac{\partial \zeta}{\partial x} - b_{uz} \right) \frac{1}{\mu_0} [B_x] = 0, \tag{87}$$

$$\left[p + \frac{B_x b_x + b_z B_z}{\mu_0} \right] = 0, \tag{88}$$

$$[W b_x] - B_z [u] + \left(B_{ux} \frac{\partial \zeta}{\partial x} - b_{uz} \right) [U] = 0. \tag{89}$$

The quantities on both sides of the flame front $z = \zeta$ is evaluated at $z = \pm 0$, because the difference of the values at $z = \pm\zeta$ and those at $z = \pm 0$ add only second-order corrections.

After substituting the solution (68)-(72) and (74)-(78) into (85)-(89) and eliminating ζ by use of (84), we obtain a system of linear algebraic equations $\vec{M} \cdot (C_1, C_2, C_3, C_4, C_5)^T = \vec{0}$. We notice, by an analysis of the 5×5 matrix $\vec{M}$, that, without specifying $\bar{n}$, the rank of of $\vec{M}$ is four and that one of (85), (86) are (89) may be discarded. Retaining (86)-(89), we are left with 4 equations for 5 constants C_1, C_2, C_3, C_4 and C_5, amplitude of the waves. A separate treatment is needed, depending on whether W_u (W_b) is larger or smaller than the Alfvén speed a_{uz} (a_{bz}).

4.5 Growth rate

By removing one of the constants C_1, C_2, C_3, C_4 and C_5, on the physical ground, from the system (86)-(89) of linear algebraic equations, we coin 4×4 non-singular matrix $\hat{\vec{M}}$ from $\vec{M}$. The growth rate $\bar{n}$ is determined by requiring $\det \hat{\vec{M}} = 0$. We have to separately deal with four cases specified by $W_u \lessgtr a_{uz}$ and $W_b \lessgtr a_{bz}$ and, in addition, with the marginal cases specified by $W_u = a_{uz}$ or $W_b = a_{bz}$.

With a view to seeing how the KHI enters the DLI, we concentrate on two cases, super-Alfvénic and sub-Alfvénic in the both regions, with the detailed classification of the results left for a future paper. We begin with the both super-Alfvénic case, the case of smaller magnetic field, as a natural extension of the original DLI. It is to be remembered that $a_{bz} = \sqrt{\alpha} a_{uz}$ because of $B_{bz} = B_{uz}$.

4.5.1 Super-Alfvénic flame: $W_u > a_{uz}, W_b > a_{bz}$

The wave with amplitude C_2 diverges as $z \to -\infty$ because $\mathrm{Re}[A_2] < 0$, and we have to set $C_2 = 0$. The situation becomes the same as that of the classical DLI in the sense that the flow in the unburned region becomes irrotational. In the burned region, the vorticity, emerging from the flame front, is carried by the two Alfvén waves with their propagating velocity $W_b \pm a_{bz} (> 0)$. The coupled system (86)-(89) of linear algebraic equations is the matrix equation with 4×5 matrix $\{\vec{m}_1, \vec{m}_2, \vec{m}_3, \vec{m}_4, \vec{m}_5\}$ represented in the form of an array of columnar vectors $\vec{m}_i$ $(i = 1, \cdots, 5)$. When $C_2 = 0$, $\vec{m}_2$ is excluded, and, with $\bar{n}$ being unspecified, we are left with a non-singular square matrix $\hat{\bar{M}} = \{\vec{m}_1, \vec{m}_3, \vec{m}_4, \vec{m}_5\}$. The requirement of $\det \hat{\bar{M}} = 0$ produces a 5th-order polynomial equation for $\bar{n}$.

$$
\begin{aligned}
&\left\{\alpha^3 - 2\alpha^2(\bar{a}_{uz}^2 - i\bar{a}_{uz}\bar{a}_{ux} + \bar{n}) + \alpha(\bar{a}_{uz}^2 - i\bar{a}_{uz}\bar{a}_{ux} + \bar{n})^2 + (\bar{a}_{ux} + i\bar{a}_{uz}\bar{n})\right\} \\
&\times \Big\{(1+\bar{n})[\alpha^3 + (-i\bar{a}_{ux} + \bar{a}_{uz}\bar{n})^2] \\
&\quad + \alpha[2\bar{a}_{ux}^2 - 2i\bar{a}_{uz}\bar{a}_{ux}(-2 + \bar{a}_{uz}^2 - \bar{n})\bar{n} - \bar{n}(2\bar{a}_{uz}^4 + \bar{n} + \bar{n}^2) \\
&\qquad + \bar{a}_{uz}^2(1 + (3 + 2\bar{a}_{ux}^2)\bar{n} + 3\bar{n}^2 + \bar{n}^3)] \\
&\quad + \alpha^2[\bar{a}_{uz}^2(-1 + \bar{n}) + 2i\bar{a}_{uz}\bar{a}_{ux}\bar{n} - (1+\bar{n})(\bar{a}_{ux}^2 + (1+\bar{n})^2)]\Big\} = 0. \qquad (90)
\end{aligned}
$$

Fortunately, this 5th-order equation is factorized into second-order and third-order equations, in the same way as the classical DLI. The roots of the second-order equation are both trivial, with vanishing eigenfunctions $(C_1, C_3, C_4, C_5) = \vec{0}$. Thus the eigen-value equation is simplified into the third-order polynomial equation. If we take the limit $\bar{a}_{ux} \to 0$ of (90), we reproduce the result of Dursi [7] for the magnetic DLI subject only to the normal magnetic field,

$$
\begin{aligned}
&\left\{-\alpha^2 - \bar{n}^2 + \alpha(\bar{a}_{uz}^2 + 2\bar{n})\right\} \\
&\times \left\{(1+\bar{n})(\alpha^2 - \bar{n}^2) - \alpha(1 + 3\bar{n} - 2\bar{a}_{uz}^2\bar{n} + 3\bar{n}^2 + \bar{n}^3)\right\} = 0, \qquad (91)
\end{aligned}
$$

and if we take the limit $\bar{a}_{uz} \to 0$, we reproduce (42) with $\nu = 1$ and $\bar{g} = \bar{\sigma} = 0$, supporting for (90).

Figure 2 depicts the stability boundary of the magnetic DLI in the $\bar{a}_{uz}\alpha$-plane ($\bar{a}_{uz} < 1$, $1 < \alpha < 30$) for typical values $2, 3, 4$ and 6 of $\bar{a}_{ux}$. For a polynomial equation with complex constants, the Bilharz criterion applies to determine the neutral stability condition $\mathrm{Re}[\bar{n}] = 0$ [2, 8, 12]. The solid and broken lines correspond to the neutral stability curves for $\bar{a}_{ux} = 6, 4, 3, 2$ from above, except for the case of $\bar{a}_{ux} = 2$ where the neutral stability curve consists two lines. The region below the neutral curve gives parameter values for which the magnetic DLI is suppressed. For $\bar{a}_{ux} = 2$, the region between the two lines is the stability region. Given $\bar{a}_{uz}$, the stability range in the thermal expansion α is widened as $\bar{a}_{ux}$ or B_{ux} is increased. In the super-Alfvénic case in two dimensions, the magnetic DLI can be suppressed by imposing larger tangential magnetic field. When we turn off the tangential magnetic

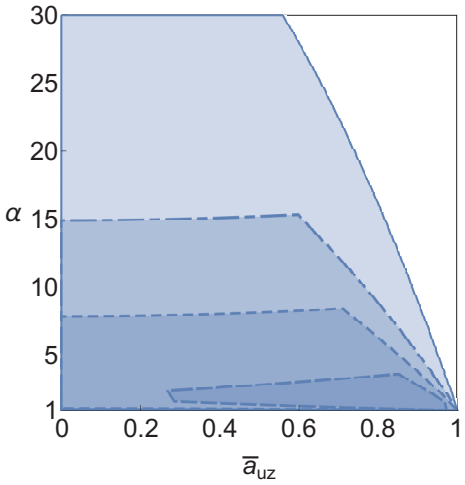


Fig. 2 Variation of stability boundary of the magnetic DLI in the $\bar{a}_{uz}\alpha$-plane ($1 < \alpha < 30$), with obliqueness of the external magnetic field, for the super-Alfvénic flame ($\bar{a}_{uz} = a_{uz}/W_u < 1$). The solid and broken lines correspond to the neutral stability curves for $\bar{a}_{ux} = 6, 4, 3, 2$ from above, except for the case of $\bar{a}_{ux} = 2$ where the neutral curve consists of two lines. The region below the neutral stability curve (the dark side) gives parameter values for which the magnetic DLI is suppressed. The stability region disappears for $\bar{a}_{ux} = 0$.

field ($\bar{a}_{ux} = 0$), the stability region disappears, a result being acceptable as a natural continuation of the DLI for a neutral fluid.

4.5.2 Sub-Alfvénic flame: $W_u < a_{uz}, W_b < a_{bz}$

The effect of the external magnetic field will clearly show up by increasing $B_{uz} = B_{bz}$ so that the Alfvén speed goes beyond the flow speed in each region: $a_{uz} > W_u$ and $a_{bz} > W_b$. In this case, $\mathrm{Re}[A_2] > 0$, and the Alfvén wave with amplitude C_2 is permitted, but, because of $\mathrm{Re}[A_5] > 0$, the Alfvén wave with amplitude C_5 is prohibited. We have to take $C_5 = 0$, while keeping C_2. With this, we reduce the connecting conditions across the flame front to the Matrix equation with non-singular matrix $\hat{\vec{M}} = \{\vec{m}_1, \vec{m}_2, \vec{m}_3, \vec{m}_4\}$. The Alfvén waves traveling away from the flame front are incorporated in both the unburned and burned regions. The vorticity baroclinically created in the flame front is carried, by these Alfvén waves, to both $z \to -\infty$ and $z \to \infty$.

Enforcement of $\det \hat{\vec{M}} = 0$ yields the dispersion relation determining the growth rate $\bar{n}$. As in the super-Alfvénic case, the resulting relation takes the form of 5th-order polynomial equation in $\bar{n}$, which is factorized into a second-order and a third-order polynomial equations. The two roots of the first factor

$$(\bar{n} + 1 - \bar{a}_{uz} - i\bar{a}_{ux}) \left\{ \alpha\sqrt{\alpha} - \sqrt{\alpha}(\bar{a}_{uz}^2 - i\bar{a}_{uz}\bar{a}_{ux} + \bar{n}) - i\bar{a}_{ux} + \bar{a}_{uz}\bar{n} \right\},$$

turn out to be trivial. As a consequence, we have only to solve the third-order polynomial equation. Compared with the the super-Alfvénic case, this equation is lengthy. Below we write down the coefficients of the same power of $\bar{n}$ order by order. The coefficient of $\bar{n}^3$ is

$$(1+\alpha)(\sqrt{\alpha}-\bar{a}_{uz})^2(\sqrt{\alpha}+\bar{a}_{uz}),$$

with no imaginary part. The real part of the coefficient of $\bar{n}^2$ is

$$-(1+\sqrt{\alpha})\sqrt{\alpha}(\sqrt{\alpha}-\bar{a}_{uz})\left\{\alpha^{3/2}-\alpha-2\alpha\bar{a}_{uz}+(1-\sqrt{\alpha})\bar{a}_{uz}^2+2\bar{a}_{uz}^3\right\},$$

and its imaginary part is

$$(1-\sqrt{\alpha})(\sqrt{\alpha}-\bar{a}_{uz})\left\{\alpha^{3/2}+(3+\alpha)\bar{a}_{uz}+\sqrt{\alpha}(1+2\bar{a}_{uz})\right\}\bar{a}_{ux}.$$

The real part of the coefficient of $\bar{n}$ is

$$\begin{aligned}&\alpha^{7/2}-\alpha^3(4+3\bar{a}_{uz})-3\bar{a}_{uz}\bar{a}_{ux}^2+\sqrt{\alpha}(1+2\bar{a}_{uz})\bar{a}_{ux}^2+\alpha^2\bar{a}_{uz}(-1+2\bar{a}_{uz}+\bar{a}_{uz}^2-\bar{a}_{ux}^2)\\&+\alpha^{5/2}(1+6\bar{a}_{uz}+3\bar{a}_{uz}^2+\bar{a}_{ux}^2)-\alpha^{3/2}\bar{a}_{uz}\left\{\bar{a}_{uz}(1+2\bar{a}_{ux}^2)+6\bar{a}_{uz}^2+4\bar{a}_{uz}^3+2\bar{a}_{ux}^2\right\}\\&+\alpha\left\{2\bar{a}_{uz}^4+2\bar{a}_{uz}^5+\bar{a}_{uz}^3(1+2\bar{a}_{ux}^2)+2\bar{a}_{uz}\bar{a}_{ux}^2\right\},\end{aligned}$$

and its imaginary part is

$$-2(1+\sqrt{\alpha})\sqrt{\alpha}\left\{\alpha^{3/2}+\alpha\bar{a}_{uz}^2+\bar{a}_{uz}^2(1+2\bar{a}_{uz})-\sqrt{\alpha}\bar{a}_{uz}(1+3\bar{a}_{uz}+\bar{a}_{uz}^2)\right\}\bar{a}_{ux}.$$

Finally, the real part of the coefficient of $\bar{n}^0$ is

$$(\sqrt{\alpha}-1)(\sqrt{\alpha}+\alpha)(\sqrt{\alpha}-\bar{a}_{uz})\left\{\alpha^2-\alpha(\bar{a}_{uz}^2+\bar{a}_{ux}^2)-(1+2\bar{a}_{uz})\bar{a}_{ux}^2\right\},$$

and its imaginary part is

$$\begin{aligned}&(\sqrt{\alpha}-1)\bar{a}_{ux}\left\{\alpha^3+2\alpha^{5/2}(1+\bar{a}_{uz})-2\alpha^{3/2}\bar{a}_{uz}(1+\bar{a}_{uz})-\bar{a}_{ux}^2\right.\\&\left.+\alpha^2(1-2\bar{a}_{uz}-5\bar{a}_{uz}^2-\bar{a}_{ux}^2)+\alpha\bar{a}_{uz}^2(1+2\bar{a}_{uz}+2\bar{a}_{uz}^2+2\bar{a}_{ux}^2)\right\}.\end{aligned}$$

Figure 3 depicts the stability boundary of the magnetic DLI in the $\bar{a}_{uz}\alpha$-plane ($\bar{a}_{uz}>1$, $1<\alpha<30$) for typical values $2,4$ and 6 of $\bar{a}_{ux}$. The solid line draws $\bar{a}_{uz}=\sqrt{\alpha}$. This line coincides with the critical line $\bar{a}_{ux}=1$, the left-hand side of which is the trans-Alfvénic regime with $W_u<a_{uz}, W_b>a_{bz}$ and the left-hand side of which is the sub-Alfvénic regime, and happens to give the neutral stability curve for all values of $\bar{a}_{ux}(>1)$. The broken lines draw the neutral stability curves for $\bar{a}_{ux}=2,4$ and 6 from inside. The region bounded by the solid line and the broken gives parameter values for which the magnetic DLI is suppressed. This is the region complementary to the banana-shaped region encircled by the dashed line, on the right-hand side of the solid line. For $\bar{a}_{ux}=1$, the whole right-hand side is the stability

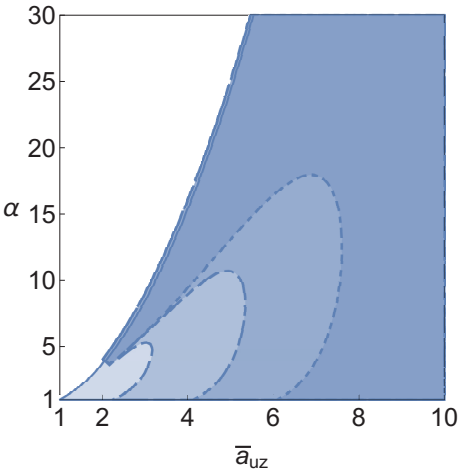


Fig. 3 Variation of stability boundary of the magnetic DLI in the $\bar{a}_{uz}\alpha$-plane ($1 < \alpha < 30$), with obliqueness of the external magnetic field, for the sub-Alfvénic flame ($\bar{a}_{uz} = a_{uz}/W_u > 1$). The solid line $\bar{a}_{uz} = \sqrt{\alpha}$ is the critical line $\bar{a}_{ux} = 1$ dividing the trans-Alfvénic regime (left) from the sub-Alfvénic regime (right) and gives the neutral stability curve for all values of $\bar{a}_{ux}(> 1)$, and the broken lines enclosing banana-shaped regions, with their vertices located at $(\bar{a}_{uz}, \alpha) = (1, 1)$, correspond to the neutral stability curves for $\bar{a}_{ux} = 2, 4, 6$ from inside. The stability parameters lie in the exterior to the banana-shaped region on the right hand side of the solid line. For $\bar{a}_{ux} = 1$, the whole region on the right-hand side of the solid line corresponds to the stability region.

region. The stability region shrinks as $\bar{a}_{ux}$ is increased. Comparing Figs. 2 and 3, the stability boundary exhibits opposite behavior near $(\bar{a}_{uz}, \alpha) = (1, 1)$. In the super-Alfvénic case (Fig. 2), given a moderate value of α, $\alpha = 3$ say, smaller values of $\bar{a}_{uz}(< 1)$ is required for stability, with its critical value larger for a larger value of $\bar{a}_{ux}$. The predominance of B_{ux} over B_{uz} is vital to stability, and the stability region expands as $\bar{a}_{ux}$ is increased. By contrast, in the sub-Alfvénic case (Fig. 3), for $\alpha = 3$, larger values of $\bar{a}_{uz}(> 1)$ is required for stability, with its critical value larger for a larger value of $\bar{a}_{ux}$.

The banana-shaped instability region is a peculiar feature intrinsic to the case of large imposed magnetic field. Given the value of $\bar{a}_{ux}$, it emanates from $(\bar{a}_{uz}, \alpha) = (\bar{a}_{ux}, 1)$. Figure 3 admits an interpretation that, for a moderate value of α, a distinct species of instability with $\bar{a}_{uz} \approx \bar{a}_{ux}$ parasitizes in the stability region of the DLI. Requisite for this instability is simultaneous application of $B_{ux}(\neq 0)$ and $B_{uz}(\neq 0)$, namely, of an oblique external magnetic field, a result drastically different from the case of the tangential magnetic field alone as discussed in section 3 (see also [7]). As emphasized in section 4.1, the tangential velocity discontinuity $[U]$ is induced only in the simultaneous presence of B_{ux} and B_{uz}. It is probable that this instability has an origin of the KHI. A scrutiny of the eigenfunction is required for convincing this. By increasing $\bar{a}_{uz}$ beyond the critical value depending on $\bar{a}_{uz}$, this instability disappears.

The tans-Alfvénic regime ($W_u < a_{uz}, W_b > a_{bz}$), located on the left-handed side of the solid line in Fig. 3, poses a difficult problem of shortage of the jump conditions [7]. As is readily seen from (66) and (67), all the three Alfvén waves with coefficients C_2, C_4 and C_5 are excitable, yet the number of the jump conditions (85)-(89) at the flame front remains the same. Dursi [7] somehow identified an unstable mode over the whole tras-Alfvénic regime, and we rely on this result.

5 Conclusion

In general, the magnetic field is expected to be an agent for stabilizing the instability of an interface, across which the density and/or the velocity are discontinuous (*cf*. [10, 17]). We have explored the influence of the external magnetic field on the Darrieus-Landau instability of a front of a premixed flame, a less investigated problem. Dursi [7] made a pioneering theoretical work on this problem. We have revisited this and have extended to include the effect of the surface tension and of the difference of the magnetic permeability between the unburned and burned gases. Furthermore, we have tackled the case of the oblique magnetic field, a problem left untouched.

To extend the analysis of [7], we have derived the jump conditions from the first principle of the magnetohydrodynamics following [1, 11], whereby we have incorporated the effect of the magnetic permeability disparity, in addition to the surface tension. Section 3 considered the situation in which only the tangential magnetic field is externally imposed. In section 3.4, we have disclosed that the DLI is enhanced for a diamagnetic fuel. For improving Landau's assumption, we have reconsidered the Markstein effect, and have corrected the previous result [7]. The analysis described in section 3 is limited to two dimensions and it is shown that sufficient strong tangential magnetic field is able to subside down the magnetic DLI. We have also carried out the analysis of three-dimensional stability, that is, the stability of a flat flame to disturbances with wavenumber $\vec{k} = (k_x, k_y)$. We can verify that the stabilizing effect of tangential magnetic field $\vec{B}$ is completely lost when it is orthogonal to the wavenumber: $\vec{B} \cdot \vec{k} = 0$.

In section 4, we have dealt with the oblique external magnetic field, the simultaneous application of both the normal and the tangential magnetic field. Only by the existence of the both fields, the discontinuity of the tangential velocity is induced, an unusual situation when the basic flow penetrates the interface. The presence of the tangential-velocity discontinuity offers the situation where the KHI coexists with the DLI. In section 4.5.2, we have captured this symptom for sufficiently strong magnetic field that the Alfvén speed is faster than that of the basic normal flow on the both sides.

This paper has treated only the limited cases, and a substantial effort will be required to grasp an overall perspective of the magnetic field effect. A separate treatment is necessary for the tras-Alfvénic flame, as touched upon at the end of section 4.5.2, and for the marginal cases where the Alfvén speed coincides with the normal-

flow speed in the unburned and the burned regions. Dependence of the magnetic DLI on the effect of the gravity force, the surface tension and the Markstein effect is left for a future study. For astrophysical phenomena as exemplified by supernova explosions [9], the compressibility effect may be called into play. The Markstein effect corrected by the compressibility effect [19] is worth testing.

Acknowledgements We are grateful to Snezhana Abarzhi, Alexander Klimenko, Kaname Matsue, Saleh Tanveer and Keigo Wada for helpful advices and invaluable comments. Y.F. was supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (grant no.19K03672).

References

1. Abarzhi, S.I., Fukumoto, Y., Kadanoff, L.P.: Stability of a hydrodynamic discontinuity. Phys. Scr. **90**, 018002 (2015)
2. Bilharz, H.: Bemerkung zu einem Satze von Hurwitz. Z. Angew. Math. Mech. **24**, 77-82 (1944)
3. Class, A.G., Matkowsky, B.J., Klimenko, A.Y.: A unified model of flames as gasdynamic discontinuities. J. Fluid Mech. **491**, 11-49 (2003)
4. Class, A.G., Matkowsky, B.J., Klimenko, A.Y.: Stability of planar flames as gasdynamic discontinuities. J. Fluid Mech. **491**, 51-63 (2003).
5. Clavin, P., Searby, G.: Combustion Waves and Fronts in Flows. Cambridge University Press (2016)
6. Darrieus, G.: unpublished works presented at La Technique Moderne (1938)
7. Dursi, L.J.: The linear instability of astrophysical flames in magnetic fields. Astrophys. J. **606**, 1039-1056 (2004)
8. Frank, E.: On the real parts of the zeros of complex polynomials and applications to continued fraction expansions of analytic functions. Trans. Amer. Math. Soc. **62**, 272-283 (1947)
9. Hillebrandt, W., Niemeyer, J.C.: Type IA supernova explosion models. Ann. Rev. Astron. Astrophys. **38**, 191-230 (2000)
10. Hosking, R.J., Dewar, R.L.: Fundamental Fluid Mechanics and Magnetohydrodynamics. Springer Verlag (2016)
11. Ilyin, D.V., Fukumoto, Y., Goddard III, W.A., Abarzhi, S.I.: Analysis of dynamics, stability, and flow fields'structure of an accelerated hydrodynamic discontinuity with interfacial mass flux by a general matrix method. Phys. Plasmas **25**, 112105 (2018)
12. Kirillov, O.N.: Nonconservative Stability Problems of Modern Physics. Walter de Gruyter GmbH, Berlin/Boston (2013)
13. Landau, L.D.: On the theory of slow combustion. Acta Phys. (USSR) **19**, 77-85 (1944)
14. Landau, L.D., Lifshitz, E.M.: Fluid Mechanics: Course of Theoretical Physics Vol. 6, 2nd edn., p. 488 Butterworth-Heinemann (1987)
15. Markstein, G.H.: Experimental and theoretical studies of flame-front stability. J. Aero. Sci. **18**, 199–209 (1951)
16. Matalon, M., Matkowsky, B.J.: Flames as gasdynamic discontinuities. J. Fluid Mech. **124**, 239-259 (1982)
17. Matsuoka, C., Nishihara, K., Sano, T.: Nonlinear interfacial motion in magnetohydrodynamic flows High Energy Density Phys. **31**, 19-23 (2019)
18. Shu, F.H.: Gas Dynamics, Mill Valley: University Science Books (1992)
19. Wada. K., Fukumoto, Y.: Compressibility effect on Markstein number in long-wavelength approximation. 2019 Matrix annals., to appear (2020)

Numerical Study of Crystal Growth in Reaction-Diffusion Systems using Front Tracking

Saurabh Joglekar and Xiaolin Li

Abstract We study the crystal growth in a Reaction-Diffusion System for the generic reaction $A + B \longrightarrow C$. Reactants A and B react to form the product C which then undergoes phase transition. We have used the Lagrangian Front Tracking to explicitly track the crystal surface. The evolution of the concentrations of A, B and C is described by a system of three partial differential equations. This system is solved using finite difference method. Main focus of the study is on observing the effects of different parameters on the crystal growth, namely the diffusion coefficients, homogeneous reaction constant, heterogeneous reaction constant and the equilibrium concentration.

Key words: Reaction-Diffusion Equations, Crystal Growth, Lagrangian Front Tracking

1 Introduction

Xiaolin Li et al [1] have studied a single component Reaction-Diffusion system through front tracking without consideration of advection, in which the reaction term is replaced by precipitation term at the fluid-solid interface. The governing equations for the solute concentration $C = C(\vec{x}, t)$ are as follows:

$$\frac{\partial C}{\partial t} = D\nabla^2 C, \qquad \text{for } \vec{x} \in \Omega \tag{1}$$

Here Ω is the ambient region containing solute and D is diffusion coefficient. At the fluid-solid interface $\partial\Omega$, the front growth is governed by:

Saurabh Joglekar and Xiaolin Li
Dept. of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA
e-mail: saurabh.joglekar@gmail.com, e-mail: xiaolin.li@stonybrook.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_27

$$D\frac{dC}{dn}(\vec{x_s}) = k(C(\vec{x_s}) - C_e) \tag{2}$$

where k is the reaction rate per unit area for the solute from the liquid phase to precipitate onto the solid phase at the interface, C_e is the equilibrium concentration and $C(\vec{x_s})$ is the local concentration of solute at the interface. Interface is propagated with the normal velocity,

$$v_n = \frac{D}{\rho_s}\frac{dC}{dn}(\vec{x_s}) \tag{3}$$

where ρ_s is the density of solid phase. Interface growth and the dendritic structure of the precipitate have been studied at different Damkohler numbers. Front tracking is well suited to dendritic structures at large Damkohler numbers where high resolution is necessary.

Tartakovsky et al [3] have studied multi-component Reaction-Diffusion systems for the chemical reaction $A+B \longrightarrow C_{(aq+solid)}$, on two different spatial scales, Pore-scale and Darcy-scale. Smoothed particle hydrodynamics (SPH) has been applied to carry out hybrid simulations on two different spatial scales. Let $A(\vec{x},t)$, $B(\vec{x},t)$, $C(\vec{x},t)$ and D_a, D_b, D_c be the concentrations and diffusion coefficients of components A, B and C in solute phase. Let k and k_{AB} be heterogeneous and homogeneous reaction rates, and ρ_s be the density of solid phase. Then the Pore scale model satisfies following system of equations:

$$\frac{\partial A}{\partial t} = \nabla\cdot(D_a\nabla A) - k_{AB}AB \tag{4}$$

$$\frac{\partial B}{\partial t} = \nabla\cdot(D_b\nabla B) - k_{AB}AB \tag{5}$$

$$\frac{\partial C}{\partial t} = \nabla\cdot(D_c\nabla C) + k_{AB}AB - k\int_F H(C-C_{eq})\delta(\vec{x}-\vec{x}_f)d\vec{x}_f \tag{6}$$

where $H(x)$ is the Heaviside step function and the integration is taken over the whole fluid-solid interface. Soluble precipitate C follows the first order kinetic reaction model on the fluid-solid interface,

$$D_c\frac{dC}{dn} = k(C-C_{eq}) \tag{7}$$

The interface advances into the liquid with normal velocity,

$$v_n(\vec{x_s}) = \frac{D_c}{\rho_s}\nabla C\cdot\vec{n} \tag{8}$$

Simulations start with a crystal seed already present in the domain. Hence, although the same equations govern the formation of Liesegang Patterns, nucleation theories have not been considered in Tartakovsky et al [3].

2 Front Tracking

The "Front" is defined as the boundary point between two regions containing a sharp discontinuity of a physical variable, e.g. density, concentration, viscocity etc. Theoretically, the function representing the physical variable is not differentiable at a point of discontinuity. This problem can be handled by using the integral form of the governing equations. However, if the numerical scheme is of low order, then the front diffuses quickly losing its sharpness. On the other hand, a high order numerical scheme may cause numerical oscillations near the front and reduce the high order of accuracy near the region[40]. To solve these diffculties, there exist two main strategies, namely front-capturing and front-tracking.

The main idea of front capturing is to use a high order scheme and use artificial viscosity around the front to diffuse it slightly to avoid oscillations. Front capturing works well for shocks but does not work very well for contact discontinuities[40]. It also requires high resolution.

Second appoach is front tracking in which the front is represented by hypersurface elements (line segments in 2D and triangles in 3D). This approach is best suited to sharp discontinuities.

We apply the front tracking method and the FronTier code to study crystal formation in a generic 3 component reaction-diffusion system. We use front tracking to track the position of the front where there is a discontinuity in solute concentration. We then use finite difference scheme (Crank-Nicolson) to update the concentrations of the reactants and the product which are still in the liquid phase.

The front tracking method treats the moving interface as an interior boundary and applies finite difference method to each subdomain where concentration fields are smooth.

We use the FronTier library to implememnt the front tracking and crystal growth. The functions implemented in the library can be classified as follows: [1]

1. **Initialization:** Initialization functions are capable of initializing the problem parameters as well as geometrical parameters for the computations such dimension, domain, computational grid and boundary conditions. This is done through the input routines. Initialization of the interface is also done through these functions as well as the front velocity initialization.
2. **Query Functions:** Query functions are used to obtain information about the front interface such as vertex coordinates, hypersurface elements (bonds in 2D and triangular surface elements in 3D), access to the manifold (hypersurface), tangents and normals to the surface elements etc.
3. **Propagation Control Functions:** These functions include advancement of the front interface, redistribution and bifurcation.
4. **Front and Subdomain Interaction Functions:** These include the functions which couple the PDE solvers with the front interface functions. These functions can be used to obtain information like the nearest grid points, values of the physical variables in a cell/grid point near the interface etc.

5. **Output and Data Saving Functions:** These functions mainly deal with the data output which is used for visualization of the simulations. The compatible file types include VTK for VisIt, Paraview, Geomview, HDF and GD packages. These functions also have the capability to halt and/or restart the program run from a specific time or time-step.

3 Numerical Method

Consider a reaction-diffusion system given by $nA + mB \longrightarrow C$ and let a seed be already present inside the computational domain. The evolution of concentrations is governed by the following system of equations:

$$\frac{\partial A}{\partial t} = \nabla \cdot (D_A \nabla A) - k_{AB} A^n B^m \tag{9}$$

$$\frac{\partial B}{\partial t} = \nabla \cdot (D_B \nabla B) - k_{AB} A^n B^m \tag{10}$$

$$\frac{\partial C}{\partial t} = \nabla \cdot (D_C \nabla C) + k_{AB} A^n B^m - k \int_F H(C - C_{eq}) \delta(\vec{x} - \vec{x}_f) d\vec{x}_f \tag{11}$$

where $A(\vec{x},t)$, $B(\vec{x},t)$, $C(\vec{x},t)$ are normalized concentrations, D_A, D_B, D_C are diffusion coefficients, $k_{AB} > 0$ is the rate coefficient of homogeneous reaction (liquid phase), $k > 0$ is the rate coefficient of heterogeneous reaction (precipitation), $\vec{x_f}$ is a point on fluid-solid interface and C_{eq} is the equilibrium concentration. $H(\cdot)$ represents the Heaviside step function and $\delta(\cdot)$ represents the dirac-delta function. The integration is taken over the whole fluid-solid interface.

The fluid-solid interface propagates with the normal velocity:

$$v_n(\vec{x_s}) = -\frac{1}{\rho_s} D_C \frac{dC}{dn} \tag{12}$$

where ρ_s is the crystal density and $\frac{dC}{dn}$ is the normal derivative of the concentration $C(\vec{x},t)$.

The FronTier code has the ability to detect if a cell contains liquid phase or solid phase. We use this capability, and for purely liquid phase, we note that $k \int_F H(C - C_{eq}) \delta(\vec{x} - \vec{x_f}) d\vec{x_f} = 0$. Assume that the diffusion coefficients stay constant throughout the liquid phase. Then, for the computational cells containing only the liquid phase, the equations are reduced to:

$$\frac{\partial A}{\partial t} = D_A \nabla^2 A - k_{AB} A^n B^m \tag{13}$$

$$\frac{\partial B}{\partial t} = D_B \nabla^2 B - k_{AB} A^n B^m \tag{14}$$

$$\frac{\partial C}{\partial t} = D_C \nabla^2 C + k_{AB} A^n B^m \tag{15}$$

Any high order finite difference scheme may be used to solve this system of equations. In the present work, we use Crank-Nicolson scheme.

When a cell contains purely solid phase, we assume that there is neither reaction nor diffusion taking place. Thus there is no need to solve the system for the cells containing purely solid phase.

When a cell contains fluid-solid interface, both liquid and solid phases are present inside the cell. At a point on the interface, $k \int_F H(C - C_{eq}) \delta(\vec{x} - \vec{x_f}) d\vec{x_f} = kH(C - C_{eq})$. To update the concentrations at the grid point of a cell containing the interface, we introduce ghost points in the direction opposite to that of the interface and then solve the system using finite differences. The ghost points are introduced to maintain second order accuracy of the finite difference scheme.

Once the concentrations are updated, we propagate the fluid-solid interface by the methods described by Li et al.[1]. To update the concentrations at a point on the interface, we assume that the solute concentrations at the fluid-solid interface are $A_s = B_s = 0$. Thus, for the $(n+1)$-th time step, the discretized equation for C_s is given by:

$$\frac{C_s^{(n+1)} - C_s^{(n)}}{\Delta t} = \left(D_C \frac{C_{s+h}^{(n+1)} - C_s^{(n+1)}}{h} - kH(C_s^{(n+1)} - C_{eq}) \right) \cdot \frac{2}{h} \tag{16}$$

where h is the spatial step in normal direction. The superscripts denote time step.

Once the concentration of C is updated, it is also necessary to update the concentrations of A and B in the region near the front. To do this, we first approximate the area swept by the moving front. The situation is shown more precisely in the following figure.

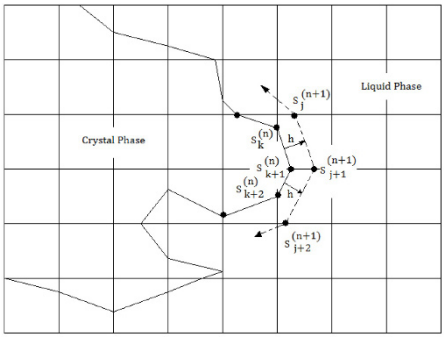


Fig. 1: Advancing the Front and updating concentations

For each segment of the front at time step (n), we approximate the area swept by that segment by the length of the segment times the spatial step in the normal

direction, h. i.e. $\Delta V_k^{(n)} = |S_k^{(n)} - S_{k+1}^{(n)}| \cdot h$. The mass of A and B contained in this area is given by $A \cdot \Delta V_k^{(n)}$ and $B \cdot \Delta V_k^{(n)}$ respectively. A and B in this case are taken to be the interpolated concentrations at the center of the computational cell in which the segment of the front is located. In case the center already lies inside the solid phase, we approximate A and B to be the concentrations at the nearest grid point. We then redistribute this mass equally among the nearest grid points at time step $(n+1)$.

4 Numerical Results

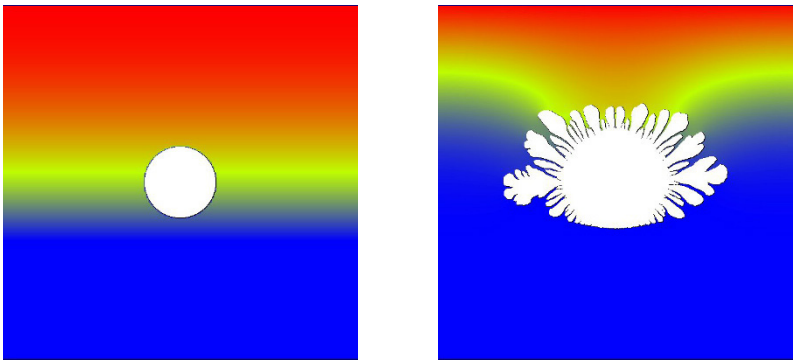
In this section, we present the numerical results which show the effects of different parameters on the crystal growth. The parameters which control the reaction-diffusion system described by equations (9) to (11) are D_A, D_B, D_C and k_{AB}, k.

We set the computational domain to be the square $[0,1] \times [0,1]$. The boundary conditions used for testing are $A(x,0,t) = B(x,1,t) = 0$ and $A(x,1,t) = B(x,0,t) = 1$.

4.1 Effects of k_{AB} and k

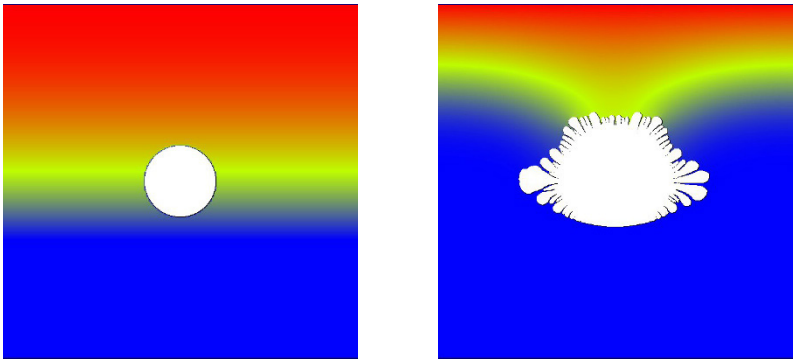
We first explore the effects of k_{AB} and k on the crystal growth. Initial concentrations are assumed to be uniformy distributed along the y-axis.

It can be observed from the following tests that the dendritic growth is pronounced when k is high. k_{AB} has negligible effect on the dendritic growth. It will also be observed that in general, the direction in which dendrites grow is controlled by $\frac{D_A}{D_B}$. This point will be further explored in the next section.



(a) $t = 0$ (b) $t = 2.755$

Fig. 2: Parameters are $k_{AB} = 150$, $k = 800$ and $D_A = 0.3$, $D_B = 0.7$, $D_C = 0.5$

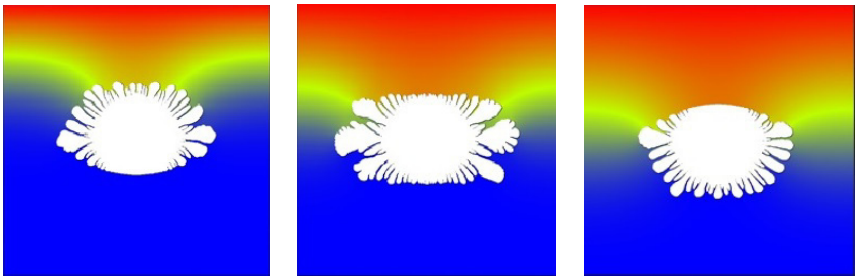


(a) $t = 0$ (b) $t = 2.755$

Fig. 3: Parameters are $k_{AB} = 1500$, $k = 200$ and $D_A = 0.3$, $D_B = 0.7$, $D_C = 0.5$

4.2 Effect of the diffusivities

As mentioned in the previous section, the direction in which the dendrites grow is controlled by the value of $\frac{D_A}{D_B}$. The effects are explored in this section. The initial conditions vary for each simulation. However, initial conditions are found to have negligible effect on the direction of growth.



(a) $D_A = 0.3$, $D_B = 0.7$ (b) $D_A = 0.5$, $D_B = 0.5$ (c) $D_A = 0.7$, $D_B = 0.3$

Fig. 4: Parameters are $k_{AB} = 150$, $k = 800$ and $D_C = 0.5$ Each image is taken at $t = 2.68$

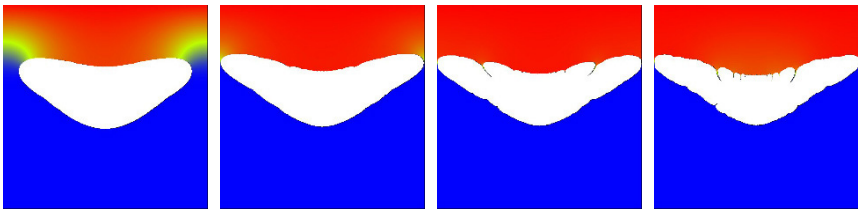
4.3 Effect of the Damkohler Number

The Damkohler Number, d, is defined by $d = \frac{kL}{D_C}$ [1, 2] where k is the heterogeneous reaction constant, L is the characteristic length and D_C is the diffusion

coefficient for the product C. Damkohler number is closely tied with the dendritic growth of the crystal. High Damkohler number produces high dendritic structure and vice-versa. In this section, we provide numerical results which show that the dendritic growth in a Reaction-Diffusion System for $A + B \longrightarrow C$ occurs only when the Damkohler number is higher than a threshold value. In most of our simulations, the threshold was in the range of 60 to 80. Although it is diffcult to predict the exact value of the threshold, we mention that traces of dendritic structures started to appear for $d = 60$ and they were well formed for $d = 80$. For lower Damkohler numbers, the crystal growth was smooth without any dendrites. The direction of the growth was still controlled by $\frac{D_A}{D_B}$. The initial shape of the seed had no effect on the threshold value.

The computational domain is $[0,1] \times [0,1]$. Reactants A and B are initially separated at $y = 0.5$. Other parameters are as follows: $k_{AB} = 1500$, $D_A = 0.3$, $D_B = 0.7$. We wish to mention that the simulations were carried out for a range of Damkohler Numbers, in particular for $d = 0.1, 0.5, 1, 5, 10, 20, 40, 60, 80, 160$. For small values of d, the crystal growth was not qualitatively different, the only signi

cant difference being the amount of growth in a given time. We also carried out the simulations for a range of values of D_A and D_B. Numerical results showed difference in the direction of crystal growth.

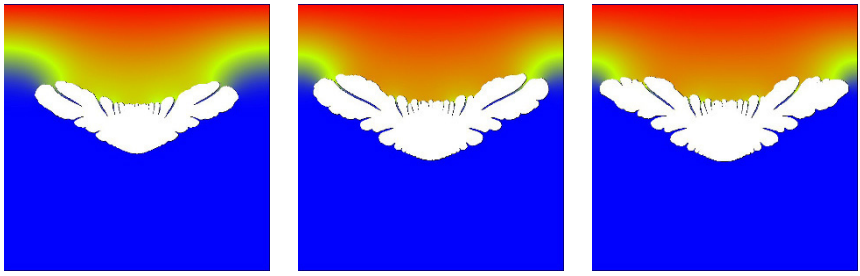


(a) $d = 20$ (b) $d = 40$ (c) $d = 60$ (d) $d = 80$

Fig. 5: Effects of the Damkohler Number. Parameters are $k_{AB} = 1500$, $D_A = 0.3$ and $D_B = 0.7$. Each image is taken at $t = 2.57$. Initial seed is circular.

4.4 Effect of the equilibrium concentration

The supersaturation theory asserts that the deposition of mass occurs only when $C_s > C_{eq}$ where C_s is the concentration of C at a point on the fluid-solid interface and C_{eq} is the equilibrium concentration. The theory also asserts that once the concentration C attains the equilibrium concentration, the deposition occurs instantaneously. It is natural to expect that this process will have effect on the dendritic growth. This is confirmed by the following numerical results. Lower equilibrium concentration produces more dendritic growth and vice-versa, when all other parameters are held constant. Parameters used are $k_{AB} = 1500$, $k = 100$, $D_A = 0.3$, $D_B = 0.7$, and $D_C = 0.5$. Other computational setup is the same as previous sections.



(a) $C_{eq} = 0.1$ (b) $C_{eq} = 0.01$ (c) $C_{eq} = 0.001$

Fig. 6: Effects of the Equilibrium Concentration. Parameters are $k_{AB} = 1500$, $k = 100$, $D_A = 0.3$, $D_B = 0.7$, and $D_C = 0.5$. Each image is taken at $t = 2.629$. Initial seed is circular.

5 Summary and Conclusions

In the present work involving the crystal growth in a reaction-diffusion system containing three components, we examined the effects of parameters D_A, D_B, D_C and k_{AB}, k. We found that the dendritic growth is controlled predominantly by the heterogeneous reaction constant k, and the homogeneous reaction constant k_{AB} has no effect on the dendrites. We also observed that the direction in which the dendrites grow is controlled by $\frac{D_A}{D_B}$. Numerical simulations show that the Damkohler number produces dendritic growth only if it's value is higher than some threshold. The threshold for the tests done in the presenst study appears to be in the range of $d = 60$ to $d = 80$. The equilibrium concentration, C_{eq} also has effect on the dendrites with lower C_{eq} being responsible for higher dendritic growth and vice-versa.

Acknowledgements This work is supported in part by the US Army Research Office under the award W911NF-14-1-0428 and the ARO-DURIP Grant W911NF-15-1-0403. We would also like to thank Roman Samulyak and Yangang Liu for their suggestions. Also, thanks to Yijing Hu for helpful discussions and earlier work on the problem of crystal growth.

References

1. Study of crystal growth and solute precipitation through front tracking method, *Acta Mathematica Scientia, Volume 30, Issue 2, March 2010, Pages 377-390,* Xiaolin Li, James Glimm, Xiangmin Jiao, Charles Peyser, Yanhong Zhao
2. Numerical simulations of phase transition problems with explicit interface tracking, *Chemical Engineering Science 128(2015) 92-108,* Yijing Hu, Qiangqiang Shi, Valmor F. De Almeida, Xiaolin Li
3. Hybrid Simulations of Reaction-Diffusion Systems in Porous Media, *SIAM J. Sci. Comput. 30 2799 (2008),* A. M. Tartakovsky, D. M. Tartakovsky, T. D. Scheibe, and P. Meakin

4. Reaction-Diffusion Cellular Automata Model for the Formation of Liesegang Patterns, *Phys. Rev. Lett. 72, 1384-1387 (1994),* B. Chopard, P. Luthi and M. Droz
5. Wilhelm Ostwald, *Lehrbuch der Allgemeinen Chemie II/2,* W. Engelmann, Leipzig (1896-1902)
6. Patterns Produced by Precipitation at a Moving Reaction Front, *Phys. Rev. Lett. 57, 275-278 (1986),* G. T. Dee
7. Measurements and Hypothesis on Periocic Precipitation Processes, *J. Phys. Chem. 91, 6300-6308 (1987),* M. E. LeVan and J. Ross
8. Derivation of the Matalon-Packter law for Liesegang paterns, *J. Chem. Phys. 109, 9479-9486 (1998),* T. Antal, M. Droz, J. Magnin, Z. Racz and M. Zrnyi
9. The Liesegang Phenomenon. I. Sol Protection and Diffusion, *J. Colloid Sci. 10, 46-61 (1955),* R. Matalon and A. Packter
10. Pattern Formation in a New Class of Precipitation Reactions, *Ph.D. Thesis, Peter Hantz,* UNIVERSITÉ DE GENÈVE
11. Chemistry and Crystal Growth, *Angew. Chem. Int. Ed. Engl. 33, 143-162 (1994),* Jurg Hulliger
12. Dendrites, Viscous Fingers, and the Theory of Pattern Formation, *Science 243, 1150-1156 (1989),* J. S. Langer
13. Competition between kinetic and surface tension anisotropy in dendritic growth, *Eur. Phys. J. B 16, 337-344 (2000),* T. Ihle
14. Diffusion-Limited Aggregation: A Model for Pattern Formation, *Physics Today 53, 36-41 (2000),* Thomas C. Halsey
15. Mathematical Analysis of the Formation of Periodic Precipitates, *J. Colloid Sci. 5, 85-97 (1950),* C. Wagner
16. Periodic Precipitation Patterns in the Presence of Concentration Gradients 1., *J. Phys. Chem. 86, 4078-4087 (1982),* Stefan C. Muller, Shoichi Kai and John Ross
17. The Formation of Liesegang Rings as a Periodic Coagulation Phenomenon, *Journal of the Chemical Society 1928/II, 2714-2727 (1928),* Ernest S. Hedges and Rosalind V. Henley
18. Mechanism of chemical instability for periodic precipitation phenomena, *J. Chem. Phys. 60, 3458-3465 (1974),* Michael Flicker and John Ross
19. The Concentration Distribution in the Gel before the Periodic Precipitation, *Memoirs of the Faculty of Science, Kyushu University, Series C Chemistry, 5, 33-42 (1962),* Hiroshige Higuchi and Ryohei Matura
20. Nucleation and Spinodal Decomposition, *Solid State Phenomena 56, 67-106 (1997),* L. Granasy
21. Phase Transitions and Critical Phenomena, *vol. 8, Academic Press, London (1989),* C. Domb and J. L. Lebowitz (editors)
22. Formation of Liesegang Patterns, *Physica A 274, 50-59 (1999),* Zoltan Racz
23. Pattern formation induced by ion-selective surfaces: Models and simulations, *J. Chem. Phys. 123, 034707 (2005),* Szabolcs Horvt and Pter Hantz
24. Properties of the reaction front in an $A + B \longrightarrow C$ type reaction-diffusion process, *Physical Review A 1988, Volume 38, Number 6: 3151-3154,* L. Gal and Z. Racz
25. Properties of the asymptotic $nA + mB \longrightarrow C$ reaction-diffusion fronts, *Eur. Phys J. B 17 (2000): 673-678,* J. Magnin
26. Dynamic multiscaling of the recation-diffusion front for $mA + nB \longrightarrow C$, *Physical Review E, October 1995, Volume 52, Number 4,* S. Cornell, Z. Koza and M. Droz
27. Reaction front for $A + B \longrightarrow C$ diffusion-reaction systems with initially separated reactants, *Physical Review A, July 1992, Volume 46, Number 2,* H. Larralde, M. Araujo, S. Havlin and H. Stanley
28. Steady-State Reaction-Diffusion Front Scaling for $mA + nB \longrightarrow C$, *Physical Review Letters, June 1993, Volume 70, Number 24,* S. Cornell and M. Droz
29. Asymptotic behaviour of initially separated $A + B_{(static)} \longrightarrow C$ reaction-diffusion systems, *Physica A 240 (1997) 622-634,* Z. Koza
30. Reaction-Diffusion fronts in systems with concentration-dependent diffusivities, *Physical Review E 74, 036103 (2006),* P. Polanowski and Z. Koza

31. Reaction fronts in reversible $A + B \rightleftharpoons C$ reaction-diffusion systems, *Physica A 330 (2003) 160-166,* Z. Koza
32. Reversible and irreversible reaction fronts in two competing reaction system, *Nuclear Instruments and Methods in Physica Research B 186 (2002) 161-165,* M. Sinder, H. Taitelbaum, J. Pelleg
33. Asymptotic expansion for reversible $A + B \longleftrightarrow C$ reaction-diffusion process, *Physical Review E 66, 011103 (2002),* Z. Koza
34. The Long-time Behavior of Initially Separated $A + B \longrightarrow C$ Reaction-Diffusion Systems with Arbitrary Diffusion Constants, *J. Stat. Phys. 85, 179-191(1996),* Z. Koza
35. Some Properties of the $A + B \longrightarrow C$ Reaction-Diffusion System with Initially Separated Components, *Journal of Statistical Physics, Vol. 65, Nos. 5/6, 1991,* H. Taitelbaum, S. Havlin, J. Kiefer, B. Trus, and G. Weiss
36. Numerical analysis of reversible $A + B \longleftrightarrow C$ reaction-diffusion systems, *Eur. Phys. J. B 32, 507-511(2003),* Z. Koza
37. Simulation study of reaction fronts, *Physical Review A, December 1990, Volume 42, Number 12,* Z. Jiang and C. Ebner
38. Refined simulations of the reaction front for diffusion-limited two-species annihilation in one dimension, *Physical Review E, May 1995, Volume 51, Number 5,* S. Cornell
39. Role of fluctuations for inhomogeneous reaction-diffusion phenomena, *Physical Review A, Volume 44, Number 8, Oct. 1991,* S. Cornell, M. Droz, B. Chopard
40. A Front-Tracking Method for Viscous, Incompressible, Multi-fluid Flows, *Journal of Computational Physics 100, 25-37 (1992),* Salih Ozen Unverdi, Gretar Tryggvason
41. Numerical simulation of dendritic solidification with convection: Two-Dimensional Geometry, *Journal of Computational Physics, Volume 180, Issue 2, 10 August 2002, Pages 471-496,* Nabeel Al-Rawahi, Gretar Tryggvason
42. Numerical simulation of dendritic solidification with convection: Three-dimensional flow, *Journal of Computational Physics, Volume 194, Issue 2, 1 March 2004, Pages 677-696,* Nabeel Al-Rawahi, Gretar Tryggvason
43. Front tracking for gas dynamics, *Journal of Computational Physics, Volume 62, Issue 1, January 1986, Pages 83-110,* I.-L Chern, J. Glimm, O. Mcbryan, B. Plohr, S. Yaniv
44. A simple package for front tracking, *Journal of Computational Physics 213:613-628, 2006,* Jian Du, Brian Fix, James Glimm, Xicheng Jia, Xiaolin Li, Yunhua Li, and Lingling Wu
45. A level set simulation of dendritic solidification with combined features of front-tracking and fixed-domain methods, *Journal of Computational Physics, 211:36-63, 2006,* Lijian Tan and Nicholas Zabaras
46. Front tracking in two and three dimensions, *Comput. Math. Appl., 35(7):1-11, 1998,* J. Glimm, M. J. Graham, J. W. Grove, X.-L. Li, T. M. Smith, D. Tan, F. Tangerman, and Q. Zhang
47. Frontier and applications to scientific and engineering problems, *Proceedings of International Congress of Industrial and Applied Mathematics, pages 1024507 - 1024508, 2008,* W. Bo, B. Fix, J. Glimm, X. L. Li, X. T. Liu, R. Samulyak, and L. L. Wu
48. Diamond crystsal growth by plasma chemical vapor deposition, *Journal of Applied Physics, 63:1744-1748, 1988,* C. P. Chang, D. L. Flamm, D. E. Ibbostson, and J. A. Mucha
49. Precipitation and dissolution of reactive solutes in fractures, *Water Resources Research, 34:457-470, 1998,* Peter Dijk and Brian Berkowitz
50. Simulation of dissolution and precipitation in porous media, *J. Geophys. Res., 108:2505, 2003,* Q. Kang, D. Zhang, and S. Chen
51. Numerical modeling of ice deposition, *Journal of the Atmospheric Sciences, 28:226-237, 1970,* L. R. Koenig
52. An experimental investigation of nonaqueous phase liquid dissolution in saturated subsurface systems: Steady state mass transfer rates, *Water Resources Research, 28:2691-2705, 1992,* Susan E. Powers, Linda M. Abriola, and Walter J. Weber JR

Numerical Study of Center of Reaction Front for Reaction-Diffusion System $nA + mB \longrightarrow C$ with Arbitrary Diffusivities

Saurabh Joglekar and Xiaolin Li

Abstract We study the movement of the center of reaction front in the reaction-diffusion system $nA + mB \longrightarrow C$ for arbitrary diffusivities ($D_a \neq D_b$). We present numerical evidence that $x_f(t) \propto \sqrt{t}$ for all $t \in (0, \infty)$. Numerical experiments are carried out for $(n,m) = (1,1), (1,2), (2,1)$ and $(2,2)$ and for various $\frac{D_a}{D_b}$. Finite difference method is used. Empahsis is not on asymptotic behaviour or scaling, rather on verifying the stated claim for all t.

Key words: Reaction-Diffusion Equations, Reaction Front

1 Introduction

In case of a reaction $nA + mB \longrightarrow C$ in which the two reactants are initially separated, the formation of a reaction front is a well studied phenomenon. Galfi and Racz [1] assume that for the case when $(n,m) = (1,1)$, the width of reaction zone is negligible to the width of depletion zone in large time limit. Magnin [2] has expanded on this assumption for the general case (n,m). 1D and effectively 1D reaction-diffusion systems for the given chemical reaction with rate k are believed to be accurately described by following equations [2, 3, 5]:

$$\frac{\partial A(X,T)}{\partial T} = D_a \frac{\partial^2 A(X,T)}{\partial X^2} - knA^n(X,T)B^m(X,T) \tag{1}$$

$$\frac{\partial B(X,T)}{\partial T} = D_b \frac{\partial^2 B(X,T)}{\partial X^2} - kmA^n(X,T)B^m(X,T) \tag{2}$$

Saurabh Joglekar and Xiaolin Li
Dept. of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA
e-mail: xiaolin.li@stonybrook.edu, e-mail: joglekar@ams.sunysb.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_28

$$\frac{\partial C(X,T)}{\partial T} = kA^n(X,T)B^m(X,T) \tag{3}$$

At $T = 0$, the reactants are separated at $X = 0$ with constant densities i.e. $A = a_0, B = 0$ for $X < 0$ and $A = 0, B = b_0$ for $X > 0$. To render the system dimensionless, characteristic length, time and concentration are introduced as [1, 2] $l = \sqrt{D_a/(ka_0)}, t = 1/(ka_0)$ and a_0. With these substitutions, equations (1), (2) and (3) become:

$$\frac{\partial a(x,t)}{\partial t} = \frac{\partial^2 a(x,t)}{\partial x^2} - na^n(x,t)b^m(x,t) \tag{4}$$

$$\frac{\partial b(x,t)}{\partial t} = \frac{D_b}{D_a}\frac{\partial^2 b(x,t)}{\partial x^2} - ma^n(x,t)b^m(x,t) \tag{5}$$

$$\frac{\partial c(x,t)}{\partial t} = a^n(x,t)b^m(x,t) \tag{6}$$

This problem, as described by J. Magnin [2], is an initial value problem over the domain $\Omega = \{(x,t)|(x,t) \in \mathbb{R} \times [0,\infty)\}$. Initial conditions are given by

$$a(x,0) = \theta(-x);\ b(x,0) = \frac{b_0}{a_0}\theta(x);\ c(x,0) = 0 \tag{7}$$

Here $\theta(x)$ is the Heaviside Step function and a_0 and b_0 are the initial concentrations of species A and B respectively. It is clear from the dimensionless equations that $\frac{D_b}{D_a}, n, m$ and $q = \frac{b_0}{a_0}$ are free parameters which completely describe the reaction-diffusion system. Many authors [1, 2, 3, 4, 9, 16] on the subject make a key assumption that $D_a = D_b$ for the sake of keeping mathematics within reach. Others [7, 11] describe the asymptotic behaviour of the system with arbitrary diffusivities as $t \longrightarrow \infty$. The center of reaction front x_f has been variably defined as the point where $a/n = b/m$ (Magnin, [2]) or as the point of maximal reaction rate (Koza et al. [11]) (which need not necessarily be the same points). We adopt the first definition. With this particular definition, it has been proved [2, 3] that for the dimensionless system in which $D_a = D_b$,

$$x_f(t) = 2\sqrt{t}\ erf^{-1}\left(\frac{a_0/n - b_0/m}{a_0/n + b_0/m}\right) = 2\sqrt{t}\ erf^{-1}\left(\frac{1 - \frac{n}{m}q}{1 + \frac{n}{m}q}\right) \tag{8}$$

To the best of our knowledge, no such analytical result yet exists in the case of $D_a \neq D_b$. Nonetheless, it seems possible that $x_f(t) \propto \sqrt{t}$ for all possible values of (D_a, D_b) and (n,m). The reason for this proposal is that the term $x/\sqrt{t}$ is the signature of diffusion process in general. Our aim in the present work is to provide compelling numerical evidence to support the claim that $x_f(t) = \eta\left(\frac{D_b}{D_a}, n, m, q\right)\sqrt{t}$ holds true for all values of $\frac{D_b}{D_a}, n, m$ and q as t runs through its domain $(0,\infty)$. $\eta\left(\frac{D_b}{D_a}, n, m, q\right)$ can be thought of as a constant of proportionality which depends on the parameters of the problem. We use finite difference method to solve the equa-

tions. To establish the accuracy of our method, we compare our numerical results with the analytical results in case of equal diffusivities ($D_a = D_b$). We also discuss the effect of grid refinement. Once the accuracy of our method has been established, we present the numerical results for the case of unequal diffusivities ($D_a \neq D_b$).

2 Numerical Method

We use Crank-Nicolson method to solve Eqn. (4) and (5). Discretization is given as follows:

$$\frac{a_j^{(k+1)} - a_j^{(k)}}{\Delta t} = \frac{1}{2}\left(\frac{a_{j-1}^{(k+1)} - 2a_j^{(k+1)} + a_{j+1}^{(k+1)}}{\Delta x^2} + \frac{a_{j-1}^{(k)} - 2a_j^{(k)} + a_{j+1}^{(k)}}{\Delta x^2}\right) - n(a_j^{(k)})^n (b_j^{(k)})^m \quad (9)$$

$$\frac{b_j^{(k+1)} - b_j^{(k)}}{\Delta t} = \frac{1}{2}\frac{D_b}{D_a}\left(\frac{b_{j-1}^{(k+1)} - 2b_j^{(k+1)} + b_{j+1}^{(k+1)}}{\Delta x^2} + \frac{b_{j-1}^{(k)} - 2b_j^{(k)} + b_{j+1}^{(k)}}{\Delta x^2}\right) - m(a_j^{(k)})^n (b_j^{(k)})^m \quad (10)$$

Here the subscript j represents spatial index and the superscript (k) represents time step. Since Eqn. (6) is decoupled from Eqn. (4) and (5), consistency and stability of it's numerical solution will have no effect on the other two. Hence we neglect the equation. Notice that the original problem is an Initial Value Problem. So ideally j runs through all non-negative integers. However, due to the finite memory constraints of any computing platform, we restrict the computational domain to $[-1,1]$ and let j run through 0,1,...,M where $2/\Delta x = M$. Since the left and right boundaries are reasonably far from the initial reaction zone, it is also reasonable to approximate the boundary conditions as follows:

$$\left.\frac{\partial a(x,t)}{\partial x}\right|_{x=-1} = \left.\frac{\partial a(x,t)}{\partial x}\right|_{x=1} = \left.\frac{\partial b(x,t)}{\partial x}\right|_{x=-1} = \left.\frac{\partial b(x,t)}{\partial x}\right|_{x=1} = 0 \quad (11)$$

We implement these boundary conditions numerically as follows:

$$a_0^{(k+1)} = a_1^{(k+1)}; \quad a_M^{(k+1)} = a_{M-1}^{(k+1)}; \quad b_0^{(k+1)} = b_1^{(k+1)}; \quad b_M^{(k+1)} = b_{M-1}^{(k+1)} \quad (12)$$

We note that the numerical boundary conditions can lead to only first order consistency at the boundary while Crank-Nicolson is expected to produce second order accuracy elsewhere in the computational domain. We overcome this problem by terminating the program run as soon as any of $|a_0^{(k)} - a_1^{(k)}|, |a_M^{(k)} - a_{M-1}^{(k)}|, |b_0^{(k)} -$

$b_1^{(k)}|, |b_M^{(k)} - b_{M-1}^{(k)}|$ is greater than $O(\Delta x^2)$. To detect the position of the center of reaction zone $x_f(t)$, we use linear interpolation between the grid points $x_0, x_1, ..., x_M$ at every time step. In particular,at every time step k, we find out the index $i \in \{0,1,2,...,M\}$ such that $(a_i^{(k)} - \frac{n}{m} b_i^{(k)})(a_{i+1}^{(k)} - \frac{n}{m} b_{i+1}^{(k)}) \leq 0$. Then we find out the location $x_f(t)$ by solving the two linear equations $\frac{y-a_i^{(k)}}{x-x_i} = \frac{a_{i+1}^{(k)} - a_i^{(k)}}{x_{i+1}-x_i}$ and $\frac{y-\frac{n}{m}b_i^{(k)}}{x-x_i} = \frac{n}{m}\frac{b_{i+1}^{(k)} - b_i^{(k)}}{x_{i+1}-x_i}$.

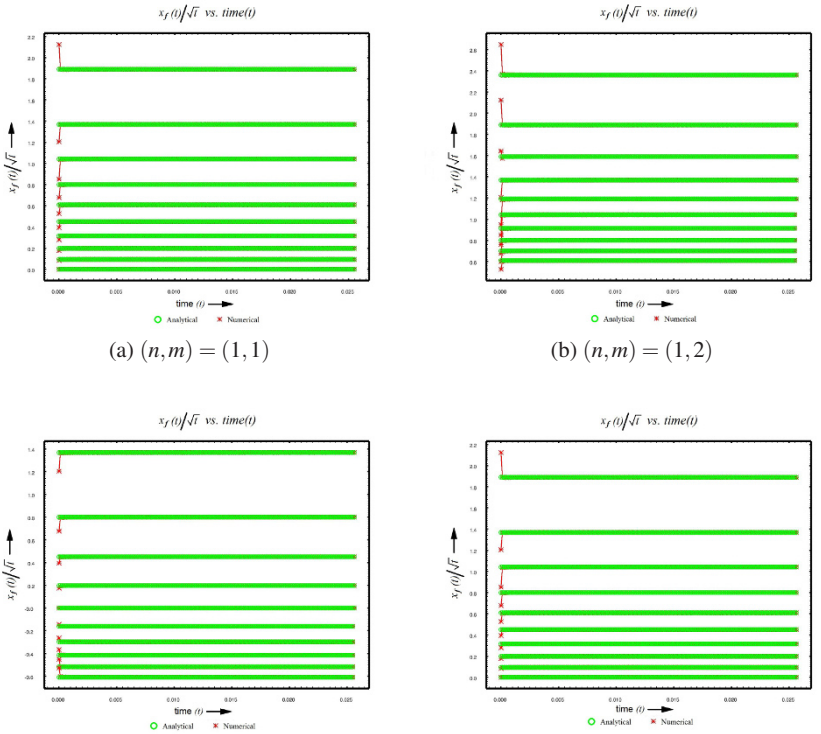
3 Results

3.1 Numerical Results for $\frac{D_b}{D_a} = 1$

In this case, $x_f(t)$ is given by $x_f(t) = 2\sqrt{t}\ erf^{-1}\left(\frac{1-\frac{n}{m}q}{1+\frac{n}{m}q}\right)$ In this section we compare the numerical results with analytical results and show that the numerical method described above indeed gives accurate results. All numerical tests have been carried out with the *FronTier* software library released and maintained by Stony Brook University, NY. The spike and a small oscillatory behaviour near $t = 0$ in Fig.1, can be explained by the fact that grid points are finite in number. Hence the Heaviside Step Function cannot be realized perfectly on any computational grid. This is illustrated in Fig.2. This fact is further verified by changing the grid-size. Fig.3 shows the effects of grid-size on the convergence of $x_f(t)/\sqrt{t}$. It can be seen that as the grid is refined, $x_f(t)/\sqrt{t}$ attains it's theoretical value at earlier time.

3.2 Numerical Results for $\frac{D_b}{D_a} \neq 1$

Previous section presents enough evidence to show that the numerical method described indeed gives results that match with theoretical results. This section presents numerical results for $D_b/D_a \neq 1$. No closed form analytical expression is available in this case. Hence, we provide numerical verification that $x_f(t)$ is proportional to $\sqrt{t}$. Due to the constraints of space, only a few results are presented here. However, tests were done for all of $D_b/D_a = 0.1, 0.2, ..., 1.0$, $q = 0.1, 0.2, ..., 1.0$ and $(n,m) = (1,1), (1,2), (2,1), (2,1)$. In each case, $x_f(t)/\sqrt{t}$ was a constant depending on parameters $D_b/D_a, n, m$ and q. As mentioned previously, q changes from 0.1 to 1.0 in steps of 0.1 as one moves from the topmost branch of each graph to the lowest branch.



(a) $(n,m) = (1,1)$ (b) $(n,m) = (1,2)$

(c) $(n,m) = (2,1)$ (d) $(n,m) = (2,2)$

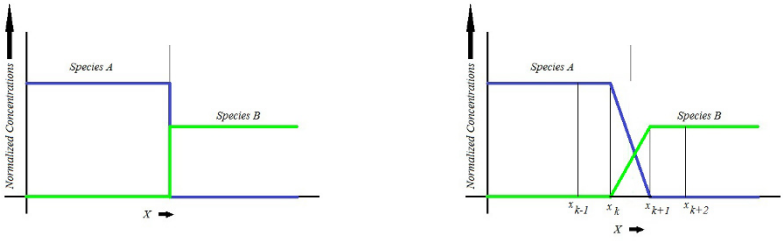
Fig. 1: $x_f(t)/sqrt(t)$ for $D_b/D_a = 1$ and different (n,m). In all figures, the topmost to lowermost branch corresponds to $q = 0.1, 0.2, ..., 1.0$ respectively.

4 Summary and Conclusion

We have presented enough numerical evidence to support the claim that $x_f(t)$ is proportional to $\sqrt{t}$ for every $D_b/D_a, n, m$ and q. Although the results are presented only for $n = 1,2$ and $m = 1,2$, and only for a few values of the stated parameters, we mention in passing that similar results were obtained for $D_b/D_a = 0.1, 0.2, ...1.0$ and for $q = 0.1, 0.2, ...1.0$. In our opinion, the results are consistent enough to support the claim.

References

1. Properties of the reaction front in an $A + B \longrightarrow C$ type reaction-diffusion process, *Physical Review A 1988, Volume 38, Number 6: 3151-3154,* L. Galfi and Z. Racz



(a) Ideal Profile (b) Numerical Implementation

Fig. 2: Implementation of Heaviside Step Function on a finite compuational grid

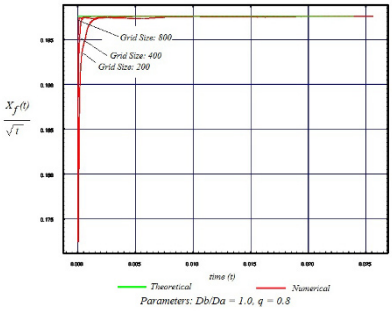
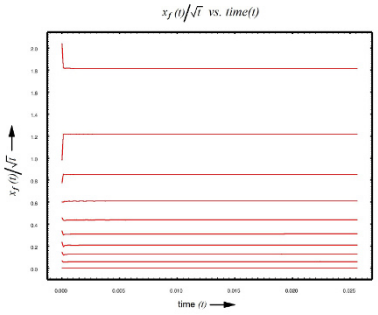
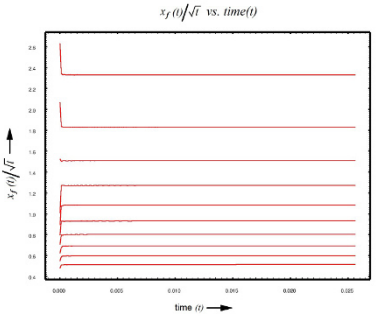


Fig. 3: $x_f(t)/\sqrt{t}$ vs t for varying grid size.

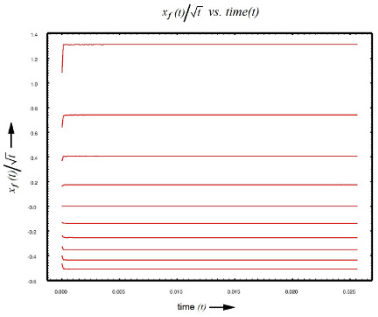
2. Properties of the asymptotic $nA + mB \longrightarrow C$ reaction-diffusion fronts, *Eur. Phys J. B 17 (2000): 673-678,* J. Magnin
3. Dynamic multiscaling of the recation-diffusion front for $mA + nB \longrightarrow C$, *Physical Review E, October 1995, Volume 52, Number 4,* S. Cornell, Z. Koza and M. Droz
4. Reaction front for $A + B \longrightarrow C$ diffusion-reaction systems with initially separated reactants, *Physical Review A, July 1992, Volume 46, Number 2,* H. Larralde, M. Araujo, S. Havlin and H. Stanley
5. Steady-State Reaction-Diffusion Front Scaling for $mA + nB \longrightarrow C$, *Physical Review Letters, June 1993, Volume 70, Number 24,* S. Cornell and M. Droz
6. Asymptotic behaviour of initially separated $A + B_{(static)} \longrightarrow C$ reaction-diffusion systems, *Physica A 240 (1997) 622-634,* Z. Koza
7. Reaction-Diffusion fronts in systems with concentration-dependent diffusivities, *Physical Review E 74, 036103 (2006),* P. Polanowski and Z. Koza
8. Reaction fronts in reversible $A + B \rightleftharpoons C$ reaction-diffusion systems, *Physica A 330 (2003) 160-166,* Z. Koza
9. Reversible and irreversible reaction fronts in two competing reaction system, *Nuclear Instruments and Methods in Physica Research B 186 (2002) 161-165,* M. Sinder, H. Taitelbaum, J. Pelleg
10. Asymptotic expansion for reversible $A + B \longleftrightarrow C$ reaction-diffusion process, *Physical Review E 66, 011103 (2002),* Z. Koza



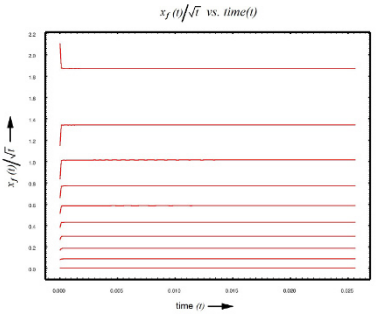
(a) $D_b/D_a = 0.2$, $(n,m) = (1,1)$



(b) $D_b/D_a = 0.4$, $(n,m) = (1,2)$



(c) $D_b/D_a = 0.6$, $(n,m) = (2,1)$



(d) $D_b/D_a = 0.8$, $(n,m) = (2,2)$

Fig. 4: $x_f(t)/\sqrt{t}$ vs. t for various D_b/D_a and (n,m)

11. The Long-time Behavior of Initially Separated $A+B \longrightarrow C$ Reaction-Diffusion Systems with Arbitrary Diffusion Constants, *J. Stat. Phys. 85, 179-191(1996),* Z. Koza
12. Some Properties of the $A+B \longrightarrow C$ Reaction-Diffusion System with Initially Separated Components, *Journal of Statistical Physics, Vol. 65, Nos. 5/6, 1991,* H. Taitelbaum, S. Havlin, J. Kiefer, B. Trus, and G. Weiss
13. Numerical analysis of reversible $A+B \longleftrightarrow C$ reaction-diffusion systems, *Eur. Phys. J. B 32, 507-511(2003),* Z. Koza
14. Simulation study of reaction fronts, *Physical Review A, December 1990, Volume 42, Number 12,* Z. Jiang and C. Ebner
15. Refined simulations of the reaction front for diffusion-limited two-species annihilation in one dimension, *Physical Review E, May 1995, Volume 51, Number 5,* S. Cornell
16. Role of fluctuations for inhomogeneous reaction-diffusion phenomena, *Physical Review A, Volume 44, Number 8, Oct. 1991,* S. Cornell, M. Droz, B. Chopard

Chapter 7

Structural Graph Theory Downunder

Subdivided Claws and the Clique-Stable Set Separation Property

Maria Chudnovsky* and Paul Seymour†

Abstract Let $\mathscr{C}$ be a class of graphs closed under taking induced subgraphs. We say that $\mathscr{C}$ has the *clique-stable set separation property* if there exists $c \in \mathbb{N}$ such that for every graph $G \in \mathscr{C}$ there is a collection $\mathscr{P}$ of partitions (X,Y) of the vertex set of G with $|\mathscr{P}| \leq |V(G)|^c$ and with the following property: if K is a clique of G, and S is a stable set of G, and $K \cap S = \emptyset$, then there is $(X,Y) \in \mathscr{P}$ with $K \subseteq X$ and $S \subseteq Y$. In 1991 M. Yannakakis conjectured that the class of all graphs has the clique-stable set separation property, but this conjecture was disproved by M. Göös in 2014. Therefore it is now of interest to understand for which classes of graphs such a constant c exists. In this paper we define two infinite families $\mathscr{S}, \mathscr{K}$ of graphs and show that for every $S \in \mathscr{S}$ and $K \in \mathscr{K}$, the class of graphs with no induced subgraph isomorphic to S or K has the clique-stable set separation property.

1 Introduction

All graphs in this paper are finite and simple. Let G be a graph. A *clique* in G is a set of pairwise adjacent vertices, and a *stable set* is a set of pairwise non-adjacent vertices. Let $\mathscr{C}$ be a class of graphs closed under taking induced subgraphs. We say that $\mathscr{C}$ has the *clique-stable set separation property* if there exists $c \in \mathbb{N}$ such that for every graph $G \in \mathscr{C}$ there is a collection $\mathscr{P}$ of partitions (X,Y) of the vertex set of G with $|\mathscr{P}| \leq |V(G)|^c$ and with the following property: if K is a clique of G, and S is a stable set of G, and $K \cap S = \emptyset$, then there is $(X,Y) \in \mathscr{P}$ with $K \subseteq$

Maria Chudnovsky and Paul Seymour
Department of Mathematics, Princeton University, Princeton, NJ 08544, USA
e-mail: mchudnov@math.princeton.edu, e-mail: pds@math.princeton.edu

* This material is based upon work supported in part by the U. S. Army Research Office under grant number W911NF-16-1-0404, and by NSF grant DMS-1763817.

† Partially supported by NSF grant DMS-1800053 and AFOSR grant A9550-19-1-0187.

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_29

X and $S \subseteq Y$. This property plays an important role in a large variety of fields: communication complexity, combinatorial optimization, constraint satisfaction and others (for a comprehensive survey of these connections see [3]).

In 1991 Mihalis Yannakakis conjectured that the class of all graphs has the clique-stable set separation property [5], but this conjecture was disproved by Mika Göös in 2014 [2]. Therefore it is now of interest to understand for which classes of graphs such a constant c exists; our main result falls into that category.

Let G be a graph and let X, Y be disjoint subsets of $V(G)$. We denote by $G[X]$ the subgraph of G induced by X, by $N(X)$ the set of all vertices of $V(G) \setminus X$ with a neighbor in X, and by $N[X]$ the set $N(X) \cup X$. We say that X is *complete* to Y if every vertex of X is adjacent to every vertex of Y, and that X is *anticomplete* to Y if every vertex of X is non-adjacent to every vertex of Y. We say that X and Y are *matched* if every vertex of X has exactly one neighbor in Y, and every vertex of Y has exactly one neighbor in X (and therefore $|X| = |Y|$). For a graph H, we say that G is *H-free* if no induced subgraph of G is isomorphic to H.

Next we define two types of graphs. Let $p, q \in \mathbb{N}$. We define the graph $F_S^{p,q}$ as follows:

- $V(F_S^{p,q}) = K \cup S_1 \cup S_2 \cup S_3$ where K is a clique, S_1, S_2, S_3 are stable sets, and the sets K, S_1, S_2, S_3 are pairwise disjoint;
- $|K| = |S_1| = p$, and K and S_1 are matched;
- $|S_2| = |S_3| = q$, and S_2 and S_3 are matched;
- K is complete to S_2;
- there are no other edges in $F_S^{p,q}$.

The graph $F_K^{p,q}$ is obtained from $F_S^{p,q}$ by making all pairs of vertices of S_3 adjacent.

Fig. 1 The graphs $F_S^{3,3}$ and $F_K^{3,3}$

Let $\mathscr{F}^{p,q}$ be the class of all graphs that are both $F_S^{p,q}$-free and $F_K^{p,q}$-free. We can now state our main result:

Theorem 1. *For all $p, q > 0$ the class $\mathscr{F}^{p,q}$ has the clique-stable set separation property.*

Since the clique-stable set separation property is preserved under taking complements, we immediately deduce:

Theorem 2. *For all $p,q > 0$ the class of graphs whose complements are in $\mathscr{F}^{p,q}$ has the clique-stable set separation property.*

2 The Proof

In this section we prove 1. The idea of the proof comes from [1]. Let $G \in \mathscr{F}^{p,q}$. Define $\mathscr{P}_1$ to be the set of all partitions $(N[X], V(G) \setminus N[X])$ and $(N(X), V(G) \setminus N(X))$ where X is a subset of $V(G)$ with $|X| < p$. Clearly $|\mathscr{P}_1| \leq 2|V(G)|^p$.

Write $R = R(q,q)$ to mean the smallest positive integer R such that every 2-coloring of the edges of the complete graph on R vertices contains a monochromatic complete graph on q vertices. Ramsey's Theorem [4] implies:

Theorem 3. $R(q,q) \leq 2^{2q}$.

For $a,b \in \mathbb{N}$ let the graph $F_{a,b}$ be defined as follows:

- $V(F_{a,b}) = K_1 \cup S_1 \cup S_2 \cup W$ where K_1 is a clique, S_1, S_2 are stable sets, and the sets K_1, S_1, S_2, W are pairwise disjoint;
- $|K_1| = |S_1| = a$, and K_1 and S_1 are matched;
- $|S_2| = |W| = b$, and S_2 and W are matched;
- K_1 is complete to S_2;
- there is no restriction on the adjacency of pairs of vertices of W;
- there are no other edges in $F_{a,b}$.

From the definition of R we immediately deduce:

Theorem 4. *G is $F_{p,R}$-free.*

For every triple $X = (K_1, S_1, S_2)$ of pairwise disjoint non-emtpy subsets of $V(G)$ such that $|K_1| = |S_1| = p$ and $|S_2| < R$ we define the partition P_X of $V(G)$ as follows. Let Z be the set of all vertices of G that are anticomplete to $K_1 \cup S_1$. Let A_X be the set of all vertices v of G such that

- either $v \in K_1$, or v is complete to K_1, and
- either v has a neighbor in S_1, or v has a neighbor in $Z \setminus N(S_2)$.

Note that, since S_1 is a stable set and Z is anticomplete to S_1, A_X is disjoint from $S_1 \cup Z$. Define $P_X = (A_X, V(G) \setminus A_X)$, and let $\mathscr{P}_2$ be the set of all such partitions P_X. Since $|K_1 \cup S_1 \cup S_2| \leq 2p + R - 1$, and since by 3 $R \leq 2^{2q}$, we deduce that $|\mathscr{P}_2| < |V(G)|^{2p+2^{2q}}$.

In order to complete the proof of 1 we will prove the following:

Theorem 5. *For every clique K and stable set S of G such that $K \cap S = \emptyset$, there exists $(X,Y) \in \mathscr{P}_1 \cup \mathscr{P}_2$ with $K \subseteq X$ and $S \subseteq Y$.*

Proof. Let K and S be as in the statement of 5.

(1) *We may assume that K is a maximal clique of G, and S is a maximal stable set of G.*

Let K' be a maximal clique of G with $K \subseteq K'$, and let S' be a maximal stable set of G with $S \subseteq S'$. If $K' \cap S' = \emptyset$, then the existence of the desired partition for K, S follows from the existence of such a partition for K', S'; thus we may assume that $K' \cap S' \neq \emptyset$. Since K' is a clique and S' is a stable set, it follows that $|K' \cap S'| = 1$, say $K' \cap S' = \{v\}$. But now the partitions $(N[\{v\}], V(G) \setminus N[\{v\}])$ and $(N(\{v\}), V(G) \setminus N(\{v\}])$ are both in $\mathscr{P}_1$, and at least one of them has the desired property. This proves (1).

In view of (1) from now on we assume that K is a maximal clique of G, and S is a maximal stable set of G. Consequently every vertex of K has a neighbor in S. Let $S'_1 \subseteq S$ be a minimal subset of S such that every vertex of K has a neighbor in S'_1. It follows from the minimality of S'_1 that there is a subset K'_1 of K such that S'_1 and K'_1 are matched. If $|S'_1| < p$, then the partition $(N(S'_1), V(G) \setminus N(S'_1)) \in \mathscr{P}_1$ has the desired property, so we may assume that $|S'_1| \geq p$.

Let S_1 be a subset of S'_1 with $|S_1| = p$, and let $K_1 = N(S_1) \cap K'_1$. Then S_1 and K_1 are matched, and so $|K_1| = p$. Let Z be the set of vertices of G that are anticomplete to $S_1 \cup K_1$. Then $S'_1 \setminus S_1 \subseteq Z \cap S$, and in particular every vertex of K has a neighbor either in S_1 or in $Z \cap S$. Let S' be the subset of vertices of $S \setminus S_1$ that are complete to K_1. Note that $S' \cap Z = \emptyset$. Let S_2 be a minimal subset of S' such that $N(S_2) \cap Z = N(S') \cap Z$. It follows from the minimality of S_2 that there is a subset $W \subseteq Z \cap N(S')$ such that W and S_2 are matched. Observe that $G[K_1 \cup S_1 \cup S_2 \cup W]$ is isomorphic to $F_{p,|S_2|}$ (with K_1, S_1, S_2, W as in the definition of $F_{a,b}$). It follows from 4 that $|S_2| < R$.

Let $X = (K_1, S_1, S_2)$. We claim that the partition $P_X \in \mathscr{P}_2$ has the desired property for the pair K, S. Recall that $P_X = (A_X, V(G) \setminus A_X)$, where A_X is the set of all vertices v of G such that

- either $v \in K_1$, or v is complete to K_1, and
- either v has a neighbor in S_1, or v has a neighbor in $Z \setminus N(S_2)$.

We need to show that $K \subseteq A_X$, and $S \cap A_X = \emptyset$.

(2) $K \subseteq A_X$.

Let $k \in K$. Clearly either $k \in K_1$ or k is complete to K_1. Moreover, k has a neighbor in S'_1, and $S'_1 \subseteq S_1 \cup (Z \cap S)$. Since S is a stable set, it follows that $Z \cap S \subseteq Z \setminus N(S_2)$, and thus k has a neighbor either in S_1, or in $Z \setminus N(S_2)$. This proves (2).

(3) $S \cap A_X = \emptyset$.

Suppose that $s \in S \cap A_X$. Then $s \notin K_1$; therefore s is complete to K_1, and so $s \in S'$. Since S is a stable set, it follows that s is anticomplete to S_1, and therefore s has a

neighbor in $Z \setminus N(S_2)$. But $N(S') \cap Z = N(S_2) \cap Z$, a contradiction. This proves (3).

Now 5 follows from (2) and (3). □

This completes the proof of 1.

Acknowledgements This work was done during the Structural Graph Theory Downunder MATRIX Program. The authors express their gratitude to the MATRIX Institute for the funding it provided and the use of its facilities, and to the organizers of the program for the invitation.

References

1. T. Abrishami, M. Chudnovsky, M. Pilipczuk, P. Rzazewski and P. Seymour: Induced subgraphs of bounded tree-width and the container method, *in preparation*.
2. M. Göös: Lower bounds for clique vs. independent Set. In: Proc. 56th Foundations of Computer Science (FOCS), 2015: 1066–1077.
3. A. Lagoutte: Interactions entre les cliques et les stables dans un graphe. PhD thesis, ENS de Lyon, 2015.
4. F.P. Ramsey: On a problem of formal logic. Proc. London Math. Soc. **30**, 264–286 (1930).
5. M. Yannakakis: Expressing combinatorial optimization problems by linear programs. J. Comput. Syst. Sci. **43**, 441–466 (1991).

Notes on tree- and path-chromatic number

Tony Huynh, Bruce Reed, David R. Wood, and Liana Yepremyan

Abstract *Tree-chromatic number* is a chromatic version of treewidth, where the cost of a bag in a tree-decomposition is measured by its chromatic number rather than its size. *Path-chromatic number* is defined analogously. These parameters were introduced by Seymour [JCTB 2016]. In this paper, we survey all the known results on tree- and path-chromatic number and then present some new results and conjectures. In particular, we propose a version of Hadwiger's Conjecture for tree-chromatic number. As evidence that our conjecture may be more tractable than Hadwiger's Conjecture, we give a short proof that every K_5-minor-free graph has tree-chromatic number at most 4, which avoids the Four Colour Theorem. We also present some hardness results and conjectures for computing tree- and path-chromatic number.

1 Introduction

Tree-chromatic number is a hybrid of the graph parameters treewidth and chromatic number, recently introduced by Seymour [17]. Here is the definition.

Tony Huynh and David R. Wood
School of Mathematics, Monash University, Australia
e-mail: {tony.huynh2, david.wood}@monash.edu

Bruce Reed
School of Computer Science, McGill University, Canada
e-mail: breed@cs.mcgill.ca

Liana Yepremyan
Department of Mathematics, Statistics, and Computer Science, UIC, USA
and Department of Mathematics, London School of Economics, UK
e-mail: {lyepre2@uic.edu, L.Yepremyan@lse.ac.uk}

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_30

A *tree-decomposition* of a graph G is a pair $(T,\mathscr{B})$ where T is a tree and $\mathscr{B} := \{B_t \mid t \in V(T)\}$ is a collection of subsets of vertices of G, called *bags*, satisfying:

- for each $uv \in E(G)$, there exists $t \in V(T)$ such that $u,v \in B_t$, and
- for each $v \in V(G)$, the set of all $t \in V(T)$ such that $v \in B_t$ induces a non-empty subtree of T.

A graph G is *k-colourable* if each vertex of G can be assigned one of k colours, such that adjacent vertices are assigned distinct colours. The *chromatic number* of a graph G is the minimum integer k such that G is k-colourable.

For a tree-decomposition $(T,\mathscr{B})$ of G, the *chromatic number* of $(T,\mathscr{B})$ is $\max\{\chi(G[B_t]) \mid t \in V(T)\}$. The *tree-chromatic number* of G, denoted $\mathsf{tree}\text{-}\chi(G)$, is the minimum chromatic number taken over all tree-decompositions of G. The *path-chromatic number* of G, denoted $\mathsf{path}\text{-}\chi(G)$, is defined analogously, where we insist that T is a path instead of an arbitrary tree. Henceforth, for a subset $B \subseteq V(G)$, we will abbreviate $\chi(G[B])$ by $\chi(B)$. For $v \in V(G)$, let $N_G(v)$ be the set of neighbours of v and $N_G[v] := N_G(v) \cup \{v\}$.

The purpose of this paper is to survey the known results on tree- and path-chromatic number, and to present some new results and conjectures.

Clearly, $\mathsf{tree}\text{-}\chi$ and $\mathsf{path}\text{-}\chi$ are monotone under the subgraph relation, but unlike treewidth, they are not monotone under the minor relation. For example, $\mathsf{tree}\text{-}\chi(K_n) = n$, but the graph G obtained by subdividing each edge of K_n is bipartite and so $\mathsf{tree}\text{-}\chi(G) \leq \chi(G) = 2$.

By definition, for every graph G,

$$\mathsf{tree}\text{-}\chi(G) \leq \mathsf{path}\text{-}\chi(G) \leq \chi(G).$$

Section 2 reviews results that show that each of these inequalities can be strict and in fact, both of the pairs $(\mathsf{tree}\text{-}\chi(G), \mathsf{path}\text{-}\chi(G))$ and $(\mathsf{path}\text{-}\chi(G), \chi(G))$ can be arbitrarily far apart.

We present our new results and conjectures in Sections 3-5. In Section 3, we propose a version of Hadwiger's Conjecture for tree-chromatic number and show how it is related to a 'local' version of Hadwiger's Conjecture. In Section 4, we prove that K_5-minor-free graphs have tree-chromatic number at most 4, without using the Four Colour Theorem. We finish in Section 5, by presenting some hardness results and conjectures for computing $\mathsf{path}\text{-}\chi$ and $\mathsf{tree}\text{-}\chi$.

2 Separating χ, $\mathsf{path}\text{-}\chi$ and $\mathsf{tree}\text{-}\chi$

Complete graphs are a class of graphs with unbounded tree-chromatic number. Are there more interesting examples? The following lemma of Seymour [17] leads to an answer. A *separation* (A,B) of a graph G is a pair of edge-disjoint subgraphs whose union is G.

Lemma 1. *For every graph G, there is a separation (A,B) of G such that $\chi(A\cap B) \leq$ tree-$\chi(G)$ and*

$$\chi(A-V(B)), \chi(B-V(A)) \geq \chi(G) - \text{tree-}\chi(G).$$

Seymour [17] noted that Lemma 1 shows that the random construction of Erdős [6] of graphs with large girth and large chromatic number also have large tree-chromatic number with high probability.

Interestingly, it is unclear if the known *explicit* constructions of large girth, large chromatic graphs also have large tree-chromatic number. For example, *shift graphs* are one of the classic constructions of triangle-free graphs with unbounded chromatic number, as first noted in [7]. The vertices of the n-th shift graph S_n are all intervals of the form $[a,b]$, where a and b are integers satisfying $1 \leq a < b \leq n$. Two intervals $[a,b]$ and $[c,d]$ are adjacent if and only if $b = c$ or $d = a$. The following lemma (first noted in [17]) shows that the gap between χ and path-χ is unbounded on the class of shift graphs.

Lemma 2. *For all $n \in \mathbb{N}$, path-$\chi(S_n) = 2$ and $\chi(S_n) \geq \lceil \log_2 n \rceil$.*

Proof. The fact that $\chi(S_n) \geq \lceil \log_2 n \rceil$ is well-known; we include the proof for completeness. Let $\ell = \chi(S_n)$ and $\phi : V(S_n) \to [\ell]$ be a proper ℓ-colouring of S_n. For each $j \in [n]$ let $C_j = \{\phi([i,j]) \mid i < j\}$. We claim that for all $j < k$, $C_j \neq C_k$. By definition, $\phi([j,k]) \in C_k$. If $C_j = C_k$, then $\phi([i,j]) = \phi([j,k])$ for some $i < j$. But this is a contradiction, since $[i,j]$ and $[j,k]$ are adjacent in S_n. Since there are 2^ℓ subsets of $[\ell]$, $2^\ell \geq n$, as required.

We now show that path-$\chi(S_n) = 2$. For each $i \in [n]$, let $B_i = \{[a,b] \in V(S_n) \mid a \leq i \leq b\}$. Let P_n be the path with vertex set $[n]$ (labelled in the obvious way). We claim that $(P_n, \{B_i \mid i \in [n]\})$ is a path-decomposition of S_n. First observe that $[a,b] \in B_i$ if and only if $a \leq i \leq b$. Next, for each edge $[a,b][b,c] \in E(S_n)$, $[a,b],[b,c] \in B_b$. Finally, observe that for all $i \in [n]$, $X_i = \{[a,b] \in B_i \mid b = i\}$ and $Y_i = \{[a,b] \in B_i \mid b > i\}$ is a bipartition of $S_n[B_i]$. Therefore, S_n has path-chromatic number 2, as required.

Given that shift graphs contain large complete bipartite subgraphs, the following question naturally arises.

Open Problem 1 *Does there exist a function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that for all $s \in \mathbb{N}$ and all $K_{s,s}$-free graphs G, $\chi(G) \leq f(s,$ tree-$\chi(G))$?*

It is not obvious that the parameters path-χ and tree-χ are actually different. Indeed, Seymour [17] asked if path-$\chi(G) =$ tree-$\chi(G)$ for all graphs G? Huynh and Kim [10] answered the question in the negative by exhibiting for each $k \in \mathbb{N}$, an infinite family of k-connected graphs for which tree-$\chi(G) + 1 =$ path-$\chi(G)$. They also prove that the Mycielski graphs [14] have unbounded path-chromatic number.

However, can tree-$\chi(G)$ and path-$\chi(G)$ be arbitrarily far apart? Seymour [17] suggested the following family as a potential candidate. Let T_n be the complete binary rooted tree with 2^n leaves. A path P in T_n is called a $\vee$ if the vertex of P

closest to the root (which we call the *low point* of the V) is an internal vertex of P. Let G_n be the graph whose vertices are the Vs of T_n, where two Vs are adjacent if the low point of one is an endpoint of the other.

Lemma 3 ([17]). *For all $n \in \mathbb{N}$, $\mathsf{tree}\text{-}\chi(G_n) = 2$ and $\chi(G_n) \geq \lceil \log_2 n \rceil$.*

Proof. For each $t \in V(T_n)$, let B_t be the set of Vs in T_n which contain t. We claim that $(T_n, \{B_t \mid t \in V(T_n)\})$ is a tree-decomposition of G_n with chromatic number 2. First observe that if P is a V, then $\{t \in V(T_n) \mid P \in B_t\} = V(P)$, which induces a non-empty subtree of T_n. Next, if P_1 and P_2 are adjacent Vs with $V(P_1) \cap V(P_2) = \{t\}$, then $P_1, P_2 \in B_t$. Finally, for each $t \in B_t$, let X_t be the elements of B_t whose low point is t and let $Y_t := B_t \setminus X_t$. Then (X_t, Y_t) is a bipartition of $G_n[B_t]$, implying that $\mathsf{tree}\text{-}\chi(G_n) = 2$.

For the second claim, it is easy to see that G_n contains a subgraph isomorphic to the n-th shift graph S_n. Thus, $\chi(G_n) \geq \chi(S_n) \geq \lceil \log_2 n \rceil$, by Lemma 2.

Barrera-Cruz, Felsner, Mészáros, Micek, Smith, Taylor, and Trotter [1] subsequently proved that $\mathsf{path}\text{-}\chi(G_n) = 2$ for all $n \in \mathbb{N}$. However, with a slight modification of the definition of G_n, they were able to construct a family of graphs with tree-chromatic number 2 and unbounded path-chromatic number.

Theorem 2 ([1]). *For each integer $n \geq 2$, there exists a graph H_n with $\mathsf{tree}\text{-}\chi(H_n) = 2$ and $\mathsf{path}\text{-}\chi(H_n) = n$.*

The definition of H_n is as follows. A subtree of the complete binary tree T_n is called a Y if it has three leaves and the vertex of the Y closest to the root of T_n is one of its three leaves. The vertices of H_n are the Vs and Ys of T_n. Two Vs are adjacent if the low point of one is an endpoint of the other. Two Ys are adjacent if the lowest leaf of one is an upper leaf of the other. A V is adjacent to a Y if the low point of the V is an upper leaf of the Y. The proof that $\mathsf{path}\text{-}\chi(H_n) = n$ uses Ramsey theoretical methods for trees developed by Milliken [13].

3 Hadwiger's Conjecture for $\mathsf{tree}\text{-}\chi$ and $\mathsf{path}\text{-}\chi$

One could hope that difficult conjectures involving χ might become tractable for $\mathsf{tree}\text{-}\chi$ or $\mathsf{path}\text{-}\chi$, thereby providing insightful intermediate results. Indeed, the original motivation for introducing $\mathsf{tree}\text{-}\chi$ was a conjecture of Gyárfás [8] from 1985, on χ-boundedness of triangle-free graphs without long holes [1].

Conjecture 1 (Gyárfás's Conjecture [8]). For every integer ℓ, there exists c such that every triangle-free graph with no hole of length greater than ℓ has chromatic number at most c.

Seymour [17] proved that Conjecture 1 holds with χ replaced by $\mathsf{tree}\text{-}\chi$.

[1] A *hole* in a graph is an induced cycle of length at least 4.

Theorem 3 ([17]). *For all integers $d \geq 1$ and $\ell \geq 4$, if G is a graph with no hole of length greater than ℓ and $\chi(N_G(v)) \leq d$ for all $v \in V(G)$, then* $\mathsf{tree}\text{-}\chi(G) \leq d(\ell-2)$.

Note that Theorem 3 with $d = 1$ implies that $\mathsf{tree}\text{-}\chi(G) \leq \ell - 2$ for every triangle-free graph G with no hole of length greater than ℓ. A proof of Gyárfás's Conjecture [8] (among other results) was subsequently given by Chudnovsky, Scott, and Seymour [3].

The following is another famous conjectured upper bound on χ, due to Hadwiger [9]; see [16] for a survey.

Conjecture 2 ([9]). If G is a graph without a K_{t+1}-minor, then $\chi(G) \leq t$.

We propose the following weakenings of Hadwiger's Conjecture.

Conjecture 3. If G is a graph without a K_{t+1}-minor, then $\mathsf{tree}\text{-}\chi(G) \leq t$.

Conjecture 4. If G is a graph without a K_{t+1}-minor, then $\mathsf{path}\text{-}\chi(G) \leq t$.

By Theorem 2, $\mathsf{tree}\text{-}\chi(G)$ and $\mathsf{path}\text{-}\chi(G)$ can be arbitrarily far apart, so Conjecture 3 may be easier to prove than Conjecture 4. By Theorem 3, χ and $\mathsf{tree}\text{-}\chi$ can be arbitrarily far apart, so Conjecture 3 may be easier to prove than Hadwiger's Conjecture. We give further evidence of this in the next section, by proving Conjecture 3 for $t = 5$, without using the Four Colour Theorem.

Robertson, Seymour, and Thomas [15] proved that every K_6-minor-free graph is 5-colourable. Their proof uses the Four Colour Theorem and is 83 pages long. Thus, even if we are allowed to use the Four Colour Theorem, it would be interesting to find a short proof that every K_6-minor-free graph has tree-chromatic number at most 5.

Conjectures 3 and 4 are also related to a 'local' version of Hadwiger's Conjecture via the following lemma.

Lemma 4. *Let $(T, \{B_t \mid t \in V(T)\})$ be a* $\mathsf{tree}\text{-}\chi$*-optimal tree-decomposition of G, with $|V(T)|$ minimal. Then there are vertices $v \in V(G)$ and $\ell \in V(T)$ such that $N_G[v] \subseteq B_\ell$.*

Proof. Let ℓ be a leaf of T and u be the unique neighbour of ℓ in T. If $B_\ell \subseteq B_u$, then $T - \ell$ contradicts the minimality of T. Therefore, there is a vertex $v \in B_\ell$ such that $v \notin B_t$ for all $t \neq \ell$. It follows that $N_G[v] \subseteq B_\ell$, as required.

Lemma 4 immediately implies that the following 'local version' of Hadwiger's Conjecture follows from Conjecture 3.

Conjecture 5. If G is a graph without a K_{t+1}-minor, then there exists $v \in V(G)$ such that $\chi(N_G[v]) \leq t$.

It is even open whether Conjectures 3, 4, or 5 hold with an upper bound of $10^{100}t$ instead of t. Finally, the following apparent weakening of Hadwiger's Conjecture (and strengthening of Conjecture 5) is actually equivalent to Hadwiger's Conjecture.

Conjecture 6. If G is a graph without a K_{t+1}-minor, then $\chi(N_G[v]) \leq t$ for all $v \in V(G)$.

Proof (Proof of equivalence to Hadwiger's Conjecture). Clearly, Hadwiger's Conjecture implies Conjecture 6. For the converse, let G be a graph without a K_{t+1}-minor. Let G^+ be the graph obtained from G by adding a new vertex v adjacent to all vertices of G. Since G^+ has no K_{t+2}-minor, Conjecture 6 yields $\chi(N_{G^+}[v]) \leq t+1$. Since $\chi(N_{G^+}[v]) = \chi(G)+1$, we have $\chi(G) \leq t$, as required.

4 K_5-minor-free graphs

As evidence that Conjecture 3 may be more tractable than Hadwiger's Conjecture, we now prove it for K_5-minor-free graphs without using the Four Colour Theorem. We begin with the planar case.

Theorem 4. *For every planar graph G,* tree-$\chi(G) \leq 4$.

Proof. We use the same tree-decomposition previously used by Eppstein [5] and Dujmović, Morin, and Wood [4].

Say G has n vertices. We may assume that $n \geq 3$ and that G is a plane triangulation. Let $F(G)$ be the set of faces of G. By Euler's formula, $|F(G)| = 2n-4$ and $|E(G)| = 3n-6$. Let r be a vertex of G. Let $(V_0, V_1, \ldots, V_t)$ be the bfs layering of G starting from r. Let T be a bfs tree of G rooted at r. Let T^* be the subgraph of the dual G^* with vertex set $F(G)$, where two vertices are adjacent if the corresponding faces share an edge not in T. Thus

$$|E(T^*)| = |E(G)| - |E(T)| = (3n-6) - (n-1) = 2n-5 = |F(G)| - 1 = |V(T^*)| - 1.$$

By the Jordan Curve Theorem, T^* is connected. Thus T^* is a tree.

For each vertex u of T^*, if u corresponds to the face xyz of G, let $C_u := P_x \cup P_y \cup P_z$, where P_v is the vertex set of the vr-path in T, for each $v \in V(G)$. See [5, 4] for a proof that $(T^*, \{C_u : u \in V(T^*)\})$ is a tree-decomposition of G.

We now prove that $G[C_u]$ is 4-colourable. Let ℓ be the largest index such that $\{x,y,z\} \cap V_\ell \neq \emptyset$. For each $k \in \{0, \ldots, \ell\}$, let $G_k = G[C_u \cap (\bigcup_{j=0}^{k} V_j)]$. Note that $G_\ell = G[C_u]$. We prove by induction on k that G_k is 4-colourable. This clearly holds for $k \in \{0,1\}$, since $|V(G_1)| \leq 4$.

For the inductive step, let $k \geq 2$. For each $i \in \{0, \ldots, \ell\}$, let $W_i = C_u \cap V_i$. Since W_i contains at most one vertex from each of P_x, P_y, and P_z, $|W_i| \leq 3$.

First suppose $|W_i| \leq 2$ for all $i \leq k$. Since all edges of G are between consecutive layers or within a layer, we can 4-colour G_k by using the colours $\{1,2\}$ on the even layers and $\{3,4\}$ on the odd layers.

Next suppose $|W_k| \leq 2$. We are done by the previous case unless $k = \ell, |W_\ell| \in \{1,2\}$, and $|W_{\ell-1}| = 3$. By induction, let $\phi' : V(G_{\ell-2}) \to [4]$ and $\phi : V(G_{\ell-1}) \to [4]$ be 4-colourings of $G_{\ell-2}$ and $G_{\ell-1}$, respectively. If $|W_\ell| = 1$, then clearly we can extend ϕ to a 4-colouring of G_ℓ. So, we may assume $|W_\ell| = 2$.

Note that ϕ extends to a 4-colouring of G_ℓ unless every vertex of $W_{\ell-1}$ is adjacent to every vertex of W_ℓ and the two vertices of W_ℓ are adjacent. If $G[W_{\ell-1}]$ is a triangle, then $G[W_{\ell-1} \cup W_\ell] = K_5$, which contradicts planarity. If $G[W_{\ell-1}]$ is a path, say abc, then we obtain a K_5-minor in G by contracting all but one edge of the a–c path in T. If $W_{\ell-1}$ is a stable set, then ϕ' can be extended to a 4-colouring of $G_{\ell-1}$ such that all vertices in $W_{\ell-1}$ are the same colour. This colouring can clearly be extended to a 4-colouring of G_ℓ. The remaining case is if $G[W_{\ell-1}]$ is an edge ab together with an isolated vertex c. It suffices to show that there is a colouring of $G_{\ell-1}$ that uses at most two colours on $W_{\ell-1}$, since such a colouring can be extended to a 4-colouring of G_ℓ. Note that ϕ' can be extended to such a colouring unless ϕ' uses three colours on $W_{\ell-2}$ and a and b are adjacent to all vertices of $W_{\ell-2}$. Since ϕ is a 4-colouring, this implies that ϕ uses at most two colours on $W_{\ell-2}$. Thus we may recolour ϕ so that only two colours are used on $W_{\ell-1}$, as required.

Henceforth, we may assume $|W_k| = 3$. By induction, let $\phi : V(G_{k-1}) \to [4]$ be a 4-colouring of G_{k-1}. Let $\phi_{k-1} = \phi(W_{k-1})$.

If $|\phi_{k-1}| = 1$, then we can extend ϕ to a 4-colouring of G_k by using $[4] \setminus \phi_{k-1}$ to 3-colour W_k.

Suppose $|\phi_{k-1}| = 2$. By induction, G_{k-2} has a 4-colouring ϕ'. If W_{k-1} is a stable set, then we can extend ϕ' to a 4-colouring of G_{k-1} such that all vertices of W_{k-1} are the same colour. Thus, $|\phi'_{k-1}| = 1$, and we are done by the previous case. Let $a,b \in W_{k-1}$ such that $ab \in E(G_{k-1})$. Let c be the other vertex of W_{k-1} (if it exists). By relabeling, we may assume that $\phi(a) = 1, \phi(b) = 2$, and $\phi(c) = 2$. Let $N(a)$ be the set of neighbours of a in W_k and $N(b,c)$ be the set of neighbours of $\{b,c\}$ in W_k. Observe that ϕ extends to a 4-colouring of G_k unless $N(a) = N(b,c) = W_k$. However, if, $N(a) = N(b,c) = W_k$, then we obtain a K_5-minor in G by using T to contract W_k onto $\{x,y,z\}$ and c onto b (if c exists). This contradicts planarity.

The remaining case is $|\phi_{k-1}| = 3$. In this case, ϕ extends to a 4-colouring of G_k, unless there exist distinct vertices $a,b \in W_{k-1}$ such that a and b are both adjacent to all vertices of W_k. Again we obtain a K_5-minor in G by using T to contract W_k onto $\{x,y,z\}$ and contracting all but one edge of the a–b path in T.

We finish the proof by using Wagner's characterization of K_5-minor-free graphs [19], which we now describe. Let G_1 and G_2 be two graphs with $V(G_1) \cap V(G_2) = K$, where K is a clique of size k in both G_1 and G_2. The *k-sum* of G_1 and G_2 (along K) is the graph obtained by gluing G_1 and G_2 together along K (and keeping all edges of K). The *Wagner graph* V_8 is the graph obtained from an 8-cycle by adding an edge between each pair of antipodal vertices.

Theorem 5 (Wagner's Theorem [19]). *Every edge-maximal K_5-minor-free graph can be obtained from* 1*-,* 2*-, and* 3*-sums of planar graphs and V_8.*

Theorem 6. *For every K_5-minor-free graph G,* tree-$\chi(G) \leq 4$.

Proof. Let G be a K_5-minor-free graph. We proceed by induction on $|V(G)|$. We may assume that G is edge-maximal. First note that if $G = V_8$, then tree-$\chi(G) \leq \chi(G) = 4$. Next, if G is planar, then tree-$\chi(G) \leq 4$ by Theorem 4 (whose proof

avoids the Four Colour Theorem). By Theorem 5, we may assume that G is a k-sum of two graphs G_1 and G_2, for some $k \in [3]$. Let K be the clique in $V(G_1) \cap V(G_2)$ along which the k-sum is performed. Since G_1 and G_2 are both K_5-minor-free graphs with $|V(G_1)|, |V(G_2)| < |V(G)|$, we have $\mathsf{tree}\text{-}\chi(G_1) \leq 4$ and $\mathsf{tree}\text{-}\chi(G_2) \leq 4$ by induction. For $i \in [2]$, let $(T^i, \{B^i_t \mid t \in V(T^i)\})$ be a tree-decomposition of G_i with chromatic number at most 4. Since K is a clique in G_i, $K \subseteq B^1_x \cap B^2_y$ for some $x \in V(T^1)$ and $y \in V(T^2)$. Let T be the tree obtained from the disjoint union of T^1 and T^2 by adding an edge between x and y. Then $(T, \{B^1_t \mid t \in V(T^1)\} \cup \{B^2_t \mid t \in V(T^2)\})$ is a tree-decomposition of G with chromatic number at most 4.

5 Computing $\mathsf{tree}\text{-}\chi$ and $\mathsf{path}\text{-}\chi$

We finish by showing some hardness results for computing $\mathsf{tree}\text{-}\chi$ and $\mathsf{path}\text{-}\chi$. We need some preliminary results. For a graph G, let K^G_t be the graph consisting of t disjoint copies of G and all edges between distinct copies of G.

Lemma 5. *For all $t \in \mathbb{N}$ and all graphs G without isolated vertices,*

$$(t-1)\chi(G)+2 \leq \mathsf{tree}\text{-}\chi(K^G_t) \leq \mathsf{path}\text{-}\chi(K^G_t) \leq t\,\chi(G).$$

Proof. Let $(T, \{B_t \mid t \in V(T)\})$ be a $\mathsf{tree}\text{-}\chi$-optimal tree-decomposition of $K := K^G_t$, with $|V(T)|$ minimal. By Lemma 4, there exists $\ell \in V(T)$ and $v \in V(K)$ such that $N_K[v] \subseteq B_\ell$. Since G has no isolated vertices, v has a neighbour in the same copy of G in which it belongs. Therefore,

$$\mathsf{tree}\text{-}\chi(K) \geq \chi(B_\ell) \geq \chi(N_K[v]) \geq 2+(t-1)\chi(G).$$

For the other inequalities, $\mathsf{tree}\text{-}\chi(K) \leq \mathsf{path}\text{-}\chi(K) \leq \chi(K) = t\,\chi(G)$.

We also require the following hardness result of Lund and Yannakakis [12].

Theorem 7 ([12]). *There exists $\varepsilon > 0$, such that it is NP-hard to correctly determine $\chi(G)$ within a multiplicative factor of n^ε for every n-vertex graph G.*

Our first theorem is a hardness result for approximating $\mathsf{tree}\text{-}\chi$ and $\mathsf{path}\text{-}\chi$.

Theorem 8. *There exists $\varepsilon' > 0$, such that it is NP-hard to correctly determine $\mathsf{tree}\text{-}\chi(G)$ within a multiplicative factor of $n^{\varepsilon'}$ for every n-vertex graph G. The same hardness result holds for $\mathsf{path}\text{-}\chi$ with the same ε'.*

Proof. We show the proof for $\mathsf{tree}\text{-}\chi$. The proof for $\mathsf{path}\text{-}\chi$ is identical. Let $\varepsilon' = \frac{\varepsilon}{3}$, where ε is the constant from Theorem 7. Let G be an n-vertex graph.

Note that K^G_n has n^2 vertices, and $(n^2)^{\varepsilon'} = n^{\frac{2\varepsilon}{3}}$. If $k \in [\frac{\mathsf{tree}\text{-}\chi(K^G_n)}{n^{\frac{2\varepsilon}{3}}}, n^{\frac{2\varepsilon}{3}}\,\mathsf{tree}\text{-}\chi(K^G_n)]$, then $\frac{k}{n} \in [\frac{\chi(G)}{n^\varepsilon}, n^\varepsilon\chi(G)]$ by Lemma 5. Therefore, if we can approximate $\mathsf{tree}\text{-}\chi(K^G_n)$ within a factor of $(n^2)^{\varepsilon'}$, then we can approximate $\chi(G)$ within a factor of n^ε.

For the decision problem, we use the following hardness result of Khanna, Linial, and Safra [11].

Theorem 9 ([11]). *Given an input graph G with $\chi(G) \neq 4$, it is NP-complete to decide if $\chi(G) \leq 3$ or $\chi(G) \geq 5$.*

As a corollary of Theorem 9, we obtain the following.

Theorem 10. *It is NP-complete to decide if* tree-$\chi(G) \leq 6$. *It is also NP-complete to decide if* path-$\chi(G) \leq 6$.

Proof. Let G be a graph without isolated vertices and $\chi(G) \neq 4$. By Lemma 5, if tree-$\chi(K_2^G) \leq 6$, then $\chi(G) \leq 3$ and if tree-$\chi(K_2^G) \geq 7$, then $\chi(G) \geq 5$. Same for path-χ. Finally, a tree- or path-decomposition and a 6-colouring of each bag is a certificate that tree-$\chi(G) \leq 6$ or path-$\chi(G) \leq 6$.

Combining the standard $O(2^n)$-time dynamic programming for computing pathwidth exactly (see Section 3 of [18]) and the $2^n n^{O(1)}$-time algorithm of Björklund, Husfeldt, and Koivisto [2] for deciding if $\chi(G) \leq k$, yields a $4^n n^{O(1)}$-time algorithm to decide to path-$\chi(G) \leq k$. As far as we know, there is no faster algorithm for deciding path-$\chi(G) \leq k$ (except for small values of k, where faster algorithms for deciding k-colourability can be used instead of [2]).

Finally, unlike for $\chi(G)$, we conjecture that it is still NP-complete to decide if tree-$\chi(G) \leq 2$.

Conjecture 7. It is NP-complete to decide if tree-$\chi(G) \leq 2$. It is also NP-complete to decide if path-$\chi(G) \leq 2$.

References

1. Barrera-Cruz, F., Felsner, S., Mészáros, T., Micek, P., Smith, H., Taylor, L., Trotter, W.T.: Separating tree-chromatic number from path-chromatic number. J. Combin. Theory Ser. B **138**, 206–218 (2019). doi:10.1016/j.jctb.2019.02.003.
2. Björklund, A., Husfeldt, T., Koivisto, M.: Set partitioning via inclusion-exclusion. SIAM J. Comput. **39**(2), 546–563 (2009). doi:10.1137/070683933.
3. Chudnovsky, M., Scott, A., Seymour, P.: Induced subgraphs of graphs with large chromatic number. III. Long holes. Combinatorica **37**(6), 1057–1072 (2017). doi:10.1007/s00493-016-3467-x.
4. Dujmović, V., Morin, P., Wood, D.R.: Layered separators in minor-closed graph classes with applications. J. Combin. Theory Series B **127**, 111–147 (2017). doi:10.1016/j.jctb.2017.05.006
5. Eppstein, D.: Subgraph isomorphism in planar graphs and related problems. J. Graph Algorithms Appl. **3**(3), 1–27 (1999). doi:10.7155/jgaa.00014
6. Erdős, P.: Graph theory and probability. Canadian J. Math. **11**, 34–38 (1959). doi:10.4153/CJM-1959-003-9.
7. Erdős, P., Hajnal, A.: On chromatic number of infinite graphs. In: Theory of Graphs (Proc. Colloq., Tihany, 1966), pp. 83–98. Academic Press, New York (1968)
8. Gyárfás, A.: Problems from the world surrounding perfect graphs. Tanulmányok—MTA Számitástech. Automat. Kutató Int. Budapest (177), 53 (1985)

9. Hadwiger, H.: Über eine Klassifikation der Streckenkomplexe. Vierteljschr. Naturforsch. Ges. Zürich **88**, 133–142 (1943)
10. Huynh, T., Kim, R.: Tree-chromatic number is not equal to path-chromatic number. J. Graph Theory **86**(2), 213–222 (2017)
11. Khanna, S., Linial, N., Safra, S.: On the hardness of approximating the chromatic number. Combinatorica **20**(3), 393–415 (2000). doi:10.1007/s004930070013.
12. Lund, C., Yannakakis, M.: On the hardness of approximating minimization problems. J. Assoc. Comput. Mach. **41**(5), 960–981 (1994). doi:10.1145/185675.306789.
13. Milliken, K.R.: A Ramsey theorem for trees. J. Combin. Theory Ser. A **26**(3), 215–237 (1979). doi:10.1016/0097-3165(79)90101-8.
14. Mycielski, J.: Sur le coloriage des graphs. Colloq. Math. **3**, 161–162 (1955)
15. Robertson, N., Seymour, P., Thomas, R.: Hadwiger's conjecture for K_6-free graphs. Combinatorica **13**(3), 279–361 (1993). doi:10.1007/BF01202354.
16. Seymour, P.: Hadwiger's conjecture. In: J.F.N. Jr., M.T. Rassias (eds.) Open Problems in Mathematics, pp. 417–437. Springer (2015). doi:10.1007/978-3-319-32162-2
17. Seymour, P.: Tree-chromatic number. J. Combin. Theory Ser. B **116**, 229–237 (2016). doi: 10.1016/j.jctb.2015.08.002.
18. Suchan, K., Villanger, Y.: Computing pathwidth faster than 2^n. In: Parameterized and exact computation, *Lecture Notes in Comput. Sci.*, vol. 5917, pp. 324–335. Springer, Berlin (2009). doi:10.1007/978-3-642-11269-0_27.
19. Wagner, K.: Über eine Eigenschaft der ebenen Komplexe. Math. Ann. **114**(1), 570–590 (1937). doi:10.1007/BF01594196.

Note on Hedetniemi's conjecture and the Poljak-Rödl function

Xuding Zhu

Abstract Hedetniemi conjectured in 1966 that $\chi(G \times H) = \min\{\chi(G), \chi(H)\}$ for any graphs G and H. Here $G \times H$ is the graph with vertex set $V(G) \times V(H)$ defined by putting (x,y) and (x',y') adjacent if and only if $xx' \in E(G)$ and $yy' \in V(H)$. This conjecture received a lot of attention in the past half century. It was disproved recently by Shitov. The Poljak-Rödl function is defined as $f(n) = \min\{\chi(G \times H) : \chi(G) = \chi(H) = n\}$. Hedetniemi's conjecture is equivalent to saying $f(n) = n$ for every integer n. Shitov's result shows that $f(n) < n$ when n is sufficiently large. Using Shitov's result, Tardif and Zhu showed that $f(n) \leq n - (\log n)^{1/4-o(1)}$ for sufficiently large n. Using Shitov's method, He and Wigderson showed that for $\varepsilon \approx 10^{-9}$ and n sufficiently large, $f(n) \leq (1-\varepsilon)n$. In this note we observe that a slight modification of the proof in the paper of Zhu and Tardif shows that $f(n) \leq (\frac{1}{2} + o(1))n$ for sufficiently large n. On the other hand, it is unknown whether $f(n)$ is bounded by a constant. However, we do know that if $f(n)$ is bounded by a constant, then the smallest such constant is at most 9. This note gives self-contained proofs of the above mentioned results.

1 Introduction

The *product* $G \times H$ of graphs G and H has vertex set $V(G) \times V(H)$ and has (x,y) adjacent to (x',y') if and only if $xx' \in E(G)$ and $yy' \in E(H)$. Many names for this product are used in the literature, including the *categorical product*, the *tensor product* and the *direct product*. It is the most important product in this note. We just call it *the product*. We may write $x \sim y$ (in G) to denote $xy \in E(G)$.

A proper colouring ϕ of G induces a proper colouring Φ of $G \times H$ defined as $\Phi(x,y) = \phi(x)$. So $\chi(G \times H) \leq \chi(G)$. Symmetrically, we also have $\chi(G \times H) \leq$

Xuding Zhu
Zhejiang Normal University, Jinhua, Zhejiang, China, e-mail: xdzhu@zjnu.edu.cn
Grant number: NSFC: 11971438

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_31

$\chi(H)$. Therefore $\chi(G\times H)\le \min\{\chi(G),\chi(H)\}$. In 1966, Hedetniemi conjectured in [5] that $\chi(G\times H)=\min\{\chi(G),\chi(H)\}$ for all graphs G and H. This conjecture received a lot of attention in the past half century (see [1, 6, 10, 13, 18, 19]). Some special cases are confirmed. In particular, it is known that if $\min\{\chi(G),\chi(H)\}\le 4$, then the conjecture holds [1]. Also, a fractional version of Hedetniemi's conjecture is true [19]. However, Shitov recently refuted Hedetniemi's conjecture [11]. He proved that for sufficiently large n, there are n-chromatic graphs G and H with $\chi(G\times H)<n$.

The Poljak-Rödl function [9] is defined as

$$f(n)=\min\{\chi(G\times H):\chi(G)=\chi(H)=n\}.$$

Hedetniemi's conjecture is equivalent to saying $f(n)=n$ for all positive integer n. Shitov's result shows that $f(n)<n$ for sufficiently large n. Right after Shitov put his result on arxiv, using his result, Tardif and Zhu [16] showed that the difference $n-f(n)$ can be arbitrarily large. Indeed, they proved that $f(n)\le n-(\log n)^{1/4-o(1)}$ for sufficiently large n. It is also shown in [16] that if a special case of Stahl's conjecture in [12] on the multi-chromatic number of Kneser graphs is true, then $\lim_{n\to\infty} f(n)/n\le 1/2$. He and Wigderson, using Shitov's method, proved that $f(n)\le(1-\varepsilon)n$ for $\varepsilon\approx 10^{-9}$ and suffciently large n. Very recently, Zhu observed that the conclusion $\lim_{n\to\infty} f(n)/n\le 1/2$ holds without assuming Stahl's conjecture.

2 Exponential graph

One of the standard tools used in the study of Hedetniemi's conjecture is the concept of *exponential graphs*. Let c be a positive integer. We denote by $[c]$ the set $\{1,2,\dots,c\}$. For a graph G, the exponential graph K_c^G has vertex set

$$\{f: f \text{ is a mapping from } V(G)\to[c]\},$$

with $fg\in E(K_c^G)$ if and only if for any edge $xy\in E(G)$, $f(x)\ne g(y)$. In particular, $f\sim f$ is a loop in K_c^G if and only if f is a proper c-colouring of G. So if $\chi(G)>c$, then K_c^G has no loop.

For convenience, when we study properties of K_c^G, vertices in K_c^G will be called *maps*. The term "vertices" is reserved for vertices of G. That is, to refer to a vertex of K_c^G, we will say that it is a map in K_c^G or a map from G to $[c]$.

For two graphs G and H, a *homomorphism from G to H* is a mapping $\phi: V(G)\to V(H)$ that preserves edges, i.e., for every edge xy of G, $\phi(x)\phi(y)$ is an edge of H. We say G is *homomorphic* to H, and write $G\to H$, if there is a homomorphism from G to H. The "homomorphic" relation "$\to$" is a quasi-order. It is reflexive and transitive: if $G\to H$ and $H\to Q$ then $G\to Q$. The composition $\psi\circ\phi$ of a homomorphism ϕ from G to H and a homomorphism ψ from H to Q is a homomorphism from G to Q.

Note that a homomorphism from a graph G to K_c is equivalent to a proper c-colouring of G. Thus if $G \to H$, then $\chi(G) \leq \chi(H)$.

Lemma 1. *For any graph F, $\chi(G \times F) \leq c$ if and only if F is homomorphic to K_c^G.*

Proof. Assume $\chi(G \times F) \leq c$ and $\Psi : V(G \times F) \to [c]$ is a proper colouring of $G \times F$. For any vertex $u \in V(F)$, let $f_u \in K_c^G$ be defined as $f_u(v) = \Psi(u,v)$. Then the mapping sending u to f_u is a homomorphism from F to K_c^G. Indeed, if $uv \in E(F)$, then for any edge $xy \in E(G)$, $(u,x) \sim (v,y)$ in $G \times F$. Therefore $f_u(x) = \Psi(u,x) \neq \Psi(v,y) = f_v(y)$. Thus $f_u \sim f_v$ in K_c^G.

Conversely, the mapping $\Psi : V(G \times K_c^G) \to [c]$ defined as $\Psi(x,f) = f(x)$ is a proper colouring of $G \times K_c^G$. Indeed, if $(x,f) \sim (y,g)$ in $G \times K_c^G$, then $xy \in E(G)$ and $fg \in E(K_c^G)$. Therefore $\Psi(x,f) = f(x) \neq g(y) = \Psi(g,y)$.

If F is homomorphic to K_c^G, then $G \times F$ is homomorphic to $G \times K_c^G$ and hence $\chi(G \times F) \leq c$.

In this sense, K_c^G is the largest graph H in the order of homomorphism with the property that $\chi(G \times H) \leq c$. Thus Hedetniemi's conjecture is equivalent to the following statement:

If $\chi(G) > c$, then $\chi(K_c^G) = c$.

The concept of exponential graphs was first used by El-Zahar and Sauer in [1], where it is shown that if $\chi(G) \geq 4$, then K_3^G is 3-colourable. Hence the product of two 4-chromatic graphs has chromatic number 4.

The result of El-Zahar and Sauer is still the best result in the positive direction of Hedetniemi's conjecture. We do not know whether or not the product of two 5-chromatic graphs equals 5. On the other hand, there is a nice strengthening of this result by Tardif [14] in the study of multiplicative graphs. We say a graph Q is *multiplicative* if for any two graphs G,H, $G \not\to Q$ and $H \not\to Q$ implies that $G \times H \not\to Q$. Hedetniemi's conjecture is equiavelnt to say that K_n is multiplicative for any positive integer n. El-Zahar and Sauer proved that K_3 is multiplicative. Häggkvist, Hell, Miller and Neumann Lara [3] proved that odd cycles are multiplicative and Tardif [14] proved that circular cliques $K_{p/q}$ for $p/q < 4$ are multiplicative, where $K_{p/q}$ has vertex set $[p]$ with $i \sim j$ if and only if $q \leq |i-j| \leq p-q$. (So $K_{p/1} = K_p$ and $K_{(2k+1)/k} = C_{2k+1}$).

3 Shitov's Theorem

To disprove Hedetniemi's conjecture, it suffices to find a graph G and a positive integer c so that $\chi(G) > c$ and $\chi(K_c^G) > c$.

For a map $f \in K_c^G$, the image set of f is $Im(f) = \{f(v) : v \in V(G)\}$. Note that for $f,g \in K_c^G$, if $Im(f) \cap Im(g) = \emptyset$, then $f \sim g$. For $i \in [c]$, we denote by $g_i \in V(K_c^G)$ the constant map $g_i(v) = i$ for all $v \in V(G)$. So $Im(g_i) = \{i\}$. Thus for any graph G and any positive integer c, $\{g_i : i \in [c]\}$ induces a c-clique in K_c^G and $\chi(K_c^G) \geq c$.

We denote by $G[K_q]$ the graph obtained from G by *blowing up* each vertex of G into a q-clique. The vertices of $G[K_q]$ are denoted by (x,i), where $x \in V(G)$ and $i \in [q]$. So (x,i) and (y,j) are adjacent in $G[K_q]$ if and only if either $x \sim y$ or $x = y$ and $i \neq j$. For a graph G, the *independence number* $\alpha(G)$ of G is the size of a largest independent set in G. This section proves the following result of Shitov:

Theorem 1 (Shitov). *Let G be a graph with $|V(G)| = p$, $\alpha(G) \leq \frac{p}{4.1}$ and* girth$(G) \geq 6$. *Let $q \geq 2^{p-1}p^2$ and $c = 4q+2$. Then $\chi(G[K_q]) > c$ and $\chi\left(K_c^{G[K_q]}\right) > c$.*

The above formulation of the theorem is slightly different from the formulation in [11]. The proof also seems different. But all the claims and lemmas are either stated in [11] or hidden in the text in [11].

It is a classical result of Erdős [2] that there are graphs of arbtirary large girth and large chromatic number. This result is included in most graph theory textbooks (see [17]). The probabilistic proof of this result actually shows that there are graphs G of arbitrary large girth and arbtirary small independence ratio $\alpha(G)/|V(G)|$. What we need here is a graph of girth 6 and with $\alpha(G) \leq |V(G)|/4.1$.

Proof of Theorem 1 Since $G[K_q]$ has the same independence number as G, we have

$$\chi(G[K_q]) \geq \frac{|V(G[K_q])|}{\alpha(G[K_q])} = \frac{|V(G)|q}{\alpha(G)} \geq 4.1q > c.$$

It remains to show that $\chi(K_c^{G[K_q]}) > c$.

Assume to the contrary that $\chi(K_c^{G[K_q]}) = c$ (recall that $K_c^{G[K_q]}$ has a c-clique and hence has chromatic number at least c), and Ψ is c-colouring of K_c^G. We may assume that the constant map g_i is coloured by colour i. Thus for any map $\phi \in K_c^{G[K_q]}$, if $i \notin Im(\phi)$, then $\phi \sim g_i$ and hence $\Psi(\phi) \neq i$. Thus we have the following lemma.

Lemma 2. *For any map $\phi \in K_c^{G[K_q]}$, $\Psi(\phi) \in Im(\phi)$.*

Definition 1. A map $\phi \in K_c^{G[K_q]}$ is called *simple* if ϕ is constant on each copy of K_q that is a blow-up of a vertex of G, i.e., for any $x \in V(G), i,j \in [q]$, $\phi(x,i) = \phi(x,j)$.

For simplicity, we shall write $\phi(x)$ for $\phi(x,i)$ when ϕ is a simple map.

Note that in $K_c^{G[K_q]}$, two simple maps ϕ and ψ are adjacent if and only if for each edge xy of G, $\phi(x) \neq \psi(y)$, and moreover, for each vertex x, $\phi(x) \neq \psi(x)$. This is so, because for $i \neq j \in [q]$, $(x,i)(x,j)$ is an edge of $G[K_q]$ and $\phi(x)$ is a shorthand for $\phi(x,i)$ and $\psi(x)$ is a shorthand for $\psi(x,j)$.

In this sense, the subgraph of $K_c^{G[K_q]}$ induced by simple maps is isomorphic to $K_c^{G^o}$, where G^o is obtained from G by adding a loop to each vertex of G. We shall just treat $K_c^{G^o}$ as an induced subgraph of $K_c^{G[K_q]}$ and write $\phi \in V(K_c^{G^o})$ to mean that ϕ is a simple map in $K_c^{G[K_q]}$. Most of our argument is about properties of the subgraph $K_c^{G^o}$ of $K_c^{G[K_q]}$.

The graph $K_c^{G[K_q]}$ is a huge graph. As G has girth 6 and fractional chromatic number at least 4.1, $p = |V(G)|$ is probably about 200. The number in $K_c^{G[K_q]}$ is c^{pq}, which is roughly $(2^{200})^{2^{200}}$. The subgraph $K_c^{G^o}$ has c^p vertices, which is roughly $(2^{200})^{200}$. So $K_c^{G^o}$ is huge, but it is a very tiny fraction of $K_c^{G[K_q]}$.

Definition 2. For $v \in V(G)$ and $b \in [c]$, let

$$I(v,b) = \{\phi \in K_c^{G^o} : \Psi(\phi) = b = \phi(v)\}.$$

By Observation 2, $\Psi(\phi) \in Im(\phi)$ for any $\phi \in K_c^{G^o}$. Therefore

$$V(K_c^{G^o}) = \bigcup_{v \in V(G), b \in [c]} I(v,b).$$

As $K_c^{G^o}$ has c^p vertices, the average size of $I(v,b)$ is

$$\frac{c^p}{pc} = \frac{c^{p-1}}{p}.$$

Definition 3. We say $I(v,b)$ is *large* if $|I(v,b)| \geq 2pc^{p-2}$.

Observe that, by hypothesis, c is much larger than p. The power of c is the dominating factor. So $2pc^{p-2}$ is much smaller than the average size of $I(v,b)$. Thus intuitively, "most" of the $I(v,b)$'s should be large. So the next lemma is not a surprise.

Lemma 3. *There exists a vertex v of G such that*

$$|\{b \in [c] : I(v,b) \text{ is large }\}| > c/2.$$

Proof. For each vertex v of G, let $S(v) = \{b : I(v,b) \text{ is small}\}$. Assume to the contrary that for each v, $|S(v)| \geq c/2$. Let

$$\mathscr{L} = \{\phi \in K_c^{G^o} : \forall v \in V(G), \phi(v) \in S(v)\}.$$

Then

$$|\mathscr{L}| = \prod_{v \in V(G)} |S(v)| \geq \left(\frac{c}{2}\right)^p.$$

For any $\phi \in \mathscr{L}$, if $\phi \in I(v,b)$, then $I(v,b)$ is small. Thus

$$\mathscr{L} \subset \bigcup_{v \in V(G), b \in [c], I(v,b) \text{ is small}} I(v,b).$$

Therefore $|\mathscr{L}| < p \cdot c \cdot 2pc^{p-2} = 2p^2c^{p-1}$. But then

$$\left(\frac{c}{2}\right)^p < 2p^2c^{p-1}$$

which implies that $c < 2^{p+1}p^2$. But by our choice of c, we have $c = 4q+2 > 4q \geq 2^{p+1}p^2$, a contradiction. □

For two vertices x, y of G, denote by $d_G(x,y)$ the distance between x and y. Let v be a vertex of G for which $|\{b \in [c] : I(v,b) \text{ is large }\}| > c/2$. For $t \in \{2q+1, 2q+2, \ldots, 4q+2\}$, let $\mu_t \in K_c^{G[K_q]}$ be defined as

$$\mu_t(x,i) = \begin{cases} i, & \text{if } d_G(x,v) = 0,2, \\ q+i, & \text{if } d_G(x,v) = 1, \\ t, & \text{if } d_G(x,v) \geq 3. \end{cases}$$

Observe that μ_t are not simple maps. These will be the only non-simple maps used in the proof.

Claim. The set of maps $\{\mu_t : t \in \{2q+1, 2q+2, \ldots, 4q+2\}\}$ induces a clique in $K_c^{G[K_q]}$.

Proof. Assume to the contrary that for some $t \neq t'$, $\mu_t \not\sim \mu_{t'}$. Then there is an edge $(x,i)(y,j)$ of $G[K_q]$ such that $\mu_t(x,i) = \mu_{t'}(y,j)$. Let $\alpha = \mu_t(x,i) = \mu_{t'}(y,j)$.

Then $\alpha \in Im(\mu_t) \cap Im(\mu_{t'}) \subseteq \{i, q+i, t\} \cap \{j, q+j, t'\}$. As $t \neq t'$, we conclude that $i = j$ and $\alpha = i$ or $q+i$. Since $(x,i),(y,i)$ are distinct adjacent vertices, we conclude that $x \neq y$ and $xy \in E(G)$. If $\alpha = i$, then $d_G(x,v), d_G(y,v) \in \{0,2\}$ implies that G has a 3-cycle or a 5-cycle, contrary to the assumption that G has girth 6. If $\alpha = q+i$, then $d_G(v,x) = d_G(v,y) = 1$, and G has a 3-cycle, again a contradiction. This completes the proof of Claim 3.

So maps $\{\mu_t : t = 2q+1, 2q+2, \ldots, 4q+2\}$ are coloured by distinct colours, and hence there exists t such that $\Psi(\mu_t) \notin \{1,2,\ldots,2q\}$. As $\Psi(\mu_t) \in Im(\mu_t) = \{1,2,\ldots,q,t\}$, we have $\Psi(\mu_t) = t$.

Since $|\{b \in [c] : I(v,b) \text{ is large }\}| > c/2 = 2q+1$, there is a colour $b \in [c] - \{1,2,\ldots,2q,t\}$ such that $I(v,b)$ is large. Let $\theta \in K_c^{G^o}$ be defined as follows:

$$\theta(x) = \begin{cases} b, & \text{if } d_G(x,v) \geq 2, \\ t, & \text{if } d_G(x,v) \leq 1. \end{cases}$$

Claim. For $t \in \{2q+1, 2q+1, \ldots, 4q+2\}$, $\theta \sim \mu_t$.

Proof. Assume to the contrary that $\theta \not\sim \mu_t$. Then there is an edge $(x,i)(y,j) \in E(G[K_q])$ such that $\theta(x) = \theta(x,i) = \mu_t(y,j)$. (Note that $\theta(x,i) = \theta(x)$ as θ is a simple map). As $Im(\theta) \cap Im(\mu_t) = \{t\}$, we conclude that $\theta(x) = \mu_t(y,j) = t$. But then $d_G(x,v) \leq 1$ and $d_G(y,v) \geq 3$, and hence $x \neq y$ and $xy \notin E(G)$, contrary to the assumption that $(x,i)(y,j) \in E(G[K_q])$.

Thus $\Psi(\theta) \neq \Psi(\mu_t) = t$. As $\Psi(\theta) \in Im(\theta)$, we conclude that $\Psi(\theta) = b$.

Claim. For any $\phi \in I(v,b)$, there exists a vertex $x \neq v$ such that $\phi(x) \in \{b,t\}$.

Proof. Let $\phi \in I(v,b)$. By definition $\Psi(\phi) = b = \phi(v)$. So $\Psi(\phi) = \Psi(\theta)$. Hence $\phi \not\sim \theta$. So there is an edge $xy \in E(G^o)$ such that $\phi(x) = \theta(y)$. If $x = v$, then $\theta(y) = \phi(v) = b$. By definition of θ, we have $d_G(y,v) \geq 2$. Hence xy cannot be an edge in G^o, a contradiction. So $x \neq v$. As $\phi(x) = \theta(y) \in \{b,t\}$, this completes the proof of the claim.

For each $x \neq v$, let

$$J_x = \{\phi \in I(v,b) : \phi(x) \in \{b,t\}\}.$$

For a map $\phi \in J_x$, the image $\phi(v)$ of v is fixed, i.e., $\phi(v) = b$. The image $\phi(x)$ of x has two choices: b and t. For each of other $n-2$ vertices y of G, $\phi(y)$ has c choices. Therefore $|J_x| \leq 2c^{n-2}$. By Claim 3, $I(v,b) = \cup_{x \in V(G)-\{v\}} J_x$. So $|I(v,b)| \leq 2(n-1)c^{n-2}$, contrary to the assumption that $I(v,b)$ is large. This completes the proof of Theorem 1.

Remark 1. The key part of the proof of Theorem 1 is to show that $K_c^{G[K_q]}$ is not c-colourable. For each vertex v of G, for $t \in \{2q+1, 2q+2, \ldots, 4q+2\}$, let

$$\mu_{v,t}(x,i) = \begin{cases} i, & \text{if } d_G(x,v) = 0,2, \\ q+i, & \text{if } d_G(x,v) = 1, \\ t, & \text{if } d_G(x,v) \geq 3; \end{cases}$$

Let H be the subgraph of $K_c^{G[K_q]}$ induced by

$$V(K_c^{G^o}) \cup \{\mu_{v,t} : v \in V(G), t \in \{2q+1, 2q+2, \ldots, 4q+2\}\}.$$

What we have proved is that the subgraph H of $K_c^{G[K_q]}$ is not c-colourable. Note that H is a very tiny fraction of $K_c^{G[K_q]}$, although H by itself is a huge graph.

The reviewer of this note asks if there is an intuition as to why this subgraph H is the right thing to be thinking about. Also, once you have the intuition that this subgraph should have high chromatic number, why are the sets $I(b,v)$ the right things to look at to analyse this?

This is also a question in my mind. Reading Shitov's paper, one naturally wonders how did he come up with this proof? I am not the right person to answer this question. However, since one main purpose of this note is to explain Shitov's proof, I will give it a try.

The maps $\{g_i : i \in [c]\}$ is already a c-clique. So all the c colours are used by these maps in a proper c-colouring Ψ of $K_c^{G[K_q]}$, where we assume that $\Phi(g_i) = i$ for $i \in [c]$. To derive a contradiction, it is natural to consider maps that have many neighbors in this c-clique, namely, maps ϕ with a small image set $Im(\phi)$. The smallest image set has size 2 (for otherwise it is one of these constant maps). For each vertex v of G, for any two colors $b,t \in [c]$, let $\theta_{v,b,t} \in K_c^{G^o}$ be defined as

$$\theta_{v,b,t}(x) = \begin{cases} b, & \text{if } d_G(x,v) \geq 2, \\ t, & \text{if } d_G(x,v) \leq 1. \end{cases}$$

Now $\Psi(\theta_{v,b,t}) = b$ or t, as it is adjacent to every g_i with $i \neq b,t$. If we can somehow fix the colour of $\theta_{v,b,t}$ to be b, that is very useful. The maps $\mu_{v,t}$ are used to forces $\theta_{v,b,t}$ to be colored by b.

Now it remains to find a map ϕ with $\Psi(\phi) = b$ which is adjacent to $\theta_{v,b,t}$, so that we obtain a contradiction to the assumption that Ψ is a proper colouring of $K_c^{G[K_q]}$. The candidates are those maps $\phi \in K_c^{G^o}$ such that $\Psi(\phi) = \phi(v) = b$, because if $\Psi(\phi) = b$, then there is a vertex $x \in V(G)$ such that $\phi(x) = b$. If $x \neq v$, then ϕ is not adjacent to $\theta_{v,b,t}$. Indeed, the definition of $\theta_{v,b,t}$ is chosen in such a way that its neighors coloured with colour b has a simple structure.

This is why we have the definition of $I(v,b)$.

Can we find such a map in $I(v,b)$? Intuitively, this is promising: For a map $\phi \in I(v,b)$ to be a neighbor of $\theta_{v,b,t}$, one just need to avoid assigning color b to any other vertex, and avoid assigning color t to the neighbours of v. If $I(v,b)$ is large enough, then such a map shall exists. Once we have shown that for approriate v,b,t, $I(v,b)$ is large enough and by using maps $\mu_{v,t}$, we can force $\theta_{v,b,t}$ be coloured by b, we arrive at a contradiction.

4 The Poljak-Rödl function

The Poljak-Rödl function is defined in [9]:

$$f(n) = \min\{\chi(G \times H) : \chi(G), \chi(H) \geq n\}.$$

Hedetniemi's conjecture is equivalent to saying that $f(n) = n$ for all positive integer n. Shitov's Theorem says that for sufficiently large n, $f(n) \leq n-1$. Using Shitov's result, Tardif and Zhu [16] proved that $f(n) \leq n - (\log n)^{1/4-o(1)}$. Tardif and Zhu asked in [16] if there is a positive constant ε such that $f(n) \leq (1-\varepsilon)n$ for sufficiently large n. This question was answered in affirmative by He and Wigderson [4] with $\varepsilon \approx 10^{-9}$. On the other hand, in [16], Tardif and Zhu proved that if a special case of a conjecture of Stahl [12] conserning the multi-chromatic number of Kneser graph is true, then we have $\limsup_{n\to\infty} \frac{f(n)}{n} \leq \frac{1}{2}$.

Recently, I proved in [20] that the conclusion $\limsup_{n\to\infty} \frac{f(n)}{n} \leq \frac{1}{2}$ holds without assuming Stahl's conjecture.

Theorem 2. *For $d \geq 1$, let G be a p-vertex graph of girth 6 and with $\alpha(G) \leq \frac{p}{8.1d}$. Let $q \geq 2^{p-1}p^2$ and $c = 4q+2$. Then $\chi(G[K_q]) \geq 2dc - 2c + 2$ and $\chi(K_{dc}^{G[K_q]}) \geq 2dc - 2c + 2$. Consequently, $f(2dc-2c+2) \leq dc$.*

Proof. As explained before, the existence of a graph G described above was proved by Erdős. Similarly as in the proof of Theorem 1, $\chi(G[K_q]) \geq \frac{|V(G[K_q])}{\alpha(G[K_q])} \geq \frac{pq}{p/8.1d} = 8.1dq \geq 2dc > 2dc - 2c + 2$. Now we show that $\chi(K_{dc}^{G[K_q]}) \geq 2dc - 2c + 2$.

Assume Ψ is a $(dc+t)$-colouring of $K_{dc}^{G[K_q]}$ with colour set $[dc+t]$. We shall show that $dc + t \geq 2dc - 2c + 2$, i.e., $t \geq dc - 2c + 2$. Let $S = [dc+t] - [dc]$. The colours in $[dc]$ are called *primary colours* and colours in S are called *secondary colours*. So we have $t = |S|$ secondary colours.

Similarly as before, we may assume that $\Psi(g_i) = i$ for $i \in [dc]$. Then for any map $\phi \in K_{dc}^{G[K_q]}$, if $i \notin Im(\phi)$, then $\phi \sim g_i$ and $\Psi(\phi) \neq i$. Thus for any $\phi \in K_{dc}^{G[K_q]}$, $\Psi(\phi) \in Im(\phi) \cup S$.

For positive integers $m \geq 2k$, let $K(m,k)$ be the Kneser graph whose vertices are k-susbets of $[m]$, and for two k-subsets A, B of $[m]$, $A \sim B$ if $A \cap B = \emptyset$. It was proved by Lovász in [7] that $\chi(K(m,k)) = m - 2k + 2$.

For a c-subset A of $[cd]$, let H_A be the subgraph of $K_{cd}^{G[K_q]}$ induced by

$$\{\phi \in V(K_{cd}^{G[K_q]}) : Im(\phi) \subseteq A\}.$$

Then H_A is isomorphic to $K_c^{G[K_q]}$. By Theorem 1, $|\Psi(H_A)| \geq c+1$. As $Im(\phi) \subseteq A$ and $|A| = c$, $\Phi(H_A)$ contains at least one secondary colour. Let $\tau(A)$ be an arbitrary secondary colour contained in $\Psi(H_A)$.

If A, B are c-subsets of $[dc]$ and $A \cap B = \emptyset$, then every vertex in H_A is adjacent to every vertex in H_B. Hence $\Psi(H_A) \cap \Psi(H_B) = \emptyset$. In particular, $\tau(A) \neq \tau(B)$. Thus τ is a proper colouring of the Kneser graph $K(dc,c)$. As $\chi(K(dc,c)) = dc - 2c + 2$, we conclude that $t = |S| \geq dc - 2c + 2$. This completes the proof of Theorem 2.

For a positive integer d, let $p = p(d)$ be the minimum number of vertices of a graph G with girth 6 and $\chi_f(G) \geq 8.1d$. It follows from Theorem 2 that for any integer $q \geq p^2 2^{p-1}$, $f(2(d-1)(4q+2)+2) \leq (4q+2)d$. As $f(n)$ is non-decreasing, for integers n in the interval $[2(d-1)(4q+2)+2, 2(d-1)(4q+6)+2]$, we have $f(n) \leq (4q+6)d$.

Hence for all integers $n \geq 2(4q+2)(d-1)+2$,

$$\frac{f(n)}{n} \leq \frac{(4q+6)d}{2(4q+2)(d-1)+3} = \frac{1}{2} + \frac{4q+4d+1}{2(d-1)(4q+2)+2}.$$

Note that if $d \to \infty$, then $p = p(d)$ goes to infinity, and $q \geq p^3 2^p$ goes to infinity. Therefore

$$\limsup_{n\to\infty} \frac{f(n)}{n} \leq \frac{1}{2}.$$

Theorem 2 improves the result of He and Wigderson [4]. However, He and Wigderson use a modification of Shitov's method, which might be of independent interest.

In the proof of Theorem 2, we actually showed that a tiny subgraph of $K_{dc}^{G[K_q]}$ has chromatic number close to $2dc$. It is not clear if the remaining part of the graph

$K_{dc}^{G[K_q]}$ can be used to show that this graph actually has a much larger chromatic number. We observe that if one can show that the chromatic number of $K_{dc}^{G[K_q]}$ is more than kdc for some positive integer k, then Stahl's conjecture implies that $\limsup_{n\to\infty} \frac{f(n)}{n} \leq \frac{1}{k+1}$.

5 Lower bound for $f(n)$

The breakthrough result of Shitov leads to an improvement of the upper bound for the function $f(n)$. On the other hand, the only known lower bound for $f(n)$ is that $f(n) \geq 4$ for $n \geq 4$. We do not know if $f(n)$ is bounded by a constant or not. What we do know is that if $f(n)$ is bounded by a constant, then the smallest such constant is at most 9.

To prove this result, we need to consider the product of digraphs. For a digraph D, we use $A(D)$ to denote the set of arcs of D. An arc in D is either denoted by an ordered pair (x,y), or by an arrow $x \to y$. Digraphs are allowed to have digons, i.e., a pair of opposite arcs.

Assume D_1, D_2 are digraphs. The product $D_1 \times D_2$ has vertex set $V(D_1) \times V(D_2)$, where $(x,y) \to (x',y')$ is an arc if and only if (x,x') is an arc in D_1 and (y,y') is an arc in D_2. The chromatic number of a digraph D is defined to be $\chi(\underline{D})$, where $\underline{D}$ is the underlying graph of D, i.e., obtained from D by replacing each arc (x,y) with an edge xy. Given a digraph D, let D^{-1} be the digraph obtained from D by reversing the direction of all its arcs. It is easy to see that for any digraphs D_1, D_2,

$$\underline{D_1} \times \underline{D_1} = \underline{(D_1 \times D_2)} \cup \underline{(D_1 \times D_2^{-1})}.$$

Hence

$$\chi(\underline{D_1} \times \underline{D_1}) \leq \chi(D_1 \times D_2) \times \chi(D_1 \times D_2^{-1}).$$

Let

$$g(n) = \min\{\chi(D_1 \times D_2) : \chi(D_1), \chi(D_2) \geq n\},$$
$$h(n) = \min\{\max\{\chi(D_1 \times D_2), \chi(D_1 \times D_2^{-1})\} : \chi(D_1), \chi(D_2) \geq n\}.$$

Since $E(\underline{D}_1 \times \underline{D}_2) = E(\underline{D_1 \times D_2}) \cup E(\underline{D_1 \times D_2^{-1}})$, we have

$$g(n) \leq h(n) \leq f(n) \leq h(n)^2.$$

The following result was proved by Poljak and Rödl in [9].

Theorem 3. *If $g(n)$ (respectively $h(n)$) is bounded by a constant, then the smallest such constant is at most* 4. *Consequently, if $f(n)$ is bounded by a constant, then the smallest such constant is at most* 16.

Proof. For a graph D, let $\partial(D)$ be the digraph with vertex set $A(D)$, where $(x,y) \to (x',y')$ is an arc of $\partial(D)$ if and only if $y = x'$. In particular, if $(x,y),(y,x)$ is a digon in D, then $(x,y) \to (y,x)$ and $(y,x) \to (x,y)$ is a digon in $\partial(D)$.

Lemma 4. *For any digraph D,*

$$\min\{k : 2^k \geq \chi(D)\} \leq \chi(\partial(D)) \leq \min\{k : \binom{k}{\lceil k/2 \rceil} \geq \chi(D)\}.$$

Proof. If $\phi : V(\partial(D)) \to [k]$ is a proper colouring of $\partial(D)$, then for each vertex v of D, let $\psi(v) = \{\phi(e) : e \in A^+(v)\}$, where $A^+(v)$ is the set of out-arcs at v. Then ψ is a proper colouring of D (with subsets of $[k]$ as colours). Indeed, if $e = (x,y)$ is an arcs of D, then $\phi(e) \in \psi(x) - \psi(y)$. So $\psi(x) \neq \psi(y)$. The number of colours used by ψ is at most the number of subsets of $[k]$, which is 2^k.

If $\psi : V(D) \to \binom{k}{\lceil k/2 \rceil}$ is a proper colouring of D (where the colours are $\lceil k/2 \rceil$-subsets of $[k]$), then for any arc $e = (x,y)$ of D, let $\phi(e)$ be any integer in $\psi(y) - \psi(x)$ (as $\psi(y) \neq \psi(x)$, such an integer exists). Then if $(x,y) \to (y,z)$ is an arc in $\partial(D)$, then $\phi(x,y) \in \psi(y)$ and $\phi(y,z) \notin \psi(y)$. Hence $\phi(x,y) \neq \phi(y,z)$. I.e., ϕ is a proper colouring of $\partial(D)$. This completes the proof of Lemma 4.

It follows easily from the definition that

$$\partial(D_1 \times D_2) = \partial(D_1) \times \partial(D_2),$$
$$\partial(D^{-1}) = (\partial(D))^{-1}.$$

Suppose $g(n)$ is bounded and C is the smallest upper bound. As $g(n)$ is non-decreasing, there is an integer n_0 such that $g(n) = C$ for all $n \geq n_0$. Let $n_1 = 2^{n_0}$, and let D_1, D_2 be digraphs with $\chi(D_1), \chi(D_2) \geq n_1$ and $\chi(D_1 \times D_2) = C$. It follows from Lemma 4 that $\chi(\partial(D_1)), \chi(\partial(D_2)) \geq n_0$ and hence $\chi(\partial(D_1) \times \partial(D_2)) = \chi(\partial(D_1 \times D_2)) \geq C$. By Lemma 4 again, we have

$$C \geq \chi(D_1 \times D_2) > \binom{C-1}{\lceil (C-1)/2 \rceil}.$$

This implies that $C \leq 4$.

The same argument shows that if $h(n)$ is bounded by a constant, then the smallest such constant is at most 4. Since $h(n) \leq f(n) \leq h(n)^2$, if $f(n)$ is bounded by a constant, then the smallest such a constant is at most 16.

Next we show that if $g(n)$ (respectively, $h(n)$) is bounded by a constant, then the smallest such constant cannot be 4. Assume to the contrary that the smallest constant bound for $g(n)$ is 4. Let n_0 be the integer given above, and let $n_1 = 2^{n_0}, n_2 = 2^{n_1}$. Then $g(n_2) = g(n_1) = g(n_0) = 4$. Let D_1, D_2 be two digraphs with $\chi(D_1), \chi(D_2) \geq n_2$ and $\chi(D_1 \times D_2) = 4$. The same argument as above shows that

$$\chi(\partial(\partial(D_1 \times D_2))) = 4.$$

However, we shall show that if $\chi(D) \leq 4$, then $\chi(\partial(\partial(D))) \leq 3$. Let $\vec{K}_4$ be the complete digraph with vertex set $\{1,2,3,4\}$, where (i,j) is an arc for any distinct $i,j \in \{1,2,3,4\}$. If $\chi(D) = 4$, then D admits a homomorphism to $\vec{K}_4$. Hence $\partial(\partial(D))$ admits a homomorphism to $\partial(\partial(\vec{K}_4))$. So it suffices to show that $\partial(\partial(\vec{K}_4)) \leq 3$. In 1990, I was a Ph.D. student at The University of Calgary. After reading the paper by Poljak and Rödl [9], I found a 3-colouring of $\partial(\partial(\vec{K}_4))$ by brute force. I was happy to tell this to my supervisor Professor Norbert Sauer, who then told the result to Duffus. Then I learned from Duffus the following elegant 3-colouring of $\partial(\partial(\vec{K}_4))$, given earlier by Schelp that was not published.

Each vertex of $\partial(\partial(\vec{K}_4))$ is a sequence ijk with $i,j,k \in [k]$, $i \neq j$, $j \neq k$ (but i may equal to k). Let

$$c(ijk) = \begin{cases} j, & \text{if } j \neq 4, \\ s, & \text{if } j = 4 \text{ and } s \in \{1,2,3\} - \{i,k\} \end{cases}$$

Then it is easy to verify that c is a proper 3-colouring of $\partial(\partial(\vec{K}_4))$. This completes the proof that $g(n)$ is either bounded by 3 or goes to infinity. Similarly, $h(n)$ is either bounded by 3 or goes to infinity, and consequently, $f(n)$ is either bounded by 9 or goes to infinity.

Later I learned from Hell that Poljak also obtained this strengthening independently and that was published later (in 1992) [8].

Tardif and Wehlau [15] proved that $f(n)$ is bounded if and only if $g(n)$ is bounded.

The fractional version of Hedetniemi's conjecture was proved in [19]: For any two graphs G and H, $\chi_f(G \times H) = \min\{\chi_f(G), \chi_f(H)\}$. Thus if $f(n)$ is bounded by 9, and G and H are n-chromatic graphs with $\chi(G \times H) \leq 9$, then at least one of G and H has fractional chromatic number at most 9.

In [19], I defined the following Poljak-Rödl type function:

$$\psi(n) = \min\{\chi(G \times H) : \chi_f(G), \chi(H) \geq n\}.$$

I proposed a weaker version of Hedetniemi's conjecture, which is equivalent to the statement that $\psi(n) = n$ for all positive integer n. However, Shitov's proof actually refutes this weaker version of Hedetniemi's conjecture, as the graph G used in the proof of Theorem 1 has large fractional chromatic number. The proof of Theorem 2 shows that

$$\limsup_{n \to \infty} \frac{\psi(n)}{n} \leq \frac{1}{2}.$$

On the other hand, it follows from the definition that $f(n) \leq \psi(n)$. A natural question is the following:

Question 1. Is $\psi(n)$ bounded by a constant? If $\psi(n)$ is bounded by a constant, what could be the smallest such constant?

Remark Very recently, I constructed relatively small counterexample to Hedetniemi's conjecture in [21]: There are graphs G and H with $3,403$ and $10,501$ vertices respetively such that $\chi(G), \chi(H) \geq 126$ and $\chi(G \times H) \leq 125$.

References

1. M. El-Zahar, N.W. Sauer, *The chromatic number of the product of two 4- chromatic graphs is 4*, Combinatorica 5 (1985) 121–126.
2. P. Erdős, *Graph theory and probability*, Canad. J. Math. 11 (1959), 34– 38.
3. R. Häggkvist, P. Hell, J. Miller and V. Neumann Lara, *On multiplicative graphs and the product conejcture*, Combiantorica 8(1988), 71-81.
4. X. He and Y. Wigderson, *Hedetniemi's conjecture is asympototically false*, J. Combin. Theory, Seri. B, 2020. https://doi.org/10.1016/j.jctb.2020.03.003, arXiv:1906.06/83v2.
5. S. Hedetniemi, *Homomorphisms of graphs and automata*, Technical Report 03105-44-T, University of Michigan, 1966.
6. S. Klavžar, *Coloring graph products—a survey*, Discrete Math. 155 (1996) 135–145.
7. L. Lovász, *Kneser's conjecture, chromatic number, and homotopy*, J. Combin. Theory Ser. A 25 (1978), 319–324.
8. S. Poljak, *Coloring digraphs by iterated antichains*, Comment. Math. Univ. Carolin. 32(1991), 209-212.
9. S. Poljak and V. Rödl, *On the arc-chromatic number of a digraph*, J. Combin. Theory Ser. B 31(1981), 339-350.
10. N. Sauer, *Hedetniemi's conjecture—a survey*, Discrete Math. 229 (2001) 261–292.
11. Y. Shitov, *Counterexamples to Hedetniemi's Conjecture*, Ann. of Math. (2) 190(2019), no. 2, 663–667.
12. S. Stahl, *n-tuple colorings and associated graphs*, J. Combinatorial Theory Ser. B 20 (1976), no. 2, 185–203.
13. C. Tardif, *Hedetniemi's conjecture, 40 years later*, Graph Theory Notes N. Y. 54 (2008) 46–57.
14. C. Tardif, *Multiplicative graphs and semi-lattice endomorphisms in the category of graphs*, Journal of Combinatorial Theory Ser. B 95 (2005), 338-345.
15. C. Tardif and D. Wehlau, *Chromatic numbers of products of graphs: the directed and undirected versions of the Poljak-Rödl function*, J. Graph Theory 51(2006), 33-36.
16. C. Tardif and X. Zhu, *A note on Hedetniemi's conjecture, Stahl'sconjecture and the Poljak-Rödl function*, Electronic J. Combin. 26(4)(2019), #P4.321
17. D. B. West. *Introduction to graph theory*. Prentice Hall, Inc., Upper Saddle River, NJ, 1996.
18. X. Zhu, *A survey on Hedetniemi's conjecture*, Taiwanese J. Math. 2 (1998) 1–24.
19. X. Zhu, *The fractional version of Hedetniemi's conjecture is true*, European J. Combin. 32 (2011), 1168–1175.
20. X. Zhu, *A note on Poljak-Rödl function*, Electronic J. Combin. 27(3)(2020), #P3.2. https://doi.org/10.37236/9371.
21. X. Zhu, *Relatively small counterexamples to Hedetniemi's conjecture*, manuscript, 2020, arXiv:2004.09028.

Notes on Graph Product Structure Theory

Zdeněk Dvořák, Tony Huynh, Gwenaël Joret, Chun-Hung Liu, and David R. Wood

Abstract It was recently proved that every planar graph is a subgraph of the strong product of a path and a graph with bounded treewidth. This paper surveys generalisations of this result for graphs on surfaces, minor-closed classes, various non-minor-closed classes, and graph classes with polynomial growth. We then explore how graph product structure might be applicable to more broadly defined graph classes. In particular, we characterise when a graph class defined by a cartesian or strong product has bounded or polynomial expansion. We then explore graph product structure theorems for various geometrically defined graph classes, and present several open problems.

Zdeněk Dvořák
Charles University, Prague, Czech Republic
Supported by project 17-04611S (Ramsey-like aspects of graph coloring) of Czech Science Foundation
e-mail: rakdver@iuuk.mff.cuni.cz

Tony Huynh and David R. Wood
School of Mathematics, Monash University, Melbourne, Australia
Research supported by the Australian Research Council
e-mail: tony.bourbaki@gmail.com, david.wood@monash.edu

Gwenaël Joret
Département d'Informatique, Université libre de Bruxelles, Brussels, Belgium
Research supported by the Wallonia-Brussels Federation of Belgium, and by the Australian Research Council
e-mail: gjoret@ulb.ac.be

Chun-Hung Liu
Department of Mathematics, Texas A&M University, College Station, Texas, USA
Partially supported by the NSF under Grant No. DMS-1929851
e-mail: chliu@math.tamu.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_32

1 Introduction

Studying the structure of graphs is a fundamental topic of broad interest in combinatorial mathematics. At the forefront of this study is the Graph Minor Theorem of Robertson and Seymour [47], described by Diestel [8] as "among the deepest theorems mathematics has to offer". At the heart of the proof of this theorem is the Graph Minor Structure Theorem, which shows that any graph in a minor-closed family[1] can be constructed using four ingredients: graphs on surfaces, vortices, apex vertices, and clique-sums. Graphs of bounded genus, and in particular planar graphs are basic building blocks in graph minor structure theory. Indeed, the theory says nothing about the structure of planar graphs. So it is natural to ask whether planar graphs can be described in terms of some simpler graph classes. In a recent breakthrough, Dujmović, Joret, Micek, Morin, Ueckerdt, and Wood [15, 16] provided an answer to this question by showing that every planar graph is a subgraph of the strong product[2] of a graph of bounded treewidth[3] and a path.

Theorem 1 ([15, 16])**.** *Every planar graph is a subgraph of:*
(a) $H \boxtimes P$ *for some graph* H *of treewidth at most* 8 *and for some path* P*;*
(b) $H \boxtimes P \boxtimes K_3$ *for some graph* H *of treewidth at most* 3 *and for some path* P*.*

This *graph product structure theorem* is attractive since it describes planar graphs in terms of graphs of bounded treewidth, which are considered much simpler than planar graphs. For example, many NP-complete problem remain NP-complete on planar graphs but are polynomial-time solvable on graphs of bounded treewidth.

Despite being only 10 months old, Theorem 1 is already having significant impact. Indeed, it has been used to solve two major open problems and make additional progress on two other longstanding problems:

- Dujmović et al. [15, 16] use Theorem 1 to show that planar graphs have queue layouts with a bounded number of queues, solving a 27 year old problem of Heath, Leighton, and Rosenberg [33].

[1] A graph H is a *minor* of a graph G if a graph isomorphic to H can be obtained from a subgraph of G by contracting edges. A class of graphs $\mathscr{G}$ is *minor-closed* if for every graph $G \in \mathscr{G}$ every minor of G is in $\mathscr{G}$, and some graph is not in $\mathscr{G}$. A graph G is *H-minor-free* if H is not a minor of G.

[2] The *cartesian product* of graphs A and B, denoted by $A \square B$, is the graph with vertex set $V(A) \times V(B)$, where distinct vertices $(v,x),(w,y) \in V(A) \times V(B)$ are adjacent if: $v = w$ and $xy \in E(B)$; or $x = y$ and $vw \in E(A)$. The *strong product* of graphs A and B, denoted by $A \boxtimes B$, is the graph with vertex set $V(A) \times V(B)$, where distinct vertices $(v,x),(w,y) \in V(A) \times V(B)$ are adjacent if: $v = w$ and $xy \in E(B)$; or $x = y$ and $vw \in E(A)$; or $vw \in E(A)$ and $xy \in E(B)$. If X is a subgraph of $A \square B$, then the *projection* of X into A is the set of vertices $v \in V(A)$ such that $(v,w) \in V(X)$ for some $w \in V(B)$.

[3] A *tree decomposition* of a graph G is a collection $(B_x \subseteq V(G) : x \in V(T))$ of subsets of $V(G)$ (called *bags*) indexed by the nodes of a tree T, such that (i) for every edge $uv \in E(G)$, some bag B_x contains both u and v, and (ii) for every vertex $v \in V(G)$, the set $\{x \in V(T) : v \in B_x\}$ induces a non-empty (connected) subtree of T. The *width* of a tree decomposition is the size of the largest bag minus 1. The *treewidth* of a graph G, denoted by $\mathrm{tw}(G)$, is the minimum width of a tree decomposition of G. See [3, 4, 32, 45, 46] for surveys on treewidth. A *path decomposition* is a tree decomposition where the underlying tree is a path. The *pathwidth* of a graph G, denoted by $\mathrm{pw}(G)$, is the minimum width of a path decomposition of G.

- Dujmović, Eppstein, Joret, Morin, and Wood [12] use Theorem 1 to show that planar graphs can be nonrepetitively coloured with a bounded number of colours, solving a 17 year old problem of Alon, Grytczuk, Hałuszczak, and Riordan [1].
- Dębski, Felsner, Micek, and Schröder [10] use Theorem 1 to prove the best known results on p-centred colourings of planar graphs, reducing the bound from $O(p^{19})$ to $O(p^3 \log p)$.
- Bonamy, Gavoille, and Pilipczuk [5] use Theorem 1 to give more compact graph encodings of planar graphs. In graph-theoretic terms, this implies the existence of a graph with $n^{4/3+o(1)}$ vertices that contains each planar graph with at most n vertices as an induced subgraph, This work improves a sequence of results that goes back 27 years to the introduction of implicit labelling schemes by Kannan, Naor, and Rudich [34].

The first goal of this paper is to introduce several product structure theorems that have been recently established, most of which generalise Theorem 1. First Section 2 considers minor-closed classes. Then Section 3 considers several examples of non-minor-closed classes. Section 4 introduces the notion of graph classes with polynomial growth and their characterisation in terms of strong products of paths due to Krauthgamer and Lee [36]. We prove an extension of this result for strong products of graphs of given pathwidth.

The remaining sections explore how graph product structure might be applicable to more broadly defined graph classes. The following definition by Nešetřil and Ossona de Mendez [40] provides a setting for this study[4]. A graph class $\mathscr{G}$ has *bounded expansion* with *expansion function* $f : \mathbb{Z}^+ \to \mathbb{R}$ if, for every graph $G \in \mathscr{G}$ and for all disjoint subgraphs $B_1, \dots, B_t$ of radius at most r in G, every subgraph of the graph obtained from G by contracting each B_i into a vertex has average degree at most $f(r)$. When $f(r)$ is a constant, $\mathscr{G}$ is contained in a proper minor-closed class. As $f(r)$ is allowed to grow with r we obtain larger and larger graph classes. A graph class $\mathscr{G}$ has *linear expansion* if $\mathscr{G}$ has bounded expansion with an expansion function in $O(r)$. A graph class $\mathscr{G}$ has *polynomial expansion* if $\mathscr{G}$ has bounded expansion with an expansion function in $O(r^c)$, for some constant c.

We characterise when a graph class defined by a cartesian or strong product has bounded or polynomial expansion. For $\star \in \{\boxtimes, \square\}$ and for hereditary[5] graph classes $\mathscr{G}_1$ and $\mathscr{G}_2$, let

$$\mathscr{G}_1 \star \mathscr{G}_2 := \{G : G \subseteq G_1 \star G_2, G_1 \in \mathscr{G}_1, G_2 \in \mathscr{G}_2\}.$$

Note that $\mathscr{G}_1 \star \mathscr{G}_2$ is hereditary. Sections 5 and 6 characterise when $\mathscr{G}_1 \star \mathscr{G}_2$ has bounded or polynomial expansion. In related work, Wood [53] characterised when

[4] Let $d_G(u,v)$ be the distance between vertices u and v in a graph G. For a vertex v in a graph G and $r \in \mathbb{N}$, let $N_G^r(v)$ be the set of vertices of G at distance exactly r from v, and let $N_G^r[v]$ be the set of vertices at distance at most r from v. The set $N_G^r[v]$ is called an *r-ball*. We drop the subscript G when the graph is clear from the context.

[5] A class of graphs is *hereditary* if it is closed under induced subgraphs.

$\mathscr{G}_1 \square \mathscr{G}_2$ has bounded Hadwiger number, and Pecaninovic [43] characterised when $\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has bounded Hadwiger number.

Section 7 explores graph product structure theorems for various geometrically defined graph classes. We show that multi-dimensional unit-disc graphs have a product structure theorem, and discusses whether two other naturally defined graph classes might have product structure theorems. We finish with a number of open problems in Section 8.

2 Minor-Closed Classes

Here we survey results generalising Theorem 1 for minor-closed classes. First consider graphs embeddable on a fixed surface[6].

Theorem 2 ([15, 16])**.** *Every graph of Euler genus g is a subgraph of:*

(a) $H \boxtimes P \boxtimes K_{\max\{2g,1\}}$ *for some graph H of treewidth at most* 9 *and for some path P;*

(b) $H \boxtimes P \boxtimes K_{\max\{2g,3\}}$ *for some graph H of treewidth at most* 4 *and for some path P.*

(c) $(K_{2g} + H) \boxtimes P$ *for some graph H of treewidth at most* 8 *and some path P.*

Here $A + B$ is the complete join of graphs A and B. The proof of Theorem 2 uses an elegant cutting lemma to reduce to the planar case.

Theorem 2 is generalised as follows. A graph X is *apex* if $X - v$ is planar for some vertex v.

Theorem 3 ([15, 16])**.** *For every apex graph X, there exists $c \in \mathbb{N}$ such that every X-minor-free graph G is a subgraph of $H \boxtimes P$ for some graph H of treewidth at most c and some path P.*

The proof of Theorem 3 is based on the Graph Minor Structure Theorem of Robertson and Seymour [49] and in particular a strengthening of it by Dvořák and Thomas [21].

For an arbitrary proper minor-closed class, apex vertices are unavoidable; in this case Dujmović et al. [15, 16] proved the following product structure theorem.

Theorem 4 ([15, 16])**.** *For every proper minor-closed class $\mathscr{G}$ there exist $k, a \in \mathbb{N}$ such that every graph $G \in \mathscr{G}$ can be obtained by clique-sums of graphs $G_1, \dots, G_n$ such that for $i \in \{1, \dots, n\}$,*

$$G_i \subseteq (H_i \boxtimes P_i) + K_a,$$

for some graph H_i with treewidth at most k and some path P_i.

[6] The *Euler genus* of the orientable surface with h handles is $2h$. The *Euler genus* of the non-orientable surface with c cross-caps is c. The *Euler genus* of a graph G is the minimum Euler genus of a surface in which G embeds (with no crossings). See [39] for background on embeddings of graphs on surfaces.

If we assume bounded maximum degree, then apex vertices in the Graph Minor Structure Theorem can be avoided, which leads to the following theorem of Dujmović, Esperet, Morin, Walczak, and Wood [14].

Theorem 5 ([14])**.** *For every proper minor-closed class $\mathscr{G}$, every graph in $\mathscr{G}$ with maximum degree Δ is a subgraph of $H \boxtimes P$ for some graph H of treewidth $O(\Delta)$ and for some path P.*

It is worth highlighting the similarity of Theorem 5 and the following result of Ding and Oporowski [9] (refined in [52]).Theorem 6 says that graphs of bounded treewidth and bounded degree are subgraphs of the product of a tree and a complete graph of bounded size, whereas Theorem 5 says that graphs excluding a minor and with bounded degree are subgraphs of the product of a bounded treewidth graph and a path.

Theorem 6 ([9, 52])**.** *Every graph with maximum degree $\Delta \geqslant 1$ and treewidth at most $k \geqslant 1$ is a subgraph of $T \boxtimes K_{18k\Delta}$ for some tree T.*

3 Non-Minor Closed Classes

A recent direction pursued by Dujmović, Morin, and Wood [17] studies graph product structure theorems for various non-minor-closed graph classes. First consider graphs that can be drawn on a surface of bounded genus and with a bounded number of crossings per edge. A graph is (g,k)-*planar* if it has a drawing in a surface of Euler genus at most g such that each edge is involved in at most k crossings. Even in the simplest case, there are $(0,1)$-planar graphs that contain arbitrarily large complete graph minors [13].

Theorem 7 ([17])**.** *Every (g,k)-planar graph is a subgraph of $H \boxtimes P$, for some graph H of treewidth $O(gk^6)$ and for some path P.*

Map and string graphs provide further examples of non-minor-closed classes that have product structure theorems.

Map graphs are defined as follows. Start with a graph G_0 embedded in a surface of Euler genus g, with each face labelled a 'nation' or a 'lake', where each vertex of G_0 is incident with at most d nations. Let G be the graph whose vertices are the nations of G_0, where two vertices are adjacent in G if the corresponding faces in G_0 share a vertex. Then G is called a (g,d)-*map graph*. A $(0,d)$-map graph is called a (plane) d-*map graph*; see [7, 26] for example. The $(g,3)$-map graphs are precisely the graphs of Euler genus at most g; see [13]. So (g,d)-map graphs generalise graphs embedded in a surface, and we now assume that $d \geqslant 4$ for the remainder of this section.

Theorem 8 ([17])**.** *Every (g,d)-map graph is a subgraph of:*

- *$H \boxtimes P \boxtimes K_{O(d^2g)}$, where H is a graph with treewidth at most 14 and P is a path,*
- *$H \boxtimes P$, where H is a graph with treewidth $O(gd^2)$ and P is a path.*

A *string graph* is the intersection graph of a set of curves in the plane with no three curves meeting at a single point; see [27, 28, 42] for example. For $\delta \in \mathbb{N}$, if each curve is in at most δ intersections with other curves, then the corresponding string graph is called a δ-*string graph*. A (g,δ)-*string* graph is defined analogously for curves on a surface of Euler genus at most g.

Theorem 9 ([17])**.** *Every (g,δ)-string graph is a subgraph of $H \boxtimes P$, for some graph H of treewidth $O(g\delta^7)$ and some path P.*

Theorems 7 to 9 all follow from a more general result of Dujmović et al. [17]. A collection $\mathscr{P}$ of paths in a graph G is a (k,d)-*shortcut system* (for G) if:

- every path in $\mathscr{P}$ has length at most k, and
- for every $v \in V(G)$, the number of paths in $\mathscr{P}$ that use v as an internal vertex is at most d.

Each path $P \in \mathscr{P}$ is called a *shortcut*; if P has endpoints v and w then it is a vw-*shortcut*. Given a graph G and a (k,d)-shortcut system $\mathscr{P}$ for G, let $G^{\mathscr{P}}$ denote the supergraph of G obtained by adding the edge vw for each vw-shortcut in $\mathscr{P}$.

Theorem 10 ([17])**.** *Let G be a subgraph of $H \boxtimes P$, for some graph H of treewidth at most t and for some path P. Let $\mathscr{P}$ be a (k,d)-shortcut system for G. Then $G^{\mathscr{P}}$ is a subgraph of $J \boxtimes P'$ for some graph J of treewidth at most $d(k^3+3k)\binom{k+t}{t}-1$ and some path P'.*

Theorems 7 to 9 are then proved by simply constructing a shortcut system. For example, by adding a dummy vertex at each crossing, Dujmović et al. [17] noted that every (g,k)-planar graph is a subgraph of $G^{\mathscr{P}}$ for some graph G of Euler genus at most g and for some $(k+1,2)$-shortcut system $\mathscr{P}$ for G.

Powers of graphs can also be described by a shortcut system. The *k-th power* of a graph G is the graph G^k with vertex set $V(G^k) := V(G)$, where $vw \in E(G^k)$ if and only if $d_G(v,w) \leqslant k$. Dujmović et al. [17] noted that if a graph G has maximum degree Δ, then $G^k = G^{\mathscr{P}}$ for some $(k,2k\Delta^k)$-shortcut system $\mathscr{P}$. Theorem 10 then implies:

Theorem 11 ([17])**.** *For every graph G of Euler genus g and maximum degree Δ, the k-th power G^k is a subgraph of $H \boxtimes P$, for some graph H of treewidth $O(g\Delta^k k^8)$ and some path P.*

4 Polynomial Growth

This section discusses graph classes with polynomial growth. A graph class $\mathscr{G}$ has *polynomial growth* if for some constant c, for every graph $G \in \mathscr{G}$, for each $r \geqslant 2$ every r-ball in G has at most r^c vertices. For example, every r-ball in an $n \times n$ grid graph is contained in a $(2r+1) \times (2r+1)$ subgrid, which has size $(2r+1)^2$; therefore the class of grid graphs has polynomial growth. More generally, let $\mathbb{Z}^d$ be the strong product of d infinite two-way paths. That is, $V(\mathbb{Z}^d) = \{(x_1,\dots,x_d) : x_1,\dots,x_d \in \mathbb{Z}\}$ where distinct vertices $(x_1,\dots,x_d)$ and $(y_1,\dots,y_d)$ are adjacent in

$\mathbb{Z}^d$ if and only if $|x_i - y_i| \leqslant 1$ for each $i \in \{1,\dots,d\}$. Then every r-ball in $\mathbb{Z}^d$ has size at most $(2r+1)^d$. Krauthgamer and Lee [36] characterised the graph classes with polynomial growth as the subgraphs of $\mathbb{Z}^d$.

Theorem 12 ([36])**.** *Let G be a graph such that for some constant c and for every integer $r \geqslant 2$, every r-ball in G has at most r^c vertices. Then $G \subseteq \mathbb{Z}^{O(c \log c)}$.*

We show that a seemingly weaker condition also characterises graph classes with polynomial growth. (We emphasise that in Theorem 13, H_1 does not necessarily have bounded maximum degree.)

Theorem 13. *The following are equivalent for a class of graphs $\mathscr{G}$:*

(1) $\mathscr{G}$ has polynomial growth,

(2) there exists $d \in \mathbb{N}$ such that every graph in $\mathscr{G}$ is a subgraph of $\mathbb{Z}^d$,

(3) there exist $d, k, \ell, \Delta \in \mathbb{N}$ such that for every graph $G \in \mathscr{G}$ there exist graphs $H_1, \dots, H_d$ such that:

- *G has maximum degree Δ,*
- *$\mathrm{pw}(H_i) \leqslant k$ for each $i \in \{1,\dots,k\}$,*
- *H_i has maximum degree at most ℓ for each $i \in \{2,\dots,d\}$,*
- *$G \subseteq H_1 \boxtimes H_2 \boxtimes \cdots \boxtimes H_d$.*

Proof. Krauthgamer and Lee [36] proved that (1) and (2) are equivalent. It is immediate that (2) implies (3) with $k = 1$ and $\ell = 2$ and $\Delta = 3^d - 1$. So it suffices to show that (3) implies (1). Consider graphs $G \in \mathscr{G}$ and $H_1, \dots, H_d$ satisfying (3). For $i \in \{2,\dots,d\}$, by Lemma 14 below (with $d = 0$), every r-ball in H_i has at most $(1+\ell)^k(2r+1)^{k+1}$ vertices. By the result of Krauthgamer and Lee [36], $H_i \subseteq \mathbb{Z}^c$ for some $c = c(k,\ell)$. Thus

$$G \subseteq H_1 \boxtimes \mathbb{Z}^{c(d-1)}.$$

By Lemma 14 again, every r-ball in G has size at most

$$(1+\Delta)^k(2r+1)^{(k+1)(c(d-1)+1)},$$

which is at most $r^{c'}$ for some $c' = c'(c,\Delta,k)$ and $r \geqslant 2$. Hence (1) holds. □

Lemma 14. *For every graph H with pathwidth at most $k \in \mathbb{N}_0$, for every connected subgraph G of $H \boxtimes \mathbb{Z}^d$ with radius at most r and maximum degree at most Δ,*

$$|V(G)| \leqslant (1+\Delta)^k(2r+1)^{(k+1)(d+1)}.$$

Proof. The BFS spanning tree of G rooted at the centre of G has radius at most r. So it suffices to prove the result when G is a tree. We proceed by induction on $k \geqslant 0$ with the following hypothesis: For every graph H with pathwidth at most $k \in \mathbb{N}_0$, for every subtree T of $H \boxtimes \mathbb{Z}^d$ with radius at most r and maximum degree at most Δ,

$$|V(T)| \leqslant (1+\Delta)^k(2r+1)^{(k+1)(d+1)}.$$

Since T is connected, we may assume that H is connected. Since T has radius at most r,

$$T \subseteq H \boxtimes P_1 \boxtimes \cdots \boxtimes P_d,$$

where each P_i is a path on $2r+1$ vertices.

In the base case $k = 0$, we have $H = K_1$ and $T \subseteq P_1 \boxtimes \cdots \boxtimes P_d$, implying

$$|V(T)| \leqslant (2r+1)^d \leqslant (1+\Delta)^0(2r+1)^{(0+1)(d+1)}.$$

Now assume that $k \geqslant 1$ and the claim holds for $k-1$. Let $\widetilde{T}$ be the projection of $V(T)$ into H. Let $(X_1,\dots,X_n)$ be a path decomposition of H with width $\mathrm{pw}(H)$. We may delete any bag X_j such that $X_j \cap \widetilde{T} = \emptyset$. Now assume that $X_1 \cap \widetilde{T} \neq \emptyset$ and $X_n \cap \widetilde{T} \neq \emptyset$. Let x be a vertex in $X_1 \cap \widetilde{T}$, and let y be a vertex in $X_n \cap \widetilde{T}$. Thus $(x,x_1,\dots,x_d) \in V(T)$ and $(y,y_1,\dots,y_d) \in V(T)$ for some $x_i, y_i \in V(P_i)$. Let P be the path in T with endpoints $(x,x_1,\dots,x_d)$ and $(y,y_1,\dots,y_d)$. Since T has radius at most r, P has at most $2r+1$ vertices. Let $\tilde{P}$ be the set of vertices $v \in V(H)$ such that $(v,z_1,\dots,z_d) \in V(P)$ where $z_i \in V(P_i)$. Thus $|\tilde{P}| \leqslant 2r+1$. By the choice of x and y, we have $\tilde{P} \cap X_j \neq \emptyset$ for each $j \in \{1,\dots,n\}$. Let $H' := H - \tilde{P}$. Thus $(X_1 \setminus \tilde{P},\dots,X_n \setminus \tilde{P})$ is a path decomposition of H' with width at most $\mathrm{pw}(H)-1$. Let $R := \{(v,z_1,\dots,z_d) : v \in \tilde{P}, z_i \in V(P_i), i \in \{1,\dots,d\}\}$. Thus $|R| \leqslant (2r+1)^{d+1}$. Let $T' := T - R$. Hence T' is a subgraph of $H' \boxtimes P_1 \boxtimes \cdots \boxtimes P_d$. Each component of T' has a neighbour in R, implying that T' has at most $\Delta\,|R|$ components. Every subtree of T has radius at most r (centred at the vertex closest to the centre of T). By induction, each component of T' has at most $(1+\Delta)^{k-1}(2r+1)^{k(d+1)}$ vertices. Thus

$$\begin{aligned}
|V(T)| &\leqslant |R| + \Delta|R|\,(1+\Delta)^{k-1}(2r+1)^{k(d+1)} \\
&= |R|\,(1+\Delta(1+\Delta)^{k-1}(2r+1)^{k(d+1)}) \\
&\leqslant |R|\,(1+\Delta)(1+\Delta)^{k-1}(2r+1)^{k(d+1)} \\
&\leqslant (2r+1)^{d+1}(1+\Delta)^k(2r+1)^{k(d+1)} \\
&= (1+\Delta)^k(2r+1)^{(k+1)(d+1)},
\end{aligned}$$

as desired. □

Property (3) in Theorem 13 is best possible in a number of respects. First, note that we cannot allow H_1 and H_2 to have unbounded maximum degree. For example, if H_1 and H_2 are both $K_{1,n}$, then H_1 and H_2 both have pathwidth 1, but $K_{1,n} \boxtimes K_{1,n}$ contains $K_{n,n}$ as a subgraph, which contains a complete binary tree of $\Omega(\log n)$ height, which is a bounded-degree graph with exponential growth. Also, bounded pathwidth cannot be replaced by bounded treewidth, again because of the complete binary tree.

5 Polynomial Expansion

This section characterises when $\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has polynomial expansion. Separators are a key tool here. A *separation* in a graph G is a pair (G_1, G_2) of subgraphs of G such that $G = G_1 \cup G_2$ and $E(G_1) \cap E(G_2) = \emptyset$. The *order* of (G_1, G_2) is $|V(G_1) \cap V(G_2)|$. A separation (G_1, G_2) is *balanced* if $|V(G_1) \setminus V(G_2)| \leqslant \frac{2}{3}|V(G)|$ and $|V(G_2) \setminus V(G_1)| \leqslant \frac{2}{3}|V(G)|$. A graph class $\mathscr{G}$ admits *strongly sublinear separators* if there exists $c \in \mathbb{R}^+$ and $\beta \in [0,1)$ such that for every graph $G \in \mathscr{G}$, every subgraph H of G has a balanced separation of order at most $c|V(H)|^\beta$. Dvořák and Norin [20] noted that a result of Plotkin, Rao, and Smith [44] implies that graph classes with polynomial expansion admit strongly sublinear separators. Dvořák and Norin [20] proved the converse (see [18, 22, 25] for more results on this theme).

Theorem 15 ([20])**.** *A hereditary class of graphs admits strongly sublinear separators if and only if it has polynomial expansion.*

Robertson and Seymour [48] established the following connection between treewidth and balanced separations.

Lemma 16 ([48, (2.6)])**.** *Every graph G has a balanced separation of order at most* $\mathrm{tw}(G)+1$.

Dvořák and Norin [23] proved the following converse.

Lemma 17 ([23])**.** *If every subgraph of a graph G has a balanced separation of order at most s, then* $\mathrm{tw}(G) \leqslant 15s$.

We have the following strongly sublinear bound on the treewidth of graph products.

Lemma 18. *Let G be an n-vertex subgraph of $\mathbb{Z}^d \boxtimes H$ for some graph H. Then*

$$\mathrm{tw}(G) \leqslant 2(\mathrm{tw}(H)+1)^{1/(d+1)}(dn)^{d/(d+1)} - 1.$$

Proof. Let $t := \mathrm{tw}(H)$. For $i \in \{1,\dots,d\}$, let $\langle V_0^i, V_1^i, \dots\rangle$ be the layering of G determined by the i-th dimension. Let

$$m := \left\lceil \left(\frac{dn}{t+1}\right)^{1/(d+1)} \right\rceil.$$

For $i \in \{1,\dots,d\}$ and $\alpha \in \{0,\dots,m-1\}$, let $V^{i,\alpha} := \bigcup\{V_j^i : j \equiv \alpha \pmod m\}$. Thus $V^{i,0},\dots,V^{i,m-1}$ is a partition of $V(G)$. Hence $|V^{i,\alpha_i}| \leqslant \frac{n}{m}$ for some $\alpha_i \in \{0,\dots,m-1\}$. Let $X := \bigcup_{i=1}^d V^{i,\alpha_i}$. Thus $|X| \leqslant \frac{dn}{m}$. Note that each component of $G - X$ is a subgraph of $Q^d \boxtimes H$, where Q is the path on $m-1$ vertices. Since $\mathrm{tw}(G)$ equals the maximum treewidth of the connected components of G, we have $\mathrm{tw}(G) \leqslant \mathrm{tw}(Q^d \boxtimes H) + |X|$. To obtain a tree decomposition of $Q^d \boxtimes H$ with width $(t+1)(m-1)^d - 1$, start with an optimal tree decomposition of H, and replace each instance of a vertex of H by the corresponding copy of Q^d. Thus

$$\mathrm{tw}(G) \leqslant (t+1)(m-1)^d - 1 + \frac{dn}{m} \leqslant 2(t+1)^{1/(d+1)}(dn)^{d/(d+1)} - 1. \qquad \square$$

Lemma 18 is generalised by our next result, which characterises when a graph product has polynomial expansion. The following definition is key. Say that graph classes $\mathscr{G}_1$ and $\mathscr{G}_2$ have *joint polynomial growth* if there exists a polynomial function p such that for every $r \in \mathbb{N}$, there exists $i \in \{1,2\}$ such that for every graph $G \in \mathscr{G}_i$ every r-ball in G has size at most $p(r)$.

Theorem 19. *The following are equivalent for hereditary graph classes $\mathscr{G}_1$ and $\mathscr{G}_2$:*
(1) $\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has polynomial expansion,
(2) $\mathscr{G}_1 \square \mathscr{G}_2$ has polynomial expansion,
(3) $\mathscr{G}_1$ has polynomial expansion, $\mathscr{G}_2$ has polynomial expansion, and $\mathscr{G}_1$ and $\mathscr{G}_2$ have joint polynomial growth.

Proof. (1) implies (2) since $\mathscr{G}_1 \square \mathscr{G}_2 \subseteq \mathscr{G}_1 \boxtimes \mathscr{G}_2$.

We now show that (2) implies (3). Assume that $\mathscr{G}_1 \square \mathscr{G}_2$ has polynomial expansion. That is, for some polynomial g, for every graph $G \in \mathscr{G}_1 \square \mathscr{G}_2$, every r-shallow minor of G has average degree at most $g(r)$. Since $\mathscr{G}_1 \cup \mathscr{G}_2 \subseteq \mathscr{G}_1 \square \mathscr{G}_2$, both $\mathscr{G}_1$ and $\mathscr{G}_2$ have polynomial expansion.

Assume for the sake of contradiction that $\mathscr{G}_1$ and $\mathscr{G}_2$ do not have joint polynomial growth. Thus for every polynomial p there exists $r \in \mathbb{N}$ such that for each $i \in \{1,2\}$ some r-ball of some graph $G_i \in \mathscr{G}_i$ has at least $p(r)$ vertices. Apply this where p is a polynomial with $p(r) \geqslant \max\{1+rn, \binom{n}{2}\}$, where $n := \lceil g(2r)+2 \rceil$. Since $\mathscr{G}_1$ and $\mathscr{G}_2$ are hereditary, there exists $r \in \mathbb{N}$ such that there is a graph $G_1 \in \mathscr{G}_1$ with radius at most r and at least $1+rn$ vertices, and there is a graph $G_2 \in \mathscr{G}_2$ with radius at most r and at least $\binom{n}{2}$ vertices.

Let z be the central vertex in G_1. Since $|V(G_1)| \geqslant 1+rn$, for some $i \in \{1,\dots,r\}$, there is a set A of n vertices in G_1 at distance exactly i from z. For all $\{v,w\} \in \binom{A}{2}$, let $P_{v,w}$ be the shortest vw-path contained with the union of a shortest vz-path and a shortest wz-path in G_1. Thus P_{vw} has length at most $2r$ and $V(P_{v,w}) \cap A = \{v,w\}$. Let B be a set of $\binom{n}{2}$ vertices in G_2. Fix an arbitrary bijection $\sigma : \binom{A}{2} \to B$.

Let $G := G_1 \square G_2$. For each $v \in A$, let $X_v := G[\{(v,x) : x \in V(G_2)\}]$; note that X_v is isomorphic to G_2, and thus has radius at most r. Moreover, X_v and X_w are disjoint for distinct $v,w \in A$. For $\{v,w\} \in \binom{A}{2}$, let $Y_{v,w} := G[\{(x,\sigma((v,w))) : x \in V(G_1)\}]$; note that $Y_{v,w}$ is isomorphic to G_1. Let $Q_{v,w}$ be the copy of the path $P_{v,w}$ within $Y_{v,w}$. Since $V(P_{v,w}) \cap A = \{v,w\}$, the only vertices of $Q_{v,w}$ in $\bigcup_{u \in A} X_u$ are $(v,\sigma((v,w)))$ and $(w,\sigma((v,w)))$, which are the endpoints of $Q_{v,w}$ in X_v and X_w respectively. Since $P_{v,w}$ has length at most $2r$, so does $Q_{v,w}$.

By construction, $Q_{v,w}$ and $Q_{p,q}$ are disjoint for distinct $\{v,w\},\{p,q\} \in \binom{A}{2}$. Contract X_v to a vertex for each $v \in A$, and contract $Q_{v,w}$ to an edge for each $\{v,w\} \in \binom{A}{2}$. We obtain the complete graph K_n as a minor in G. Moreover, the minor is $2r$-shallow. This is a contradiction, since K_n has average degree greater than $g(2r)$.

We prove that (3) implies (1) by a series of lemmas below (culminating in Lemma 23 below). $\square$

For a graph G, a set $X \subseteq V(G)$ is *r-localising* if for every component C of $G-X$, there exists a vertex $v \in V(G)$ such that $d_G(u,v) < r$ for every $u \in C$ (note that the distance is in G, not in $G-X$).

The following is a variation on Lemma 5.2 of [36]. For $r \in \mathbb{N}$ and $p,q \in \mathbb{R}$ with $0 < p,q < 1$, consider the following function $f_{r,p,q}$ defined on $\{0,1,\dots,r\}$. First, let $f_{r,p,q}(r) := p$. Now, for every integer $s \in \{0,1,\dots,r-1\}$, inductively define

$$f_{r,p,q}(s) := \min(q\, f_{r,p,q}(\{s+1,\dots,r\}), 1 - f_{r,p,q}(\{s+1,\dots,r\})),$$

where $f(S) := \sum_{i \in S} f(i)$.

Lemma 20. *Fix $r \in \mathbb{N}$ and $p,q \in \mathbb{R}$ with $0 < p,q < 1$, such that $f_{r,p,q}(\{0,1,\dots,r\}) = 1$ (so $f_{r,p,q}$ defines a probability distribution on $\{0,1,\dots,r\}$). For every graph G, there exists a probability distribution over the r-localising subsets of $V(G)$ such that the set X drawn from this distribution satisfies $\mathbb{P}[v \in X] \leqslant p|N^r(v)| + q$ for every $v \in V(G)$.*

Proof. Let $V(G) = \{v_1,\dots,v_n\}$. For $i \in \{1,\dots,n\}$, choose $r_i \in \{0,1,\dots,r\}$ independently at random such that $\mathbb{P}[r_i = s] = f_{r,p,q}(s)$. For each $x \in V(G)$, let $i(x)$ be the minimum index i such that $d(x,v_i) \leqslant r_i$, and let $X := \{x \in V(G) : d(x,v_{i(x)}) = r_{i(x)}\}$.

First we argue that X is r-localising. Consider any component C of $G-X$, and let z be the vertex of C with $i(z)$ minimum. Suppose for the sake of contradiction that C contains a vertex u at distance at least r from $v_{i(z)}$, and let P be a path from z to u in C. Then P contains a vertex x at distance exactly $r_{i(z)}$ from $v_{i(z)}$. However, since $i(x) \geqslant i(z)$, we conclude $i(x) = i(z)$ and $x \in X$, which is a contradiction.

Next, we bound the probability that a vertex v of G is in X. Consider any $i \in \{1,\dots,n\}$. If $d(v,v_i) > r$, then $\mathbb{P}[i(v) = i] = 0$. If $d(v,v_i) = r$, then $\mathbb{P}[i(v) = i] \leqslant \mathbb{P}[r_{i(v)} = r] = p$. If $d(v,v_i) < r$, then letting $s := d(v,v_i)$ we have

$$\begin{aligned}
\mathbb{P}[v \in X | i(v) = i] &= \mathbb{P}[r_i = s \,|\, r_1 < d(v,v_1),\dots,r_{i-1} < d(v,v_{i-1}), r_i \geqslant s] \\
&= \mathbb{P}[r_i = s \,|\, r_i \geqslant s] \\
&= \frac{f_{r,p,q}(s)}{f_{r,p,q}(\{s,\dots,r\})} \\
&\leqslant q.
\end{aligned}$$

Therefore,

$$\mathbb{P}[v \in X] = \sum_{i=1}^{n} \mathbb{P}[i(v) = i] \cdot \mathbb{P}[v \in X | i(v) = i] \leqslant p|N^r(v)| + q. \qquad \square$$

Corollary 21. *For every polynomial g, there exists r_0 such that the following holds. Let $r \geqslant r_0$ be a positive integer and let G be a graph such that $|N^r(v)| \leqslant g(r)$ for every $v \in V(G)$. Then there exists a probability distribution over the r-localising subsets of $V(G)$ such that the set X drawn from this distribution satisfies $\mathbb{P}[v \in X] \leqslant 2r^{-1/2}$ for every $v \in V(G)$.*

Proof. Let c be the degree of g plus one, so that $g(r) \leqslant r^c$ for every sufficiently large r. Let $p := r^{-c-1/2}$ and $q := r^{-1/2}$. Note that for sufficiently large r,

$$p(1+q)^r \geqslant pe^{qr/2} = \exp(\sqrt{r}/2 - (c+1/2)\log r) > 1.$$

Hence $f_{r,p,q}(r) = p > 1/(q+1)^r$. It follows by induction that $f_{r,p,q}(\{s,\dots,r\}) \geqslant 1/(q+1)^s$ for each $s \in \{1,\dots,r\}$. Thus $f_{r,p,q}(0) = 1 - f_{r,p,q}(\{1,\dots,r\})$ and $f_{r,p,q}(\{0,\dots,r\}) = 1$. The claim follows from Lemma 20. □

Corollary 22. *For every polynomial g, there exists r_0 such that the following holds. Let $r \geqslant r_0$ be a positive integer and let G be a graph such that $|N^r[v]| \leqslant g(r)$ for every $v \in V(G)$. Then for every function $w : V(G) \to \mathbb{R}_0^+$, there exists $X \subseteq V(G)$ such that $w(X) \leqslant 2r^{-1/2}w(V(G))$ and each component of $G - X$ has at most $g(r)$ vertices.*

Proof. Choose an r-localising set $X \subseteq V(G)$ using Corollary 21. Since X is r-localising and $|N^r[v]| \leqslant g(r)$ for every $v \in V(G)$, each component of $G - X$ has at most $g(r)$ vertices. Furthermore,

$$\mathbb{E}[|w(X)|] = \sum_{v \in V(G)} \mathbb{P}[v \in X]w(v) \leqslant 2r^{-1/2}w(V(G)).$$

Hence there is a choice for X such that $w(X) \leqslant 2r^{-1/2}w(V(G))$. □

Lemma 23. *Suppose $\mathscr{G}_1$ and $\mathscr{G}_2$ are classes with strongly sublinear separators and of joint polynomial growth (bounded by a polynomial g). Then $\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has strongly sublinear separators.*

Proof. Let $\varepsilon > 0$ be such that every subgraph F of a graph from $\mathscr{G}_1 \cup \mathscr{G}_2$ has a balanced separator of order at most $\lceil |V(F)|^{1-\varepsilon} \rceil$. Let $\beta > 0$ be sufficiently small (depending on ε and g).

Suppose $G_1 \in \mathscr{G}_1$, $G_2 \in \mathscr{G}_2$, and H is a subgraph of $G_1 \boxtimes G_2$. Let π_1 and π_2 be the projections from H to G_1 and G_2. Let $n := |V(H)|$ and $r := n^\beta$. By symmetry, we may assume $|N^r[v]| \leqslant g(r)$ for every vertex v of G_1. Let $w(v) := |\pi_1^{-1}(v)|$ for each $v \in V(G_1)$. By Corollary 22, there exists $X \subseteq V(G_1)$ such that $w(X) = O(r^{-1/2}n) = O(n^{1-\beta/2})$ and each component of $G_1 - X$ has at most $g(r) = g(n^\beta) = O(n^{\varepsilon/2})$ vertices. Let $A := \pi_1^{-1}(X)$; then $|A| = w(X) = O(n^{1-\beta/2})$.

The graph G_2 has treewidth $O(n^{1-\varepsilon})$, and thus the product of G_2 with $G_1 - X$ (as well as its subgraph $H - A$) has treewidth $O(n^{1-\varepsilon}g(r)) = O(n^{1-\varepsilon/2})$. Consequently, $H - A$ has a balanced separator B of order $O(n^{1-\varepsilon/2})$, and $A \cup B$ is a balanced separator of H of order $O(n^{1-\min(\varepsilon,\beta)/2})$. □

6 Bounded Expansion

This section characterises when $\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has bounded expansion. The following definition by Kierstead and Yang [35] is the key tool. For a graph G, linear ordering $\preceq$ of $V(G)$, vertex $v \in V(G)$, and integer $r \geqslant 1$, a vertex x is $(r,\preceq)$-*reachable* from v if there is a path $v = v_0, v_1, \dots, v_{r'} = x$ of length $r' \in \{0,1,\dots,r\}$ such that $x \preceq v \prec v_i$ for all $i \in \{1,2,\dots,r'-1\}$. For a graph G and $r \in \mathbb{N}$, the *r-colouring number* $\mathrm{col}_r(G)$ of G, also known as the *strong r-colouring number*, is the minimum integer k such that there is a linear ordering $\preceq$ of $V(G)$ such that at most k vertices are $(r,\preceq)$-reachable from each vertex v of G. For example, van den Heuvel, Ossona de Mendez, Quiroz, Rabinovich, and Siebertz [50] proved that every planar graph G satisfies $\mathrm{col}_r(G) \leqslant 5r+1$, and more generally, that every K_t-minor-free graph G satisfies $\mathrm{col}_r(G) \leqslant \binom{t-1}{2}(2r+1)$. Most generally, Zhu [54] showed that these r-colouring numbers characterise bounded expansion classes.

Theorem 24 ([54])**.** *A graph class $\mathscr{G}$ has bounded expansion if and only if for each $r \in \mathbb{N}$ there exists $c \in \mathbb{N}$ such that $\mathrm{col}_r(G) \leqslant c$ for all $G \in \mathscr{G}$.*

We now show that if G has bounded r-colouring number and H has bounded maximum degree, then $G \boxtimes H$ has bounded r-colouring number.

Lemma 25. *If G is a graph with $\mathrm{col}_r(G) \leqslant c$ and H is a graph with maximum degree at most Δ, then $\mathrm{col}_r(G \boxtimes H) < c(\Delta+2)^r$.*

Proof. Let G^+ and H^+ be the pseudographs obtained from G and H by adding a loop at every vertex. Let $\preceq_G$ be a vertex-ordering of G witnessing that $\mathrm{col}_r(G) \leqslant c$. Let $\preceq$ be an ordering of $V(G \boxtimes H)$ where $v_1 \prec_G v_2$ implies $(v_1,w_1) \prec (v_2,w_2)$ for all $v_1, v_2 \in V(G)$ and $w_1, w_2 \in V(H)$. We now bound the number of vertices of $G \boxtimes H$ that are $(r,\preceq)$-reachable from a fixed vertex $(v,w) \in V(G \boxtimes H)$. Say (x,y) is $(r,\preceq)$-reachable from (v,w). Thus there is a path $(v,w) = (v_0,w_0),(v_1,w_1),\dots,(v_{r'},w_{r'}) = (x,y)$ of length $r' \in \{0,\dots,r\}$, such that $(v_{r'},w_{r'}) \preceq (v,w) \prec (v_i,w_i)$ for each $i \in \{1,\dots,r'-1\}$. Charge (x,y) to the pair $(x,(w_0,w_1,\dots,w_{r'}))$. By the definition of $\boxtimes$, the sequence $(v_0,v_1,\dots,v_{r'})$ is a walk in G^+, and the sequence $(w_0,w_1,\dots,w_{r'})$ is a walk in H^+. By the definition of $\preceq$, we have $v_{r'} \preceq v_0 \preceq v_i$ for each $i \in \{1,\dots,r'-1\}$. Thus $v_{r'}$ is $(r,\preceq_G)$-reachable from v_0 in G. By assumption, at most c vertices are $(r,\preceq_G)$-reachable from v_0 in G. The number of walks of length at most r in H^+ starting at w_0 is at most $\sum_{i=0}^{r}(\Delta+1)^i < (\Delta+2)^r$. Thus for each vertex $x \in V(G)$, less than $(\Delta+2)^r$ vertices $(r,\preceq)$-reachable from (v,w) are charged to some pair (x,W). Hence, less than $c(\Delta+2)^r$ vertices in $G \boxtimes H$ are $(r,\preceq_G)$-reachable from (v,w) in $\preceq$. Therefore $\mathrm{col}_r(G \boxtimes H) < c(\Delta+2)^r$. □

The next theorem is the main contribution of this section.

Theorem 26. *The following are equivalent for hereditary graph classes $\mathscr{G}_1$ and $\mathscr{G}_2$:*

1. *$\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has bounded expansion,*
2. *$\mathscr{G}_1 \square \mathscr{G}_2$ has bounded expansion,*
3. *both $\mathscr{G}_1$ and $\mathscr{G}_2$ have bounded expansion, and at least one of $\mathscr{G}_1$ and $\mathscr{G}_2$ has bounded maximum degree.*

Proof. (1) implies (2) since $G_1 \square G_2 \subseteq G_1 \boxtimes G_2$.

We now show that (2) implies (3). Assume that $\mathscr{G}_1 \square \mathscr{G}_2$ has bounded expansion. Since $\mathscr{G}_1 \cup \mathscr{G}_2 \subseteq \mathscr{G}_1 \boxtimes \mathscr{G}_2$, both $\mathscr{G}_1$ and $\mathscr{G}_2$ also have bounded expansion. If neither $\mathscr{G}_1$ nor $\mathscr{G}_2$ have bounded maximum degree, then for every $n \in \mathbb{N}$, the star graph $K_{1,n}$ is a subgraph of some graph $G_1 \in \mathscr{G}_1$ and of some graph $G_2 \in \mathscr{G}_2$. Observe that $K_{1,n} \square K_{1,n}$ contains a 1-subdivision of $K_{n,n}$. Thus $G_1 \square G_2$ contains a graph with average degree n (namely, $K_{n,n}$) as a 1-shallow minor, which is a contradiction since $\mathscr{G}_1 \square \mathscr{G}_2$ has bounded expansion. Hence at least one of $\mathscr{G}_1$ and $\mathscr{G}_2$ has bounded maximum degree.

We now show that (3) implies (1). Assume that both $\mathscr{G}_1$ and $\mathscr{G}_2$ have bounded expansion, and every graph in $\mathscr{G}_2$ has maximum degree at most Δ. By Theorem 24, for each $r \in \mathbb{N}$ there exists $c_r \in \mathbb{N}$ such that $\mathrm{col}_r(G) \leqslant c_r$ for all $G_1 \in \mathscr{G}_1$. Let $G_2 \in \mathscr{G}_2$. By Lemma 25, we have $\mathrm{col}_r(G_1 \boxtimes G_2) \leqslant c_r(\Delta+2)^r$, and the same bound holds for every subgraph of $G_1 \boxtimes G_2$. By Theorem 24, $\mathscr{G}_1 \boxtimes \mathscr{G}_2$ has bounded expansion. □

7 Geometric Graph Classes

The section explores graph product structure theorems for various geometrically defined graph classes.

The *unit disc* graph of a finite set $X \subseteq \mathbb{R}^d$ is the graph G with $V(G) = X$ where $vw \in E(G)$ if and only if $\mathrm{dist}(v,w) \leqslant 1$. Here dist is the Euclidean distance in $\mathbb{R}^d$. Let $\mathbb{Z}^d$ be the strong product of d paths $P \boxtimes \cdots \boxtimes P$ (the d-dimensional grid graph with crosses). The next result describes unit discs in terms of strong products, which implies that the class of unit disc graphs with bounded dimension and bounded maximum clique size has polynomial growth.

Theorem 27. *Every unit disc graph G in $\mathbb{R}^d$ with no $(k+1)$-clique is a subgraph of $\mathbb{Z}^d \boxtimes K_{k\lceil\sqrt{d}\rceil^d}$.*

Proof. Let $t := k\lceil\sqrt{d}\rceil^d$. Let $x_i(v)$ be the i-th coordinate of each vertex $v \in V(G)$. For $p_1, \dots, p_d \in \mathbb{Z}$, let $V\langle p_1, \dots, p_d\rangle$ be the set of vertices $v \in V(G)$ such that $p_i \leqslant x_i(v) < p_i + 1$ for each $i \in \{1, \dots, d\}$. Thus the sets $V\langle p_1, \dots, p_d\rangle$ partition $V(G)$. Each set $V\langle p_1, \dots, p_d\rangle$ consists of the set of vertices in a particular unit cube. Note that the unit cube can be partitioned into $\lceil\sqrt{d}\rceil^d$ sub-cubes, each with side length at most $\frac{1}{\sqrt{d}}$ and thus with diameter at most 1. The set of vertices in a sub-cube with diameter at most 1 is a clique in G. Thus at most k vertices lie in a single sub-cube, and $|V\langle p_1, \dots, p_d\rangle| \leqslant t$. Injectively label the vertices in $V\langle p_1, \dots, p_d\rangle$ by $1, 2, \dots, t$. Map each vertex v in $V\langle p_1, \dots, p_d\rangle$ labelled $\ell(v)$ to the vertex $(p_1, \dots, p_d, \ell(v))$ of $\mathbb{Z}^d \boxtimes K_t$. Thus the vertices of G are mapped to distinct vertices of $\mathbb{Z}^d \boxtimes K_t$. For each edge $vw \in E(G)$, if $v \in V\langle p_1, \dots, p_d\rangle$ and $w \in V\langle q_1, \dots, q_d\rangle$, then $|p_i - q_i| \leqslant 1$ for each $i \in \{1, \dots, d\}$, and if $p_i = q_i$ for each $i \in \{1, \dots, d\}$, then $\ell(v) \neq \ell(w)$. Thus v and w are mapped to adjacent vertices in $\mathbb{Z}^d \boxtimes K_t$. □

By a volume argument, every covering of the unit cube by balls of diameter 1 uses at least $(\frac{d}{18})^{d/2}$ balls. So the $k\lceil\sqrt{d}\rceil^d$ term in the above theorem cannot be drastically improved.

The *k-nearest neighbour graph* of a finite set $P \subset \mathbb{R}^d$ has vertex set P, where two vertices v and w are adjacent if w is the one of the k points in P closest to v, or v is the one of the k points in P closest to w. Miller, Teng, Thurston, and Vavasis [38] showed that such graphs admit separators of order $O(n^{1-1/d})$.

Can we describe the structure of k-nearest-neighbour graphs using graph products?

Conjecture 28. Every k-nearest neighbour graph in $\mathbb{R}^d$ is a subgraph of $H \boxtimes \mathbb{Z}^{d-1}$ for some graph H with treewidth at most $f(k,d)$.

This conjecture is trivial for $d = 1$ and true for $d = 2$, as proved by Dujmović et al. [17]. Note that "treewidth at most $f(k,d)$" cannot be replaced by "pathwidth at most $f(k,d)$" for $d \geqslant 2$ because complete binary trees are 2-dimensional 2-nearest neighbour graphs without polynomial growth (see Theorem 13).

Here is a still more general example: Miller et al. [38] defined a (d,k)-*neighbourhood system* to consist of a collection $\mathscr{C}$ of n balls in $\mathbb{R}^d$ such that no point in $\mathbb{R}^d$ is covered by more than k balls. Consider the associated graph with one vertex for each ball, where two vertices are adjacent if the corresponding balls intersect. Miller et al. [38] showed that such graphs admit balanced separators of order $O(n^{1-1/d})$. Note that by the Koebe circle packing theorem, every planar graph is associated with some $(2,2)$-neighbourhood system. Thus the result of Miller et al. [38] is a far-reaching generalisation of the Lipton-Tarjan Separator Theorem [37]. Is there a product structure theorem for these graphs? Might the structure in Theorem 4 be applicable here?

Open Problem 29. *If G is the graph associated with a (d,k)-neighbourhood system, can G be obtained from clique-sums of graphs $G_1,\dots,G_n$ such that $G_i \subseteq (H_i \boxtimes P^{(d-1)}) + K_a$, for some graph H_i with treewidth at most k, where a is a constant that depends only on k and d. The natural choice is $a = k-1$.*

One can ask a similar question for graphs embeddable in a finite-dimensional Euclidean space with bounded distortion of distances. Dvořák [22] showed that such graphs have strongly sublinear separators.

8 Open Problems

We finish the paper with a number of open problems.

It is open whether the treewidth 4 bound in Theorem 2(b) can be improved.

Open Problem 30. *For every $g \in \mathbb{N}$, does there exist $t \in \mathbb{N}$ such that every graph of Euler genus g is a subgraph of $H \boxtimes P \boxtimes K_t$ for some graph H of treewidth at most 3?*

The proofs of Theorems 3 and 4 both use the Graph Minor Structure Theorem of Robertson and Seymour [49].

Open Problem 31. *Is there a proof of Theorem 3 or Theorem 4 that does not use the graph minor structure theorem?*

The following problem asks to minimise the treewidth in Theorem 7.

Open Problem 32 ([17])**.** *Does there exist a constant c such that for every $k \in \mathbb{N}$ there exists $t \in \mathbb{N}$ such that every k-planar graph is a subgraph of $H \boxtimes P \boxtimes K_t$ for some graph H of treewidth at most c?*

Open Problem 33. *Can any graph class with linear or polynomial expansion be described as a product of simpler graph classes along with apex vertices, clique sums, and other ingredients.*

Such a theorem would be useful for proving properties about such classes. Recent results say that the "other ingredients" in Open Problem 33 are needed, as we now explain. Let G' be the $6\mathrm{tw}(G)$-subdivision of a graph G. Let $\mathscr{G} := \{G' : G \text{ is a graph}\}$. Grohe, Kreutzer, Rabinovich, Siebertz, and Stavropoulos [30] proved that $\mathscr{G}$ has linear expansion. On the other hand, Dubois, Joret, Perarnau, Pilipczuk, and Pitois [11] proved that there are graphs G such that every p-centred colouring of G' has at least $2^{cp^{1/2}}$ colours, for some constant $c > 0$. We do not define "p-centred colouring" here since we do not need the definition. All we need to know is that subgraphs of $H \boxtimes P$, where H has bounded treewidth and P is a path, have p-centred colourings with $f(p)$ colours, where f is a polynomial function (see [10, 17]). This result is easily extended to allow for apex vertices and clique sums. This shows that there are graphs with linear expansion that cannot be described solely in terms of products of bounded treewidth graphs and paths (plus apex vertices and clique sums). For related results, see [19].

One way to test the quality of such a structure theorem is whether they resolve the following questions about queue-number and nonrepetitive chromatic number mentioned in Section 1:

Open Problem 34. *Do graph classes with linear or polynomial expansion have bounded queue-number?*

Open Problem 35. *Do graphs classes with linear or polynomial (or even single exponential) expansion have bounded nonrepetitive chromatic number?*

Note that bounded degree graphs are an example with exponential expansion and unbounded queue-number [51]. Similarly, subdivisions of complete graphs K_n with $o(\log n)$ division vertices per edge are an example with super-exponential expansion and unbounded nonrepetitive chromatic number [41]. Thus the graph classes mentioned in the above open problems are the largest possible with bounded queue-number or bounded nonrepetitive chromatic number.

8.1 Algorithmic Questions

Do product structure theorems have algorithmic applications? Consider the method of Baker [2] for designing polynomial-time approximation schemes for problems on planar graphs. This method partitions the graph into BFS layers, such that the problem can be solved optimally on each layer (since the induced subgraph has bounded treewidth), and then combines the solutions from each layer. Theorem 1 gives a more precise description of the layered structure of planar graphs. It is conceivable that this extra structural information is useful when designing algorithms for planar graphs (and any graph class that has a product structure theorem).

Some NP-complete problems can be solved efficiently on planar graphs. Can these results be extended to any subgraph of the strong product of a bounded treewidth graph and a path? For example, can max-cut be solved efficiently for graphs that are subgraphs of $H \boxtimes P$, where H is a bounded treewidth graph and P is a path, such as apex-minor-free graphs? This would be a considerable generalisation of the known polynomial-time algorithm for max-cut on planar graphs [31] and on graphs of bounded genus [29].

Some problems can be solved by particularly fast algorithms on planar graphs. Can such results be generalised for any subgraph of the strong product of a bounded treewidth graph and a path? For example, can shortest paths be computed in $O(n)$ time for n-vertex subgraphs of $H \boxtimes P$, where H is a bounded treewidth graph and P is a path? Can maximum flows be computed in $n\log^{O(1)}(n)$ time for n-vertex subgraphs of $H \boxtimes P$? See [6, 24] for analogous results for planar graphs.

References

[1] Noga Alon, Jarosław Grytczuk, Mariusz Hałuszczak, and Oliver Riordan. Nonrepetitive colorings of graphs. *Random Structures Algorithms*, 21(3–4):336–346, 2002. doi: 10.1002/rsa.10057. MR: 1945373.

[2] Brenda S. Baker. Approximation algorithms for NP-complete problems on planar graphs. *J. ACM*, 41(1):153–180, 1994. doi: 10.1145/174644.174650. MR: 1369197.

[3] Hans L. Bodlaender. A tourist guide through treewidth. *Acta Cybernet.*, 11(1-2):1–21, 1993.

[4] Hans L. Bodlaender. A partial k-arboretum of graphs with bounded treewidth. *Theoret. Comput. Sci.*, 209(1-2):1–45, 1998. doi: 10.1016/S0304-3975(97)00228-4. MR: 1647486.

[5] Marthe Bonamy, Cyril Gavoille, and Michał Pilipczuk. Shorter labeling schemes for planar graphs. In Shuchi Chawla, ed., *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA '20)*, pp. 446–462. 2020. doi: 10.1137/1.9781611975994.27. arXiv: 1908.03341.

[6] Sergio Cabello, Erin W. Chambers, and Jeff Erickson. Multiple-source shortest paths in embedded graphs. *SIAM J. Comput.*, 42(4):1542–1571, 2013. doi: 10.1137/120864271.

[7] Zhi-Zhong Chen, Michelangelo Grigni, and Christos H. Papadimitriou. Map graphs. *J. ACM*, 49(2):127–138, 2002. doi: 10.1145/506147.506148. MR: 2147819.

[8] Reinhard Diestel. *Graph theory*, vol. 173 of *Graduate Texts in Mathematics*. Springer, 4th edn., 2010. http://diestel-graph-theory.com/. MR: 2744811.

[9] Guoli Ding and Bogdan Oporowski. Some results on tree decomposition of graphs. *J. Graph Theory*, 20(4):481–499, 1995. doi: 10.1002/jgt.3190200412. MR: 1358539.

[10] Michał Dębski, Stefan Felsner, Piotr Micek, and Felix Schröder. Improved bounds for centered colorings. In Shuchi Chawla, ed., *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA '20)*, pp. 2212–2226. 2020. doi: 10.1137/1.9781611975994.136. arXiv: 1907.04586.

[11] Loïc Dubois, Gwenaël Joret, Guillem Perarnau, Marcin Pilipczuk, and François Pitois. Two lower bounds for p-centered colorings. *Discrete Math. Theor. Comput. Sci.*, 22(4), 2020. https://dmtcs.episciences.org/6867

[12] Vida Dujmović, David Eppstein, Gwenaël Joret, Pat Morin, and David R. Wood. Minor-closed graph classes with bounded layered pathwidth. *SIAM J. Disc. Math.*, to appear. arXiv: 1810.08314.

[13] Vida Dujmović, David Eppstein, and David R. Wood. Structure of graphs with locally restricted crossings. *SIAM J. Discrete Math.*, 31(2):805–824, 2017. doi: 10.1137/16M1062879.

[14] Vida Dujmović, Louis Esperet, Pat Morin, Bartosz Walczak, and David R. Wood. Clustered 3-colouring graphs of bounded degree. 2020. arXiv: 2002.11721.

[15] Vida Dujmović, Gwenaël Joret, Piotr Micek, Pat Morin, Torsten Ueckerdt, and David R. Wood. Planar graphs have bounded queue-number. In *Proc. 60th Annual Symp. Foundations Comput. Sci. (FOCS '19)*, pp. 862–875. IEEE, 2019. doi: 10.1109/FOCS.2019.00056.

[16] Vida Dujmović, Gwenaël Joret, Piotr Micek, Pat Morin, Torsten Ueckerdt, and David R. Wood. Planar graphs have bounded queue-number. *J. ACM*, 67(4):22, 2020. http://dx.doi.org/10.1145/3385731

[17] Vida Dujmović, Pat Morin, and David R. Wood. Graph product structure for non-minor-closed classes, 2019. arXiv: 1907.05168.

[18] Zdeněk Dvořák. Sublinear separators, fragility and subexponential expansion. *European J. Combin.*, 52(A):103–119, 2016. doi: 10.1016/j.ejc.2015.09.001.

[19] Zdeněk Dvořák, Rose McCarty, and Sergey Norin. Sublinear separators in intersection graphs of convex shapes. 2020. arXiv: 2001.01552.

[20] Zdeněk Dvořák and Sergey Norin. Strongly sublinear separators and polynomial expansion. *SIAM J. Discrete Math.*, 30(2):1095–1101, 2016. doi: 10.1137/15M1017569. MR: 3504982.

[21] Zdeněk Dvořák and Robin Thomas. List-coloring apex-minor-free graphs. 2014. arXiv: 1401.1399.

[22] Zdeněk Dvořák. On classes of graphs with strongly sublinear separators. *European J. Combin.*, 71:1–11, 2018. doi: 10.1016/j.ejc.2018.02.032. MR: 3802229.

[23] Zdeněk Dvořák and Sergey Norin. Treewidth of graphs with balanced separations. *J. Combin. Theory Ser. B*, 137:137–144, 2019. doi: 10.1016/j.jctb.2018.12.007. MR: 3980088.

[24] Jeff Erickson. Maximum flows and parametric shortest paths in planar graphs. In Moses Charikar, ed., *Proc. 21st Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010*, pp. 794–804. SIAM, 2010. doi: 10.1137/1.9781611973075.65.

[25] Louis Esperet and Jean-Florent Raymond. Polynomial expansion and sublinear separators. *European J. Combin.*, 69:49–53, 2018. doi: 10.1016/j.ejc.2017.09.003.

[26] Fedor V. Fomin, Daniel Lokshtanov, and Saket Saurabh. Bidimensionality and geometric graphs. In *Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1563–1575. 2012. doi: 10.1137/1.9781611973099.124. MR: 3205314.

[27] Jacob Fox and János Pach. A separator theorem for string graphs and its applications. *Combin. Probab. Comput.*, 19(3):371–390, 2010. doi: 10.1017/S0963548309990459.

[28] Jacob Fox and János Pach. Applications of a new separator theorem for string graphs. *Combin. Probab. Comput.*, 23(1):66–74, 2014. doi: 10.1017/S0963548313000412.

[29] Anna Galluccio and Martin Loebl. On the theory of Pfaffian orientations. II. T-joins, k-cuts, and duality of enumeration. *Electron. J. Combin.*, 6(1):#R7, 1999. doi: 10.37236/1439.

[30] Martin Grohe, Stephan Kreutzer, Roman Rabinovich, Sebastian Siebertz, and Konstantinos Stavropoulos. Coloring and covering nowhere dense graphs. *SIAM J. Discrete Math.*, 32(4):2467–2481, 2018. doi: 10.1137/18M1168753.

[31] Frank Hadlock. Finding a maximum cut of a planar graph in polynomial time. *SIAM J. Comput.*, 4(3):221–225, 1975. doi: 10.1137/0204019.

[32] Daniel J. Harvey and David R. Wood. Parameters tied to treewidth. *J. Graph Theory*, 84(4):364–385, 2017. doi: 10.1002/jgt.22030. MR: 3623383.

[33] Lenwood S. Heath, F. Thomson Leighton, and Arnold L. Rosenberg. Comparing queues and stacks as mechanisms for laying out graphs. *SIAM J. Discrete Math.*, 5(3):398–412, 1992. doi: 10.1137/0405031. MR: 1172748.

[34] Sampath Kannan, Moni Naor, and Steven Rudich. Implicit representation of graphs. *SIAM J. Discrete Math.*, 5(4):596–603, 1992. doi: 10.1137/0405049.

[35] HAL A. KIERSTEAD AND DAQING YANG. Orderings on graphs and game coloring number. *Order*, 20(3):255–264, 2003. doi: 10.1023/B:ORDE.0000026489.93166.cb.

[36] ROBERT KRAUTHGAMER AND JAMES R. LEE. The intrinsic dimensionality of graphs. *Combinatorica*, 27(5):551–585, 2007. doi: 10.1007/s00493-007-2183-y.

[37] RICHARD J. LIPTON AND ROBERT E. TARJAN. A separator theorem for planar graphs. *SIAM J. Appl. Math.*, 36(2):177–189, 1979. doi: 10.1137/0136016. MR: 0524495.

[38] GARY L. MILLER, SHANG-HUA TENG, WILLIAM THURSTON, AND STEPHEN A. VAVASIS. Separators for sphere-packings and nearest neighbor graphs. *J. ACM*, 44(1):1–29, 1997. doi: 10.1145/256292.256294. MR: 1438463.

[39] BOJAN MOHAR AND CARSTEN THOMASSEN. *Graphs on surfaces*. Johns Hopkins University Press, 2001. MR: 1844449.

[40] JAROSLAV NEŠETŘIL AND PATRICE OSSONA DE MENDEZ. *Sparsity*, vol. 28 of *Algorithms and Combinatorics*. Springer, 2012. doi: 10.1007/978-3-642-27875-4. MR: 2920058.

[41] JAROSLAV NEŠETŘIL, PATRICE OSSONA DE MENDEZ, AND DAVID R. WOOD. Characterisations and examples of graph classes with bounded expansion. *European J. Combin.*, 33(3):350–373, 2011. doi: 10.1016/j.ejc.2011.09.008. MR: 2864421.

[42] JÁNOS PACH AND GÉZA TÓTH. Recognizing string graphs is decidable. *Discrete Comput. Geom.*, 28(4):593–606, 2002. doi: 10.1007/s00454-002-2891-4.

[43] TARIK PECANINOVIC. Complete graph minors in strong products. 2019. Honours thesis, School of Mathematics, Monash University.

[44] SERGE PLOTKIN, SATISH RAO, AND WARREN D. SMITH. Shallow excluded minors and improved graph decompositions. In DANIEL DOMINIC SLEATOR, ed., *Proc. 5th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA '94)*, pp. 462–470. ACM, 1994. `http://dl.acm.org/citation.cfm?id=314464.314625`.

[45] BRUCE A. REED. Tree width and tangles: a new connectivity measure and some applications. In *Surveys in combinatorics*, vol. 241 of *London Math. Soc. Lecture Note Ser.*, pp. 87–162. Cambridge Univ. Press, 1997. doi: 10.1017/CBO9780511662119.006. MR: 1477746.

[46] BRUCE A. REED. Algorithmic aspects of tree width. In *Recent advances in algorithms and combinatorics*, vol. 11, pp. 85–107. Springer, 2003. doi: 10.1007/0-387-22444-0_4.

[47] NEIL ROBERTSON AND PAUL SEYMOUR. Graph minors I–XXIII. *J. Combin. Theory Ser. B*, 1983–2010.

[48] NEIL ROBERTSON AND PAUL SEYMOUR. Graph minors. II. Algorithmic aspects of tree-width. *J. Algorithms*, 7(3):309–322, 1986. doi: 10.1016/0196-6774(86)90023-4. MR: 0855559.

[49] NEIL ROBERTSON AND PAUL SEYMOUR. Graph minors. XVI. Excluding a non-planar graph. *J. Combin. Theory Ser. B*, 89(1):43–76, 2003. doi: 10.1016/S0095-8956(03)00042-X. MR: 1999736.

[50] JAN VAN DEN HEUVEL, PATRICE OSSONA DE MENDEZ, DANIEL QUIROZ, ROMAN RABINOVICH, AND SEBASTIAN SIEBERTZ. On the generalised colouring numbers of graphs that exclude a fixed minor. *European J. Combin.*, 66:129–144, 2017. doi: 10.1016/j.ejc.2017.06.019.

[51] DAVID R. WOOD. Bounded-degree graphs have arbitrarily large queue-number. *Discrete Math. Theor. Comput. Sci.*, 10(1):27–34, 2008. `http://dmtcs.episciences.org/434`. MR: 2369152.

[52] DAVID R. WOOD. On tree-partition-width. *European J. Combin.*, 30(5):1245–1253, 2009. doi: 10.1016/j.ejc.2008.11.010. MR: 2514645.

[53] DAVID R. WOOD. Clique minors in cartesian products of graphs. *New York J. Math.*, 17:627–682, 2011. `http://nyjm.albany.edu/j/2011/17-28.html`.

[54] XUDING ZHU. Colouring graphs with bounded generalized colouring number. *Discrete Math.*, 309(18):5562–5568, 2009. doi: 10.1016/j.disc.2008.03.024.

Chapter 8

Harmonic Analysis and Dispersive PDEs: Problems and Progress

A note on bilinear wave-Schrödinger interactions

Timothy Candy

Abstract We consider bilinear restriction estimates for wave-Schrödinger interactions and provide a sharp condition to ensure that the product belongs to $L_t^q L_x^r$ in the full bilinear range $\frac{2}{q} + \frac{d+1}{r} < d+1$, $1 \leqslant q, r \leqslant 2$. Moreover, we give a counterexample which shows that the bilinear restriction estimate can fail, even in the transverse setting. This failure is closely related to the lack of curvature of the cone. Finally we mention extensions of these estimates to adapted function spaces. In particular we give a general transference type principle for U^2 type spaces that roughly implies that if an estimate holds for homogeneous solutions, then it also holds in U^2. This transference argument can be used to obtain bilinear and multilinear estimates in U^2 from the corresponding bounds for homogeneous solutions.

1 Introduction

Let $u = e^{it|\nabla|} f$ be a free wave, and let $v = e^{it\Delta} g$ be a homogeneous solution to the Schrödinger equation. Our goal is to understand for which $1 \leqslant q, r \leqslant \infty$ we have the bilinear estimate

$$\|uv\|_{L_t^q L_x^r(\mathbb{R}^{1+d})} \lesssim \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)}. \tag{1}$$

As a first step in this direction, assuming for instance that we have the support condition $\operatorname{supp} \widehat{f}, \operatorname{supp} \widehat{g} \subset \{|\xi| \approx 1\}$, then for any $\frac{2}{q_1} + \frac{d-1}{r_1} \leqslant \frac{d-1}{2}$ with $(q_1, r_1, d) \neq (2, \infty, 3)$, and any $\frac{2}{q_2} + \frac{d}{r_2} \leqslant \frac{d}{2}$ with $(q_2, r_2, d) \neq (2, \infty, 2)$ we have the linear Strichartz estimates

$$\|u\|_{L_t^{q_1} L_x^{r_1}(\mathbb{R}^{1+d})} \lesssim \|f\|_{L^2(\mathbb{R}^d)}, \qquad \|v\|_{L_t^{q_2} L_x^{r_2}(\mathbb{R}^{1+d})} \lesssim \|g\|_{L^2(\mathbb{R}^d)}.$$

T. Candy
Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin 9054, New Zealand, e-mail: tcandy@maths.otago.ac.nz

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_33

Consequently an application of Hölder's inequality and a short computation shows that the bilinear estimate (1) holds provided that

$$\frac{2}{q}+\frac{d}{r}\leqslant d,\qquad \frac{2}{q}+\frac{d-1}{r}\leqslant d-1+\frac{1}{d},\qquad (q,d)\neq\Big(\frac{4}{3},2\Big),(1,3). \tag{2}$$

The first condition in (2) is stronger in the region $q\geqslant 2$ and follows by simply placing $u\in L_t^\infty L_x^2$ and using the Strichartz estimate for v. Note that this explains the Schrödinger scaling of the first condition in (2). The second condition in (2) dominates in the region $1\leqslant q\leqslant 2$, where we are forced to use the Strichartz estimates on both u and v.

A natural question now arises, is it possible to improve on the conditions (2)? This question is particularly relevant in applications to nonlinear PDE, where bilinear estimates such as (2) with q,r as small as possible, are extremely useful in controlling nonlinear interactions. Note that the wave-Schrödinger interactions occur naturally in important models, see for instance the Zakharov system [21]. In the case of wave-wave interactions, it is possible to improve significantly on the range given by simply applying Hölder's inequality and the Strichartz estimate for the wave equation provided an additional transversality assumption is made.

Theorem 1 (Endpoint bilinear restriction for wave [20, 16, 18, 19]). *Let $d\geqslant 2$ and $1\leqslant q,r\leqslant 2$ with $\frac{2}{q}+\frac{d+1}{r}\leqslant d+1$ and $\frac{1}{q}<\frac{d+1}{4}$. If $f,g\in L^2(\mathbb{R}^d)$ and $\omega,\omega'\in\mathbb{S}^{d-1}$ with[1] $\measuredangle(\omega,\omega')\approx 1$ and*

$$\begin{aligned}\operatorname{supp}\widehat{f}&\subset\big\{\xi\in\mathbb{R}^d\;\big|\;|\xi|\approx 1,\measuredangle(\xi,\omega)\ll 1\big\},\\ \operatorname{supp}\widehat{g}&\subset\big\{\xi\in\mathbb{R}^d\;\big|\;|\xi|\approx 1,\measuredangle(\xi,\omega')\ll 1\big\}\end{aligned} \tag{3}$$

then

$$\|e^{it|\nabla|}fe^{it|\nabla|}g\|_{L_t^qL_x^r(\mathbb{R}^{1+d})}\lesssim\|f\|_{L^2(\mathbb{R}^d)}\|g\|_{L^2(\mathbb{R}^d)}.$$

The first result beyond the linear Strichartz theory was obtained in [5]. The full non-endpoint range when $q=r$ was obtained in [20], and extended to $q\neq r$ in [18, 13, 14]. The extension of Theorem 1 to more general frequency interactions is also known [18, 13, 14, 9]. The range for (q,r) is sharp (except possibly the case $q=\frac{4}{3}$ when $d=2$) [9, 17] , and was originally conjectured by Klainerman-Machedon. The case $q=1$ of Theorem 1 can be found in [6]. Theorem 1 is closely related to the restriction conjecture for the cone, as the free wave $e^{it|\nabla|}f$ is essentially the extension operator for the cone. In particular, bilinear estimates of the form (1) were originally used to obtain restriction estimates for the cone, see for instance [15].

Theorem 1 is truly a bilinear estimate as it relies crucially on the support assumption (3). This assumption implies that the two subsets of the cone, $\operatorname{supp}\mathscr{F}[e^{it|\nabla|}f]\subset\mathbb{R}^{1+d}$ and $\operatorname{supp}\mathscr{F}[e^{it|\nabla|}g]\subset\mathbb{R}^{1+d}$, are transverse, where $\mathscr{F}$ denotes the space-time Fourier transform. Since the waves $e^{it|\nabla|}f$ and $e^{it|\nabla|}g$ propagate in the normal di-

[1] Here $\measuredangle(x,y)=(1-\frac{x\cdot y}{|x||y|})^{\frac{1}{2}}$ is the angle between $x,y\in\mathbb{R}^d\setminus\{0\}$.

rections to these surfaces, the two waves can only interact strongly for short times. Thus we should expect the product $e^{it|\nabla|}fe^{it|\nabla|}g$ to decay faster than say $(e^{it|\nabla|}f)^2$.

If we apply the above discussion to the bilinear estimate (1), since the normal direction to the cone is $(1,-\frac{\xi}{|\xi|})$, and the normal direction to the paraboloid is $(1,2\xi)$, we should expect to improve on the range (2) obtained via the linear Strichartz estimates, by imposing a transversality condition of the form

$$\Big|\frac{\xi}{|\xi|}+2\eta\Big| \gtrsim 1 \tag{4}$$

for all $\xi \in \operatorname{supp}\widehat{f}$ and $\eta \in \widehat{g}$ (here $\widehat{f}$ denotes the spatial Fourier transform). Unfortunately, the simple transversality condition (4) is not sufficient to obtain the full range in Theorem 1 due to the lack of curvature of the cone along the surface of intersection

$$\Sigma_{wave}(a,z) = \big\{(\tau,\xi) \in \operatorname{supp}\mathscr{F}[e^{it|\nabla|}f] \;\big|\; (a,z)-(\tau,\xi) \in \operatorname{supp}\mathscr{F}[e^{it\Delta}g]\big\},$$

where $(a,z) \in \mathbb{R}^{1+d}$. In fact it is well known that for certain surfaces, transversality alone is not sufficient to obtain the full bilinear range, see for instance [12] for the example of the hyperbolic paraboloid, and the related discussion in [3, 6]. However, imposing a stronger support condition gives the following.

Theorem 2 (Wave-Schrödinger bilinear restriction [6]). *Let $d \geqslant 2$, $1 \leqslant q,r \leqslant 2$, and $\frac{2}{q}+\frac{d+1}{r} < d+1$. Let $\xi_0,\eta_0 \in \mathbb{R}^d$ such that*

$$\Big|\Big(\frac{\xi_0}{|\xi_0|}+2\eta_0\Big)\cdot\frac{\xi_0}{|\xi_0|}\Big| \gtrsim \Big|\frac{\xi_0}{|\xi_0|}+2\eta_0\Big| \tag{5}$$

and define $\lambda = |\eta_0|$, and $\alpha = |\frac{\xi_0}{|\xi_0|}+2\eta_0|$. If

$$\operatorname{supp}\widehat{f} \subset \big\{|\xi| \approx \lambda, \measuredangle(\xi,\xi_0) \ll \min\{1,\alpha\}\big\}, \qquad \operatorname{supp}\widehat{g} \subset \{|\xi-\eta_0| \ll \alpha\}$$

then we have

$$\big\|e^{it|\nabla|}fe^{it\Delta}g\big\|_{L^q_t L^r_x(\mathbb{R}^{1+d})} \lesssim (\min\{\alpha,\lambda,\alpha\lambda\})^{d+1-\frac{d+1}{r}-\frac{2}{q}}\alpha^{\frac{1}{r}-1}\lambda^{\frac{1}{q}-\frac{1}{2}}\|f\|_{L^2_x}\|g\|_{L^2_x}.$$

Theorem 2 is a consequence of a bilinear restriction estimate for general phases obtained in [6]. The special case $q=r$ and $\alpha=\lambda=1$ could also be deduced from [3]. As the precise conditions in [6] are complicated, the derivation is slightly nontrival and we give the details below in Section 2. The dependence on the parameters α and λ is sharp, and this is particularly useful in applications to nonlinear PDE where α and λ roughly correspond to a derivative loss/gain. Clearly, applying Sobolev embedding and interpolating with the trivial case $q=\infty$, $r=1$ can extend the range to $q,r \geqslant 2$ and $\frac{2}{q}+\frac{d+1}{r} < d+1$. However the dependence on α and λ would no longer be sharp (i.e. losses may occur).

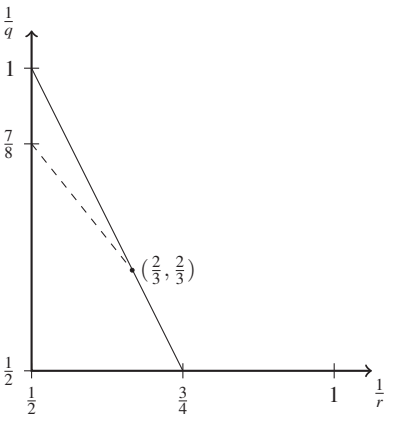


Fig. 1 The range of $1 \leqslant q, r \leqslant 2$ in $d = 3$. The line corresponds to the sharp bilinear line $\frac{2}{q} + \frac{d+1}{r} = d + 1$ given by Theorem 2. If (4) holds but (5) fails, then Theorem 3 states that the bilinear estimate (1) can only hold to the left of the dotted line.

The condition (5) is necessary to obtain the full bilinear range $\frac{2}{q} + \frac{d+1}{r} \leqslant d + 1$.

Theorem 3 (Transverse counter example). *Suppose that the estimate* (1) *holds for all* $f, g \in L^2(\mathbb{R}^d)$ *with*[2]

$$\operatorname{supp} \widehat{f} \subset \{|\xi - e_1| \ll 1\}, \qquad \operatorname{supp} \widehat{g} \subset \{|2\xi + e_1 + e_2| \ll 1\}.$$

Then

$$\frac{2}{q} + \frac{d-1}{r} + \frac{1}{2r} \leqslant d. \tag{6}$$

Note that if we let $\xi_0 = e_1$ and $\eta_0 = -\frac{1}{2}e_1 - \frac{1}{2}e_2$, then $|\frac{\xi_0}{|\xi_0|} + 2\eta_0| = 1$ but $(\frac{\xi_0}{|\xi_0|} + 2\eta_0) \cdot \frac{\xi_0}{|\xi_0|} = 0$. In other words the transversality condition (4) holds, but the stronger condition (5) fails. The range (6) is stronger than the bilinear range in Theorem 2 when q is close to 1 and $d \leqslant 5$, see figure 1. If we drop the transversality condition completely, then a similar counter example can be used to prove the following.

Theorem 4 (Non-transverse counter example). *Suppose that the estimate* (1) *holds for all* $f, g \in L^2(\mathbb{R}^d)$ *with* $\operatorname{supp} \widehat{f}, \operatorname{supp} \widehat{g} \subset \{|\xi| \approx 1\}$. *Then*

$$\frac{2}{q} + \frac{d-1}{r} \leqslant d - \frac{1}{2}, \qquad \frac{1}{q} \leqslant \frac{d+1}{4}. \tag{7}$$

We give the proof of Theorem 3 and Theorem 4 in Section 3 below. In the positive direction, if (4) holds, but (5) fails, a naive adaption of the proof of Theorem 2 should give the region $\frac{2}{q} + \frac{d}{r} < d$. The loss of dimension corresponds to the lack of curvature in the radial direction (i.e. the cone only has $d - 1$ directions of non-vanishing curvature). Note that there is a large gap between the potential range $\frac{2}{q} + \frac{d}{r} < d$ and the counter example given by Theorem 3. Similarly, even in the "linear" case when

[2] Here $e_j \in \mathbb{R}^d$, $j = 1, \ldots, n$ denote the standard basis vectors.

the transversality condition (4) is dropped, there is a gap between the counter example in Theorem 4 and the linear range given via Strichartz estimates (2). It is an interesting open question to determine the precise range of (q,r) once the general condition (5) is dropped. In particular, it is not clear to the author what the optimal range for (q,r) should be. Presumably the counter examples used in the proof of Theorem 3 and Theorem 4 can be improved.

In applications to nonlinear PDE, typically the homogeneous estimate in Theorem 2 is not sufficient, and it is more useful to have a version in suitable function spaces. One option is to work with $X^{s,b}$ type spaces. However, recently bilinear restriction estimates in the U^p type spaces have proven useful, see for instance [7] and the discussion within. In the following we wish to give a general argument which can allow multilinear estimates for homogeneous solutions, to be upgraded to estimates in the adapted function spaces U^2. The underlying idea is straight forward. The first step is use the classical theorem of Marcinkiewicz-Zygmund that given a bound for a linear operator, a standard randomisation argument via Khintchine's inequality implies that a vector valued operator bound also holds. The second step is use the observation that a vector valued estimate immediately implies a U^2 bound, see for instance [6, Section 1.2] or [8, Remark 5.2]. As an example, we extend Theorem 2, and the multilinear restriction theorem [4] to U^2.

We start with the definition of U^2. A function $\phi \in L_t^\infty L_x^2$ is an *atom* if we can write $\phi(t) = \sum_I \mathbb{1}_I(t) g_I$, with the intervals $I \subset \mathbb{R}$ forming a partition of $\mathbb{R}$, and the $g_I : \mathbb{R}^d \to \mathbb{C}$ satisfying the bound

$$\Big(\sum_I \|g_I\|_{L_x^2}^2 \Big)^{\frac{1}{2}} \leqslant 1.$$

The atomic space U^2 is then defined as

$$U^2 = \Big\{ \sum_j c_j \phi_j \;\Big|\; \phi_j \text{ an atom and } (c_j) \in \ell^1 \Big\}$$

with the induced norm

$$\|u\|_{U^2} = \inf_{u = \sum_j c_j \phi_j} \sum_j |c_j|$$

where the inf is over all representations of u in terms of atoms. These spaces were introduced in unpublished work of Tataru, and studied in detail in [11, 10]. To obtain the adapted function spaces $U^2_{|\nabla|}$ and U^2_Δ adapted to the wave and Schrödinger flows respectively, we define

$$U^2_{|\nabla|} = \{u : \mathbb{R}^{1+d} \to \mathbb{C} \mid e^{-it|\nabla|} u \in U^2\}, \qquad U^2_\Delta = \{v : \mathbb{R}^{1+d} \to \mathbb{C} \mid e^{-it\Delta} v \in U^2\}.$$

Note that since $\mathbb{1}_{\mathbb{R}}(t) f \in U^2$, we clearly have $e^{it|\nabla|} f \in U^2_{|\nabla|}$ and $e^{it\Delta} f \in U^2_\Delta$. Thus the adapted function spaces contain all homogeneous solutions. Running the argument sketched above implies the following U^2 version of Theorem 2.

Theorem 5 (Wave-Schrödinger bilinear restriction in U^2). *Let $d \geqslant 2$, $1 \leqslant q, r \leqslant 2$, and $\frac{2}{q} + \frac{d+1}{r} < d+1$. Let $\xi_0, \eta_0 \in \mathbb{R}^d$ such that (5) holds and define $\lambda = |\eta_0|$, and $\alpha = |\frac{\xi_0}{|\xi_0|} + 2\eta_0|$. If*

$$\operatorname{supp} \widehat{u} \subset \big\{ |\xi| \approx \lambda, \measuredangle(\xi, \xi_0) \ll \min\{1, \alpha\} \big\}, \qquad \operatorname{supp} \widehat{v} \subset \{ |\xi - \eta_0| \ll \alpha \} \tag{8}$$

then we have

$$\|uv\|_{L_t^q L_x^r(\mathbb{R}^{1+d})} \lesssim (\min\{\alpha, \lambda, \alpha\lambda\})^{d+1-\frac{d+1}{r}-\frac{2}{q}} \alpha^{\frac{1}{r}-1} \lambda^{\frac{1}{q}-\frac{1}{2}} \|u\|_{U^2_{|\nabla|}} \|v\|_{U^2_\Delta}.$$

Proof. Let $u = \sum_{I \in \mathscr{I}} e^{it|\nabla|} f_I$ be a $U^2_{|\nabla|}$ atom, and let $v_0 = e^{it\Delta} g$ be a homogeneous solution to the Schrödinger equation. Assume that the support conditions (8) hold. Let $(\varepsilon_I)_{I \in \mathscr{I}}$ be a family of independent, identically distributed random variables with $\varepsilon_I = 1$ with probability $\frac{1}{2}$, and $\varepsilon_I = -1$ with probability $\frac{1}{2}$. The since the intervals I are disjoint, we have via Khintchine's inequality

$$|u| \leqslant \Big(\sum_I |e^{it|\nabla|} f_I|^2 \Big)^{\frac{1}{2}} \approx \mathbf{E}\Big[\Big| \sum_I \varepsilon_I e^{it|\nabla|} f_I \Big| \Big].$$

Therefore, since $q, r \geqslant 1$, applying Theorem 2 gives

$$\begin{aligned}
\|uv_0\|_{L_t^q L_x^r} &\lesssim \Big\| \mathbf{E}\Big[\Big| \sum_I \varepsilon_I e^{it|\nabla|} f_I \Big| \Big] v_0 \Big\|_{L_t^q L_x^r} \\
&\lesssim \mathbf{E}\Big[\Big\| e^{it|\nabla|} \Big(\sum_I \varepsilon_I f_I \Big) v_0 \Big\|_{L_t^q L_x^r} \Big] \\
&\lesssim (\min\{\alpha, \lambda, \alpha\lambda\})^{d+1-\frac{d+1}{r}-\frac{2}{q}} \alpha^{\frac{1}{r}-1} \lambda^{\frac{1}{q}-\frac{1}{2}} \mathbf{E}\Big[\Big\| \sum_I \varepsilon_I f_I \Big\|_{L_x^2} \Big] \|g\|_{L_x^2}.
\end{aligned}$$

We now observe that Hölder's inequality, together with another application of Khintchine's inequality, implies that

$$\mathbf{E}\Big[\Big\| \sum_I \varepsilon_I f_I \Big\|_{L_x^2} \Big] \leqslant \Big(\mathbf{E}\Big[\Big\| \sum_I \varepsilon_I f_I \Big\|_{L_x^2}^2 \Big] \Big)^{\frac{1}{2}} = \Big(\sum_I \|f_I\|_{L^2}^2 \Big)^{\frac{1}{2}}$$

and consequently, applying the definition of the $U^2_{|\nabla|}$ norm, we obtain

$$\|uv_0\|_{L_t^q L_x^r} \lesssim (\min\{\alpha, \lambda, \alpha\lambda\})^{d+1-\frac{d+1}{r}-\frac{2}{q}} \alpha^{\frac{1}{r}-1} \lambda^{\frac{1}{q}-\frac{1}{2}} \|u\|_{U^2_{|\nabla|}} \|g\|_{L_x^2}. \tag{9}$$

To replace the homogeneous solution v_0 with a general U^2_Δ function follows by essentially repeating the above argument. In slightly more detail, suppose that $v = \sum_{J \in \mathscr{J}} e^{it\Delta} g_J$ is a U^2_Δ atom, and let $(\varepsilon_J)_{J \in \mathscr{J}}$ be a family of i.i.d. random variables with $\varepsilon_J = \pm 1$ with equal probability. Then as above, but replacing Theorem 2 with (9), we see that

$$\begin{aligned}\|uv\|_{L_t^q L_x^r} &\lesssim \Big\| u \mathbf{E}\Big[\Big|\sum_J \varepsilon_J e^{it\Delta} g_J\Big|\Big]\Big\|_{L_t^q L_x^r}\\ &\lesssim \mathbf{E}\Big[\Big\| u \sum_J e^{it\Delta}\varepsilon_J g_J\Big\|_{L_t^q L_x^r}\Big]\\ &\lesssim (\min\{\alpha,\lambda,\alpha\lambda\})^{d+1-\frac{d+1}{r}-\frac{2}{q}} \alpha^{\frac{1}{r}-1}\lambda^{\frac{1}{q}-\frac{1}{2}} \|u\|_{U^2_{|\nabla|}} \mathbf{E}\Big[\Big\|\sum_J \varepsilon_J g_J\Big\|_{L^2}\Big]\\ &\lesssim (\min\{\alpha,\lambda,\alpha\lambda\})^{d+1-\frac{d+1}{r}-\frac{2}{q}} \alpha^{\frac{1}{r}-1}\lambda^{\frac{1}{q}-\frac{1}{2}} \|u\|_{U^2_{|\nabla|}} \Big(\sum_J \|g_J\|_{L^2}^2\Big)^{\frac{1}{2}}.\end{aligned}$$

Applying the definition the U^2_Δ norm, the required bound follows.

Strictly speaking the above theorem can also be obtain via the vector valued version of Theorem 2 from [6], see for instance [6, Section 1.2]. However the above alternative argument is more direct, and has the distinct advantage that it can be applied in more general situations. As an example, consider the following special case of the multilinear restriction theorem [4].

Theorem 6 (Multilinear restriction for Schrödinger [4]). *Let $d \geqslant 2$ and $\varepsilon > 0$. Then for any $R \geqslant 1$ and any $f_j \in L^2(\mathbb{R}^d)$, $j = 1,\dots,d$ with $\operatorname{supp} \widehat{f_j} \subset \{|\xi - e_j| \ll 1\}$ we have*

$$\Big\| \Pi_j e^{it\Delta} f_j \Big\|_{L^{\frac{2}{d-1}}_{t,x}(\{|t|+|x|<R\})} \lesssim R^{\varepsilon} \Pi_j \|f_j\|_{L^2}.$$

It is conjectured that the R^ε loss can be removed, but this is currently an open question (however see [1] and [2] for recent progress). The U^2 version of Theorem 6 is then the following.

Theorem 7 (Multilinear restriction for Schrödinger in U^2). *Let $d \geqslant 2$ and $\varepsilon > 0$. Then for any $R \geqslant 1$ and any $u_j \in U^2_\Delta$, $j = 1,\dots,d$ with $\operatorname{supp} \widehat{u} \subset \{|\xi - e_j| \ll 1\}$ we have*

$$\Big\| \Pi_j u_j \Big\|_{L^{\frac{2}{d-1}}_{t,x}(\{|t|+|x|<R\})} \lesssim R^{\varepsilon} \Pi_j \|u_j\|_{U^2_\Delta}.$$

Proof. Let $B_R = \{|t| + |x| < R\}$. We proceed as in the proof of Theorem 5. Thus suppose that $u_1 = \sum_I e^{it\Delta} f_I$ is U^2_Δ atom, and let $u_j^0 = e^{it\Delta} f_j$ for $j = 2,\dots,d$. Let ε_I be a family of i.i.d. random variables with $\varepsilon_I = \pm 1$ with equal probability. An application of Khintchine's inequality implies that

$$|u_1| \leqslant \Big(\sum_I |e^{it\Delta} f_I|^2\Big)^{\frac{1}{2}} \approx \Big(\mathbf{E}\Big[\Big|\sum_I \varepsilon_I e^{it|\nabla|} f_I\Big|^{\frac{2}{d-1}}\Big]\Big)^{\frac{d-1}{2}}$$

and hence Theorem 6 together with Hölder's inequality gives

$$\begin{aligned}
\Big\| u_1 \Pi_{j=2}^d u_j^0 \Big\|_{L_{t,x}^{\frac{2}{d-1}}(B_R)} &\lesssim \Big\| \Big(\mathbf{E}\Big[\Big| \sum_I \varepsilon_I e^{it\Delta} f_I \Big|^{\frac{2}{d-1}} \Big] \Big)^{\frac{d-1}{2}} \Pi_{j=2}^d u_j^0 \Big\|_{L_{t,x}^{\frac{2}{d-1}}(B_R)} \\
&\lesssim \Big(\mathbf{E}\Big[\Big\| \sum_I \varepsilon_I e^{it\Delta} f_I \Pi_{j=2}^d u_j \Big\|_{L_{t,x}^{\frac{2}{d-1}}(B_R)}^{\frac{2}{d-1}} \Big] \Big)^{\frac{d-1}{2}} \\
&\lesssim R^{\varepsilon} \Big(\mathbf{E}\Big[\Big\| \sum_I \varepsilon_I f_I \Big\|_{L^2}^{\frac{2}{d-1}} \Big] \Big)^{\frac{d-1}{2}} \Pi_{j=2}^d \|f_j\|_{L^2} \\
&\lesssim R^{\varepsilon} \Big(\mathbf{E}\Big[\Big\| \sum_I \varepsilon_I f_I \Big\|_{L^2}^{2} \Big] \Big)^{\frac{1}{2}} \Pi_{j=2}^d \|f_j\|_{L^2} \\
&\approx R^{\varepsilon} \Big(\sum_I \|f_I\|_{L^2}^2 \Big)^{\frac{1}{2}} \Pi_{j=2}^d \|f_j\|_{L^2}.
\end{aligned}$$

Applying the definition of the U_Δ^2 norm, we conclude that

$$\Big\| u_1 \Pi_{j=2}^d u_j^0 \Big\|_{L_{t,x}^{\frac{2}{d-1}}(B_R)} \lesssim R^{\varepsilon} \|u_1\|_{U_\Delta^2} \Pi_{j=2}^d \|f_j\|_{L^2}. \tag{10}$$

As in the proof of Theorem 5, repeating this argument with Theorem 6 replaced with (10) and u_1 replaced with u_2 gives

$$\Big\| u_1 u_2 \Pi_{j=3}^d u_j^0 \Big\|_{L_{t,x}^{\frac{2}{d-1}}(B_R)} \lesssim R^{\varepsilon} \|u_1\|_{U_\Delta^2} \|u_2\|_{U_\Delta^2} \Pi_{j=3}^d \|f_j\|_{L^2}.$$

The required bound follows by continuing in this manner.

We have not attempted to write down the most general transference type argument that can be deduced from the above arguments. However the underlying idea is simple; if a estimate holds for free solutions, then via randomisation it should hold in the vector valued case, and consequently it will also hold in U^2. Of course proving U^p bounds, with $p \neq 2$ is substantially more challenging.

2 Proof of Theorem 2

It suffices to check the conditions in [6]. Suppose that $\xi_0, \eta_0 \in \mathbb{R}^d$ such that (5) holds, and define $\lambda = |\eta_0|$, and $\alpha = |\frac{\xi_0}{|\xi_0|} + 2\eta_0|$. Let

$$\Lambda_1 = \big\{ |\xi| \approx \lambda, \measuredangle(\xi, \xi_0) \ll \min\{1, \alpha\} \big\}, \qquad \Lambda_2 = \{ |\xi - \eta_0| \ll \alpha \}$$

and

$$\Phi_1(\xi) = |\xi|, \qquad \Phi_2(\xi) = -|\xi|^2, \qquad \mathscr{H}_1 = \lambda^{-1}, \qquad \mathscr{H}_2 = 1.$$

In view of [6, Lemma 2.1 and Theorem 1.2], for $\{j,k\} = \{1,2\}$ and $\xi \in \Lambda_j$, $\eta \in \Lambda_k$, it suffices to check the following conditions:

(i) for all $v \in \mathbb{R}^d$ we have

$$v\cdot(\nabla\Phi_j(\xi)-\nabla\Phi_k(\eta))=0 \qquad \Longrightarrow \qquad \big|\nabla^2\Phi_j(\xi)v\wedge\big(\nabla\Phi_j(\xi)-\nabla\Phi_k(\eta)\big)\big| \gtrsim \mathscr{H}_j\alpha|v|,$$

(ii) for $\xi' \in \Lambda_j$ and $\eta' \in \Lambda_k$ we have

$$|\nabla\Phi_j(\xi)-\nabla\Phi_j(\xi')|+|\nabla\Phi_k(\eta)-\nabla\Phi_k(\eta')| \ll \alpha,$$

(iii) the Hessian's satisfy

$$|\nabla\Phi_j(\xi)-\nabla\Phi_j(\xi')-\nabla^2\Phi_j(\xi)(\xi-\xi')| \ll \mathscr{H}_j|\xi-\xi'|,$$

(iv) for $2 < m \leqslant 5d$ we have the derivative bounds

$$\|\nabla^m\Phi_j\|_{L^\infty(\Lambda_j)}(\min\{\alpha,\lambda,\alpha\lambda\})^{m-2} \lesssim \mathscr{H}_j, \qquad \mathscr{H}_j\min\{\alpha,\lambda,\alpha\lambda\} \lesssim \alpha.$$

(v) we have the surface measure condition

$$\sup_{(a,h)\in\mathbb{R}^{1+d}} \sigma_{d-1}\big(\{\xi\in\Lambda_2\cap(h-\Lambda_1)\big|\ \Phi_2(\xi)+\Phi_1(h-\xi)=a\}\big) \lesssim \big(\min\{\alpha,\lambda,\alpha\lambda\}\big)^{d-1}$$

where σ_{d-1} is the induced Lebesgue surface measure.

To check the first property (i), by unpacking the definition, our goal is to show that for any $\xi \in \Lambda_1$ and $\eta \in \Lambda_2$ we have

$$z\cdot(\omega+2\eta)=0 \qquad \Longrightarrow \qquad \big|(z-(\omega\cdot z)\omega)\wedge(\omega+2\eta)\big| \gtrsim |z||\omega+2\eta|,$$

where $\omega = \frac{\xi}{|\xi|}$. In view of the definition of the sets Λ_j we have

$$|(\omega+2\eta)\cdot\omega| \gtrsim |\omega+2\eta|$$

and hence as $(z\cdot\omega)(\omega+2\eta)\cdot\omega = -2z\cdot(\eta-(\eta\cdot\omega)\omega)$ we get

$$|z\cdot\omega| \leqslant \frac{2|z\cdot(\eta-(\eta\cdot\omega)\omega)|}{|(\omega+2\eta)\cdot\omega|} \lesssim |z-(\omega\cdot z)\omega|.$$

Therefore

$$\big|\big(z-(\omega\cdot z)\omega\big)\wedge(\omega+2\eta)\big| \geqslant \big|z-(\omega\cdot z)\omega\big||(\omega+2\eta)\cdot\omega| \gtrsim |z||\omega+2\omega|$$

as required.

The properties (ii), ... , (iv) follow by direct computation. Finally, to check the surface measure condition (v), we note that the vector $N = \frac{\xi_0}{|\xi_0|}+2\eta_0$ is essentially normal to the surface. On the other hand, from (5), N is roughly pointing in the direction $\frac{\xi_0}{|\xi_0|}$. Hence the surface measure can be bounded by projecting onto the

plane orthogonal to $\frac{\xi_0}{|\xi_0|}$. Since this projection is contained in a ball of radius $\lesssim \min\{\alpha, \lambda, \alpha\lambda\}$, the bound follows.

3 Counter Examples

We first observe that by a randomisation argument, if the estimate (1) holds for all $f, g \in L^2$ with $\operatorname{supp} \widehat{f} \subset \Lambda_1 \subset \mathbb{R}^n$ and $\operatorname{supp} \widehat{g} \subset \Lambda_2$, then in fact we also have the vector valued version

$$\Big\| \Big(\sum_j |e^{it|\nabla|} f_j|^2 \Big)^{\frac{1}{2}} \Big(\sum_k |e^{it\Delta} g_k|^2 \Big)^{\frac{1}{2}} \Big\|_{L_t^q L_x^r} \lesssim \Big(\sum_j \|f_j\|_{L^2}^2 \Big)^{\frac{1}{2}} \Big(\sum_k \|g_k\|_{L_x^2}^2 \Big)^{\frac{1}{2}} \tag{11}$$

for all $\operatorname{supp} \widehat{f_j} \subset \Lambda_1$, $\operatorname{supp} \widehat{g_k} \subset \Lambda_2$. This follows by noting that if ε_j is an i.i.d. family of random variables with $\varepsilon_j = \pm$ with equal probability, then as in the proof of Theorem 5, we have via Khintchine's inequality and (1)

$$\begin{aligned}
\Big\| \Big(\sum_j |e^{it|\nabla|} f_j|^2 \Big)^{\frac{1}{2}} e^{it\Delta} g \Big\|_{L_t^q L_x^r} &\approx \Big\| \mathbf{E} \Big[\Big| \sum_j \varepsilon_j e^{it|\nabla|} f_j \Big| \Big] e^{it\Delta} g \Big\|_{L_t^q L_x^r} \\
&\lesssim \mathbf{E} \Big[\Big\| \sum_j \varepsilon_j e^{it|\nabla|} f_j e^{it\Delta} g \Big\|_{L_t^q L_x^r} \Big] \\
&\lesssim \mathbf{E} \Big[\Big\| \sum_j \varepsilon_j e^{it|\nabla|} f_j \Big\|_{L_x^2} \Big] \|g\|_{L_x^2} \\
&\lesssim \mathbf{E} \Big[\Big\| \sum_j \varepsilon_j e^{it|\nabla|} f_j \Big\|_{L_x^2}^2 \Big]^{\frac{1}{2}} \|g\|_{L_x^2} \approx \Big(\sum_j \|f_j\|_{L^2}^2 \Big)^{\frac{1}{2}} \|g\|_{L^2}.
\end{aligned}$$

Repeating this argument for the Schrödinger component then gives (11). Consequently, we see that the scalar version (1) holds, if and only if the vector valued version (11) holds. Thus to prove Theorem 3 and Theorem 4, it suffices to obtain vector valued counter examples.

3.1 Proof of Theorem 3

Let $N \geqslant 1$ and $\widehat{f}, \widehat{g} \in C_0^\infty$ with

$$\operatorname{supp} \widehat{f} \subset \{|\xi_1 - 1| \ll 1, |\xi'| \ll N^{-1}\}, \qquad \operatorname{supp} \widehat{g} \subset \{|2\xi - e_1 - e_2| \ll N^{-\frac{1}{2}}\}$$

and $\|f\|_{L^2} \approx N^{\frac{d-1}{2}}$, $\|g\|_{L^2} \approx N^{\frac{d}{4}}$. A short computation using integration by parts shows that we can choose f such that

$$|u(t,x)| = |e^{it|\nabla|} f(x)| \geqslant 1$$

for all $|t| \leqslant N^2$, $|x_1 + t| \leqslant 1$, and $|x'| \leqslant N$. Similarly we can choose g such that

$$|v(t,x)| = |e^{it\Delta} g(x)| \geqslant 1$$

for all $|t| \leqslant N$, $|x_1 + t| \leqslant N^{\frac{1}{2}}$, $|x_2 + t| \leqslant N^{\frac{1}{2}}$, and $|x''| \leqslant N^{\frac{1}{2}}$, where we write $x = (x_1, x') = (x_1, x_2, x'')$. In other words the free wave satisfies $|u| \geqslant 1$ on a plate of dimension $N^2 \times 1 \times N^{d-1}$ oriented in the $(1, -e_1)$ direction, with short direction e_1, while the free Schrödinger wave satisfies $|v| \geqslant 1$ on a tube of dimensions $N \times N^{\frac{d}{2}}$ oriented in the $(1, -e_1 - e_2)$ direction. Define the set

$$\Omega = \{|t| \leqslant N^2, |x_1 + t| \leqslant N^{\frac{1}{2}}, |x'| \leqslant N\}.$$

The support properties of u, implies that for any $(t,x) \in \Omega$ we have

$$U(t,x) = \Big(\sum_{\substack{j \in \mathbb{Z} \\ |j| \leqslant N^{\frac{1}{2}}}} |u(t, x + je_1)|^2 \Big)^{\frac{1}{2}} \gtrsim 1.$$

Similarly, translating the free Schrödinger wave in both space and time gives for any $(t,x) \in \Omega$

$$V(t,x) = \Big(\sum_{\substack{(j_2,\dots,j_d) \in \mathbb{Z}^{d-1} \\ |j_2|,\dots,|j_d| \lesssim N^{\frac{1}{2}}}} \sum_{\substack{k \in \mathbb{Z} \\ |k| \leqslant N}} \big| v\big(t + Nk, x + N^{\frac{1}{2}}(j_2 e_2 + \cdots + j_d e_d)\big) \big|^2 \Big)^{\frac{1}{2}} \gtrsim 1.$$

Since the wave and Schrödinger equations are translation invariant, the bound (11) implies that

$$\begin{aligned} N^{\frac{2}{q}} N^{\frac{d-1}{r} + \frac{1}{2r}} \lesssim \|\mathbb{1}_\Omega\|_{L_t^q L_x^r} &\lesssim \|UV\|_{L_t^q L_x^r} \\ &\lesssim \Big(\sum_{|j| \leqslant N^{\frac{1}{2}}} \|f\|_{L^2}^2 \Big)^{\frac{1}{2}} \Big(\sum_{|j_2|,\dots,|j_d| \lesssim N^{\frac{1}{2}}} \sum_{|k| \leqslant N} \|g\|_{L^2}^2 \Big)^{\frac{1}{2}} \\ &\lesssim N^{\frac{d-1}{2} + \frac{1}{4}} \times N^{\frac{d}{2} + \frac{1}{4}}. \end{aligned}$$

Letting $N \to \infty$, we see that this is only possible if $\frac{2}{q} + \frac{d-1}{r} + \frac{1}{2r} \leqslant d$.

3.2 Proof of Theorem 4

Let $1 \leqslant M \leqslant N$ and $\widehat{f}, \widehat{g} \in C_0^\infty$ with

$$\operatorname{supp}\widehat{f} \subset \{|\xi_1 - 1| \ll 1, |\xi'| \ll N^{-1}\}, \qquad \operatorname{supp}\widehat{g} \subset \{|2\xi - e_1| \ll M^{-1}\}$$

and $\|f\|_{L^2} \approx N^{\frac{d-1}{2}}$, $\|g\|_{L^2} \approx M^{\frac{d}{2}}$. A short computation using integration by parts shows that we can choose f, g such that

$$|u(t,x)| = |e^{it|\nabla|} f(x)| \geqslant 1 \qquad \text{for all} \qquad |t| \leqslant N^2,\ |x_1 + t| \leqslant 1,\ |x'| \leqslant N$$

and

$$|v(t,x)| = |e^{it\Delta} g(x)| \geqslant 1 \qquad \text{for all} \qquad |t| \leqslant M^2,\ |x_1 + t| \leqslant M,\ |x'| \leqslant M.$$

In other words the free wave satisfies $|u| \geqslant 1$ on a plate of dimension $N^2 \times 1 \times N^{d-1}$ oriented in the $(1, -e_1)$ direction, with short direction e_1, while the free Schrödinger wave satisfies $|v| \geqslant 1$ on a tube of dimensions $M^2 \times M^d$ oriented in the $(1, -e_1)$ direction. Similar to the proof of Theorem 3, we consider a number of temporal translated Schrödinger waves covering the space-time set

$$\Omega = \{|t| \leqslant N^2, |x_1 + t| \leqslant 1, |x'| \leqslant M\}.$$

More precisely, we have for all $(t,x) \in \Omega$

$$V(t,x) = \Big(\sum_{\substack{j \in \mathbb{Z} \\ |j| \leqslant \frac{N^2}{M^2}}} |v(t + jM^2, x)|^2 \Big)^{\frac{1}{2}} \gtrsim 1.$$

Since we clearly have $|u| \geqslant 1$ on Ω by construction, we see that if (1) holds, then the vector valued version (11) holds, and hence

$$N^{\frac{2}{q}} M^{\frac{d-1}{r}} \lesssim \|\mathbb{1}_\Omega\|_{L_t^q L_x^r} \lesssim \|uV\|_{L_t^q L_x^r} \lesssim \|f\|_{L^2} \Big(\sum_{|j| \leqslant \frac{N^2}{M^2}} \|g\|_{L^2}^2 \Big)^{\frac{1}{2}} \lesssim N^{\frac{d-1}{2}} \times N M^{\frac{d}{2} - 1}.$$

Rearranging, we see that we must have

$$N^{\frac{2}{q} - \frac{d+1}{2}} M^{\frac{d-1}{r} - \frac{d-2}{2}} \lesssim 1.$$

Letting $M = 1$ and $N \to \infty$ gives the restriction $\frac{1}{q} \leqslant \frac{d+1}{4}$. On the other hand, letting $M = N \to \infty$, gives the condition $\frac{2}{q} + \frac{d-1}{r} \leqslant d - \frac{1}{2}$.

Acknowledgements Financial support by the Marsden Fund Council grant 19-UOO-142, and the German Research Foundation (DFG) through the CRC 1283 "Taming uncertainty and profiting from randomness and low regularity in analysis, stochastics and their applications" is acknowledged. The author would also like to thank Sebastian Herr and Kenji Nakanishi for a number of helpful discussions, as well as the University of Bielefeld and MATRIX for their kind hospitality while part of this work was conducted.

References

1. Ioan Bejenaru, *The almost optimal multilinear restriction estimate for hypersurfaces with curvature: the case of $n-1$ hypersurfaces in $\mathbb{R}^n$*, arXiv:2002.12488[math.CA].
2. _____, *The multilinear restriction estimate: almost optimality and localization*, arXiv:1912.06664[math.CA].
3. Ioan Bejenaru, *Optimal bilinear restriction estimates for general hypersurfaces and the role of the shape operator*, International Mathematics Research Notices **2017** (2017), no. 23, 7109–7147.
4. Jonathan Bennett, Anthony Carbery, and Terence Tao, *On the multilinear restriction and Kakeya conjectures*, Acta Math. **196** (2006), no. 2, 261–302. MR 2275834
5. J. Bourgain, *Estimates for cone multipliers*, Geometric aspects of functional analysis (Israel, 1992–1994), Oper. Theory Adv. Appl., vol. 77, Birkhäuser, Basel, 1995, pp. 41–60. MR 1353448
6. Timothy Candy, *Multi-scale bilinear restriction estimates for general phases*, Mathematische Annalen **375** (2019), no. 1-2, 777–843.
7. Timothy Candy and Sebastian Herr, *On the division problem for the wave maps equation*, Annals of PDE **4** (2018), no. 2, 17.
8. _____, *Transference of bilinear restriction estimates to quadratic variation norms and the Dirac-Klein-Gordon system*, Anal. PDE **11** (2018), no. 5, 1171–1240.
9. D. Foschi and S. Klainerman, *Bilinear space-time estimates for homogeneous wave equations* Ann. Sci. École Norm. Sup. (4) **33** (2000), 211-274.
10. Martin Hadac, Sebastian Herr, and Herbert Koch, *Well-posedness and scattering for the KP-II equation in a critical space*, Ann. Inst. H. Poincaré Anal. Non Linéaire **26** (2009), no. 3, 917–941. MR 2526409 (2010d:35301)
11. Herbert Koch and Daniel Tataru, *Dispersive estimates for principally normal pseudodifferential operators*, Comm. Pure Appl. Math. **58** (2005), no. 2, 217–284. MR 2094851 (2005m:35323)
12. Sanghyuk Lee, *Bilinear restriction estimates for surfaces with curvatures of different signs*, Trans. Amer. Math. Soc. **358** (2006), no. 8, 3511–3533 (electronic). MR 2218987 (2007a:42023)
13. S. Lee and A. Vargas, *Sharp null form estimates for the wave equation*, Amer. J. Math. **130** (2008), no. 5, 1279-1326.
14. S. Lee, K. Rogers, and A. Vargas, *Sharp null form estimates for the wave equation in* $\mathbb{R}^{3+1}$, Int. Math. Res. Not. **2008** (2008).
15. T. Tao and A. Vargas, *A bilinear approach to cone multipliers. I. Restriction estimates*, Geom. Funct. Anal. **10** (2000), no. 1, 185–215. MR 1748920 (2002e:42012)
16. Terence Tao, *Endpoint bilinear restriction theorems for the cone, and some sharp null form estimates*, Math. Z. **238** (2001), no. 2, 215–268. MR 1865417 (2003a:42010)
17. T. Tao, *A counterexample to an endpoint bilinear Strichartz inequality*, Electron. J. Differential Equations (2006), no. 151.
18. Daniel Tataru, *Null form estimates for second order hyperbolic operators with rough coefficients*, Harmonic analysis at Mount Holyoke (South Hadley, MA, 2001), Contemp. Math., vol. 320, Amer. Math. Soc., Providence, RI, 2003, pp. 383–409. MR 1979953
19. Faruk Temur, *An endline bilinear cone restriction estimate for mixed norms*, Math. Z. **273** (2013), no. 3-4, 1197–1214. MR 3030696
20. Thomas Wolff, *A sharp bilinear cone restriction estimate*, Ann. of Math. (2) **153** (2001), no. 3, 661–698. MR 1836285 (2002j:42019)
21. Vladimir E. Zakharov, *Collapse of Langmuir waves*, Sov. Phys. JETP **35** (1972), no. 5, 908–914.

A note on the scattering for 3D quantum Zakharov system with non-radial data in L^2

Chunyan Huang

Abstract In this note, we give a remark on the scattering for quantum Zakharov system with non-radial small initial data in L^2 with one order additional angular regularity using the generalized Strichartz estimate with wider range and the normal form transformation.

1 Introduction

We study the scattering of solutions to the 3D quantum Zakharov system

$$\begin{cases} iu_t + \Delta u - \varepsilon^2 \Delta^2 u = nu, \\ n_{tt} - \Delta n + \varepsilon^2 \Delta^2 n = \Delta(|u|^2), \\ u(0,x) = u_0, \ n(0,x) = n_0, \partial_t n(0,x) = n_1, \end{cases} \tag{1}$$

where $u(t,x) : \mathbb{R}^1 \times \mathbb{R}^3 \to \mathbb{C}$ is the envelope electric field and $n(t,x) : \mathbb{R}^1 \times \mathbb{R}^3 \to \mathbb{R}$ describes the plasma density fluctuation. The quantum parameter $0 < \varepsilon \leq 1$ is the ratio between the ion plasmon energy and the electron thermal energy. For detailed background of this system, see [6].
The solutions (u,n) of (1) preserve the mass $\|u(t)\|_{L^2}$ and the energy

$$E(u,n,\partial_t n) = \int_{\mathbb{R}^d} |\nabla u(t)|^2 + \varepsilon^2 |\Delta u(t)|^2 + n|u|^2 + \frac{1}{2}(|D^{-1} n_t|^2 + n^2 + \varepsilon^2 |\nabla n(t)|^2) dx.$$

When $\varepsilon = 0$, (1) reduces to the classical Zakharov system.

For simplicity, we change (1) to a lower order system by letting

Chunyan Huang
School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China, e-mail: hcy@cufe.edu.cn

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_34

$$N = n - \frac{in_t}{\sqrt{-\Delta + \varepsilon^2\Delta^2}}. \tag{2}$$

Then (1) is transformed to

$$\begin{cases} iu_t - (-\Delta + \varepsilon^2\Delta^2)u = (\bar{N}u + Nu)/2, \\ iN_t + \sqrt{-\Delta + \varepsilon^2\Delta^2}N = \frac{\Delta}{\sqrt{-\Delta+\varepsilon^2\Delta^2}}(|u|^2), \end{cases} \tag{3}$$

with

$$u(0) = u_0, \quad N(0) = n_0 - i(-\Delta + \varepsilon^2\Delta^2)^{-\frac{1}{2}}n_1.$$

The treatment for $\bar{N}u$ is similar to Nu, we may assume that the nonlinear term in the first equation of (3) is $\bar{N}u$. The global well-posedness of (1) in energy space was obtained in [5] when $d = 1,2,3$. As pointed out in [1] that L^2 is the most important function space in mathematics and it is also important for Zakharov type system since it measures the total electric energy in physics, to this motivation the authors studied the local well-posedness with large data($1 \le d \le 8$), global well-posedness($1 \le d \le 5$) and scattering for small initial data($4 \le d \le 8$) of (3) in $L^2(\mathbb{R}^d) \times L^2(\mathbb{R}^d)$, but for $1 \le d \le 3$, scattering is not obtained in [1]. One of the main difficulties of proving scattering for quantum Zakharov systemin in low dimensions is the quadratic nonlinearities. Recently, the scattering for 3D quantum Zakharov system in $L^2(\mathbb{R}^3) \times L^2(\mathbb{R}^3)$ with small radial initial data was proved in [7] using normal form transformation and radial improved Strichartz estimates. In this note, we explain that the radial condition can be removed if we assume additional angular regularity of degree one. The Sobolev space with one order angular regularity $H^{0,1}_{2,\sigma}$ is defined in (5), the angular derivative D_σ is defined in Subsection 1.1, the Strichartz norm S and W are defined in (1). The main result is the following

Theorem 1.1 *$d = 3$. Suppose that $\|(u_0,N_0)\|_{H^{0,1}_{2,\sigma}(\mathbb{R}^3)\times H^{0,1}_{2,\sigma}(\mathbb{R}^3)} = \varepsilon_0 > 0$ which is small enough, then there exists a unique global solution (u,N) of (3) satisfying $\|(u,N)\|_{S\times W} \le C\varepsilon_0$ and scatters in this space. Namely, there exists a solution $(u^\pm,N^\pm) \in H^{0,1}_{2,\sigma}(\mathbb{R}^3) \times H^{0,1}_{2,\sigma}(\mathbb{R}^3)$ to the linear system*

$$\begin{cases} iu_t - (-\Delta + \varepsilon^2\Delta^2)u = 0, \\ iN_t + \sqrt{-\Delta + \varepsilon^2\Delta^2}N = 0, \end{cases} \tag{4}$$

satisfying

$$\|u(t) - u^\pm(t)\|_{L^2} + \|N(t) - N^\pm(t)\|_{L^2} + \|D_\sigma(u(t) - u^\pm(t))\|_{L^2} + \|D_\sigma(N(t) - N^\pm(t))\|_{L^2}$$

$\to 0$, *as $t \to \pm\infty$.*

Next we introduce some notations used in this note.

1.1 Notation

For $x \in \mathbb{R}^n$, write $\langle x\rangle := (1+|x|^2)^{1/2}$. We use $\hat{f}$ or $\mathscr{F}f$ to denote the Fourier transform of f. Write $D := \sqrt{-\Delta} = \mathscr{F}^{-1}|\xi|\mathscr{F}$ and $\langle D\rangle^s := \mathscr{F}^{-1}(1+|\xi|^2)^{s/2}\mathscr{F}$. Let $\eta : \mathbb{R}^3 \to \mathbb{R}$ be a smooth bump function supported in $B_2(0)$ and equal to 1 in $B_1(0)$. For $k \in \mathbb{Z}$, let $\chi_k(\xi) = \eta(\xi/2^k) - \eta(\xi/2^{k-1})$ and $\chi_{\leqslant k}(\xi) = \eta(\xi/2^k)$. The Littlewood-Paley operators are defined by

$$\widehat{P_k(\xi)} = \chi_k(|\xi|)\widehat{u}(\xi), \quad \widehat{P_{\leq k}(\xi)} = \chi_{\leqslant k}(|\xi|)\widehat{u}(\xi).$$

Let Δ_σ be the Laplace-Beltrami operator on the unit sphere $\mathbb{S}^{d-1}$ endowed with standard metric g and measure $d\sigma$. Denote $D_\sigma = \sqrt{-\Delta_\sigma}$ and $\Lambda_\sigma = \sqrt{1-\Delta_\sigma}$. For $1 \leq i,j \leq d$, $X_{i,j} = x_i\partial_j - x_j\partial_i$ are rotational vector fields and for $f \in C^2(\mathbb{R}^d)$, $\Delta_\sigma(f)(x) = \sum_{1\leq i,j\leq d} X_{i,j}^2(f)(x)$.

$L^p(\mathbb{R}^d)$ denotes the usual Lebesgue space and $\mathscr{L}^p(\mathbb{R}^+) = \mathscr{L}^p(\mathbb{R}^+ : \rho^{d-1}d\rho)$. We follow the notations in [2] and write $L^p_\sigma = L^p_\sigma(\mathbb{S}^{d-1})$, $\mathscr{H}^s_p = \mathscr{H}^s_p(\mathbb{S}^{d-1}) = \Lambda_\sigma^{-s}L^p_\sigma$. $\mathscr{L}^p_\rho L^q_\sigma$ and $\mathscr{L}^p_\rho \mathscr{H}^s_q$ are Banach spaces defined by the norms $\|f\|_{\mathscr{L}^p_\rho L^q_\sigma} = \|\|f(\rho\sigma)\|_{L^q_\sigma}\|_{\mathscr{L}^p_\rho}$ and $\|f\|_{\mathscr{L}^p_\rho \mathscr{H}^s_q} = \|\|f(\rho\sigma)\|_{\mathscr{H}^s_q}\|_{\mathscr{L}^p_\rho}$.

For $s \in \mathbb{R}$, $1 \leq p \leq \infty$, H^s_p denotes Banach space of elements $u \in \mathscr{S}'(\mathbb{R}^d)$ such that $\mathscr{F}^{-1}(1+|\xi|^2)^{s/2}\hat{u} \in L^p(\mathbb{R}^d)$ and $H^s(\mathbb{R}^n) = H^s_2(\mathbb{R}^n)$. The homogeneous Sobolev space $\dot{H}^s$ is defined by $\dot{H}^s(\mathbb{R}^d) = \{u \in \mathscr{S}'(\mathbb{R}^d) : \|u\|_{\dot{H}^s} = \||\xi|^s\hat{f}(\xi)\|_{L^2_\xi} < \infty\}$.

For $s \in \mathbb{R}$ and $1 \leq p,q,r \leq \infty$, $\dot{B}^s_{p,q}(\mathbb{R}^d)$ is the standard homogeneous Besov space on $\mathbb{R}^d$ with norm $\|u\|_{\dot{B}^s_{p,q}(\mathbb{R}^d)} := \left(\sum_{k\in\mathbb{Z}} 2^{qsk}\|P_k u(x)\|_p^q\right)^{1/q}$. $\dot{B}^s_{(p,q),r}$ denotes the Besov type space with norm

$$\|u\|_{\dot{B}^s_{(p,q),r}} := \left(\sum_{k\in\mathbb{Z}} 2^{rsk}\|P_k u\|^r_{\mathscr{L}^p_\rho L^q_\sigma}\right)^{1/r}.$$

For $0 \leq \alpha \leq 1$, $H^{s,\alpha}_{p,\sigma}$ is the space with norm

$$\|f\|_{H^{s,\alpha}_{p,\sigma}} = \|\Lambda^\alpha_\sigma f\|_{H^s_p}. \tag{5}$$

$\dot{H}^{s,\alpha}_{p,\sigma}$, $\dot{B}^{s,\alpha}_{p,q,\sigma}$ and $\dot{B}^{s,\alpha}_{(p,q),r,\sigma}$ are defined similarly. For simplicity, we write $\dot{B}^{s,\alpha}_{p,\sigma} = \dot{B}^{s,\alpha}_{p,2,\sigma}$ and $B^{s,\alpha}_{p,\sigma} = B^{s,\alpha}_{p,2,\sigma}$.

Let X be any Banach space of functions on $\mathbb{R}^n$, we define $L^q_t X$ to be the space on $\mathbb{R}\times\mathbb{R}^n$ with space-time norm $\|u\|_{L^q_t X} := \left(\int_{\mathbb{R}} \|u\|_X^q dt\right)^{1/q}$.

p' denotes the conjugate of $p \in [1,\infty]$ given by $\frac{1}{p} + \frac{1}{p'} = 1$.

1.2 Normal form transform

In this subsection, we use the normal form transform technique(which was first used by Shatah[8] in quadratic Klein-Gordon equations) for the quantum Zakharov system. Normal form transform method is one of the most powerful tools to exploit nonlinear structures. Write

$$\omega_1(D) = D^2 + \varepsilon^2 D^4, \quad \omega_2(D) = D\sqrt{1+\varepsilon^2 D^2},$$

and

$$\omega_1(|\xi|) = |\xi|^2 + \varepsilon^2|\xi|^4, \quad \omega_2(|\xi|) = |\xi|\sqrt{1+\varepsilon^2|\xi|^2}.$$

Define $S(t) = e^{it\omega_1(D)} := \mathscr{F}^{-1}e^{-it\omega_1(\xi)}\mathscr{F}$ to be the fourth order Schrödinger semigroup and $W(t) = e^{it\omega_2(D)} := \mathscr{F}^{-1}e^{it\omega_2(\xi)}\mathscr{F}$ to be the wave semigroup.
For any u and v, define the low-high, high-low and high-high interactions by

$$(uv)_{LH} := \sum_{k\in\mathbb{Z}} (P_{\leq k-5}u)(P_k v), \quad (uv)_{HL} := \sum_{k\in\mathbb{Z}} (P_k u)(P_{\leq k-5}v),$$
$$(uv)_{HH} := \sum_{\substack{|k_1-k_2|\leq 4\\ k_1,k_2\in\mathbb{Z}}} (P_{k_1}u)(P_{k_2}v),$$

then $uv = (uv)_{LH} + (uv)_{HL} + (uv)_{HH}$. To make a distinction with resonant and non-resonant terms, we write

$$(uv)_{1L} := \sum_{|k|\leq 1} (P_k u)(P_{\leq k-5}v), \quad (uv)_{L1} := (vu)_{1L},$$
$$(uv)_{XL} := \sum_{|k|>1} (P_k u)(P_{\leq k-5}v), \quad (uv)_{LX} := (vu)_{XL},$$

then

$$(uv)_{HL} = (uv)_{1L} + (uv)_{XL}, \quad (uv)_{LH} = (uv)_{L1} + (uv)_{LX}.$$

We use $\varphi_{XL}, \varphi_{LX}$, etc. to denote the bilinear symbol of operators u_{XL}, u_{LX}, etc.,

$$\mathscr{F}(uv)_{XL} = \int \varphi_{XL}\hat{u}(\xi-\eta)\hat{v}(\eta)d\eta, \quad \mathscr{F}(uv)_{LX} = \int \varphi_{LX}\hat{u}(\xi-\eta)\hat{v}(\eta)d\eta. \quad (6)$$

Symbols φ_{XL}, φ_{LX}, etc., can be expressed in terms of $\chi_k(\xi)$, i.e., $\varphi_{XL} = \sum_{|k|>1}\chi_k(\xi-\eta)\chi_{\leq k-5}(\eta)$. Similarly as in [7], (3) are transformed to the following equivalent integral equations

$$u = S(t)u_0 - \Omega_1(\bar{N},u)(t) + S(t)\Omega_1(\bar{N},u)(0) - i\int_0^t S(t-s)\Omega_2(D|u|^2,u)(s)ds$$
$$- i\int_0^t S(t-s)\Omega_1(\bar{N},\bar{N}u)(s)ds - i\int_0^t S(t-s)(\bar{N}u)_{HH+LH+1L}(s)ds. \quad (7)$$

$$\bar{N} = W(t)\bar{N}_0 - D\Omega_3(u,u)(t) + W(t)D\Omega_3(u,u)(0) - i\int_0^t W(t-s)(D\Omega_3(\bar{N}u,u) - D\Omega_3(u,\bar{N}u))ds - i\int_0^t W(t-s)\frac{D}{\sqrt{1+\varepsilon^2D^2}}(u\bar{u})_{HH+L1+1L}(s)ds, \tag{8}$$

where $\Omega_j (j = 1,2,3)$ are bilinear multipliers

$$\Omega_1(f,g) = \mathscr{F}^{-1}\int \varphi_{XL}\Phi_\varepsilon^{-1}\hat{f}(\xi-\eta)\hat{g}(\eta)d\eta,$$
$$\Omega_2(f,g) = \mathscr{F}^{-1}\int \frac{\varphi_{XL}}{\Phi_\varepsilon\sqrt{1+\varepsilon^2|\xi-\eta|^2}}\hat{f}(\xi-\eta)\hat{g}(\eta)d\eta,$$
$$\Omega_3(f,g) = \mathscr{F}^{-1}\int \varphi_{XL+LX}\frac{\hat{f}(\xi-\eta)\hat{\bar{g}}(\eta)}{\tilde{\Phi}_\varepsilon\sqrt{1+\varepsilon^2|\xi|^2}}d\eta,$$

in which $\Phi_\varepsilon := \omega_1(|\xi|) - \omega_1(|\eta|) - \omega_2(|\xi-\eta|)$ and $\tilde{\Phi}_\varepsilon = \omega_2(|\xi|) + \omega_1(|\eta|) - \omega_1(|\xi-\eta|)$ are resonance functions for the Schrödinger and wave component in (3). After normal form transform, the transformed new system is:

$$\begin{aligned}(i\partial_t + \omega_1(D))(u+\Omega_1(\bar{N},u)) &= (\bar{N}u)_{HH+LH+1L} - i\Omega_2(D|u|^2,u) - i\Omega_1(\bar{N},\bar{N}u),\\ (i\partial_t + \omega_2(D))(\bar{N}+D\Omega_3(u,u)) &= \frac{D}{\sqrt{1+\varepsilon^2D^2}}(u\bar{u})_{HH+L1+1L} - iD\Omega_3(\bar{N}u,u) + iD\Omega_3(u,\bar{N}u).\end{aligned} \tag{9}$$

Remark 1.2 *In proving scattering, the most difficult terms are the high-low interaction terms $(Nu)_{XL}$, $(u\bar{u})_{XL}$ and $(u\bar{u})_{XL}$. The Schrödinger component and wave component have different propagation speed in these cases. These terms are highly non-resonant which could be observed from the resonant functions Φ_ε and $\tilde{\Phi}_\varepsilon$. After normal form transform, these quadratic terms are transformed into trilinear terms and then have more freedom of space and time integrability which is crucial to close the argument.*

2 Angular Strichartz estimates and nonlinear estimates

In this section, we first recall the generalized spherically averaged Strichartz estimate proved in [2].

Lemma 2.1 ([2]) $d = 3$, $k \in \mathbb{Z}$.
(1) Let $\frac{10}{3} < r \le +\infty$, for any initial data $\phi \in L^2_x(\mathbb{R}^3)$, we have

$$\|S(t)P_k\phi\|_{L^2_t\mathscr{L}^r_\rho L^2_\sigma} \lesssim 2^{k(\frac{1}{2}-\frac{3}{r})}\|\phi\|_{L^2_x}. \tag{1}$$

(2) Let $4 < r \le +\infty$, for any initial data $\phi \in L^2_x(\mathbb{R}^3)$, there holds

$$\|W(t)P_k\phi\|_{L_t^2\mathscr{L}_\rho^r L_\sigma^2} \lesssim 2^{k(1-\frac{3}{r})}\|\phi\|_{L_x^2}. \tag{2}$$

To state the angular Strichartz estimate, we first give a definition on angular admissible pair:

Definition 2.2 *Assume that* $2 \le q, p \le \infty$.
(1)A pair (q,p) *is called angular Schrödinger-admissible if*

$$\frac{2}{q}+\frac{5}{p} < \frac{5}{2} \text{ or } (q,p)=(\infty,2). \tag{3}$$

(2)A pair (q,p) *is called angular wave-admissible if*

$$\frac{1}{q}+\frac{2}{p} < 1 \text{ or } (q,p)=(\infty,2). \tag{4}$$

Using Lemma 2.1 and interpolating with the classical Strichartz estimate, we obtain the following:

Lemma 2.3 (Angular Strichartz estimates for the fourth order Schrödinger operator) *Assume that* $2 \le q, \tilde{q}, p, \tilde{p} \le \infty$, $(q,p), (\tilde{q},\tilde{p})$ *are both angular Schrödinger-admissible pairs and* $\tilde{q} > 2$, *then we have the homogeneous Strichartz estimate:*

$$\|S(t)u_0\|_{L_t^q \dot{B}^{\frac{2}{\tilde{q}}+\frac{3}{p}-\frac{3}{2}}_{(p,2),2}} \lesssim \|u_0\|_{L_x^2}, \tag{5}$$

and the inhomogeneous Strichartz estimate

$$\left\| \int_0^t S(t-s)F(s)ds \right\|_{L_t^q \dot{B}^{\frac{2}{\tilde{q}}+\frac{3}{p}-\frac{3}{2}}_{(p,2),2}} \lesssim \|F\|_{L_t^{\tilde{q}'} \dot{B}^{\frac{3}{2}-\frac{3}{\tilde{p}}-\frac{2}{\tilde{q}}}_{(\tilde{p}',2),2}}, \tag{6}$$

where the implicit constants are independent of ε, $\frac{1}{\tilde{q}}+\frac{1}{\tilde{q}'}=1$ *and* $\frac{1}{\tilde{p}}+\frac{1}{\tilde{p}'}=1$.

Lemma 2.4 (Angular Strichartz estimates for the wave operator) *Suppose that* $2 \le q, \tilde{q}, p, \tilde{p} \le \infty$, $(q,p), (\tilde{q},\tilde{p})$ *are angular wave-admissible pairs and* $\tilde{q} > 2$, *then there holds the homogeneous Strichartz estimate*

$$\|W(t)u_0\|_{L_t^q \dot{B}^{\frac{1}{\tilde{q}}+\frac{3}{p}-\frac{3}{2}}_{(p,2),2}} \lesssim \|u_0\|_{L_x^2}, \tag{7}$$

and the inhomogeneous Strichartz estimate

$$\left\| \int_0^t W(t-s)F(s)ds \right\|_{L_t^q \dot{B}^{\frac{1}{\tilde{q}}+\frac{3}{p}-\frac{3}{2}}_{(p,2),2}} \lesssim \|F\|_{L_t^{\tilde{q}'} \dot{B}^{\frac{3}{2}-\frac{3}{\tilde{p}}-\frac{1}{\tilde{q}}}_{(\tilde{p}',2),2}}, \tag{8}$$

where the implicit constants are independent of ε, $\frac{1}{\tilde{q}}+\frac{1}{\tilde{q}'}=1$ *and* $\frac{1}{\tilde{p}}+\frac{1}{\tilde{p}'}=1$.

For $(q,p) \neq (\infty,2)$, we have slightly stronger Strichartz estimates:

Corollary 2.5 *For $2 \le q, \tilde{q}, p, \tilde{p} \le \infty$, $(q,p) \ne (\infty, 2)$ and $q > \tilde{q}'$.*
(a)Suppose that $(q,p), (\tilde{q}, \tilde{p})$ are Schrödinger admissible pairs, then

$$\|S(t)u_0\|_{L_t^q \dot{B}^{\frac{2}{q}+\frac{3}{p}-\frac{3}{2}}_{(p,2+),2}} \lesssim \|u_0\|_{L_x^2}, \tag{9}$$

$$\left\| \int_0^t S(t-s)F(s)ds \right\|_{L_t^q \dot{B}^{\frac{2}{q}+\frac{3}{p}-\frac{3}{2}}_{(p,2+),2}} \lesssim \|F\|_{L_t^{\tilde{q}'} \dot{B}^{\frac{3}{2}-\frac{3}{\tilde{p}}-\frac{2}{\tilde{q}}}_{(\tilde{p}',2),2}}. \tag{10}$$

(b)Suppose that $(q,p), (\tilde{q}, \tilde{p})$ are wave admissible pairs, then

$$\|W(t)u_0\|_{L_t^q \dot{B}^{\frac{1}{q}+\frac{3}{p}-\frac{3}{2}}_{(p,2+),2}} \lesssim \|u_0\|_{L_x^2}, \tag{11}$$

$$\left\| \int_0^t W(t-s)F(s)ds \right\|_{L_t^q \dot{B}^{\frac{1}{q}+\frac{3}{p}-\frac{3}{2}}_{(p,2+),2}} \lesssim \|F\|_{L_t^{\tilde{q}'} \dot{B}^{\frac{3}{2}-\frac{3}{\tilde{p}}-\frac{1}{\tilde{q}}}_{(\tilde{p}',2),2}}. \tag{12}$$

3 Nonlinear Estimates

For the variables u and N in the transformed system, we use the following angular Strichartz norms with wider range as working spaces

$$\begin{aligned} u \in S &= L_t^\infty H^{0,1}_{2,\sigma} \cap L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2+),\sigma} \cap L_t^2 B^{0,1}_{6,\sigma}, \\ N \in W &= L_t^\infty H^{0,1}_{2,\sigma} \cap L_t^2 \dot{B}^{-1/4-\delta,1}_{(q(-\delta),2+),\sigma}, \end{aligned} \tag{1}$$

where $0 < \delta \ll 1$ is a fixed small enough number and q is defined by $\frac{1}{q(\delta)} = \frac{1}{4} + \frac{\delta}{3}$.

For $0 < \delta \ll 1$ small enough, there holds

$$\frac{10}{3} < q(\delta) < 4 < q(-\delta) < \infty,$$

then the norms defined in (1) are angular Strichartz admissible.

For the resonant terms containing $(\bar{N}u)_{HH+LH+1L}$ and $(u\bar{u})_{HH+L1+1L}$ in (7) and (8), we apply the inhomogeneous generalized Strichartz estimates to estimate them. We have

$$\begin{aligned} &\| \int_0^t S(t-s)(\bar{N}u)_{HH+LH+1L}(s)ds\|_S \\ &\lesssim \|(\bar{N}u)_{LH}\|_{L_t^1 H^{0,1}_{2,\sigma}} + \|(\bar{N}u)_{HH}\|_{L_t^1 H^{0,1}_{2,\sigma}} + \|(\bar{N}u)_{1L}\|_{L_t^{\tilde{r}'} \dot{B}^{\frac{3}{2}-\frac{2}{\tilde{q}}-\frac{3}{\tilde{r}},1}_{(\tilde{r}',2),\sigma}}, \end{aligned} \tag{2}$$

and

$$\begin{aligned}&\|\int_0^t W(t-s)\frac{D}{\sqrt{1+\varepsilon^2D^2}}(u\bar{u})_{HH+L1+1L}(s)ds\|_W\\&\lesssim \|\frac{D}{\sqrt{1+\varepsilon^2D^2}}(u\bar{u})_{HH}\|_{L_t^1H^{0,1}_{2,\sigma}}+\|\frac{D}{\sqrt{1+\varepsilon^2D^2}}(u\bar{u})_{1L+L1}\|_{L_t^{\tilde{q}_1'}\dot{B}^{\frac{3}{2}-\frac{1}{\tilde{q}}-\frac{3}{\tilde{r}},1}_{(\tilde{r}_1',2),\sigma}},\end{aligned}\tag{3}$$

where $(\tilde{q}',\tilde{r}')$ is the dual angular Schrödinger admissible pair and $(\tilde{q}_1',\tilde{r}_1')$ is the dual angular wave admissible pair.
To deal with the other nonlinear terms, we follow [3] to use representation theory of $SO(3)$. Let μ be Haar measure of $SO(3)$ and write $L_A^q=L^q(SO(3),\mu)$. Then

$$\|f\|_{\mathscr{L}_\rho^pL_\sigma^q}\sim\|f(Ax)\|_{L_x^pL_A^q},\quad \forall 1\le p,q\le\infty.$$

Lemma 3.1 ([10]) *For any $1<q<\infty$,*

$$\|f\|_{\mathscr{L}_\rho^p\mathscr{H}_q^1}\sim\|f\|_{\mathscr{L}_\rho^pL_\sigma^q}+\sum_{i,j}\|X_{i,j}f\|_{\mathscr{L}_\rho^pL_\sigma^q},$$

where $X_{i,j}=x_i\partial_j-x_j\partial_i$.

Let T_m be a bilinear operator on $\mathbb{R}^n$ defined as

$$T_m(f,g)(x)=\int_{\mathbb{R}^{2n}}m(\xi,\eta)\hat{f}(\xi)\hat{g}(\eta)e^{ix(\xi+\eta)}d\xi d\eta.$$

We recall a bilinear multiplier estimate proved in [3].

Lemma 3.2 ([3]) *Let $1\le p,p_1,p_2\le\infty$ and $1/p=1/p_1+1/p_2$. Assume $m(\xi,\eta)=h(|\xi|,|\eta|)$ for some function h, m is bounded and satisfies for all α,β*

$$|\partial_\xi^\alpha\partial_\eta^\beta m(\xi,\eta)|\le C_{\alpha\beta}|\xi|^{-|\alpha|}|\eta|^{-|\beta|},\quad \xi,\eta\ne0.$$

Then for $q>2$,

$$\|T_m(P_{k_1}f,P_{k_2}g)\|_{\mathscr{L}_\rho^p\mathscr{H}_q^1}\leqslant C\|f\|_{\mathscr{L}_\rho^{p_1}\mathscr{H}_q^1}\|g\|_{\mathscr{L}_\rho^{p_2}\mathscr{H}_q^1},$$

for any $k_1,k_2\in\mathbb{Z}$ with an uniform C.

With Lemma 3.1 and applying Lemma 3.2 for every bilinear dyadic piece, we have the following two lemmas following the proof of [4] and [7] with slightly modifications:

Lemma 3.3 (Bilinear Estimates) *Let δ be a small number.*
(1)For any N and u, there holds

$$\|(\bar{N}u)_{LH}\|_{L_t^1 H^{0,1}_{2,\sigma}} \lesssim \|N\|_{L_t^2 \dot{B}^{-1/4-\delta,1}_{(q(-\delta),2+),\sigma}} \|u\|_{L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2+),\sigma}},$$
$$\|(\bar{N}u)_{HH}\|_{L_t^1 H^{0,1}_{2,\sigma}} \lesssim \|N\|_{L_t^2 \dot{B}^{-1/4-\delta,1}_{(q(-\delta),2+),\sigma}} \|u\|_{L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2+),\sigma}},$$
$$\|(\bar{N}u)_{1L}\|_{L_t^{\tilde{q}'} \dot{B}^{\frac{3}{2}-\frac{2}{\tilde{q}}-\frac{3}{\tilde{r}},1}_{(\tilde{r}',2),\sigma}} \lesssim \|N\|_{L_t^2 \dot{B}^{-1/4-\delta,1}_{(q(-\delta),2+),\sigma}} \|u\|_{L_t^\infty H^{0,1}_{2,\sigma} \cap L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2),\sigma}},$$

where in the third estimate $0 \le \theta \le 1$, $\frac{1}{\tilde{q}} = \frac{1}{2} - \frac{\theta}{2}$, $\frac{1}{\tilde{r}} = \frac{1}{4} + \frac{\theta}{3} + \frac{\delta}{3}$.
(2)For any u, there holds

$$\|\frac{D}{\sqrt{1+\varepsilon^2 D^2}}(u\bar{u})_{HH}\|_{L_t^1 H^{0,1}_{2,\sigma}} \lesssim \|u\|_{L_t^2 \dot{B}^{1/4-\delta,1}_{(q(-\delta),2+),\sigma}} \|u\|_{L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2+),\sigma}},$$
$$\|\frac{D}{\sqrt{1+\varepsilon^2 D^2}}(u\bar{u})_{1L+L1}\|_{L_t^{\tilde{q}_1'} \dot{B}^{\frac{3}{2}-\frac{1}{\tilde{q}}-\frac{3}{\tilde{r}},1}_{(\tilde{r}_1',2),\sigma}} \lesssim \|u\|^2_{L_t^\infty H^{0,1}_{2,\sigma} \cap L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2+),\sigma}},$$

where in the last inequality $0 \le \theta \le 1$, $\frac{1}{\tilde{q}} = \frac{1}{2} - \frac{\theta}{2}$, $\frac{1}{\tilde{r}} = \frac{1}{4} + \frac{\theta}{3} - \frac{\delta}{3}$.

Sketch of the Proof. We briefly sketch the proof of the first estimate. By dyadic decomposition,

$$\begin{aligned}\|(\bar{N}u)_{LH}\|^2_{H^{0,1}_{2,\sigma}} &\lesssim \sum_{k_2} \| \sum_{k_1<k_2-5} \Lambda^1_\sigma (P_{k_1}\bar{N}P_{k_2}u)\|^2_{\mathscr{L}^2_\rho L^{2+}_\sigma} \\ &\lesssim \sum_{k_2} (\sum_{k_1<k_2-5} \|P_{k_1}\bar{N}P_{k_2}u\|_{\mathscr{L}^2_\rho \mathscr{H}^1_{2+}})^2.\end{aligned}$$

When $q > 2$, $\mathscr{H}^1_q$ is an algebra. Using the bilinear estimate, i.e., Lemma 3.2, we have

$$\begin{aligned}\|(\bar{N}u)_{LH}\|^2_{H^{0,1}_{2,\sigma}} &\lesssim \sum_{k_2} (\sum_{k_1<k_2-5} \|P_{k_1}\bar{N}\|_{\mathscr{L}^{q(-\delta)}_\rho \mathscr{H}^1_{2+}} \|P_{k_2}u\|_{\mathscr{L}^{q(\delta)}_\rho \mathscr{H}^1_{2+}})^2 \\ &= \sum_{k_2} (\sum_{k_1<k_2-5} 2^{k_1(-\frac{1}{4}-\delta)} \|P_{k_1}\bar{N}\|_{\mathscr{L}^{q(-\delta)}_\rho \mathscr{H}^1_{2+}} 2^{k_1(\frac{1}{4}+\delta)} \|P_{k_2}u\|_{\mathscr{L}^{q(\delta)}_\rho \mathscr{H}^1_{2+}})^2 \\ &\lesssim \|N\|^2_{\dot{B}^{-1/4-\delta,1}_{(q(-\delta),2+),\sigma}} \|u\|^2_{\dot{B}^{1/4+\delta,1}_{(q(\delta),2+),\sigma}},\end{aligned}$$

which implies the first estimate by using Hölder's inequality in time. The other terms can be estimated similarly, we skip the details.
For the boundary terms, applying homogeneous Strichartz estimates, we obtain

Lemma 3.4 (Boundary terms I) *For any N_0 and u_0,*

$$\|S(t)\Omega_1(\bar{N},u)(0)\|_S \lesssim \|\Omega_1(\bar{N}_0,u_0)\|_{H^{0,1}_{2,\sigma}} \lesssim \|N_0\|_{H^{0,1}_{2,\sigma}} \|u_0\|_{H^{0,1}_{2,\sigma}},$$
$$\|W(t)D\Omega_3(u,u)(0)\|_W \lesssim \|D\Omega_3(u_0,u_0)\|_{H^{0,1}_{2,\sigma}} \lesssim \|u_0\|^2_{H^{0,1}_{2,\sigma}}.$$

Then for any N and u, there holds

$$\|\Omega_1(\bar{N},u)\|_{L_t^\infty H^{0,1}_{2,\sigma}} \lesssim \|N\|_{L_t^\infty H^{0,1}_{2,\sigma}} \|u\|_{L_t^\infty H^{0,1}_{2,\sigma}}, \quad \|D\Omega_3(u,u)\|_{L_t^\infty H^{0,1}_{2,\sigma}} \lesssim \|u\|^2_{L_t^\infty H^{0,1}_{2,\sigma}}.$$

We need the Coifman-Meyer bilinear multiplier estimates to deal with the other nonlinear terms

Lemma 3.5 *Assume m is bounded and satisfying the following estimates:*

$$|\partial_\xi^\alpha \partial_\eta^\beta m(\xi,\eta)| \le C_{\alpha\beta} |\xi|^{-|\alpha|} |\eta|^{-|\beta|}, \quad \forall \alpha,\beta.$$

Let $1 \leqslant p,q,r \leqslant \infty$, $1/r = 1/p + 1/q$, *then for any* $k_1, k_2 \in \mathbb{Z}$, *we have*

$$\|T_m(P_{k_1} f, P_{k_2} g)\|_{L^r} \leqslant C \|f\|_{L^p} \|g\|_{L^q}.$$

Since $X_{i,j}$ commutes with the radial Fourier multiplier operator and $X_{ij}(fg) = gX_{ij}f + fX_{ij}g$, applying X_{ij} to the multiplier on dyadic piece and then estimating with Lemma 3.5 similarly as in [7], we have the following bilinear and trilinear estimates:

Lemma 3.6 (Boundary terms II) *For any N and u, there holds*

$$\|\Omega_1(\bar{N},u)\|_{L_t^2 \dot{B}^{1/4+\delta,1}_{(q(\delta),2),\sigma}} \lesssim \|N\|_{L_t^\infty H^{0,1}_{2,\sigma}} \|u\|_{L_t^2 B^{0,1}_{6,\sigma}},$$
$$\|D\Omega_3(u,u)\|_{L_t^2 \dot{B}^{-1/4-\delta}_{(q(-\delta),2),\sigma}} \lesssim \|u\|_{L_t^2 B^{0,1}_{6,\sigma}} \|u\|_{L_t^\infty H^{0,1}_{2,\sigma}},$$

where the implicit constant is independent of ε.

Sketch of the Proof. For the first estimate, using dyadic decomposition, Sobolev embedding and Lemma 3.1

$$\begin{aligned}
\|\Omega_1(\bar{N},u)\|^2_{\dot{B}^{1/4+\delta,1}_{(q(\delta),2),\sigma}} &\lesssim \|D\Omega_1(\bar{N},u)\|^2_{H^{0,1}_{2,\sigma}} \\
&\lesssim \sum_{k_2} \|\Lambda^1_\sigma P_{k_2} \langle D\rangle^{-1} \sum_{k_1 \le k_2 - 5} D\langle D\rangle \Omega_1(P_{k_2}\bar{N}, P_{k_1} u)\|^2_{L^2} \\
&\lesssim \sum_{k_2} \left(\sum_{k_1 \le k_2 - 5} \langle 2^{k_2}\rangle^{-1} \|P_{k_2} D\langle D\rangle \Omega_1(P_{k_2}\bar{N}, P_{k_1} u)\|_{L^2} \right)^2 \\
&\quad + \sum_{k_2} \left(\sum_{k_1 \le k_2 - 5} \sum_{i,j} \langle 2^{k_2}\rangle^{-1} \|P_{k_2} D\langle D\rangle \Omega_1(X_{i,j} P_{k_2}\bar{N}, P_{k_1} u)\|_{L^2} \right)^2 \\
&\quad + \sum_{k_2} \left(\sum_{k_1 \le k_2 - 5} \sum_{i,j} \langle 2^{k_2}\rangle^{-1} \|P_{k_2} D\langle D\rangle \Omega_1(P_{k_2}\bar{N}, X_{i,j} P_{k_1} u)\|_{L^2} \right)^2.
\end{aligned}$$

In which $D\langle D\rangle \Omega_1(P_{k_2}\bar{N}, P_{k_1} u)$ is a bilinear multiplier with symbol

$$m_1(\xi,\eta) = \frac{|\xi+\eta|\langle \xi+\eta\rangle \chi_{k_2}(\xi)\chi_{k_1}(\eta)}{\omega_1(|\xi+\eta|) - \omega_1(|\eta|) - \omega_2(|\xi|)}.$$

One can check that $m_1(\xi,\eta)$ satisfies the conditions in the Coifman-Meyer multiplier estimate, i.e., Lemma 3.5. Then

$$\begin{aligned}\|\Omega_1(\bar{N},u)\|^2_{\dot{B}^{1/4+\delta,1}_{(q(\delta),2),\sigma}} &\lesssim \sum_{k_2}\Big(\sum_{k_1\le k_2-5}\langle 2^{k_2}\rangle^{-1}\|\Lambda^1_\sigma P_{k_2}\bar{N}\|_{L^2}\|\Lambda^1_\sigma P_{k_1}u\|_{L^\infty}\Big)^2\\ &\lesssim \|N\|^2_{H^{0,1}_{2,\sigma}}\|u\|^2_{B^{0,1}_{6,\sigma}},\end{aligned}$$

which implies the first estimate. For the boundary term containing Ω_3, we skip the details and refer to [7] for the detailed proof.
For the cubic terms, we have

Lemma 3.7 (Trilinear estimates) *For any N and u, we get*

$$\begin{aligned}&\Big\|\int_0^t S(t-s)\Omega_2(D|u|^2,u)(s)ds\Big\|_S \lesssim \|\Omega_2(D|u|^2,u)\|_{L^1_t H^{0,1}_{2,\sigma}} \lesssim \|u\|^2_{L^2_t B^{0,1}_{6,\sigma}}\|u\|_{L^\infty_t H^{0,1}_{2,\sigma}},\\ &\Big\|\int_0^t S(t-s)\Omega_1(\bar{N},\bar{N}u)(s)ds\Big\|_S \lesssim \|\Omega_1(\bar{N},\bar{N}u)\|_{L^2_t \dot{B}^{0,1}_{(\frac{6}{5},2),\sigma}} \lesssim \|u\|_{L^2_t B^{0,1}_{6,\sigma}}\|N\|^2_{L^\infty_t H^{0,1}_{2,\sigma}},\\ &\Big\|\int_0^t W(t-s)(D\Omega_3(\bar{N}u,u)-D\Omega_3(u,\bar{N}u))ds\Big\|_W \lesssim \|D\Omega_3(\bar{N}u,u)\|_{L^1_t H^{0,1}_{2,\sigma}} \lesssim \|u\|^2_{L^2_t B^{0,1}_{6,\sigma}}\|N\|_{L^\infty_t H^{0,1}_{2,\sigma}}.\end{aligned}$$

Sketch of the Proof. We only sketch the main idea of the proof for the first trilinear estimate. Using dyadic decomposition and Bernstein's inequality(see for instance [9]),

$$\begin{aligned}\|\Omega_2(D|u|^2,u)\|^2_{H^{0,1}_{2,\sigma}} &\lesssim \sum_{k_2}\Big\|\sum_{k_1\le k_2-5}\Lambda^1_\sigma P_{k_2}\Omega_2(P_{k_2}D|u|^2,P_{k_1}u)\Big\|^2_{L^2}\\ &\lesssim \sum_{k_2}2^{2k_2}\Big\|\sum_{k_1\le k_2-5}\Lambda^1_\sigma P_{k_2}\Omega_2(P_{k_2}D|u|^2,P_{k_1}u)\Big\|^2_{L^{\frac{6}{5}}}.\end{aligned}$$

Recall that the resonant function for the Schrödinger component is

$$\Phi_\varepsilon := \omega_1(|\xi|)-\omega_1(|\eta|)-\omega_2(|\xi-\eta|).$$

In the support of the symbol of Ω_2, for the low frequency part($|\xi|\lesssim 1, |\eta|\ll|\xi|\sim|\xi-\eta|$),

$$|\Phi_\varepsilon|\sim|\xi|.$$

While for the high frequency part ($|\xi|\gg 1, |\eta|\ll|\xi|\sim|\xi-\eta|$),

$$|\Phi_\varepsilon|\sim|\xi|^4.$$

Therefore it can absorb four derivatives for the high frequency in terms of Coifman-Meyer multiplier estimate(Lemma 3.5) which helps to close the argument. In detail,

$$\|\Omega_2(D|u|^2,u)\|^2_{H^{0,1}_{2,\sigma}} \lesssim \sum_{k_2}2^{2k_2}\langle 2^{k_2}\rangle^{-6}\Big(\sum_{k_1\le k_2-5}\|\Lambda^1_\sigma P_{k_2}\langle D\rangle^3\Omega_2(P_{k_2}D|u|^2,P_{k_1}u)\|_{L^{\frac{6}{5}}}\Big)^2$$

Noticing that $\langle D\rangle^3\Omega_2(P_{k_2}D|u|^2,P_{k_1}u)$ is a bilinear multiplier with symbol

$$m_2(\xi,\eta)=\frac{\langle\xi+\eta\rangle^3|\xi|\chi_{k_2}(\xi)\chi_{k_1}(\eta)}{(\omega_1(|\xi+\eta|)-\omega_1(|\eta|)-\omega_2(|\xi|))\sqrt{1+\varepsilon^2|\xi|^2}},$$

which satisfies the conditions in Lemma 3.5. Therefore

$$\begin{aligned}\|\Omega_2(D|u|^2,u)\|^2_{H^{0,1}_{2,\sigma}}\lesssim&\sum_{k_2}\langle 2^{k_2}\rangle^{-6}2^{2k_2}\left(\sum_{k_1\le k_2-5}\|P_{k_2}|u|^2\|_{L^{\frac{3}{2}}}\|P_{k_1}u\|_{L^6}\right)^2\\&+\sum_{k_2}\langle 2^{k_2}\rangle^{-6}2^{2k_2}\left(\sum_{k_1\le k_2-5}\sum_{i,j}\|X_{i,j}P_{k_2}|u|^2\|_{L^{\frac{3}{2}}}\|P_{k_1}u\|_{L^6}\right)^2\\&+\sum_{k_2}\langle 2^{k_2}\rangle^{-6}2^{2k_2}\left(\sum_{k_1\le k_2-5}\sum_{i,j}\|P_{k_2}|u|^2\|_{L^{\frac{3}{2}}}\|X_{i,j}P_{k_1}u\|_{L^6}\right)^2\\\lesssim&\|u\|^4_{B^{0,1}_{6,\sigma}}\|u\|^2_{H^{0,1}_{2,\sigma}},\end{aligned}$$

which yields the first estimate as desired. The other terms can be estimated similarly. For instance, for the third estimate containing Ω_3, one only need to notice that the resonant function for the wave component is

$$\tilde{\Phi}_\varepsilon=\omega_2(|\xi|)+\omega_1(|\eta|)-\omega_1(|\xi-\eta|),$$

and $|\tilde{\Phi}_\varepsilon|$ behaves like $\langle\xi\rangle^3|\xi|$(when $|\eta|\ll|\xi|$) which can again absorb four derivatives in high frequency. We skip the details of proof.

Remark 3.8 *When $\varepsilon=0$, namely for the original Zakharov system, the resonant function for the Schrödinger component behaves like*

$$|\Phi_\varepsilon|\sim\langle\xi\rangle|\xi|,$$

which can only absorb two derivatives for the high frequency. This is one of the main reason that quantum Zakharov system has much better properties than the original Zakharov system.

4 Scattering in L^2

For any small initial data $(u_0,N_0)\in H^{0,1}_{2,\sigma}(\mathbb{R}^3)\times H^{0,1}_{2,\sigma}(\mathbb{R}^3)$, we define the operators

$$\begin{aligned}\Psi^1_{u_0}(u,N)=&S(t)u_0-\Omega_1(\bar{N},u)(t)+S(t)\Omega_1(\bar{N},u)(0)-i\int_0^t S(t-s)\Omega_2(D|u|^2,u)(s)ds\\&-i\int_0^t S(t-s)\Omega_1(\bar{N},\bar{N}u)(s)ds-i\int_0^t S(t-s)(\bar{N}u)_{HH+LH+1L}(s)ds,\end{aligned}$$

and

$$\Psi^2_{\bar{N}_0}(u,N) = W(t)\bar{N}_0 - D\Omega_3(u,u)(t) - i\int_0^t W(t-s)(D\Omega_3(\bar{N}u,u) - D\Omega_3(u,\bar{N}u))ds$$
$$+ W(t)D\Omega_3(u,u)(0) - i\int_0^t W(t-s)\frac{D}{\sqrt{1+\varepsilon^2 D^2}}(u\bar{u})_{HH+L1+1L}(s)ds.$$

Write $\Psi : (u,N) \to (\Psi^1_{u_0}, \Psi^2_{\bar{N}_0})$ and choose the resolution space as

$$\mathscr{D} = \{(u,N) : \|(u,N)\|_X \le \alpha\},$$

with the norm $\|(u,N)\|_X = \|u\|_S + \|N\|_W$, α is a small number to be determined. Applying the Strichartz and nonlinear estimates, for any $(u,N) \in \mathscr{D}$, we have

$$\|\Psi^1_{u_0}(u,N)\|_S \lesssim \|u_0\|_{H^{0,1}_{2,\sigma}} + \|N\|_{L^\infty_t H^{0,1}_{2,\sigma}}\|u\|_{L^\infty_t H^{0,1}_{2,\sigma}} + \|N_0\|_{H^{0,1}_{2,\sigma}}\|u_0\|_{H^{0,1}_{2,\sigma}}$$
$$+ \|N\|_{L^\infty_t H^{0,1}_{2,\sigma}}\|u\|_{L^2_t B^{0,1}_{6,\sigma}} + \|u\|^2_{L^2_t B^{0,1}_{6,\sigma}}\|u\|_{L^\infty_t H^{0,1}_{2,\sigma}} + \|u\|_{L^2_t B^{0,1}_{6,\sigma}}\|N\|^2_{L^\infty_t H^{0,1}_{2,\sigma}}$$
$$+ \|N\|_{L^2_t \dot{B}^{-1/4-\delta}_{(q(-\delta),2+),\sigma}}\|u\|_{L^2_t \dot{B}^{1/4+\delta}_{(q(\delta),2+),\sigma}},$$

and

$$\|\Psi^2_{\bar{N}_0}(u,N)\|_W \lesssim \|N_0\|_{H^{0,1}_{2,\sigma}} + \|u\|^2_{L^\infty_t H^{0,1}_{2,\sigma}} + \|u_0\|^2_{H^{0,1}_{2,\sigma}}$$
$$+ \|u\|_{L^2_t B^{0,1}_{6,\sigma}}\|u\|_{L^\infty_t H^{0,1}_{2,\sigma}} + \|u\|^2_{L^2_t B^{0,1}_{6,\sigma}}\|N\|_{L^\infty_t H^{0,1}_{2,\sigma}} + \|u\|_{L^2_t \dot{B}^{1/4-\delta}_{(q(-\delta),2+),\sigma}}\|u\|_{L^2_t \dot{B}^{1/4+\delta}_{(q(\delta),2+),\sigma}}.$$

Then

$$\|\Psi(u,N)\|_X = \|\Psi^1_{u_0}(u,N)\|_S + \|\Psi^2_{\bar{N}_0}(u,N)\|_W$$
$$\lesssim \|u_0\|_{H^{0,1}_{2,\sigma}} + \|N_0\|_{H^{0,1}_{2,\sigma}} + (\|u_0\|_{H^{0,1}_{2,\sigma}} + \|N_0\|_{H^{0,1}_{2,\sigma}})^2 + \|(u,N)\|^2_X + \|(u,N)\|^3_X.$$

If the initial data is sufficiently small, namely, $\beta_0 = \|u_0\|_{H^{0,1}_{2,\sigma}} + \|N_0\|_{H^{0,1}_{2,\sigma}} \ll 1$, we choose $\alpha = C\beta_0$, then $\Psi : \mathscr{D} \to \mathscr{D}$. Similarly Ψ is a contraction mapping on $\mathscr{D}$. Therefore there exists a unique solution on $\mathscr{D}$ with global space-time bound. By the standard techniques, we obtain that the solution $(u(t), N(t))$ to (3) scatters in $H^{0,1}_{2,\sigma} \times H^{0,1}_{2,\sigma}$.

Acknowledgements The author would like to thank the anonymous referee for their thoughtful suggestions which help improve the paper. The author is supported by the National Natural Science Foundation of China (No. 11971503), the Young Talents Program(No. QYP1809) and the disciplinary funding of Central University of Finance and Economics.

References

1. Y. Fang and K. Nakanishi, Global well-posedness and scattering for the quantum Zakharov system in L2, Proceedings of the American Mathematical Society, 2019, Series B(6), 21-32.
2. Z. Guo, Z. Hani, K. Nakanishi, Scattering for the 3D Gross-Pitaevskii equation, Communications in Mathematical Physics 359(2018), 265-295.
3. Z. Guo, S. Lee, K. Nakanishi, and C. Wang, Generalized Strichartz Estimates and Scattering for 3D Zakharov System, Communications in Mathematical Physics 331 (2014), 239–259.
4. Z. Guo, K. Nakanishi, Small energy scattering for the Zakharov system with radial symmetry, Int. Math. Res. Not. 9 (2014), 2327-2342.
5. Y. Guo, J. Zhang and B. Guo, Global well-posedness and the classical limit of the solution for the quantum Zakharov system. Z. Angew. Math. Phys. 64 (2013), 53-68.
6. F. Haas, Quantum plasmas. Springer Series on Atomics, Optical and Plasma Physics 65 (2011).
7. C. Huang, B. Guo and Y. Heng, Scattering for the 3D quantum Zakharov system in L2 with radial data. Preprint.
8. J. Shatah, Normal forms and quadratic nonlinear Klein-Gordon equations. Comm. Pure Appl. Math. 38 (1985), no. 5, 685-696.
9. T. Tao, Nonlinear dispersive equations, local and global analysis. CBMS. Regional Conference Series in Mathematics, 106. Published for the Conference Board of the Mathematical Science, Washington, DC; by the American Mathematical Society, Providence, RI, 2006. ISBN: 0-8218-4143-2.
10. M. Taylor, Partial Differential Equations, Vol. 1, Second Edition, Springer, New York(2011).

Chapter 9

Tropical Geometry and Mirror Symmetry

Algebraic and symplectic viewpoint on compactifications of two-dimensional cluster varieties of finite type

Man-Wai Mandy Cheung and Renato Vianna

Abstract In this article we explore compactifications of cluster varieties of finite type in complex dimension two. Cluster varieties can be viewed as the spec of a ring generated by theta functions and a compactification of such varieties can be given by a grading on that ring, which can be described by positive polytopes [17]. In the examples we exploit, the cluster variety can be interpreted as the complement of certain divisors in del Pezzo surfaces. In the symplectic viewpoint, they can be described via almost toric fibrations over $\mathbb{R}^2$ (after completion). Once identifying them as almost toric manifolds, one can symplectically view them inside other del Pezzo surfaces. So we can identify other symplectic compactifications of the same cluster variety, which we expect should also correspond to different algebraic compactifications. Both viewpoints are presented here and several compactifications have their corresponding polytopes compared. The finiteness of the cluster mutations are explored to provide cycles in the graph describing monotone Lagrangian tori in del Pezzo surfaces connected via almost toric mutation [34].

1 Introduction

Cluster algebras, introduced by Fomin and Zelevinsky [12], are subalgebras of rational functions in n variables. The generators of cluster algebras are called the cluster variables. Instead of being given the complete sets of generators and relations as other commutative rings, a cluster algebra is defined from an (initial) seed, which

Man-Wai Mandy Cheung
Harvard University, One Oxford Street, Cambridge, MA 02138, United States of America
e-mail: mwcheung@math.harvard.edu

Renato Vianna
Institute of Mathematics, Federal University of Rio de Janeiro; Av. Athos da Silveira Ramos, 149 - Ilha do Fundão, Rio de Janeiro - RJ, 21941-909, Brazil
e-mail: renato@im.ufrj.br

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_35

includes a set of the generators and a matrix. An iterative procedure called mutation would produce new seeds from a given seed and this process gives all the cluster variables. The cluster algebra is then defined to be the ring generated by all cluster variables.

Geometrically, the cluster varieties, described by Fock and Goncharov [11], and by Gross, Hacking, Keel in [16], are defined in a similar manner. A seed data now would be associated to an algebraic torus. The mutation procedures give the birational transformations used to glue the tori. A cluster variety is then the union of the tori under the gluing.

The compactification of the cluster varieties can be given by a Rees construction. Combinatorially, the construction can be described by 'convex' polytopes, called the positive polytopes [17]. The article [8] showed that the positive polytopes satisfy a convexity condition called 'broken line convexity'. As the seed mutates, the polytope mutates correspondingly. One can then give a mutation process to the polytopes. Note that under this type of mutation, there is no change in the compactification.

More generally, one can similarly describe the compactification of the log Calabi-Yau surfaces studied in [15]. In this case, one would construct the dual intersection complex of a given Looijenga pair. The underlying topological space of the complex will carry an affine manifold structure. The affine structures would correspond to another type of mutation for the positive polytopes.

On the other hand, in the symplectic viewpoint, mutations were exploited in four dimensional symplectic geometry [33, 34], inspired by the pioneering work of Galkin-Usnhish [13] (further developed in [1]), and being grounded on the development of almost toric fibrations (ATFs) by Symington [32]. Upon identifying an almost toric fibration of a open variety, we can symplectically identify it as a symplectic submanifold of some closed symplectic manifold. We will refer to it as a (symplectic) compactification. In the examples of this paper, we can identify the symplectic form as the Kähler form of del Pezzo surfaces. We expect that symplectic compactifications can be translated to algebraic compactifications under certain nuances discussed in Section 3.

This paper is an attempt to understand the two notions. The motivation of both sides come from the Strominger-Yau-Zaslow conjecture – the conjecture suggests there are special Lagrangian fibrations for the Calabi-Yau manifold and its mirror space over the base B. The construction of the log-CY variety from the symplectic side is via the almost toric fibration, Meanwhile, in the algebro-geometric side, the construction can be described in terms of the wall crossing structures called the scattering diagrams.

We begin with the algebro-geometric perspective in Section 2. In this section, we will discuss cluster varieties, positive polytopes, compactifications, and the mutations of the polytopes. Then in Section 3.1, we give a perspective on how cluster varieties and scattering diagrams arise from considering wall crossing corrections as one attempt to build a mirror in terms of the SYZ picture. In complex dimension two, the wall-crossing happens when we consider singular Lagrangian fibrations known as almost toric fibrations (ATFs). In particular, we illustrate the idea in terms

of the A_2 cluster variety – compactified as the del Pezzo surface of degree 5 in Section 3.1.3. Afterward, we explore compactifications of cluster varieties using the almost toric viewpoint in Section 3.2.

The symplectic geometry approach to compactification via almost toric fibrations makes no reference to the complex structure, while the algebro-geometric approach does not fix a symplectic form. Nonetheless, because a similar set of data can encode the scattering diagram as well as an ATF, we seem to always be able to relate compactifications, encoded by the same polytope in both pictures. We aim to show the correspondence between the symplectic compactification of the cluster varieties to the algebro-geometric version in our upcoming papers.

2 Mutations in algebraic geometry

2.1 Cluster varieties

We will first recall some notation used in the definition of a cluster varieties. A *fixed data* consists of a lattice N with a skew-symmetric bilinear form $\{\cdot,\cdot\} : N \times N \to \mathbb{Q}$, an index set I with $|I| = \operatorname{rank} N$, positive integers d_i for $i \in I$, a sublattice $N^\circ \subseteq N$ of finite index with some integral properties, the dual lattice $M = \operatorname{Hom}(N,\mathbb{Z})$ and the corresponding $M^\circ = \operatorname{Hom}(N^\circ,\mathbb{Z})$. One can refer to [16] for the full definition of fixed data. Consider $N_{\mathbb{R}} = N \otimes \mathbb{R}$ and $M_{\mathbb{R}} = M \otimes \mathbb{R}$.

Given this fixed data, a *seed data* for this fixed data is $\mathbf{s} := (e_i \in N \mid i \in I)$, where $\{e_i\}$ is a basis for N. The basis for M° would then be $f_i = \frac{1}{d_i} e_i^*$. One can then associate the seed tori

$$\mathscr{A}_{\mathbf{s}} = T_{N^\circ} = \operatorname{Spec} \Bbbk[M^\circ], \qquad \mathscr{X}_{\mathbf{s}} = T_M = \operatorname{Spec} \Bbbk[N].$$

We will denote the coordinates as $X_i = z^{e_i}$ and $A_i = z^{f_i}$ and they are called the *cluster variables*. Similar to the definition of cluster algebras, there is a procedure, called *mutation*, to produce a new seed data $\mu(\mathbf{s})$ from a given seed $\mathbf{s}$. The mutation formula is stated in [16, Equation 2.3] which we will skip here. The essence is that we will obtain new seed tori $\mathscr{A}_{\mu(\mathbf{s})}$, $\mathscr{X}_{\mu(\mathbf{s})}$ from the mutated seed. Between the tori, there are birational maps $\mu_{\mathscr{X}} : \mathscr{X}_{\mathbf{s}} \dashrightarrow \mathscr{X}_{\mu(\mathbf{s})}$, $\mu_{\mathscr{A}} : \mathscr{A}_{\mathbf{s}} \dashrightarrow \mathscr{A}_{\mu(\mathbf{s})}$ which are stated in [16, Equations 2.5, 2.6]. Note that those birational maps are basically the mutations of cluster variables as in Fomin and Zelevinsky [12].

Let $\mathscr{A}$ be an union of tori glued by $\mathscr{A}$-mutation $\mu_{\mathscr{A}}$. A smooth scheme V is a *cluster variety of type* $\mathscr{A}$ if there is a birational map $\mu : V \dashrightarrow \mathscr{A}$ which is an isomorphism outside codimension two subsets of the domain and range. The *cluster variety of type* $\mathscr{X}$ is defined analogously.

The $\mathscr{A}$ and $\mathscr{X}$ cluster varieties can be fit into the formalism of the cluster varieties with principal coefficients $\mathscr{A}_{\text{prin}}$. The scheme $\mathscr{A}_{\text{prin}}$ is defined similarly to the $\mathscr{A}$ by 'doubling' the fixed data, i.e. considering $\widetilde{N} = N \oplus M^\circ$ as fixed data as in [16, Construction 2.11]. Then there are two natural inclusions. The first one is

$$\widetilde{p}^*: N \to \widetilde{M}^\circ = M^\circ \oplus N,$$
$$n \mapsto (p^*(n), n),$$

where $p^*(n) = \{n, \cdot\} \in M^\circ$ in the case of no frozen variable. Then for any seed $\mathbf{s}$, note that $\mathscr{A}_{\text{prin},\mathbf{s}} = T_{\widetilde{N}^\circ}$, and $\mathscr{X}_{\mathbf{s}} = T_M$, the there is the exact sequence of tori

$$1 \to T_{N^\circ} \to \mathscr{A}_{\text{prin},\mathbf{s}} \xrightarrow{\widetilde{p}} \mathscr{X}_{\mathbf{s}} \to 1.$$

The map $\widetilde{p}$ commutes with the mutation maps and thus we get the morphism $\widetilde{p}\colon \mathscr{A}_{\text{prin}} \to \mathscr{X}$. Further the T_{N° action on $\mathscr{A}_{\text{prin},\mathbf{s}}$ extends to $\mathscr{A}_{\text{prin}}$ which makes $\widetilde{p}$ a quotient map. Thus, the $\mathscr{X}$ variety can be seen as $\mathscr{A}_{\text{prin}}/T_{N^\circ}$.

The second inclusion is

$$\pi^*: N \to M^\circ,$$
$$n \mapsto (0, n).$$

In this case, the π^* map induces a projection $\pi : \mathscr{A}_{\text{prin}} \to T_M$. Then the usual $\mathscr{A}$ variety is $\pi^{-1}(e)$, where e is the identity of T_M.

We would like to indicate another viewpoint of the cluster varieties here. The mutation maps may be described in terms of elementary transformation of $\mathbb{P}^1$ bundles. Thus the cluster varieties can also be seen as the blowups of toric varieties (up to codimension two) as well.

Given a seed data, consider the fans

$$\Sigma_{\mathbf{s},\mathscr{A}} := \{0\} \cup \{\mathbb{R}_{\geq 0} d_i e_i \mid i \in I\} \subseteq N^\circ,\ \Sigma_{\mathbf{s},\mathscr{X}} := \{0\} \cup \{-\mathbb{R}_{\geq 0} d_i v_i \mid i \in I\} \subseteq M,$$

where $v_i = p^*(e_i)$ and the i only runs over the unfrozen variables if the frozen variables exist. Let $\mathrm{TV}_{\mathbf{s},\mathscr{A}}$ and $\mathrm{TV}_{\mathbf{s},\mathscr{X}}$ be the respective toric varieties. Denote D_i to be the toric divisor corresponding to the one-dimensional ray in one of these fans. Define the closed subschemes

$$Z_{\mathscr{A},i} := D_i \cap \bar{V}(1 + z^{v_i}) \subseteq \Sigma_{\mathbf{s},\mathscr{A}},\ Z_{\mathscr{X},i} := D_i \cap \bar{V}\left((1 + z^{e_i})^{\text{ind } d_i v_i}\right) \subseteq \Sigma_{\mathbf{s},\mathscr{X}},$$

where $\bar{V}$ denote the closure of the variety V, and ind $d_i v_i$ is the greatest degree of divisibility of $d_i v_i$ in M. Then consider the pairs $(\widetilde{\mathrm{TV}}_{\mathbf{s},\mathscr{A}}, D)$ and $(\widetilde{\mathrm{TV}}_{\mathbf{s},\mathscr{X}}, D)$ consisting of the blowups of $\mathrm{TV}_{\mathbf{s},\mathscr{A}}$ and $\mathrm{TV}_{\mathbf{s},\mathscr{X}}$ respectively, with D the proper transform of the toric boundaries. Define $X_{\mathbf{s},\mathscr{A}} = \widetilde{\mathrm{TV}}_{\mathbf{s},\mathscr{A}} \setminus D$ and $X_{\mathbf{s},\mathscr{X}} = \widetilde{\mathrm{TV}}_{\mathbf{s},\mathscr{X}} \setminus D$. When the seed $\mathbf{s}$ mutates to $\mathbf{s}'$, the corresponding $X_{\mathbf{s},\mathscr{A}}$, $X_{\mathbf{s}',\mathscr{A}}$ and $X_{\mathbf{s},\mathscr{X}}$, $X_{\mathbf{s}',\mathscr{X}}$ are isomorphic outside a codimension two set. In finite type, where there are only finitely cluster variables, the $\mathscr{A}$ and $\mathscr{X}$ would then also isomorphic to $X_{\mathbf{s},\mathscr{A}}$ and $X_{\mathbf{s},\mathscr{X}}$. Note that the whole set up here is building a toric model for the cluster varieties. We will introduce the notion of toric model for log Calabi Yau surfaces later in Section 2.3.

Scattering diagrams

Scattering diagrams live in the tropicalization of the cluster varieties. One can also see the diagrams encode the structure of the cluster varieties combinatorially.

A *wall* in $M_{\mathbb{R}}$ is a pair $(\mathfrak{d}, f_{\mathfrak{d}})$ where $\mathfrak{d} \subseteq M_{\mathbb{R}}$ is a convex rational polyhedral cone of codimension one, contained in $n^{\perp}$ for some $n \in N$, and $f_{\mathfrak{d}} = 1 + \sum_{k\geq 1} c_k z^{kp^*(n)}$, where $c_k \in \mathbb{C}$. A wall $(\mathfrak{d}, f_{\mathfrak{d}})$ is called *incoming* if $p^*(n) \in \mathfrak{d}$. Otherwise it is called *outgoing*. A *scattering diagram* $\mathfrak{D}$ is then a collection of walls with certain finiteness properties. Given a seed, an $\mathscr{A}_{\text{prin}}$-cluster scattering diagram can be constructed [17] and canonically determined by this given seed data. The $\mathscr{A}$ scattering diagram can be obtained by the projection $\widetilde{M}_{\mathbb{R}} \to M_{\mathbb{R}}$ while the $\mathscr{X}$ scattering diagrams can be defined as slicing the $\mathscr{A}_{\text{prin}}$ scattering diagrams by considering $\{(m,n) \mid m = p^*(n)\}$.

It is worth addressing here that for finite type, each chamber, i.e. the maximal cone, of the scattering diagram can be associated to a torus. The wall functions $f_{\mathfrak{d}}$ are actually representing the birational maps between the tori. Thus the cluster varieties can be seen as gluing of tori associated to the chamber via the wall crossing.

In this article, we will focus on the dimension 2 cluster varieties of finite type. The $\mathscr{A}$ scattering diagrams are listed as in Figure 1 while the $\mathscr{X}$ scattering diagrams of rank 2 finite type are listed as in Figure 2.

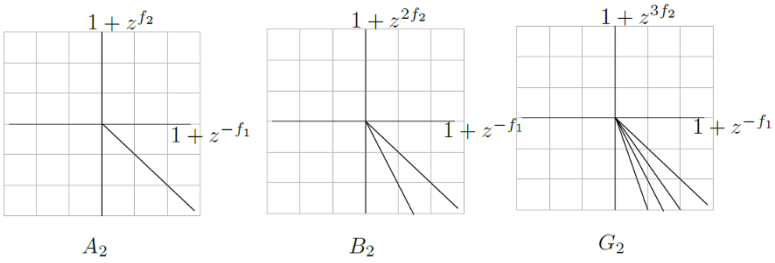


Fig. 1: $\mathscr{A}$-scattering diagrams for rank 2 finite type.

Mutation of scattering diagrams

As noted in the previous section, a seed determines canonically a scattering diagram. Two mutation-equivalent seeds would then give two different scattering diagrams. It is natural to consider 'mutation equivalent' scattering diagrams. This equivalence is given by piecewise linear maps on the lattices which are very similar to those in Section 3 and hence we will state here.

Consider two seeds $\mathbf{s}$ and $\mathbf{s}'$ which are just one mutation step apart, i.e. $\mathbf{s}' = \mu_k(\mathbf{s})$ for some $k \in I$. Then the corresponding scattering diagrams $\mathfrak{D}_{\mathbf{s}}$ and $\mathfrak{D}_{\mathbf{s}'}$ are

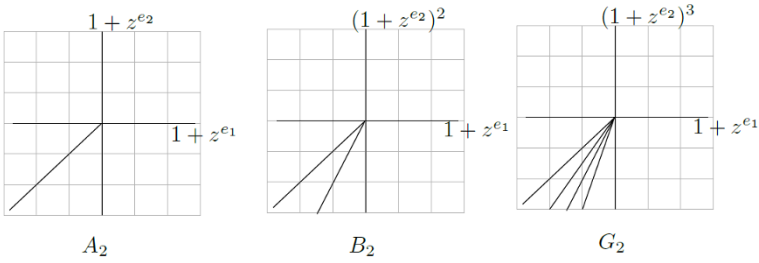


Fig. 2: $\mathscr{X}$-scattering diagrams for rank 2 finite type.

equivalent to each other by the transformation $T_k : M^\circ \to M^\circ$,

$$T_k(m) = \begin{cases} m + \langle d_k e_k, m\rangle v_k, & \text{for } m \in \mathscr{H}_{k,+} \\ m, & \text{for } m \in \mathscr{H}_{k,-} \end{cases} \tag{1}$$

for $m \in M^\circ$, $v_k = p^*(e_k)$, and $\mathscr{H}_{k,+} = \{m \in M_\mathbb{R} | \langle e_k, m\rangle \geq 0\}$, $\mathscr{H}_{k,-} = \{m \in M_\mathbb{R} | \langle e_k, m\rangle \leq 0\}$. Extending T_k to the wall functions [17, Theorem 1.24] will lead us to another consistent scattering diagram $T_k(\mathfrak{D}_s)$ which is shown to be equivalent to $\mathfrak{D}_{\mu_k(s)}$.

We can similarly define the mutation for the $\mathscr{X}$ scattering diagrams from $\mathscr{A}_{\text{prin}}$. For the scattering diagram of type A_2 in Figure 2, we can obtain the mutation process for the $\mathscr{X}$ scattering diagram as in Figure 3 and Figure 4.

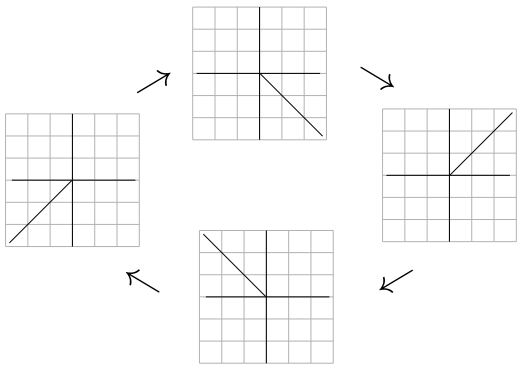


Fig. 3: Mutation of the $\mathscr{X}$ scattering diagram of type A_2 starting at the index 1

Note that the scattering diagrams are determined by seeds while the mutation of seeds are given by blow ups and blow downs of toric varieties. Thus the mutation of scattering diagrams actually represents this procedure of blowups and blowdowns.

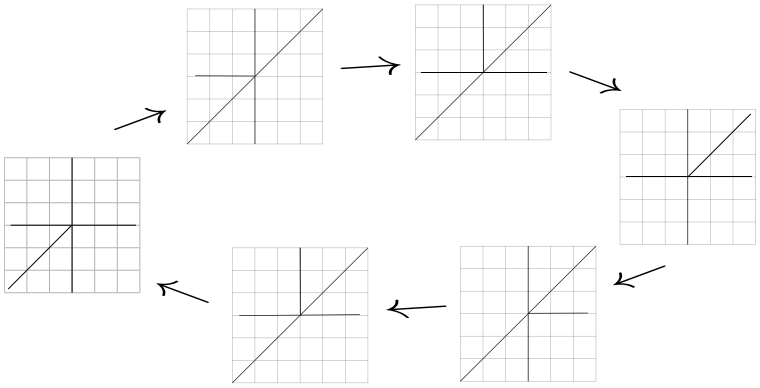


Fig. 4: Mutation of the $\mathscr{X}$ scattering diagram of type A_2 starting at the index 2

We are going to discuss a similar construction in Section 3 with a symplectic perspective.

Theta functions and the canonical algebras

Theta functions give the generators of the canonical basis of the cluster algebras. Given the cluster variety $V = \mathscr{A}, \mathscr{A}_{\text{prin}}, \mathscr{X}$, the corresponding character lattice is $L = M^\circ, \widetilde{M}^\circ$, or N. A theta function ϑ_p is associated to each point $p \in L$ by a combinatorial object – broken lines which are piecewise linear paths in $L_{\mathbb{R}}$ together with decorating monomials at each linear segment.

The free module generated by theta functions is endowed with an algebra structure from the multiplication between theta functions. Indeed the structure constants in the multiplications of theta functions can be given in terms of counting broken lines. The product of two theta functions can be expressed as

$$\vartheta_p \cdot \vartheta_q = \sum_r \alpha(p,q,r)\vartheta_r, \tag{2}$$

where the structure constants $\alpha(p,q,r)$ can be explicitly defined by counting broken lines with certain boundary conditions [17, Proposition 6.4]. In this finite type case, the structure constants α define ([17, Corollary 8.18]) the finitely generated $\mathbb{C}$-algebra structure on

$$\text{can}(V) := \bigoplus_{r \in L} \mathbb{C} \cdot \vartheta_r.$$

We will then define $X := \text{Spec}(\text{can}(V))$.

2.2 Positive polytopes

With the multiplication structure of the theta functions, we can now state the definition of a positive set– the property required for a set and its dilations to define a graded ring.

For $S \subseteq L_{\mathbb{R}} = L \otimes \mathbb{R}$ a closed subset, define the cone of S as

$$\mathbf{C}(S) = \overline{\{(p,r) \mid p \in rS, r \in \mathbb{R}_{\geq 0}\}} \subset L_{\mathbb{R}} \times \mathbb{R}_{\geq 0}.$$

Denote $dS(\mathbb{Z}) = \mathbf{C}(S) \cap (L \times \{d\})$ which is viewed as a subset of L.

A closed subset $S \subset L_{\mathbb{R}}$ is called *positive* if for any non-negative integers d_1, d_2, any $p_1 \in d_1 S(\mathbb{Z})$, $p_2 \in d_2 S(\mathbb{Z})$, and any $r \in L$ with $\alpha(p_1, p_2, r) \neq 0$, then $r \in (d_1 + d_2)S(\mathbb{Z})$.

In the ongoing example of cluster varieties of type A_2, we consider the polytope with vertices $(1,0),(0,1),(-1,0),(0,-1),(1,-1)$ as indicated in Figure 5. Note that this polytope is in the $\mathscr{X}$ diagram thus there is a flip from Figure 21. This polytope is indeed positive. In Section 3.1.3, there is a detail discussion of such a polytope in the $\mathscr{A}$ side. A similar calculation in this $\mathscr{X}$ case will still hold, thus this will correspond to the del Pezzo surface of degree 5 [17].

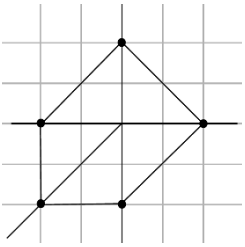


Fig. 5: Positive polytope of $\mathscr{X}$ cluster variety of type A_2.

We can apply the mutation sequences in Figures 3 and 4 to the polytope in Figure 5. Mutations of the polytopes as in Figures 6 and 7 will be obtained respectively. In the next section, we will describe mutations of the polytopes from a symplectic point of view. We observe that the mutation sequences of polytope in Figures 6 and 7 are the same as the sequences in Figures 23 and 24 respectively. The cluster mutation of the scattering diagrams comes from a change of seed data, i.e. a change of the initial variables. Thus the underlying spaces are all isomorphic.

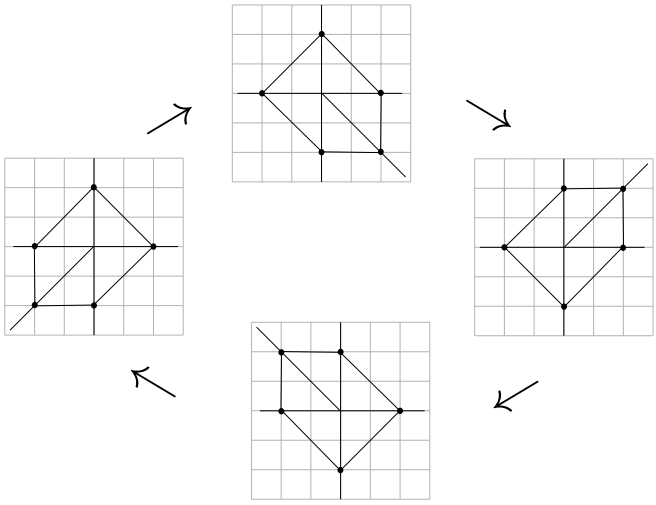


Fig. 6: Mutation of polytope starting from index 1

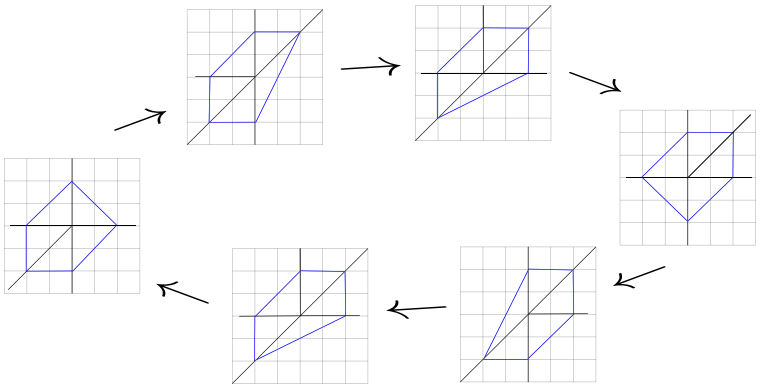


Fig. 7: Mutation of polytope starting from index 2

Compactifications from positive polytopes

We will roughly go over the geometric meaning behind the positive polytopes in this section. The motivation can be seen as the construction of projective toric varieties from the convex polytopes.

For rank 2 cluster varieties, since the $\mathscr{A}$ scattering diagrams are well defined, we can consider $\bar{S}$ the positive polytopes in the $\mathscr{A}$ scattering diagrams. For this set $\bar{S}$, define $\widetilde{S} = \bar{S} + N_{\mathbb{R}}$ which is obviously positive. Thus we can define the graded ring

$$\widetilde{R}_{\widetilde{S}} = \bigoplus_{d\geq 0} \bigoplus_{q\in d\widetilde{S}(\mathbb{Z})} \mathbb{C}\vartheta_q x^d \subset \mathrm{can}(\mathscr{A}_{\mathrm{prin}})[x],$$

with grading defined by x.

Define $Y_{\mathscr{A}_{\mathrm{prin}}} := \mathrm{Proj}(\widetilde{R}_S) \to T_M$. For the $\mathscr{A}$ variety, we take $Y_{\mathscr{A}}$ as the fiber over $e \in T_M$ in this map. More generally, for the $\mathscr{A}_t$ variety, $t \in T_M$, we can still take $Y_{\mathscr{A}_t}$ as the fiber over $t \in T_M$. Consider $X = \mathrm{Spec}(\mathrm{can}(V))$, for $V = \mathscr{A}_{\mathrm{prin}}, \mathscr{A}_t$, as in the previous subsection. Define $B = Y \setminus X$. Then [17] showed that, X is a Gorenstein scheme with trivial dualizing sheaf, in particular, for $V = \mathscr{A}_{\mathrm{prin}}, \mathscr{A}_t$, X is a K-trivial Gorenstein log canonical variety. In this finite rank 2 case, for $V = \mathscr{A}_{\mathrm{prin}}, \mathscr{A}$, $X \subseteq Y$ is a minimal model, i.e. Y is a projective normal variety, $B \subset Y$ is a reduced Weil divisor, $K_Y + B$ is trivial, and (Y, B) is log canonical.

For the case of the $\mathscr{X}$ varieties, as indicated in Section 2.1, the $\mathscr{X}$ varieties are quotients of the $\mathscr{A}_{\mathrm{prin}}$ varieties. Thus we will consider still consider $\mathscr{A}_{\mathrm{prin}}$ but instead see the lattice as $\widetilde{M}^\circ$ instead of $\widetilde{N}$ (which are actually isomorphic). We can repeat the same procedure as before and then obtain the compactification of $X = \mathrm{Spec}(\mathrm{can}(\mathscr{X}))$. The scheme X is also a K-trivial Gorenstein log canonical variety.

2.3 Canonical scattering diagrams

In the last section, we note that the cluster mutations of the scattering diagrams are not changing the underlying schemes. We are proposing another type of mutation which is given by the monodromy on B. We are going to understand the ideas behind from the mirror construction suggested by Gross, Hacking, and Keel in [15]. In Section 3.1, we will discuss the affine structure and monodromy from the SYZ perspective.

Consider a pair (Y, D), where Y is a smooth rational projective surface, and D is an anti-canonical cycle of projective lines. We will call such a pair a Looijenga pair. Let $X = Y \setminus D$. The tropicalization of (Y, D) is a pair (B, Σ), where B is an integral linear manifold with singularities, and Σ is a decomposition of B into cones. The pair (B, Σ) can be constructed by associating each node $p_{i,i+1}$ of D a rank two lattice with basis v_i, v_{i+1}. Denote the cone generated by v_i, v_{i+1} as $\sigma_{i,i+1} \subset M_{i,i+1} \otimes \mathbb{R}$. The cones $\sigma_{i,i+1}$ and $\sigma_{i-1,i}$ are glued over the ray $\rho_i = \mathbb{R}_{\geq 0} v_i$ to obtain a piecewise linear manifold B homeomorphic to $\mathbb{R}^2$ and $\Sigma = \{\sigma_{i,i+1}\} \cup \{\rho_i\} \cup \{0\}$.

The integral affine structure on $B_0 = B \setminus \{0\}$ can be defined by the charts

$$\psi_i : U_i = \mathrm{Int}(\sigma_{i-1,i} \cup \sigma_{i,i+1}) \to M_{\mathbb{R}},$$

where

$$\psi_i(v_{i-1}) = (1,0),\ \psi_i(v_i) = (0,1),\ \text{and} \quad \psi_i(v_{i+1}) = (-1, -D^2),$$

and ψ_i is linear on $\sigma_{i-1,i}$ and $\sigma_{i,i+1}$.

Now consider Y the del Pezzo surface of degree 5 and D the anti-canonical cycle of five (-1)-curves. The construction of the charts ψ will then give

$$\psi(v_1) = (1,0), \psi(v_2) = (0,1), \psi(v_3) = (-1,1), \psi(v_4) = (-1,0), \psi(v_5) = (0,-1).$$

Note however that having $\psi(v_4) = (-1,0), \psi(v_5) = (0,-1)$ will lead to

$$\psi(v_1) \rightsquigarrow (1,-1),\ \psi(v_2) \rightsquigarrow (1,0)$$

and this is NOT what we began with: $\psi(v_1) = (1,0)$, and $\psi(v_2) = (0,1)$. Thus we would like to identify the cone spanned by $(1,0)$ and $(0,1)$, and the cone spanned by $(-1,1)$ and $(1,0)$. This introduces the monodromy

$$(1,0) \mapsto (1,1), \quad (0,1) \mapsto (1,0),$$

to B_0. The affine structure is illustrated in Figure 8.

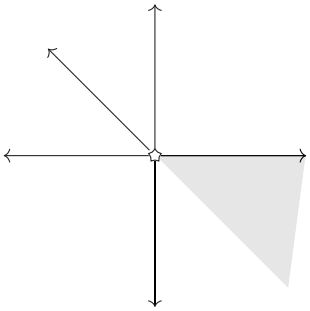


Fig. 8: The tropicalization (B,Σ) of the del Pezzo surface of degree 5

Now we would like to define the canonical scattering diagrams from (B,Σ). Rather than obtaining the diagrams by the algorithmic process with some initial data in [19] [21], the canonical scattering diagrams are defined via some Gromov-Witten type invariants. We will discuss the two types of diagrams are the 'same' later in the discussion about how to go from canonical scattering diagrams to cluster scattering diagrams. A *wall* [18] in B is a pair $(\mathfrak{d}, f_{\mathfrak{d}})$ where $\mathfrak{d} \subset \sigma_{i,i+1}$, for some i, is a ray generated by $av_i + bv_{i+1} \neq 0$, $a,b \in \mathbb{Z}$ relatively prime, and $f_{\mathfrak{d}} = 1 + \sum_{k\geq 1} c_k X_i^{-ak} X_{i+1}^{-bk} \in \mathbb{C}[[X_i^{-a} X_{i+1}^{-b}]]$ with some finiteness properties, and where c_k corresponds to the curve counting invariants. Note that the description of the wall functions $f_{\mathfrak{d}}$ indicates that all the wall are outgoing in the sense stated in the last section. Then the scattering diagrams are again the collections of walls. For example, the canonical scattering diagram associated to Figure 8 is shown in Figure 9.

Let $B(\mathbb{Z})$ be the set of points of B_0 with integral coordinates in an integral affine chart and $\{0\}$. Theta functions ϑ_q, $q \in B(\mathbb{Z})$, can similarly be defined on the scat-

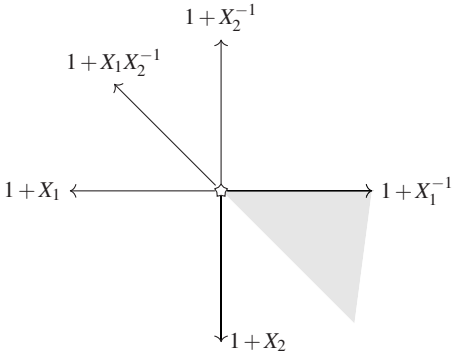


Fig. 9: Canonical scattering diagram

tering diagrams. The set of theta functions again generates an algebra structure [15] in terms of broken lines. In the finite case, we can simply consider $A = \oplus_{q\in B(\mathbb{Z})} \vartheta_q$.

Analogous to the setting in the cluster scattering diagrams, we can use the Rees construction to compactify the mirrors [20]. Positive polytopes with respect to the affine structures can be similarly defined to give a graded algebra. Using [25] or the argument in [8], the polytopes are broken line convex. In this case, since all the walls are outgoing, the positive polytopes are simply convex with respect to the affine structures.

Relation to the cluster scattering diagrams

In the case of Y a non-singular toric surface and $D = \partial Y$ the toric boundary of D, the affine structure on B extends across the origin. This identifies (B,Σ) with $M_{\mathbb{R}},\Sigma_Y$, where Σ_Y is a fan for Y.

Now given a Looijenga pair. Assume there is a toric model $p : (Y,D) \to (\bar{Y},\bar{D})$ which blows up distinct points x_{ij} on D_i. A *toric model* of (Y,D) is a birational morphism $(Y,D) \to (\bar{Y},\bar{D})$ to a smooth toric surface $\bar{Y}$ with its toric boundary $\bar{D}$ such that $D \to \bar{D}$ is an isomorphism. Consider the tropicalisation $(\bar{B},\bar{\Sigma})$ of $(\bar{Y},\bar{D})$. Thus $\bar{B} \cong M_{\mathbb{R}} = \mathbb{R}^2$ and $\bar{\Sigma}$ is the fan for $\bar{Y}$. Then there is a canonical piecewise linear map

$$\nu : B \to \bar{B}$$

which restricts to an integral affine isomorphism on the maximal cones in σ and $\bar{\Sigma}$. One can then define the scattering diagram $\mathfrak{D}$ as outlined in Section 2.1 or as in [15, Definition 3.21] for the more general setting. This step can be seen as 'pushing the singularities to infinity' or 'moving worms' [21]. By definition, the singularity of the affine structure is at $\{0\}$ as indicated in Figure 8. Then the singularity can be imagined to be pushed to the infinity of the two incoming walls. The map ν can

be extended to act on the canonical scattering diagram $\mathfrak{D}^{\text{can}}$. It is shown that [15] $\bar{\mathfrak{D}} = \nu(\mathfrak{D}^{\text{can}})$.

We have discussed in Section 2.1 that every cluster variety can be described as blow ups of a toric variety, which give the toric models for the cluster variety. Thus the cluster scattering diagrams can be seen as the diagrams arising from the canonical scattering diagrams by the map ν (as pushing singularities to infinity).

Mutation of polytopes according to the affine structures

One can imagine or with symplectic motivation as in Section 3.1, the 'pushing singularities to infinities' procedure is more general than just having singularities at the origin. For example, we can consider Figure 10 which we only push one of the singularities to infinity, resulting in a scattering diagram with one incoming wall. The

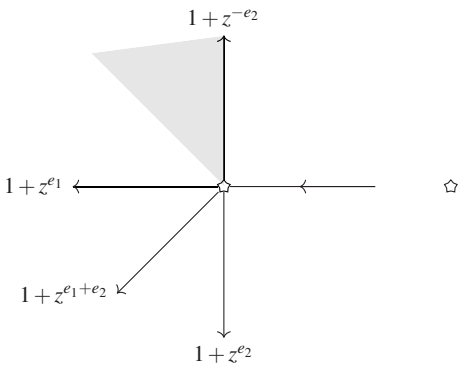


Fig. 10: Scattering diagrams with monodromy.

monodromy in Figure 10 is $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, or $(1,0) \mapsto (1,1)$, $(0,1) \mapsto (0,1)$. The polytope in Figure 5 with respect to this affine structure would then be of the form in Figure 11.

We can apply the sequence of mutations in Figure 3 to the polytope in Figure 11 and then obtain a new sequence of mutation polytopes (Figure 12). Putting the polytope in Figure 5 into the sequence (Figure 12) will get us the sequence Figure 26 which is motivated from the symplectic perspective.

The singularities can also be located on the walls instead of just at infinity or the origin. For example, one can obtain the canonical scattering diagram shown in Figure 13. Similar calculation shown in [6] indicates that the scattering diagram is consistent.

Note that the portions of the walls which go from the singularities to infinity are all outgoing. Thus using the idea in [8, Remark 6.2], we can consider convex sets in this affine structure. For example, one can construct the polytope as in Figure 14.

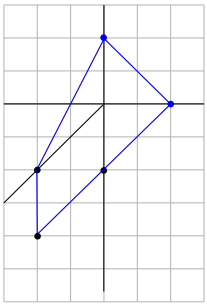


Fig. 11: Positive polytope with respect with the underlying affine structure

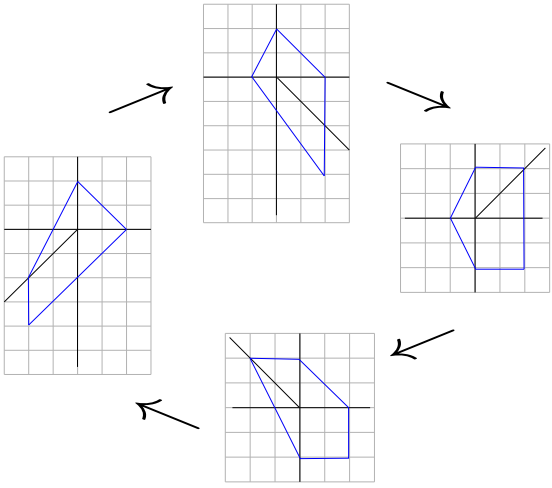


Fig. 12: Mutation of polytopes with monodromy

Applying the mutation sequence as in Figure 3, we obtain the sequence of polytopes described in Figure 15. Interestingly, this is the same sequence as in Figure 27 which is motivated from the symplectic perspective.

2.4 Type B_2 and G_2

The other types are similar and thus we only roughly go over the mutations of type B_2 and G_2. For type B_2, we will take the same skew-symmetric form with $d_1 = 1$, and $d_2 = 2$ as our fixed data. We can again take the initial seed as $\mathbf{s} = \{(1,0),(0,1)\}$. Then we will obtain the $\mathscr{A}$ and $\mathscr{X}$ scattering diagrams as in Figure 1 and 2. If we

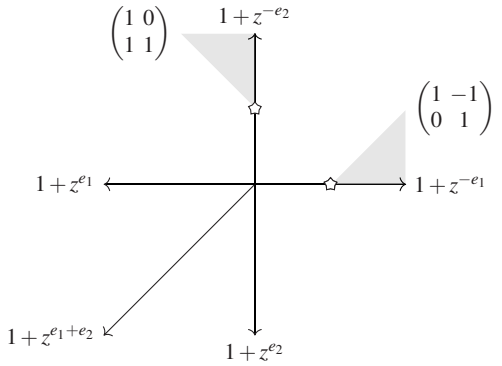


Fig. 13: Scattering diagram with monodromy on the walls.

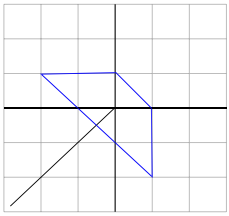


Fig. 14: Polytope lives in the affine structure indicated in Figure 13.

mutate at index 1 first, we can get the mutation of scattering diagrams very similar to the type A_2 case.

For type B_2, we can again take the primitive generators of the walls and then consider the polytope as the convex hull of those vertices. By using [8], this polytope is a positive polytope. The multiplication of the theta functions tells us [7] that the corresponding space is the del Pezzo surfaces of degree 6. The mutation sequence of the polytopes is described in Figure 16.

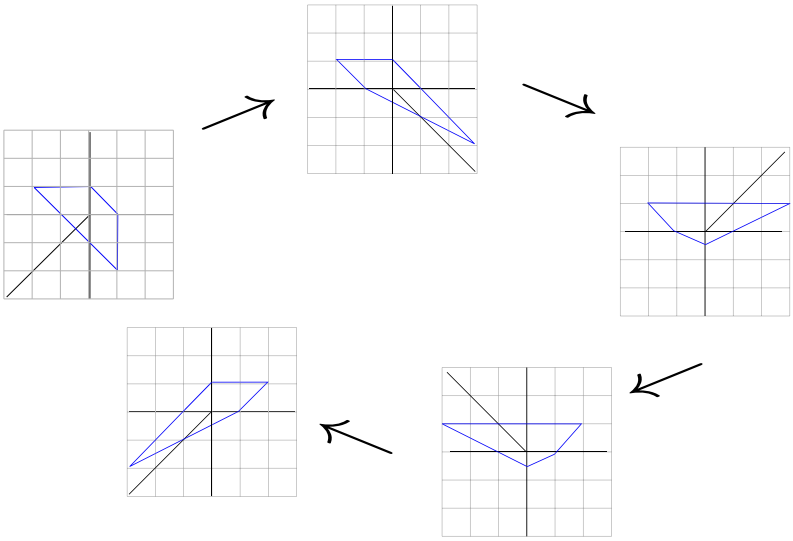


Fig. 15: Mutation for the polytope in Figure 14

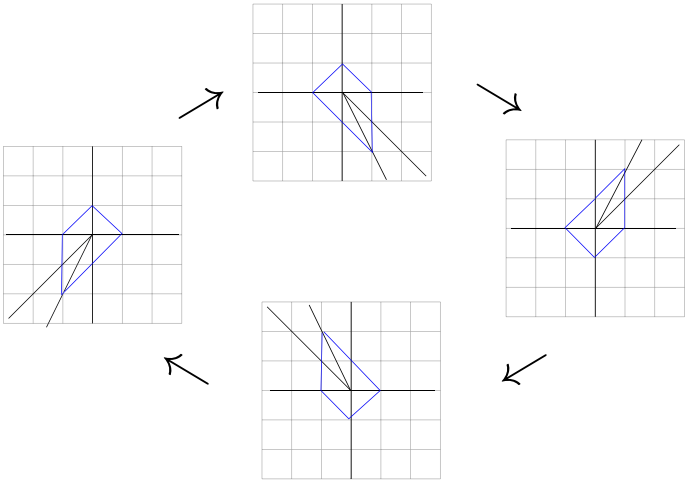


Fig. 16: Mutation of polytope for type B_2

One may want to repeat the same trick on the type G_2. The sad fact is that the if we are taking the convex hull of the primitive generators of the walls, the resulting polytope would no longer be positive. This is because the polytope is no longer broken line convex as indicated in [8]. Since we only care about the incoming walls

for broken line convexity [8], one can see that the top left polytope in Figure 29 in the next section is broken line convex. The mutation sequence of the G_2 scattering diagrams are similar to those for type A_2 and B_2. Without duplicating, one can see that the mutation indicate in Figure 29 is in fact the cluster mutation of scattering diagrams. Thus the mutation of polytopes follows correspondingly. This again tells us that the mutation sequences of the polytope with the algebro-geometric and symplectic viewpoints coincide.

3 Mutations in symplectic geometry

We begin this section giving a perspective on understanding cluster varieties, as well as scattering diagrams, as a way of building mirrors under SYZ [31] T-duality. In particular, we explain how scattering diagrams can be related to almost toric fibrations. Later, we explain how one can see compactifications of 2-dimensional cluster varieties into del Pezzo surfaces from an almost toric fibration perspective.

3.1 Cluster varieties and Mirror Symmetry

In this section we will sketch how to relate a scattering diagram data (described in Section 2), for constructing a log-CY variety X, with the base of an almost-toric fibration (ATF), describing a SYZ [31] singular Lagrangian fibration of (X, ω), with respect to a Kähler form ω in X.

3.1.1 Almost Toric Fibrations

Informally speaking an *almost toric fibrations* (ATF) in a symplectic 4-manifold X is a smooth map to a two dimensional base B, whose regular fibres are Lagrangian tori, whose allowed singular fibres are of three kinds:

- point (toric - rank 0 elliptic) – locally equivalent to the moment map at the origin in $\mathbb{C}^2$ with the standard toric action, $(e^{i\theta_1}, e^{i\theta_2}) \cdot (x, y) = (e^{i\theta_1}x, e^{i\theta_2}y)$;
- circle (toric - rank 1 elliptic) – locally equivalent to $S^1 \times \{0\} \subset \mathbb{C}^* \times \mathbb{C}$ with the standard toric action;
- nodal (a pinched torus) – with some local model described for the singular point. [See [32, 22] for precise definition, and see Section 3.1.2 for a local model of the nodal fibre.]

The toric singularities appear on the boundary of the base, while the nodal singularities project into the interior. For a precise definition of ATFs see [32].

Away from the singular fibres, by the Arnold-Liouville theorem [3], X admits locally action angle coordinates $(p_1, p_2, \theta_1, \theta_2)$ and the fibration is locally equivalent

to $(p_1,p_2,\theta_1,\theta_2)\mapsto(p_1,p_2)$, in other words, away from singular fibres X equivalent to T^*B/Λ^*, for some lattice Λ^*. Hence, B carries a natural dual lattice $\Lambda\subset TB$. The lattice has monodromy as we go around the nodal fibre, which is a shear in the direction dual to the collapsing cycle of that nodal fibre. Locally, the coordinates (p_1,p_2), can be thought as the *flux* $\mathfrak{f}\in H^1(T^2,\mathbb{R})$ relative to the Lagrangian fibre associated with $(0,0)$. The flux $\mathfrak{f}(\gamma)$ measures the symplectic area of a cylinder swept by a cycle $\gamma\in H_1(T^2,\mathbb{Z})$ as we move in a path of Lagrangian fibres connecting $(0,0)$ to (p_1,p_2). [See, for instance, [30] for a more complete understanding of flux in ATFs.]

So, in practice, we visualise the base minus a set of cuts (one for each nodal fibre) affinely embedded into $\mathbb{R}^2$ endowed with the standard affine structure. We call them almost-toric base diagrams (ATBDs) representing the ATF. The same ATF can be represented by different ATBDs, by changing the set of cuts.

Figure 17 shows the base diagram of 3 different ATFs in $\mathbb{C}^2$; the right-most diagrams on Figure 18 are different diagrams representing the same ATF in $\mathbb{C}^2\setminus\{xy=1\}$, related by a change of cut; Figures 23–33 contain examples of ATBDs in closed 4 manifolds. In these diagrams, the crosses represent the nodal fibres, the dashed lines the cuts, the edges the rank 1 and the dots rank 0 toric singularities.

Remark 1. We expect the above mentioned ATFs to be realisable as a special Lagrangian fibration in the complement of a complex divisor projecting to the boundary of the ATF, with respect to a holomorphic volume form with poles on these divisor. This is true for the fibration presented in Section 3.1.2, but we will avoid talking about the "special" condition.

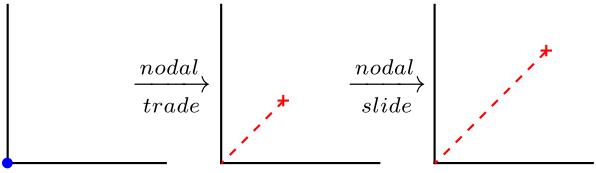


Fig. 17: Nodal trade and nodal slide operations in ATFs.

There are two ways of modifying ATFs within the same symplectic manifold X, known as nodal trade and nodal slide [32]. The diagrams in Figure 17 illustrate the change of the ATBDs after a nodal trade and a nodal slide. In del Pezzo surfaces, the monotone symplectic 4-manifolds, we defined mutation of an ATBD, the process of sliding one nodal fiber through the monotone fibre, and then redrawing the diagram by changing the direction of the cut used to slide [33, 34]. In the end, the ATBD mutates by slicing it in the direction of the cut and applying the inverse of the corresponding monodromy, which is a shear in the primitive direction associated to the cut. This transformation is the same polytope mutation as in [1, 2], and completely analogous to the mutation of seeds, and scattering diagram we will discuss later. We can extend this notion of symplectic mutation to exact almost toric manifold, for

instance, the complement of an anti-canonical divisor in a del Pezzo. In this case, the mutation corresponds to sliding a nodal fibre through the exact torus and then transferring the associated cut to the opposite side.

3.1.2 Local model for nodal fibre and wall-crossing

We briefly recall ATF presented in [4, Section 5], [5, Section 3.1.1]. This ATF appeared before in [10] and also in [14, Example 1.2], where it was shown to be a special Lagrangian fibration [with respect to certain holomorphic volume form]. We consider $X^\vee = \mathbb{C}^2 \setminus \{xy = 1\}$, with $\omega^\vee = \frac{i}{2}(dx \wedge d\bar{x} + dy \wedge d\bar{y})$ the standard symplectic form. Using $f : X^\vee \to \mathbb{C} \setminus \{1\}$, $f(x,y) = xy$, Auroux builds an ATF by parallel transport of orbits of the S^1 action $e^{i\theta} \cdot (x,y) = (e^{i\theta}x, e^{-i\theta}y)$, over circles in the base of f centred at 1. One then gets Lagrangian torus fibres, parametrised by $(r,\lambda) \in \mathbb{R}_{>0} \times \mathbb{R}$, as:

$$T_{r,\lambda} = \{(x,y) \in \mathbb{C}^2; r = |xy-1|, \lambda = |x|^2 - |y|^2\}.$$

Note that there is a nodal fibre $T_{1,0}$, that contains $(0,0)$, the fixed point of the S^1 action.

This almost toric fibration can be represented by applying a nodal trade to the standard toric fibration of $\mathbb{C}^2$, replacing the boundary divisor $\{xy = 0\}$ with the smooth divisor $\{xy = 1\}$, and then deleting this divisor living over the boundary of the base, as illustrated by Figure 18. Indeed, replacing the role of 1 by 0 in the above fibration, i.e., considering parallel transport over circles concentric at 0 (considering $r = |xy|$) one obtain precisely the standard toric fibration of $\mathbb{C}^2$. So, considering analogous fibrations by changing 0 to 1 in the definition of r constitutes a nodal trade, and moreover, varying the value of $c \in \mathbb{R}_{>0}$ in the definition of $r = |xy - c|$ provides different fibrations related by nodal slides.

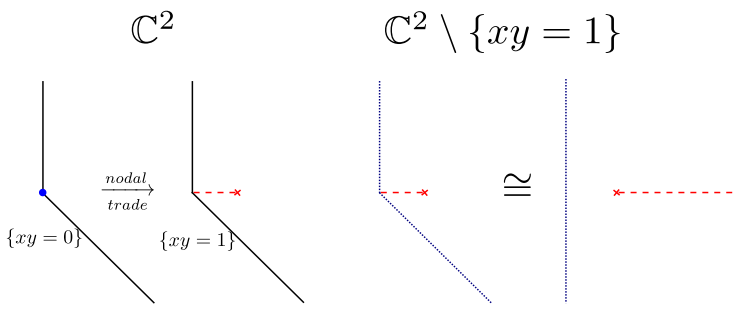


Fig. 18: Nodal trade and an ATF for the complement of a conic. The left diagrams are a shear by $(0,-1)$ of the diagrams in Figure 17. The rightmost diagrams, represent the same ATF, and differ by changing the direction of the cut.

Wall-crossing

Dualising this torus fibration, one gets the mirror variety X_Λ of $X^\vee$, over the Novikov field $\Lambda = \{\sum_{i=0}^{n} a_i T^{\sigma_i}; a_i \in \mathbb{C}, \sigma_i \in \mathbb{R}, \lim_i \sigma_i = \infty\}$, which is the moduli of almost-toric fibres (special Lagrangians), endowed with unitary Λ^*-local systems. We will be able to relate the valuation $\mathrm{val}(u) = \min\{\sigma_i; a_i \neq 0\}$ of an element $u \in \Lambda$ with the above mentioned flux, whenever $\mathrm{val}(u)$ measures the symplectic area of a disk with boundary in a varying family of the Lagrangian torus fibres. [Notation: $\Lambda_0 = \{u \in \Lambda; \mathrm{val}(u) \geq 0\}$, $\Lambda_+ = \{u \in \Lambda; \mathrm{val}(u) > 0\}$, $\Lambda^* = \{a_0 + \Lambda_+; a_0 \in \mathbb{C}^*\}$.] We will later consider the mirror of $X^\vee$ as $X = X_\mathbb{C}$ over $\mathbb{C}$, by replacing T with e^{-1}.

So we replace the Lagrangian fibre T^2, by the dual Λ-torus of unitary local systems $\hom(\pi_1(T^2); \Lambda^*) \cong (\Lambda^*)^2$. Locally identifying each relative class, $\beta \in \pi_2(\mathbb{C}, T_{r,\lambda})$, Auroux defined a function $z_\beta : X_\Lambda \to \Lambda$, for a local system ∇ in $T_{r,\lambda}$, $z_\beta(\nabla) = T^{\omega^\vee(\beta)} \nabla \cdot \partial\beta$. Choosing a basis $\{\alpha, \beta\}$ of $\pi_2(\mathbb{C}, T_{r,\lambda})$, one gets that $w := z_\alpha$, $u := z_\beta$, define local coordinates of X_Λ. After that, the idea is to define a superpotential function $W : X_\Lambda \to \Lambda_+$, which is locally defined as $W(u,w)$, and whose monomials encode the relative Gromov-Witten count of Maslov index 2 holomorphic disks in $\mathbb{C}$ with boundary on the torus fibre endowed with the respective local system determined by (u,w). [The pair (X_Λ, W) is called the Landau-Ginzburg model that is mirror dual to $\mathbb{C}$ with respect to the divisor $D = \{xy - 1\}$. We refer the reader to [4, 5] for details on mirror symmetry in the complement of divisors.]

The issue is that, in the naive definition of the mirror, the superpotential W is discontinuous. This is due to the presence of fibres $T_{1,\lambda}$, $\lambda \neq 0$, which bounds Maslov index 0 holomorphic disks. Let's denote the relative class represented by this Maslov 0 disks by α for $\lambda < 0$, and $-\alpha$ for $\lambda > 0$. In [5, Section 3.1.1], it is shown that for $r < 1$, the fibres $T_{r,\lambda}$ (called Chekanov type) bound one holomorphic disk, in a class we name β. So $W(u,w) = u$, for these fibres. The fibres $T_{r,\lambda}$, for $r > 1$, (called Clifford type) bound 2 holomorphic disks in relative classes β_1, β_2, and hence the superpotential is of the form $W(z_1, z_2) = z_1 + z_2$, where $z_i = z_{\beta_i}$.

We see in [5, Section 3.1.1] that as r approaches 1, from $r > 1$, we get $\alpha = \beta_1 - \beta_2$ (hence $w = z_1 z_2^{-1}$). Moreover, if we cross the wall at $\lambda < 0$, the class β is naturally identified with β_2, and if we cross the wall at $\lambda > 0$, the class β is naturally identified with $\beta_1 = \beta_2 + \alpha$ [which is not so surprising, as the monodromy around the nodal fibre in the ATF would fix $\partial\alpha$ and maps $\partial\beta \to \partial\beta + \partial\alpha$]. So the superpotential W should be corrected by the term $(1 + w^{\pm 1})$, representing the fact that the holomorphic disk on class β, would not only survive past the wall, but the superpotential would also acquire a holomorphic disk in class $\beta \pm \alpha$, coming from the gluing of the Maslov 2 holomorphic disk on class β with the Maslov 0 holomorphic disk on class $\pm\alpha$. Then, instead of u becoming z_2 as we cross over $\lambda > 0$, we should correct it to become $u = z_2(1+w) = z_2 + z_1$, and instead of u becoming z_1 as we cross over $\lambda < 0$, we should correct it to become $u = z_1(1 + w^{-1}) = z_1 + z_2$, and, thus, ensuring the continuity of W.

By naming $v = z_2^{-1}$, so $z_1 = v^{-1}w$, we get the corrected $u = v^{-1}(1+w) = v^{-1}w(1+w^{-1})$. We see that the corrected (and completed) mirror X_Λ, is given by

$$X_\Lambda = \{(u,v,w) \in \Lambda^2 \times (\Lambda \setminus \{0\}); uv = 1+w\}.$$

Figure 19 below describes the *mirror SYZ fibrations* on $X^\vee$ and X_Λ. We list several remarks about the diagrams in Figure 19 and the mirror X_Λ.

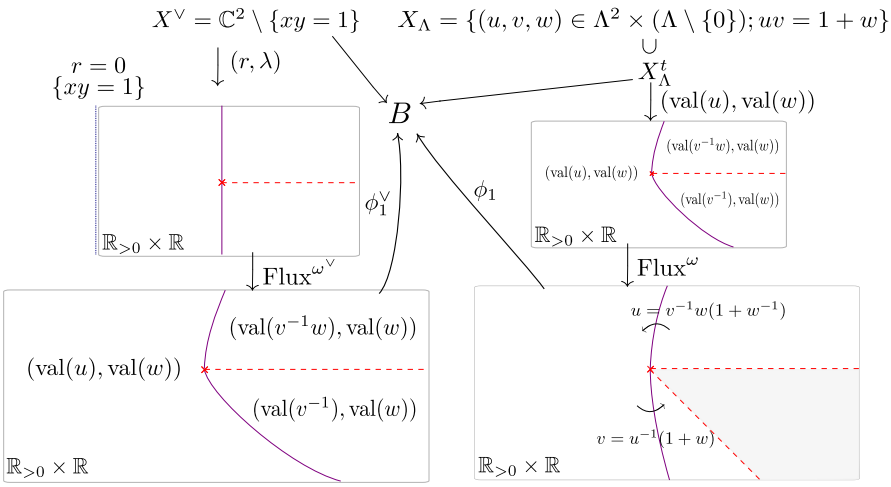


Fig. 19: SYZ fibrations for the complement of a conic, which is self-mirror, when considering $X_\mathbb{C}$.

Remark 2. We see that X_Λ is given by gluing two torus charts $(u,w) \in (\Lambda \setminus \{0\})^2$, and $(v,w) \in (\Lambda \setminus \{0\})^2$, by a rational map defined in the complement of $\{w = -1\}$. So, for instance, the case $u = 0$ would be realised as $(v,-1)$ in the (v,w)-chart.

Remark 3. Considering the symplectic form $\omega^\vee$ in $X^\vee$, the Lagrangian torus fibration would be viewed in a truncated part of the (u,w)-chart, with $0 < \mathrm{val}(u) \leq a(\mathrm{val}(w))$ or in the (v,w)-chart, with $a(\mathrm{val}(w)) \leq \mathrm{val}(v^{-1}) < \infty$ for $\mathrm{val}(w) < 0$, for instance. These bounds on the valuation would give us X^t_Λ, a truncated version of the mirror X_Λ. But in symplectic geometry, we can add to $(X^\vee, \omega^\vee)$ a contact boundary $\partial X^\vee$, and it is most natural to consider a completion procedure called the symplectization of $X^\vee$ with respect to this boundary. This endows $X^\vee$ with a different symplectic form $\omega^\vee_S$. It is equivalent to consider an infinite inflation of $(\mathbb{C}^2, \omega^\vee)$ with respect to the divisor $D = \{xy = 1\}$. In this limit we would have $\mathrm{val}(u) \to \infty$, and we would get the completed mirror X_Λ.

Remark 4. The expectation regarding the correspondence between the count of Maslov index 2 disks with boundary on a SYZ fibre and its tropical counterpart was proven in [24] for a SYZ fibration on the complement of an smooth anti-canonical divisor in a del Pezzo surface. More precisely, given a del Pezzo surface Y and a smooth anti-canonical divisor D, there exists a special Lagrangian fibration on $Y \setminus D$ with respect to the complete Ricci-flat Tian-Yau metric [9]. To understand the

Landau-Ginzburg superpotential of Y, there exists a sequence of Kähler forms ω_i on Y converging to the Tian-Yau metric pointwisely with $\int_Y \omega_i^2 \to \infty$ [24, Lemma 2.4]. Thus, these superpotentials of the special Lagrangian fibres can be defined with respect to ω_i, $i \gg 0$ and the superpotentials coincide with the tropical counterpart [24, Theorem 5.19]. This gives a geometric explanation of the renormalization procedure of taking valuation going to infinity.

Remark 5. In the (r,λ) projection of Figure 19, the singular fibre in position $(1,0)$ is depicted by an $\times$, and the wall of fibres with $r=1$ that bound Maslov index 0 disks are represented by a line. This (r,λ) coordinate does not respect the natural affine structure on the complement of the singular fibre of B. We instead consider $\mathrm{Flux}^{\omega^\vee}$, the flux with respect to a limiting fibre lying over $(0,0)$, in the next diagram. This map is then continuous, but not differentiable over the dashed ray $r \geq 1$, $\lambda = 0$, which we call the cut. Moreover, this composition represents the map to the base diagram depicted in the rightmost picture of Figure 18. The map $\phi_1^\vee$ is then an affine isomorphism to B minus the cut. The affine structure of B minus the node, is described by the gluing of the chart $\phi_1^\vee$ and a chart $\phi_2^\vee$, going from the third diagram of Figure 18, corresponding to taking the cut associated to $0 < r \leq 1$, $\lambda = 0$. The symplectic manifold $X^\vee$ can be thought then as a local model for gluing in the nodal fibre to the manifold constructed from gluing the Lagrangian torus fibres associated to $\phi_1^\vee$ and $\phi_2^\vee$. This is essentially the same model as the description of $X^\vee$ as a self-plumbing of T^*S^2 given in [32, Section 4.2].

Remark 6. The affine structure on B for the dual mirror fibration $X_\Lambda \to B$ is endowed with the dual affine structure in the complement of the node. We call the map that adjusts this affine structure in $\mathbb{R}^2$, Flux^ω. In the case we take the SYZ mirror $X_\mathbb{C}$ over $\mathbb{C}$ (by replacing T by e^{-1}), it endows a symplectic form ω as described in [5, Proposition 2.3]. In this case, Flux^ω becomes the actual flux with respect to this symplectic form.

Remark 7. The monodromy around the singular fibre of $X^\vee \to B$, represented by the bottom left diagram of Figure 19, is given by $M^{\pm 1}$, for $M = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$, fixing the cut $(1,0)$. Then, the monodromy around the node for the rightmost diagram representing $X_\Lambda \to B$ is given by $(M^T)^{\mp 1}$, with $(M^T)^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. We see that this fixes the coordinate w, associated to $(0,1)$, which we then name $\vartheta_{(0,1)}$, and it sends the coordinate $v^{-1} = \vartheta_{(1,0)}$ to $v^{-1}w = \vartheta_{(1,1)}$.

Remark 8. As described by Mikhalkin [27], we can deform the complex structure on $X^\vee$ to a limit where holomorphic curves would converge to tropical curve on the base with respect to the so-called complex affine structure, which is dual to the symplectic affine structure. So, (relative) Gromov-Witten invariants of $X^\vee$ are expected to be described by tropical curves (ϑ functions) in the base B, with the affine structure describing X_Λ (or $X_\mathbb{C}$). In particular, the wall becomes straight in this limit, as illustrated by Figure 20.

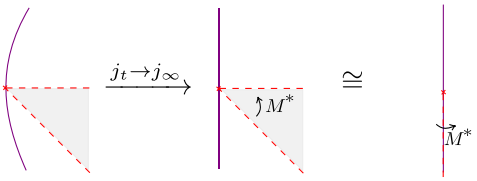


Fig. 20: In a complex structure limit, the wall becomes straight. We can move the cut to the invariant direction of the monodromy, that is the same direction as the limit straight wall.

As we mentioned before, we will now replace the formal variable T by e^{-1} in our construction, and consider the mirror as the moduli of Lagrangian fibres endowed with $U(1)$-local systems. So, after completion the mirror becomes

$$X = X_{\mathbb{C}} = \{(u,v,w) \in \mathbb{C}^2 \times \mathbb{C}^*; uv = 1+w\}$$

endowed with a completed symplectic form ω as described in [5, Proposition 2.3]. Its SYZ dual ATF (dual to the one on $X^\vee$) is then described by any of the diagrams in Figure 20.

Remark 9. As we take the completion, the val defined in X_Λ approaches $-\log|.|$ defined in $X_{\mathbb{C}}$.

We see now that the rightmost diagram in Figure 20, can describe an ATF, and once decorated with wall crossing functions [$(1+w^{\pm 1})$ accordingly] along the wall, it can algebraically determine the space $X_{\mathbb{C}}$. The variety $X_{\mathbb{C}}$ is then built out of two $(\mathbb{C}^*)^2$ charts, with coordinates (u,w) and (v,w), glued together in a cluster like transition birational map $uv = 1+w$, defined in the complement of $\{w=-1\}$. A unique wall, decorated with such wall crossing function, describing $X_{\mathbb{C}}$ is the simplest version of a scattering diagram [19, 15] (see Section 2 for more details).

3.1.3 The A_2 Cluster Variety: ATF and Scattering Diagram

As described in [17, Example 8.40] (taking the parameters X_1, X_2 to be 1), we describe the affine ($\vartheta_0 = 1$) A_2 Cluster variety by the ring in five variables ϑ_i, $i = 1,\dots,5$ satisfying the relations:

$$\vartheta_1\vartheta_3 = 1+\vartheta_2$$
$$\vartheta_2\vartheta_4 = 1+\vartheta_3$$
$$\vartheta_3\vartheta_5 = 1+\vartheta_4$$
$$\vartheta_4\vartheta_1 = 1+\vartheta_5$$
$$\vartheta_5\vartheta_2 = 1+\vartheta_1$$

We see that this variety is obtained by gluing five algebraic tori $(\mathbb{C}^*)^2$, with coordinates $(\vartheta_i, \vartheta_{i+1})$ [indices taken mod 5], according to the above cluster relations. We saw in more details in Section 2 that these relations are encoded by the data of a scattering diagram, as illustrated in the top-right picture of Figure 21.

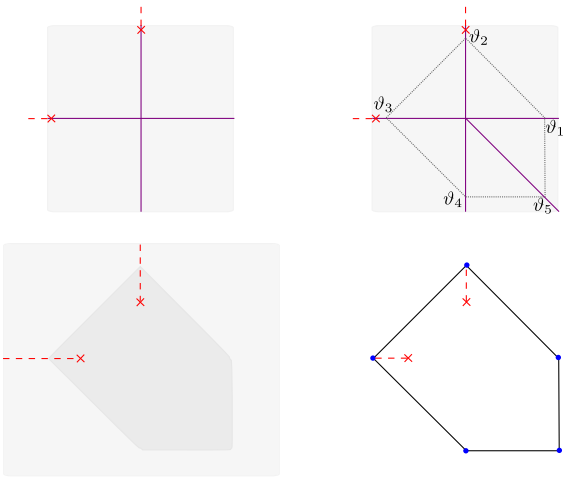


Fig. 21: Scattering diagram and ATF for the A_2 cluster variety.

Let's start with the data of an ATF describing a symplectic manifold X, with 2 nodal fibres, whose monodromies are encoded by cuts pointing away from the nodes in the directions $(0,1)$ and $(-1,0)$, respectively, as illustrated in the bottom-left picture of Figure 21. We now think think of this as endowed with the completed infinite volume symplectic form, so the base diagram covers the whole $\mathbb{R}^2$. As indicated in the previous Section, to build complex charts on this space, we add one wall for each node, represented by a line in the invariant direction of the monodromy. [These walls represent dual fibres in the mirror $X^\vee$, bounding Maslov index 0 disks with respect to a limit complex structure j_∞.] We call the chamber containing the nodes the main chamber, and we associate to it a complex torus $(\mathbb{C}^*)^2$, with coordinates $(\vartheta_2, \vartheta_3)$, and associated with the corresponding wall is a gluing function of the form $(1+\vartheta_i)$. One can check that changing coordinates around these 4 walls in a full circle, does not give you identity on the $(\vartheta_2, \vartheta_3)$ algebraic torus. To correct for that one needs to add an extra *slab*, in this case corresponding to a ray in direction $(1,-1)$, and a corresponding transition function giving you now 5 chambers, each corresponding to an algebraic torus, as illustrated in the top-right picture of Figure 21. This collection of walls and slabs is called the scattering diagram [19] [recall the details in Section 2]. This scattering diagram describe the relations of the A_2 cluster variety given in the beginning of this Section.

We can compactify this A_2 cluster variety by homogenizing its defining equations, as $\vartheta_1\vartheta_3 = \vartheta_0^2 + \vartheta_2\vartheta_0, \ldots, \vartheta_5\vartheta_2 = \vartheta_0^2 + \vartheta_1\vartheta_0$. As mentioned in [17, Exam-

ple 8.40], this gives a del Pezzo surface of degree 5 in $\mathbb{C}P^5$. Intersecting the hyperplane $\vartheta_0 = 0$, we see a chained loop of 5 divisors. Symplectically, it is natural to endow the del Pezzo surface with the monotone symplectic form given by restricting the Fubini-Study form of $\mathbb{C}P^5$. The complement of the five above mentioned divisors can be seen as a (Weinstein) subdomain of X, whose completion give X. Indeed, there is an ATF on the degree 5 del Pezzo surface, as illustrated in the bottom-right diagram of Figure 21. [We can obtain this ATF by performing a monotone blowup in a corner of [34, diagram (A_3) of Figure 16].] The chained loop of 5 divisors is identified with the boundary of this ATF, and the complement of them is a subdomain of X as illustrated by the bottom-left diagram of Figure 21.

3.2 Compactifications of Cluster varieties

We saw in the previous section how to relate the data representing an almost-toric fibration in a open symplectic manifold with a set of initial walls, out of which Gross-Siebert [19] explains how to complete to a scattering diagram that provides this manifold with complex charts given by gluing algebraic tori $(\mathbb{C}^*)^2$ along the walls.

As mentioned in Section 2.1, we can construct cluster varieties out of this data, and we will focus on the varieties of finite type A_2, B_2, G_2. These are open exact almost toric manifolds, built out of the scattering diagram with initial data given by two orthogonal walls, one of them associated to one node and the other with one, two and three nodes, respectively, as indicated in Figure 22. Recall we call the chart containing all nodes the main chart. One sees that symplectic mutation can be associated to changing the main chart, as illustrated in Figure 22. In other words, without the prior knowledge, the scattering diagram can be recovered by keeping track of the "main charts" as we apply the corresponding mutations, as illustrated in Figure 22. [This is not the case when the scattering diagram has a dense regions of slabs. For instance, when considering the scattering diagram associated to the mirror of the complement of an elliptic curve in $\mathbb{CP}^2$. Note that this case is considered in [24].]

In this section, we are interested in understanding compactifications of these cluster varieties from the symplectic perspective. Compact symplectic manifolds have finite volume, hence we will consider as X a subdomain, whose completion is the manifold described by the ATF with base diagram covering the whole $\mathbb{R}^2$. We will consider equivalent the subdomains with same completion.

All the symplectic compactifications considered here are symplectic del Pezzos, in the sense that they are endowed with a monotone symplectic form, which is unique up to scaling and symplectomorphisms [26, 23, 28, 29]. This ensures the existence of a monotone fibre, that can be detected by the intersection point of the lines in the diagrams that go through the nodes and are in the direction of the cuts. A symplectomorphism class invariant of these monotone fibres (the star-shape) is shown [30] to be given by the interior of the polytope seen in $H^1(T^2, \mathbb{R}) \cong \mathbb{R}^2$,

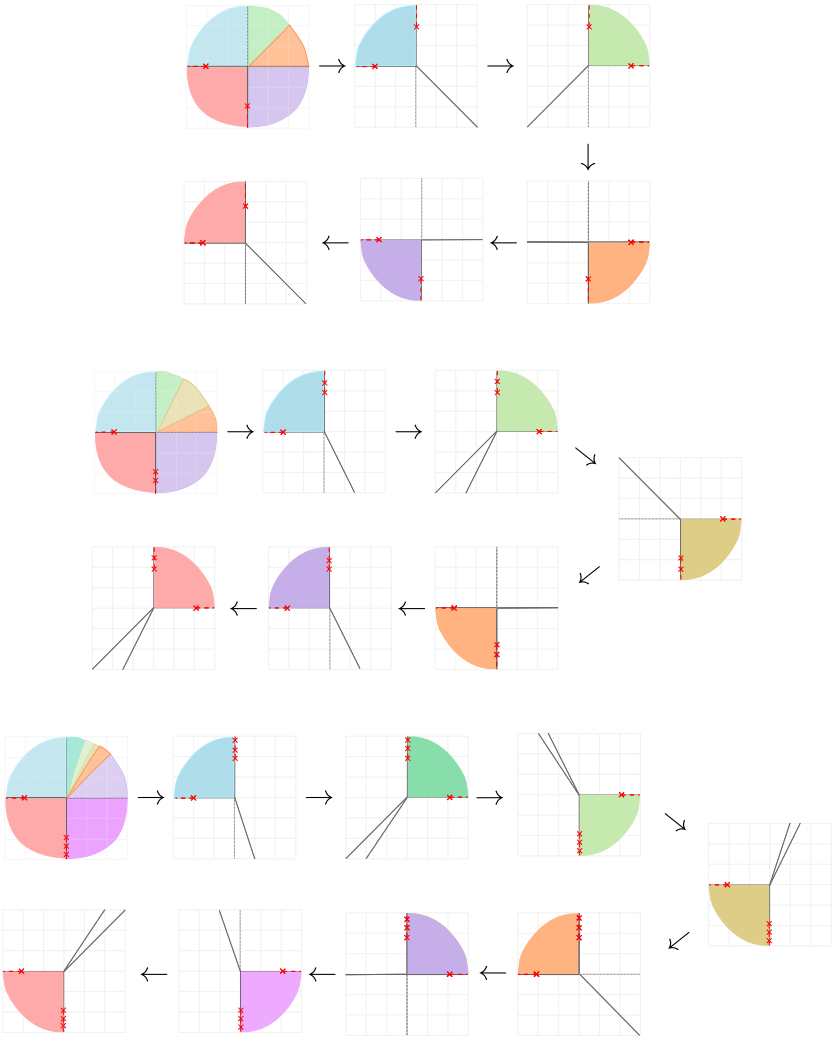


Fig. 22: Cluster charts via symplectic mutations on affine cluster varieties

as we forget the nodes and cuts. So there is a symplectomorphism identifying two monotone fibres of an ATF, if and only if, the associated polytopes are related under $\mathrm{SL}(2;\mathbb{Z})$. If there exists such ambient symplectomorphism, we say the Lagrangians are symplectomorphic.

In the definition of mutation of ATFs on del Pezzo surface [34], besides mutating the polytope by changing the direction of the cut, it is required that we slide the cut through the monotone fibre. In that sense, we say that the corresponding monotone fibres are related by mutation. We can then form a graph with vertices representing

symplectomorphism class of a Lagrangian and edges represented these Lagrangians being related by mutation. One aspect we can extract from the cyclic behaviour of these finite cluster varieties is the existence of cycles in the above mentioned graph. This behaviour does not appear in mutations of monotone almost toric fibres in $\mathbb{C}P^2$, and conjecturally in $\mathbb{C}P^1 \times \mathbb{C}P^1$.

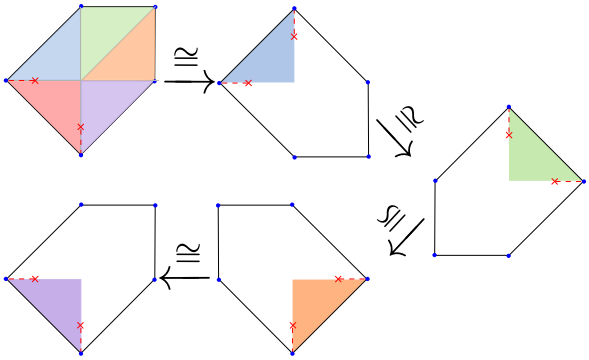


Fig. 23: Mutations on degree 5 del Pezzo – 1 torus

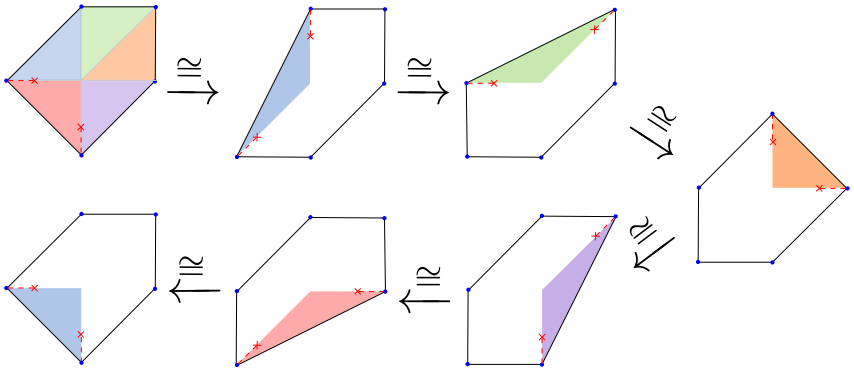


Fig. 24: Mutations on degree 5 del Pezzo – 1 torus

Let us start looking at the example from Section 3.1.3 ([17, Example 8.40], [8, Figure 1]), a compactification of the A_2 cluster variety to the degree 5 del Pezzo, by adding a chain of 5 divisors, whose union represents the anti-canonical class. This compactification and its mutations are illustrated in Figures 23, 24. Note that we get the same pattern as in Figures 6, 7, where we get back to the same picture after, respectively, 4 and 6 cycles, depending on the pattern of mutation. This is misleading, as ATFs, the mutations should not depend on which half-space is fixed, and which you decide to shear. In fact, in this example, all diagrams are $(SL(2;\mathbb{Z}))$ equiva-

lent, which in particular implies that the monotone tori in each pictures are mutually symplectomorphic. The 5-cycle pattern of the A_2 cluster appears by looking at the main charts, which we have already illustrated in Figure 22. In particular, this example does not give us a cycle of monotone Lagrangian tori, since we quotient out the graph associated to mutations by equivalence.

We want to extend a bit our notion of compactification. We will say that a (Weinstein) domain X compactifies to $\overline{Y}$, if we have $X \subset Y \subset \overline{Y}$, with X a sub-domain of Y and $Y = \overline{Y} \setminus \bigcup_i D_i$, for symplectic divisors D_i. This will be used by us to identify our domains of interest, described in Figure 22, appearing as open pieces of ATFs in del Pezzo surfaces, where we do not include all the nodes. See for instance, Figures 26, 28, 32, 33. In these cases, our domain of interest X is not the complement Y of the symplectic divisors projecting over the boundary of the ATF, but rather a subdomain of Y. The nodes not contained in the ATF describing X will be considered frozen (not used to mutate), and depicted as a blue $\times$.

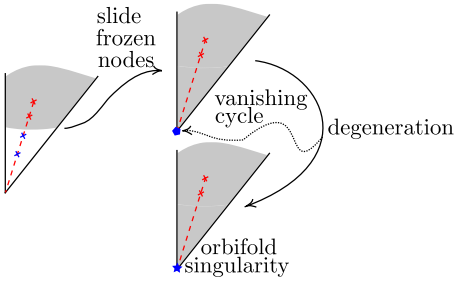


Fig. 25: Moving frozen nodes to the boundary, is equivalent to have over the vertex a possibly singular Lagrangian representing a vanishing cycle of a degeneration to a toric orbifold singularity. This vanishing cycle is represented by a pentagon over the vertex, and the orbifold singularity by a star in the above diagrams. Up to equivalence, the shaded domain can be viewed either as a subdomain of the complement of the boundary divisors in the left-picture, or the complement of the singular divisors in the orbifold diagram.

An alternative way of thinking is to disregard the frozen nodes. The total manifold $\tilde{Y}$ becomes singular, and a non-smooth compactification of X, given by adding the boundary divisor. The singularities are orbifold T-singularities [2] at each vertex, that were previously associated with the frozen nodes. Our original smooth manifold, that included the frozen nodes, is a smoothing of this orbifold. There is a continuous way of relating the Lagrangian fibrations on the orbifold with the ATF on the smoothing. We like to interpret it as a two step process, which is locally illustrated in Figure 25. The first, we keep the symplectic form on $\overline{Y}$, and consider almost-toric fibrations ATF_t, $t \in [0,1)$, so that in the limit $t \to 1$ the blue nodes slide all the way to a limit vertex at the boundary, and we are left with a singular

Lagrangian fibration SLF_1, on $\overline{Y}$, such that over the limit vertices live a possibly singular Lagrangian. The Lagrangian over each limit vertex can be recognised in each ATF_t, $t < 1$ living over the associated cut from the boundary of the ATF up to the farthest blue node and intersecting each fibre over the cut in a collapsing cycle for the corresponding nodes. This limit can be made rigorous but is beyond the scope of this article. The second step is to consider a degeneration from $\overline{Y}$ to $\tilde{Y}$, and a family of singular Lagrangian fibrations SLF_s, in the fibers corresponding to $s \in (0,1]$, where in the limit $s \to 0$ the singular Lagrangian over each vertex collapses to the corresponding orbifold singularity. The Lagrangian fibrations SLF_s are identified under symplectic parallel transport, so the singular Lagrangian over each vertex degenerating to an orbifold singularity is precisely the vanishing cycle of that orbifold singularity. In the examples presented here, the singularities associated with the frozen nodes will always be of A_n type, and hence the corresponding Lagrangian a chain of $n-1$ spheres.

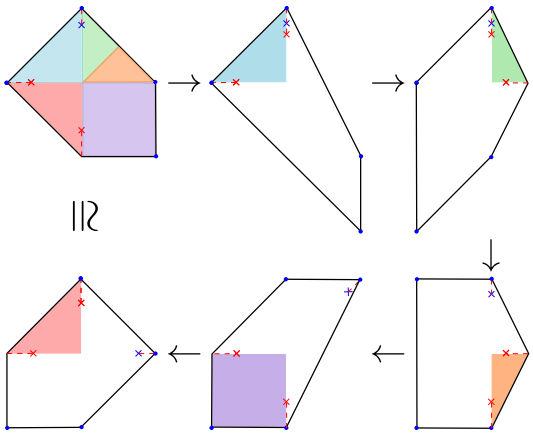


Fig. 26: Mutations on degree 5 del Pezzo – 5 tori

Let's turn our attention now to Figure 26, where we realize the degree 5 del Pezzo surface $\overline{Y}$ as a compactification of the A_2 cluster variety X, in a different way. We perform a nodal trade in a vertex at the bottom of the second diagram in Figure 23, and we freeze the top node. Now the boundary divisor represents 4 symplectic spheres, and X is a subdomain of the complement of these divisors. In this case, the mutation cycle induced by the nodal singularities in X does provides us with a 5-cycle of distinct monotone Lagrangian tori. Recall that monotone fibres of non-$SL(2,\mathbb{Z})$ related diagrams are distinct [30].

Remark 10. Disregarding the frozen node creates a double point singularity. In contrast with [17, Example 8.40], this is the same as considering $X_1 = 0$ in their setting. Now, consider the cycle of 5 divisors given by $\vartheta_0 = 0$. We claim that if one smooths

one node of this chain (represented by our nodal trade), and then delete the resulting chain of 4 divisors in this orbifold, one recovers X, the A_2 cluster variety.

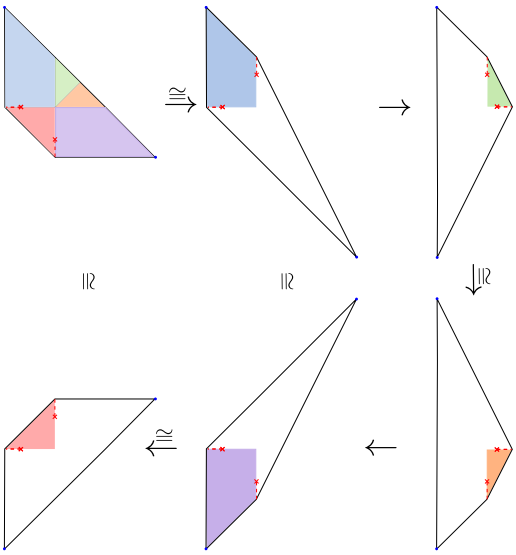


Fig. 27: Mutations on degree 8 del Pezzo – 2 tori

We can see that there is a simpler compactification of the A_2 cluster variety X by $\overline{Y} = \mathbb{C}P^2 \# \overline{\mathbb{C}P^2}$, a degree 8 del Pezzo, as illustrated in Figure 27 [which is the same obtained in Figure 14]. Here, X is the complement of two divisors in classes H and $2H - E$, where H is the class of the line, and E is the exceptional class. Note that we do not get a cycle of monotone Lagrangian tori, though not all tori are equivalent, we only see 2 tori in the whole cycle, which gives us only an edge on the unoriented graph of mutations of monotone tori, modulo equivalence.

Clearly, performing a blowup on one of the divisors of $\overline{Y}$ gives us another compactification of X. The top left diagram of Figure 28 corresponds to a toric blowup of monotone size [recall that in symplectic geometry, the blowups depend on the size of a symplectic ball one chooses to delete] in the top left diagram of either Figure 23 or Figure 26. Note in this case that the third and fourth, as well as the second and fifth, diagrams are equivalent, failing to deliver a cycle on the mutation graph of monotone Lagrangian tori in the degree 4 del Pezzo.

Let us consider now X the B_2 type cluster variety, with almost toric fibrations as in the series of diagrams in the middle of Figure 22. The first compactification $\overline{Y}$ we look at is the degree 6 del Pezzo, starting with the ATF depicted in the top-left diagram of Figure 29. [This diagram is $SL(2;\mathbb{Z})$ equivalent to [34, Diagram (A_5), Figure 16] (up to nodal trades).] In this case X is the complement of three divisors

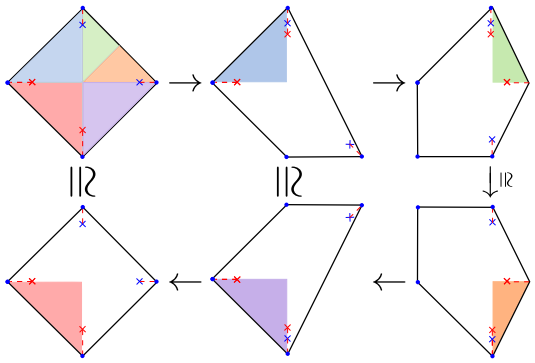


Fig. 28: Mutations on degree 4 del Pezzo – 3 tori

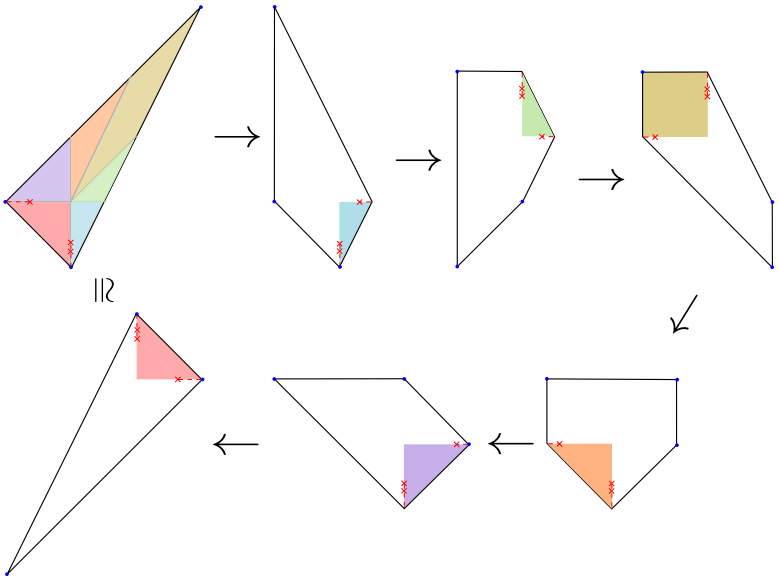


Fig. 29: Mutations on degree 6 del Pezzo – 6 tori

of $\overline{Y} = \mathbb{C}P^2 \# 3\overline{\mathbb{C}P^2}$, having symplectic areas 1, 2 and 3. We see that in this case we do get a cycle of size 6, with one torus corresponding to each cluster chart.

It is interesting to notice that the sixth diagram seems to have come from the scattering diagram Figure 16. But it is not quite the case, since that scattering diagram has a square function corresponding to the horizontal cut in the sixth diagram Figure 29, while a simple function corresponding to the vertical cut. This means that the natural compactification coming from that scattering diagram would be the same del Pezzo, but represented by the diagram of Figure 30 coming from apply-

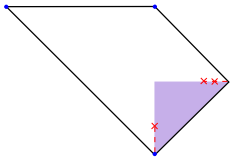


Fig. 30: This diagram differ from the sixth diagram in Figure 29, by one nodal trade and one inverse nodal trade. Mutations of the displayed nodes give equivalent polytopes.

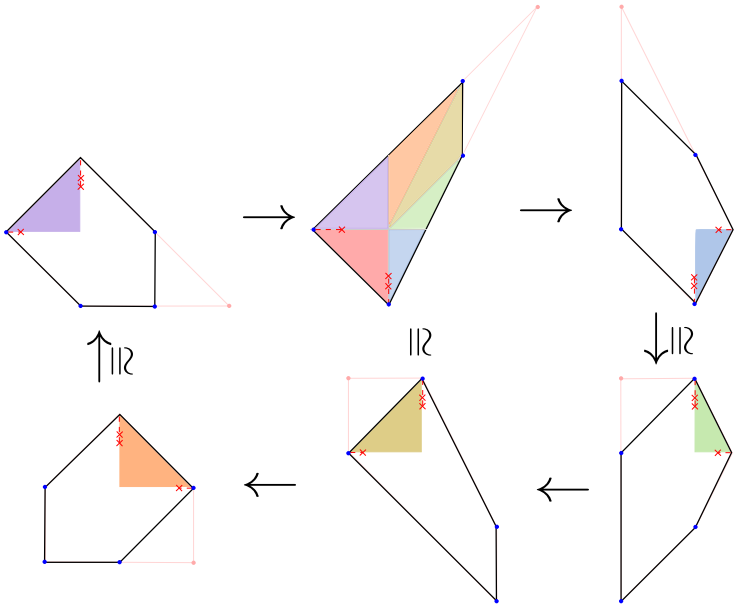


Fig. 31: Mutations on degree 5 del Pezzo – 3 tori

ing a nodal trade to the corner associated to the horizontal cut in the sixth diagram Figure 29, and an inverse nodal trade on one node at the vertical cut. In particular, X would be seen as the complement of 3 divisors in $\overline{Y} = \mathbb{C}P^2\#3\overline{\mathbb{C}P^2}$, each of symplectic area 2. The reader can check that in this case, the mutations associated to X would give equivalent monotone Lagrangian tori, analogous to the previous case depicted in Figures 23, 24.

Clearly we can also compactify the B_2 type cluster variety X to the degree 5 del Pezzo, as the complement of four divisors as depicted in Figure 31, by simply applying a nodal trade to a diagram in Figure 23. In Figure 31, we depicted segments outside the diagrams to indicate that they come from applying blowups to the dia-

grams in Figure 29. Curiously, it behaves similarly to the case in Figure 28, where we have only three non-equivalent Lagrangian tori, not providing a cycle.

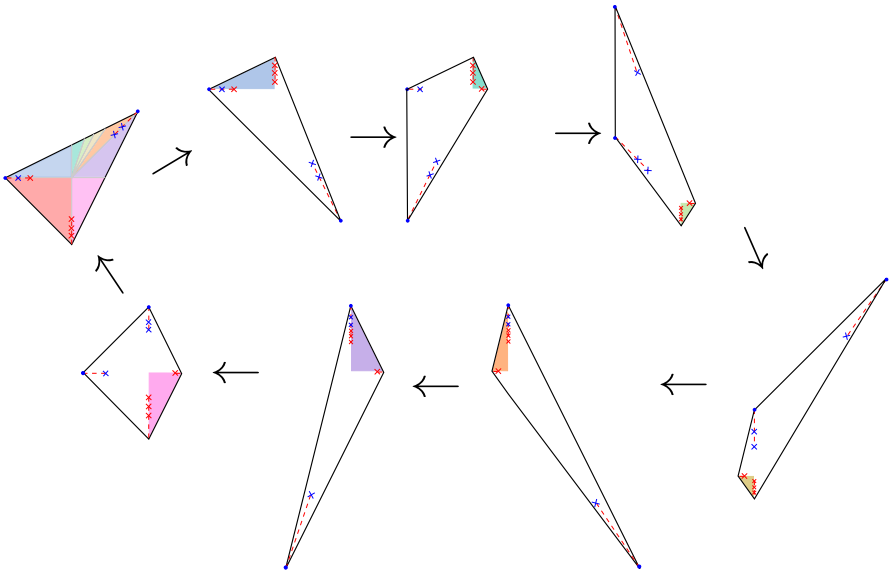


Fig. 32: Mutations on degree 3 del Pezzo – 8 tori

We now finish by presenting two compactifications of the G_2 cluster variety. We name it X, and consider it as an almost toric variety corresponding to the bottom series of diagrams in Figure 22. We start noting that the compactifications described in [8], see for instance [8, Figure 18], seems to be giving partially, but not fully, smoothable orbifolds. Here we look to two compactifications to degree 3 and 4 del Pezzo surfaces. [Both contain frozen variables, so the reader may prefer the alternative idea of seeing X compactifying to a degeneration of these surfaces.]

We start with the top left ATBD in Figure 32, which is equivalent to [34, Diagram (B_2), Figure 19], representing an ATF of the cubic $\mathbb{C}P^2\#6\overline{\mathbb{C}P^2}$. In this case, X is a subdomain of the complement of two symplectic divisors. Sliding the frozen nodes to the corresponding vertex gives one Lagrangian sphere, in the horizontal cut, and a chain of two Lagrangian spheres in $(1,1)$-cut. This indicates that disregarding the frozen nodes corresponds to considering an orbifold with one double-point singularity and one triple-point singularity. We do get one monotone Lagrangian torus for each of the 8 cluster charts in this case.

Another compactification of X is given in Figure 33. The top left diagram of Figure 33 is equivalent (up to nodal trades) to [34, Diagram (B_2), Figure 18]. Here, X is viewed as a subdomain of the complement of three symplectic divisors in $\overline{Y} = \mathbb{C}P^2\#5\overline{\mathbb{C}P^2}$. Sliding the frozen node to the vertex provides Lagrangian sphere, or

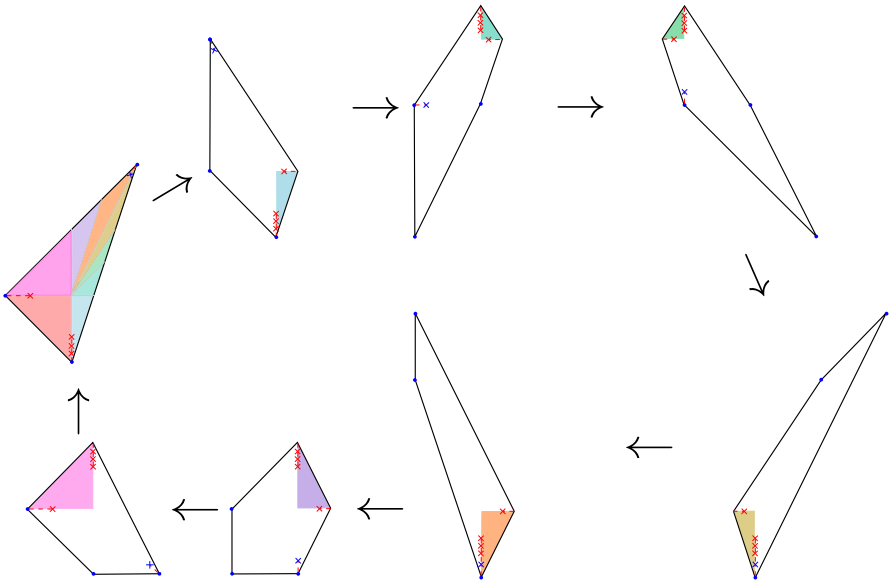


Fig. 33: Mutations on degree 4 del Pezzo – 8 tori

equivalently, disregarding the node gives a double-point orbifold singularity at the vertex. As before, we get one monotone Lagrangian torus for each cluster chart.

Acknowledgements This project initialized from discussions during the conference "Tropical Geometry And Mirror Symmetry" in the MATRIX Institute. The authors would like to thank the MATRIX institute for their hospitality.

The authors would like to thank Denis Auroux, and the referee for helpful feedback on the first version of the paper. The first author would like to thank Tim Magee, and Yu-shen Lin for helpful discussions. The first author is supported by NSF grant DMS-1854512. The second author is supported by Brazil's National Council of scientific and technological development CNPq, via the research fellowships 405379/2018-8 and 306439/2018-2, and by the Serrapilheira Institute grant Serra-R-1811-25965.

References

1. Akhtar, M., Coates, T., Galkin, S., Kasprzyk, A.M.: Minkowski polynomials and mutations. SIGMA Symmetry Integrability Geom. Methods Appl. **8**, Paper 094, 17 (2012)
2. Akhtar, M.E., Kasprzyk, A.M.: Mutations of fake weighted projective planes. Proc. Edinb. Math. Soc. (2) **59**(2), 271–285 (2016)
3. Arnold́, V.I.: Mathematical methods of classical mechanics, *Graduate Texts in Mathematics*, vol. 60. Springer-Verlag, New York (1989). Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein, Corrected reprint of the second (1989) edition

4. Auroux, D.: Mirror symmetry and T-duality in the complement of an anticanonical divisor. J. Gökova Geom. Topol. **1**, 51–91 (2007)
5. Auroux, D.: Special Lagrangian fibrations, wall-crossing, and mirror symmetry. In: Surveys in differential geometry. Vol. XIII. Geometry, analysis, and algebraic geometry: forty years of the Journal of Differential Geometry, *Surv. Differ. Geom.*, vol. 13, pp. 1–47. Int. Press, Somerville, MA (2009). DOI 10.4310/SDG.2008.v13.n1.a1
6. Cheung, M.W., Lin, Y.S.: Some examples of Family Floer mirror. In preparation
7. Cheung, M.W., Magee, T.: Towards Batyrev duality for finite-type cluster varieties. In preparation
8. Cheung, M.W., Magee, T., Nájera-Chávez, A.: Compactifications of cluster varieties and convexity. arXiv preprint arXiv:1912.13052 (2019)
9. Collins, T., Jacob, A., Lin, Y.S.: Special lagrangian submanifolds of log calabi-yau manifolds
10. Eliashberg, Y., Polterovich, L.: Unknottedness of Lagrangian surfaces in symplectic 4-manifolds. Internat. Math. Res. Notices (11), 295–301 (1993). DOI 10.1155/S1073792893000339. URL http://dx.doi.org/10.1155/S1073792893000339
11. Fock, V., Goncharov, A.B.: Cluster ensembles, quantization and the dilogarithm. Annales scientifiques de l'École Normale Supérieure **42**(6), 865–930 (2009)
12. Fomin, S., Zelevinsky, A.: Cluster algebras I: Foundations. J. Amer. Math. Soc. **15**, 497—529 (2002)
13. Galkin, S., Usnich, A.: Laurent phenomenon for Landau-Ginzburg potential (2010). Available at `http://research.ipmu.jp/ipmu/sysimg/ipmu/417.pdf`
14. Gross, M.: Special Lagrangian fibrations. I: Topology. AMS/IP Stud. Adv. Math. **23**, 65–93 (2001)
15. Gross, M., Hacking, P., Keel, S.: Mirror symmetry for log Calabi-Yau surfaces I. Publications mathématiques de l'IHÉS pp. 1–104 (2011)
16. Gross, M., Hacking, P., Keel, S.: Birational geometry of cluster algebras. Algebraic Geometry **2**(2), 137–175 (2015)
17. Gross, M., Hacking, P., Keel, S., Kontsevich, M.: Canonical bases for cluster algebras. Journal of the American Mathematical Society **31**(2), 497–608 (2018)
18. Gross, M., Hacking, P., Keel, S., Siebert, B.: The mirror of the cubic surface. arXiv preprint arXiv:1910.08427 (2019)
19. Gross, M., Siebert, B.: From real affine geometry to complex geometry. Annals of mathematics **174**(3), 1301–1428 (2011)
20. Gross, M., Siebert, B.: Intrinsic mirror symmetry. arXiv preprint arXiv:1909.07649 (2019)
21. Kontsevich, M., Soibelman, Y.: Affine structures and non-Archimedean analytic spaces. In: The unity of mathematics, *Progr. Math.*, vol. 244, pp. 321–385. Birkhäuser Boston (2006)
22. Leung, N.C., Symington, M.: Almost toric symplectic four-manifolds. J. Symplectic Geom. **8**(2), 143–187 (2010)
23. Li, T.J., Liu, A.: Symplectic structure on ruled surfaces and a generalized adjunction formula. Math. Res. Lett. **2**(4), 453–471 (1995). URL https://doi.org/10.4310/MRL.1995.v2.n4.a6
24. Lin, Y.S.: Enumerative geometry of del pezzo surfaces. arXiv preprint arXiv:2005.08681 (2020)
25. Mandel, T.: Tropical theta functions and log calabi–yau surfaces. Selecta Mathematica **22**(3), 1289–1335 (2016)
26. McDuff, D.: The structure of rational and ruled symplectic 4-manifolds. J. Amer. Math. Soc. **3**(3), 679–712 (1990). URL http://dx.doi.org/10.2307/1990934
27. Mikhalkin, G.: Amoebas of algebraic varieties and tropical geometry. In: Different faces of geometry, *Int. Math. Ser. (N. Y.)*, vol. 3, pp. 257–300. Kluwer/Plenum, New York (2004)
28. Ohta, H., Ono, K.: Notes on symplectic 4-manifolds with $b_2^+ = 1$. II. Internat. J. Math. **7**(6), 755–770 (1996). URL http://dx.doi.org/10.1142/S0129167X96000402
29. Ohta, H., Ono, K.: Symplectic 4-manifolds with $b_2^+ = 1$. In: Geometry and physics (Aarhus, 1995), *Lecture Notes in Pure and Appl. Math.*, vol. 184, pp. 237–244. Dekker, New York (1997)
30. Shelukhin, E., Tonkonog, D., Vianna, R.: Geometry of symplectic flux and Lagrangian torus fibrations. arXiv:1804.02044 (2018)

31. Strominger, A., Yau, S.T., Zaslow, E.: Mirror symmetry is T-duality. Nuclear Phys. B **479**(1-2), 243–259 (1996). URL http://dx.doi.org/10.1016/0550-3213(96)00434-8
32. Symington, M.: Four dimensions from two in symplectic topology. In: Topology and geometry of manifolds (Athens, GA, 2001), *Proc. Sympos. Pure Math.*, vol. 71, pp. 153–208. Amer. Math. Soc., Providence, RI (2003)
33. Vianna, R.: Infinitely many exotic monotone Lagrangian tori in $\mathbb{CP}^2$. J. Topol. **9**(2), 535–551 (2016)
34. Vianna, R.: Infinitely many monotone Lagrangian tori in del Pezzo surfaces. Selecta Math. (N.S.) **23**(3), 1955–1996 (2017)

Observations on disks with tropical Lagrangian boundary

Jeff Hicks

Abstract In this survey, we look at some expectations for Lagrangian submanifolds which are built as the lifts of tropical curves from the base of an Lagrangian torus fibration. In particular, we perform a first computation showing that holomorphic triangles can appear with boundary on the Lagrangian submanifold. We speculate how these holomorphic triangles can contribute to the count of holomorphic strips in the Lagrangian intersection Floer cohomology between a tropical Lagrangian submanifold and a fiber of the SYZ fibration.

1 Tropical Lagrangians and Holomorphic Disks

Mirror symmetry is a geometric duality between symplectic geometry on (X, ω) and complex geometry on a "mirror space" $(\check{X}, J)$ [2]. The spaces $X, \check{X}$ are expected to arise as the total spaces of dual Lagrangian torus fibrations over a common base space Q [9]. The base of a Lagrangian torus fibration always is equipped with an affine structure, and it is predicted that both the symplectic geometry of X and complex geometry of $\check{X}$ degenerate to tropical geometry on Q [4]. In good examples, one uses the affine structure on Q to identify lattices $T_{\mathbb{Z}}Q$ and $T^*_{\mathbb{Z}}Q$ inside the tangent and cotangent bundle respectively. The mirror spaces can be reconstructed from this data as:

$$X := T^*Q/T_{\mathbb{Z}}Q \xrightarrow{\text{val}} Q \xleftarrow{\check{\text{val}}} \check{X} = TQ/T_{\mathbb{Z}}Q$$

which are equipped with their canonical symplectic and almost complex structures. The simplest example of this (which we will focus on) is the example where $Q = \mathbb{R}^n$ and $X = \check{X} = (\mathbb{C}^*)^n$. Recently, the independent works of [7, 8, 5, 6] constructed

Jeff Hicks
Centre for Mathematical Sciences, Wilberforce Rd, Cambridge CB3 0WB, U.K.
e-mail: jh2234@cam.ac.uk

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_36

Lagrangian submanifolds $L(V) \subset X$ whose valuation projection $\mathrm{val}(L(V))$ could be made arbitrarily close to a given tropical subvariety $V \subset Q$. These Lagrangians fit in with predictions from mirror symmetry principles. For example, in [6] the number of tropical Lagrangians found was at least as many as the number of curves in the mirror quintic, and in [5] tropical Lagrangian hypersurfaces were shown to be homologically mirror to sheaves supported on divisors.

Parallel to computations in mirror symmetry, tropical geometry has been employed to understand the counts of holomorphic disks with boundary on a given Lagrangian inside of a symplectic manifold. One particularly visible instance of this method of computation is from [10], which used tropical techniques to understand the holomorphic disk count for Lagrangian tori inside of $\mathbb{CP}^2$. These holomorphic disk counts were used to distinguish Hamiltonian isotopy classes of monotone Lagrangian tori. More generally, the count of holomorphic disks with boundary on non-exact Lagrangian submanifolds $L \subset X$ provide a deformation to the homology of L. These deformations are a necessary ingredient in the mirror symmetry prediction, as they provide "corrections" to the identification of symplectic and complex geometric invariants. For example, the correspondence between the moduli space of Lagrangian tori on X and points on the mirror space $\check{X}$ is only expected to hold once these correction terms have been computed [1].

These invariants are frequently difficult to compute, so the promise of reducing them to combinatorial type computations in the setting of tropical geometry is particularly enticing. We exhibit an explicit computation for holomorphic disks for a fixed example, and speculate on what this computation means for the more general problem of computing holomorphic strips contributing to the differential in Lagrangian intersection Floer theory.

2 Disks via sections of Lagrangian fibrations

The first example we consider is the symplectic manifold $X = (\mathbb{C}^*)^2$ with base $Q = \mathbb{R}^2$ and torus fibration given by $\mathrm{val}(z_1, z_2) = (\log|z_1|, \log|z_2|)$. We consider the three tropical curves drawn in fig. 1, and consider their lift to 3 Lagrangian cylinders in X,

$$\begin{aligned} L_1 &= \{(e^r, e^{i\theta}) \mid r \in \mathbb{R}, \theta \in S^1\} \\ L_2 &= \{(e^{i\theta}, e^r) \mid r \in \mathbb{R}, \theta \in S^1\} \\ L_\lambda &= \{(e^{r+i\theta}, e^\lambda e^{-r+i\theta}) \mid r \in \mathbb{R}, \theta \in S^1\} \end{aligned}$$

where $\lambda \in \mathbb{R}$ is a parameter picked to define the third line. Although these Lagrangians are non-compact, they are conical at infinity and so we may count holomorphic disks and polygons with boundary on the L_i. For topological reasons, the L_i do not individually bound holomorphic disks. However, the collection of all 3 has a chance to bound a holomorphic triangle. We parameterize this triangle with the

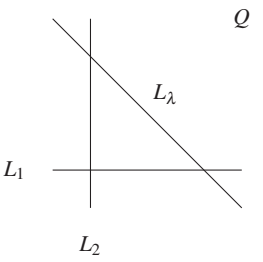


Fig. 1 Projection of three Lagrangians $L_i \subset X$ to the base of the SYZ fibration via the valuation map. These Lagrangians bound a holomorphic section of $\mathrm{val} : X \to Q$ for particular values of λ.

domain $\Delta_\lambda := \{x+iy \mid x,y \geq 0, x+y \leq \lambda\}$. We then consider the holomorphic map:

$$\begin{aligned} u_\lambda : \Delta_\lambda &\to (\mathbb{C}^*)^2 \\ (x+iy) &\mapsto (e^{x+iy}, e^{y-ix}) \end{aligned}$$

One notices that the boundary $u_\lambda(x,0) = (e^x, e^{-ix})$ is contained in L_1 and $u_\lambda(0,y)$ is contained in L_2. However, the remaining boundary $u_\lambda(t,\lambda-t) = \left(e^{t+i(\lambda-t)}, e^{\lambda-t-it}\right)$ will lie in the Lagrangian L_λ if and only if $\lambda \in 2\pi\mathbb{Z}$. This leads to the following strange behaviour: as one modifies the parameter λ, the three Lagrangians L_1, L_2, L_λ periodically bound a holomorphic section over a triangle in the base of $\mathrm{val} : (\mathbb{C}^*)^2 \to \mathbb{R}^2$. These holomorphic triangles are not regular, so it is not unexpected upon taking a generic choice of λ we see no disk. However, for families of Lagrangians parameterized by λ, these holomorphic disks do appear regularly.

This sporadic appearance of Maslov index zero disks also occurs in the descriptions of wall-crossings for Lagrangian tori [1]. In that setting, as one takes a family of Lagrangian tori interpolating between the Chekanov torus and product torus in $\mathbb{C}^2$, a non-regular Maslov index 0 disk with boundary flashes in and out of existence.

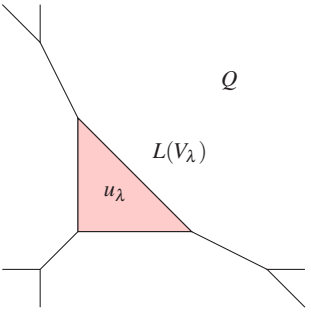


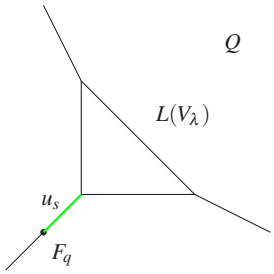
Fig. 2 The valuation projection of a tropical Lagrangian bounding a holomorphic disk. The disk is obtained by taking the holomorphic section of fig. 1 over the triangle, and rounding off the corners. It exists only for certain values of λ determining the tropical Lagrangian submanifold

We can replicate this kind of behaviour by turning our holomorphic triangles into holomorphic disks by performing Lagrangian surgery on the intersections of L_1, L_2, L_λ to obtain a embedded tropical Lagrangian submanifold $L(V_\lambda)$, whose projection to the base Q is drawn in fig. 2. After performing surgery on the Lagrangian,

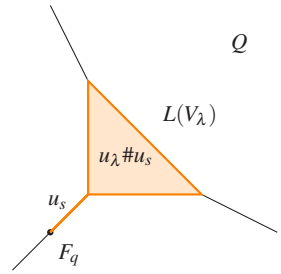
it is expected that holomorphic triangles with boundary on the L_i become holomorphic disks on the Lagrangian $L(V_\lambda)$ [3]. The holomorphic triangles described in the first computation give examples of holomorphic disks with boundary on the tropical Lagrangian $L(V_\lambda)$ when the parameter λ passes through a multiple of 2π. This kind of holomorphic disk is expected to exists: in fact, for some choice of almost complex structure, the wall crossing phenomenon for tropical Lagrangians observed in [5] proves that there is a non-regular disk with boundary on this tropical Lagrangian submanifold corresponding to the wall-crossing phenomenon for the Chekanov and Clifford tori in $\mathbb{CP}^2$.

3 Towards computing Floer Support

The presence of non-regular holomorphic disks plays an important role in homological mirror symmetry, where the flux coordinates parameterizing the space of Lagrangian submanifolds must be corrected by contributions from bubbling of disks. Additionally, the presence of a non-regular holomorphic disk u_λ can honestly modify the behaviour of a regular holomorphic disk u_s with Lagrangian boundary, as u_λ and u_s can be glued along their boundary to produce a new regular holomorphic disk. We now speculate about the Lagrangian intersection Floer theory between the $L(V_\lambda)$ and a fiber of the SYZ fibration, $F_q := \mathrm{val}^{-1}(q)$, as drawn in fig. 3a.



(a) A small holomorphic strip, observed in [5].



(b) Creating a bigger strip by attaching the non-regular disk to the regular strip.

Fig. 3: Examples of holomorphic strips contributing to the differential of $CF^\bullet(L(V_\lambda), F_q)$.

In [5], we gave an example worked out with Diego Matessi regarding an interesting holomorphic strip: $u_s : [0,1] \times \mathbb{R} \to (\mathbb{C}^*)^2$ with boundaries on F_q and V_λ. This holomorphic strip projects under the valuation to the line segment drawn in fig. 3a. u_s is a regular holomorphic strip which persists even as we modify the parameters λ governing the size of the "hole" in the tropical elliptic curve. As a result, for choices of λ which the disk u_λ appears, the strip u_s intersects u_λ, and we conjecture that it is possible to glue together u_s and u_λ to obtain a larger regular holomorphic

strip $u_\lambda \# u_s$ contributing to the Floer differential of $CF^\bullet(L(V_\lambda), F_q)$. This conjectured holomorphic strip is drawn in fig. 3b.

The Floer theoretic support, which is the set of Lagrangian branes F_q for which this Floer cohomology does not vanish, gives the equation of an algebraic curve in the mirror $\check{X}$. In order to compute this support, it is necessary to compute all of the holomorphic strips with boundary on F_q and $L(V_\lambda)$. We hope that this combinatorial description of two such holomorphic strips can be extended to produce a combinatorial model for the Lagrangian intersection Floer theory of arbitrary tropical Lagrangian submanifolds, which would significantly improve our understanding of the interplay between homological mirror symmetry and tropical geometry.

Acknowledgements This survey arose from a series of conversations with Brett Parker and Renato Vianna during the *Tropical Geometry and Mirror Symmetry* program hosted by MATRIX, and benefited from the helpful comments of an anonymous reviewer. This work was partially funded by EPSRC Grant EP/N03189X/1.

References

1. Auroux, D.: Mirror symmetry and T-duality in the complement of an anticanonical divisor. arXiv preprint arXiv:0706.3207 (2007)
2. Candelas, P., Xenia, C., Green, P.S., Parkes, L.: A pair of Calabi-Yau manifolds as an exactly soluble superconformal theory. Nuclear Physics B **359**(1), 21–74 (1991)
3. Fukaya, K., Oh, Y., Ono, K., Ohta, H.: Lagrangian intersection Floer theory-anomaly and obstruction-Chapter 10. Preprint available on K. Fukaya's homepage (2007)
4. Gross, M., Siebert, B.: Affine manifolds, log structures, and mirror symmetry. Turkish Journal of Mathematics **27**(1), 33–60 (2003)
5. Hicks, J.: Tropical Lagrangians and homological mirror symmetry. Ph.D. thesis, University of California, Berkeley (2019)
6. Mak, C.Y., Ruddat, H.: Tropically constructed Lagrangians in mirror quintic threefolds. arXiv preprint arXiv:1904.11780 (2019)
7. Matessi, D.: Lagrangian pairs of pants. arXiv:1802.02993 (2018)
8. Mikhalkin, G.: Examples of tropical-to-Lagrangian correspondence. arXiv:1802.06473 (2018)
9. Strominger, A., Yau, S.T., Zaslow, E.: Mirror symmetry is T-duality. Nuclear Physics B **479**(1-2), 243–259 (1996)
10. de Velloso Vianna, R.F.: Infinitely many exotic monotone Lagrangian tori in $\mathbb{CP}^2$. Journal of Topology **9**(2), 535–551 (2016)

Compactifying torus fibrations over integral affine manifolds with singularities

Helge Ruddat and Ilia Zharkov

Abstract This is an announcement of the following construction: given an integral affine manifold B with singularities, we build a topological space X which is a torus fibration over B. The main new feature of the fibration $X \to B$ is that it has the discriminant in codimension 2.

1 Introduction

There have been a lot of studies of half-dimensional torus fibrations and their integral affine structures on the base spaces inspired by the Strominger-Yau-Zaslow conjecture [SYZ96]. This area was very active in the beginning of the 2000's with many approaches of different flavor: topological [Zh00], [G01], symplectic [G00], [Leung], [Rua], [J03],[CBM09], [Au07], [Au09], [EM19], metric [GW00], [KS00], [LYZ], non-Archimedean [KS06], tropical [Mi04], combinatorial [HZ05], and log-geometric [GS06], [GS10], [Pa07]. For surveys on the early developments, see [T06, G09]. The toric case was considered in [CL, CLL, FLTZ12]. A more recent surge and interest is mostly tropical [Mat, SS18, AGIS, Mi19, H19, MR], non-Archimedean [NXY] or topological [AS], [P18]. For more recent surveys, we refer to [G12, Ch]. Broadly speaking, all this research developed into a new field of mathematics: tropical geometry.

In this note, we essentially follow the Gross-Siebert setup [GS06], [GS10], with some slight modifications. We replace the polyhedral decomposition of the base B

Helge Ruddat
Johannes Gutenberg-Universität Mainz, Inst. f. Mathematik, e-mail: ruddat@uni-mainz.de
Universität Hamburg, Fachbereich Mathematik, e-mail: helge.ruddat@uni-hamburg.de

Ilia Zharkov
Kansas State University, 138 Cardwell Hall, Manhattan, KS 66506, e-mail: zharkov@ksu.edu

H.R. was supported by DFG grant RU 1629/4-1 and the Department of Mathematics at Universität Hamburg. The research of I.Z. was supported by Simons Collaboration grant A20-0125-001.

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_37

by a regular CW-decomposition for the gain of flexibility, cf. the notion of "symple" in [Ru20]. Also we relax requirements for the monodromy by allowing arbitrary lattice simplices for local monodromies, not just the elementary ones. That requires a little more care for the local monodromy assumptions, but does not seem to affect the topological side of the story much. On the other hand, when we compare our model with the Kato-Nakayama space of a canonical Calabi-Yau family, we use the machinery of log-structures on toroidal crossing spaces, so we restrict ourselves back to the Gross-Siebert polyhedral base B with elementary simple singularities.

This note consists of two parts. The first three sections are devoted to the construction of the compactification of the torus bundle from over the smooth part B_0 of the base to all of B. The last section compares the topology of the total space of the compactified torus bundle with the Kato-Nakayama space obtained from a toric log Calabi-Yau space.

The primary purpose of this note is an announcement, however, we do give a precise definition of the setup, its basic notions, some discussion of these and the statement of the main results to be achieved. We carry out the compactification construction in dimension three under a unimodularity assumption for illustration. Some results may be stated only in special cases and proofs may be sketchy or omitted. All statements in full generality and rigorous proofs will appear soon in [RZ2].

2 Integral affine manifolds with singularities

Let B be a pure n-dimensional regular CW complex which is a manifold. We fix the first barycentric subdivision $\mathrm{bsd}\,B$ of B and let $\bar{D}$ be the subcomplex of $\mathrm{bsd}\,B$ which consists of simplices spanned by the barycenters of strata of B which are not vertices and not facets. That is, $\bar{D}$ is an $(n-2)$-dimensional subcomplex of $\mathrm{bsd}\,B$ which lives inside the $(n-1)$-skeleton of B and misses all vertices of B.

Suppose that we are given an integral affine structure on $B_{00} := B \setminus \bar{D}$. That is, B_{00} is given the structure of a smooth manifold and a flat connection of its tangent bundle TB_{00} with holonomy in $\mathrm{GL}_n(\mathbb{Z})$. We denote by Λ the rank n local system of flat integral vectors in TB_{00}. Similar, the local system $\check{\Lambda}$ stands for the flat integral covectors in the cotangent bundle T^*B_{00}.

Each facet of $\bar{D}$, being of codimension 2, has a small loop around it in B_{00} and we compute the monodromy of the affine structure along this loop. If the monodromy is trivial we can extend the affine structure over this facet. If the monodromy is not trivial, then this facet becomes a part of the true **discriminant** D which is a full-dimensional subcomplex of $\bar{D}$, that is still a codimension 2 subcomplex of $\mathrm{bsd}\,B$. We denote by $B_0 := B \setminus D$ the **smooth** part of the base, this is as far as the affine structure extends.

Now we describe the requirements for the monodromy of the affine structure. Let $\iota : B_0 \hookrightarrow B$ be the inclusion of the smooth part into the base. Then $\iota_*\Lambda$ and $\iota_*\check{\Lambda}$ are the constructible sheaves of locally invariant sublattices of Λ and $\check{\Lambda}$. In particular,

the stalk of $\iota_*\Lambda$ at a point x in the discriminant D extends as a constant subsheaf of Λ in a neighborhood U of x (the Λ itself is not trivializable on $U \setminus D$), and similar for $\check{\Lambda}$. We denote the restriction of $\iota_*\Lambda$ to D by L, and the restriction of $\iota_*\check{\Lambda}$ to D by $\check{L}$, both are constructible sheaves on D.

Let $x \in D$ be a point which lies in the stratum τ. Pick a nearby base point $y \in B_0$. The local fundamental group of B_0 in a neighborhood of x is generated by the loops around the maximal strata of D and we want to see its monodromy image G_x in $\mathrm{GL}(\Lambda_y)$. The minimal requirement is that G_x is an abelian subgroup of $\mathrm{GL}(\Lambda_y)$. In fact we want to require even more. Since $L, \check{L}$ are constant on the relative interior of each stratum τ of D, we simply refer to the stalk at any point in that relative interior by L_τ, respectively $\check{L}_\tau$.

Definition 1. For a stratum $\tau \subset D$, suppose there are sublattices $L_1, \dots, L_r$ in L_τ, linearly independent over $\mathbb{Q}$, and sublattices $\check{L}_1, \dots, \check{L}_r$ in $\check{L}_\tau$, also linearly independent over $\mathbb{Q}$. We call the collection of sublattices **semi-simple** if every L_i is orthogonal to every $\check{L}_j$ (including $i = j$). If every stratum $\tau \subset D$ permits a semi-simple collection of sublattices so that the monodromy group G_x for any x in the interior of τ has the form $\mathrm{id} + L_1 \otimes \check{L}_1 + \cdots + L_r \otimes \check{L}_r$ then we say that (B, D) is an integral affine manifold with **semi-simple abelian** (or for short just semi-simple) singularities.

We denote the rank of L_i by k_i and the rank of $\check{L}_i$ by $\check{k}_i$, and let $\ell := \sum_i k_i$ and $\check{\ell} := \sum_i \check{k}_i$. It holds $s := n - \ell - \check{\ell} \geq 0$. The semi-simpleness condition says that in a neighborhood U of $x \in D$ the monodromy matrices in a suitable basis of $\Lambda_y \otimes_{\mathbb{Z}} \mathbb{Q}$ when acting on column vectors have the shape

$$\begin{pmatrix}
1 & 0 & \cdots & 0 & \circledast & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & \ddots & \ddots & \vdots \\
\vdots & 0 & \ddots & \vdots & \vdots & \ddots & \ddots & 0 \\
\vdots & \vdots & \ddots & 1 & 0 & \cdots & 0 & \circledast \\
0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 \\
0 & \cdots & \cdots & 0 & 0 & 1 & \ddots & \vdots \\
0 & \cdots & \cdots & 0 & \vdots & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & 0 & 0 & \cdots & 0 & 1
\end{pmatrix}$$

where the first columns correspond to a basis of L, the last rows correspond to a basis of $\check{L}$ and the $(k_i \times \check{k}_i)$-size $\circledast$-blocks correspond to the lattices $L_i \otimes \check{L}_i$.

If $\dim B = 3$, then necessarily $r \leq 1$ and D is a graph. Following Gross, we call a vertex of D **positive** if $\dim \check{L} = 2$ and we call a vertex **negative** if $\dim L = 2$.

In fact we want even more. To every stratum τ of D we would like to associate two collections of lattice polytopes $(\Delta_1, \dots, \Delta_r)_\tau$ in L_τ and $(\check{\Delta}_1, \dots, \check{\Delta}_r)_\tau$ in $\check{L}_\tau$ such that each L_i is generated by the edge vectors of Δ_i, and similar for $\check{L}_i$. We denote by $\mathscr{P}$ the collection of $\{\Delta_i, \check{\Delta}_i\}$ for all strata of D. Next we discuss the compatibility of the collection $\mathscr{P}$ that we require for the inclusion maps $\phi \colon L_\tau \hookrightarrow L_\sigma$ and $\check{\phi} \colon \check{L}_\tau \hookrightarrow \check{L}_\sigma$ for any two incident strata $\tau \prec \sigma$ of D. Note that $r_\tau \geq r_\sigma$, and we can always

match the number r of polytopes in σ and τ by adding the origins $\{0\}$ to play the role of missing $\Delta, \check{\Delta}$ to the σ-collection.

Definition 2. The collection of polytopes $\mathscr{P}$ is **compatible** if for any incident pair $\tau \prec \sigma$ of D, after a suitable integral translation, $\phi(\Delta_{i,\tau})$ is a face of $\Delta_{i,\sigma}$ and, similarly after integral translation, $\check{\phi}(\check{\Delta}_{i,\tau})$ is a face of $\check{\Delta}_{i,\sigma}$ for all $i = 1, \dots, r$ (up to reordering the indices in $\{1, \dots, r\}$).

Next, we describe the correlation between $\mathscr{P}$ and the discriminant D. To any polytope Δ_i one can associate its normal fan. Let $Y_i \subset \mathbb{R}^{k_i}$ be the codimension 1 skeleton of that normal fan. Similarly, $\check{Y}_i \subset \mathbb{R}^{\check{k}_i}$ is the codimension 1 skeleton of the normal fan to $\check{\Delta}_i$. For a point $x \in D$ in a stratum τ we consider the codimension 2 fans in $\mathbb{R}^n = \mathbb{R}^s \times \mathbb{R}^{k_1} \times \mathbb{R}^{\check{k}_1} \times \cdots \times \mathbb{R}^{k_r} \times \mathbb{R}^{\check{k}_r}$:

$$S_{x,i} := \mathbb{R}^s \times Y_i \times \check{Y}_i \times \mathbb{R}^{\ell - k_i + \check{\ell} - \check{k}_i}. \tag{1}$$

Let S_x be their union: $S_x := \bigcup_i S_{x,i}$. Note that $\bigcap_i S_{x,i} = \mathbb{R}^s$. Maximal cones in S_x are labeled by the pairs of edges in (e, f), where e is an edge in Δ_i and f is an edge in $\check{\Delta}_i$ for some i.

Definition 3. A compatible collection of polytopes $\mathscr{P}$ is **normal** if for every point $x \in D$ there is a homeomorphism of its neighborhood $U \subset B$ to an open subset $V \subset \mathbb{R}^n$ which maps $D \cup U$ to $S_x \cup V$.

Finally we make connection between $\mathscr{P}$ and the monodromy of the affine structure. Let x be a point in D which lies in the stratum τ. Pick a nearby base point $y \in B_0$. We assume that the polytopal collection is compatible and normal. Then the local fundamental group of B_0 in a neighborhood of x is generated by the loops around the maximal strata of D, which are labeled by the pairs of edges (e, f), e in Δ_i and f in $\check{\Delta}_i$, some i. Orienting the edges determines an orientation of the loop around the corresponding stratum $\sigma_{e,f}$ of D, see [GS] for details.

Definition 4. A semi-simple integral affine manifold (B, D) with a compatible normal collection of polytopes $\mathscr{P}$ is called **semi-simple polytopal** if the local monodromy along the loop $\sigma_{e,f}$ is given by $\mathrm{id} + e \otimes f$.

The collection of polytopes $\mathscr{P}$ is reminiscent of the Batyrev-Borisov nef-partitions. The semi-simple polytopal integral affine manifolds therefore mimic local complete intersections in algebraic geometry.

We next give an example of a semi-simple affine structure which is not polytopal. The figure below shows a part of discriminant in a 3-dimensional base and the monodromy matrices around the intervals in D with respect to some base point $y \in B \setminus D$ and a suitable basis of $\Lambda_y \cong \mathbb{Z}^3$. The edge τ connects the positive vertex on the left with the negative vertex on the right. The monodromy around the middle edge τ is twice the standard focus-focus case. The point is that it is impossible to decide which of the two intervals Δ_τ or $\check{\Delta}_\tau$ has length 2. The vertex on the left requires Δ_τ to be length one while the vertex on the right requires $\check{\Delta}_\tau$ to be length one.

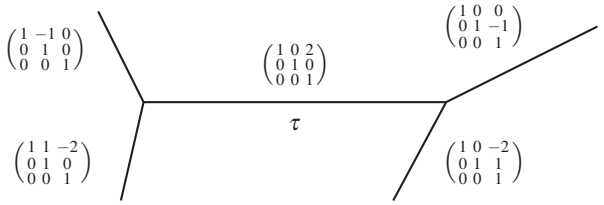


Fig. 1 A non-polytopal semi-simple affine structure.

Another feature of a polytopal affine structure is some sort of local convexity of the monodromy. Figure 2 shows an example of a negative vertex in a 3-dimensional base with the monodromy matrices around the four adjacent edges. The monodromy vectors do not form a convex polytope.

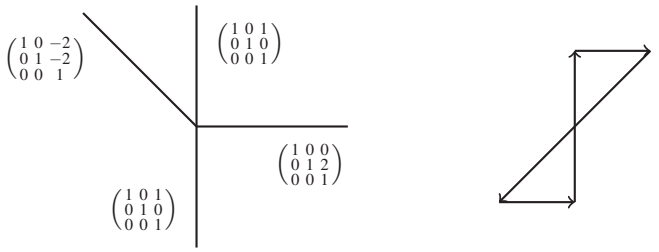


Fig. 2 Another non-polytopal semi-simple affine structure.

3 Local models

Let us consider a point x in the interior of a stratum τ in the discriminant. We will describe a local model of the torus fibration over a neighborhood of x in B. The construction of X_τ as a fiber product (the left side of the diagram in Figure 3) is pretty standard, see, e.g. [GS]. The novelty here is the rightmost column.

We explain the details now. Let Σ be the cone over the convex hull $\mathrm{Conv}\{(\check{\Delta}_i, e_i)\} \subset \check{L}_{\mathbb{R}} \oplus \mathbb{R}^r$, where $e_i = (0,\dots,1,\dots,0)$ is the i-th basis vector of $\mathbb{R}^r$, and let $\Sigma^\vee$ be its dual cone and $\Sigma^\vee_{\mathbb{Z}}$ the integral points in the dual cone. The affine toric variety $U_\Sigma = \mathrm{Spec}\,\mathbb{C}[\Sigma^\vee_{\mathbb{Z}}]$ has r monomials z^{w_i} corresponding to the integral vectors w_i in $\Sigma^\vee$ defined by

$$w_i(\check{\Delta}_i) = 1, \quad w_i(\check{\Delta}_j) = 0,\ j \neq i,$$

which gives the map $U_\Sigma \to \mathbb{C}^r$. The map $\mu_0 \colon U_\Sigma \to \Sigma^\vee$ is the moment map, and the cone $\Sigma^\vee$ projects surjectively to $\mathbb{R}^{\check{\ell}}$ by taking the quotient by the subspace spanned by all w_i's.

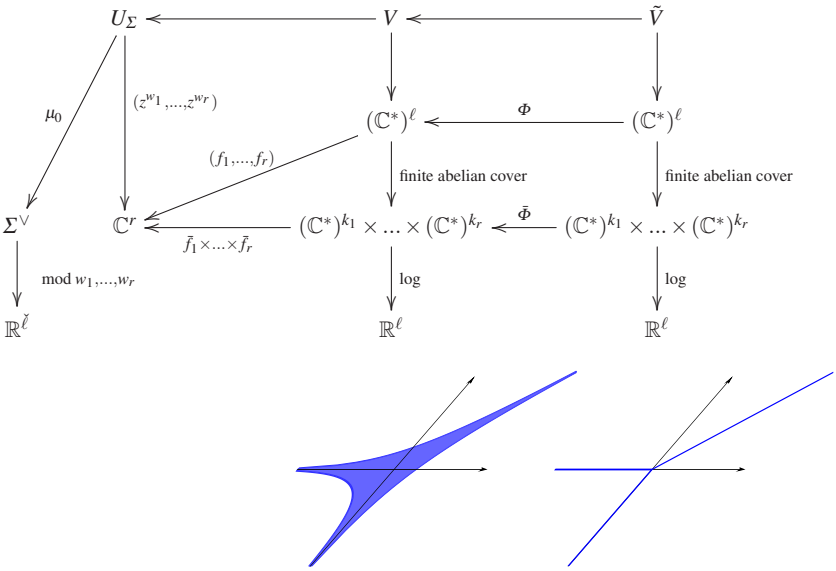


Fig. 3 Bottom pictures: $\log(f^{-1}(0))$ and $\log((f\circ\Phi)^{-1}(0))$ for $r=1,\ k=2$.

Let $L'\cong\mathbb{Z}^\ell$ be the sublattice in L generated by $L_1,\dots,L_r$ over $\mathbb{Q}$. That is $L'=L\cap((L_1\oplus\dots\oplus L_r)\otimes\mathbb{Q})$ and L' is the unique direct summand of L that has rank ℓ and contains $L_1,\dots,L_r$. The map

$$(\mathbb{C}^*)^\ell\cong\operatorname{Hom}(L',\mathbb{C}^*)\to\operatorname{Hom}(L_1\oplus\dots\oplus L_r,\mathbb{C}^*)\cong(\mathbb{C}^*)^{k_1}\times\dots\times(\mathbb{C}^*)^{k_r}$$

is the abelian cover that appears as the center vertical map in Figure 3 (and a homeomorphic version of it also on the right). The finite abelian cover takes care of the fact that the sublattice $L_1\oplus\dots\oplus L_r$ may have finite index >1 in $L'\subseteq L$. There are two ingredients for that. First, the lattice L_i may have an index in its saturation in L. Second, the direct sum $(L_1\oplus\dots\oplus L_r)\otimes\mathbb{Q}$ may not split over $\mathbb{Z}$.

Each polytope Δ_i defines a function

$$f_i=\sum_{v\in\operatorname{vert}\Delta_i}c_vz^v\quad:\quad\operatorname{Hom}(L',\mathbb{C}^*)\to\mathbb{C},\quad\text{for a general choice of } c_v\in\mathbb{C}^*.$$

The same expression also defines a function $\bar{f}_i\colon\operatorname{Hom}(L_i,\mathbb{C}^*)\to\mathbb{C}$, so that the triangle in Figure 3 commutes. The space V is given as the complete intersection $\{f_i=w_i\}$ in $U_\Sigma\times(\mathbb{C}^*)^\ell$. The full local model for the torus fibration X_τ over a neighborhood of x is given by multiplying V by the factor of $\log\colon(\mathbb{C}^*)^s\to\mathbb{R}^s$. We reserve the right to split this factor between U_Σ and $(\mathbb{C}^*)^\ell$ as needed to match the models for adjacent strata in D.

To actually attach the right column in Figure 3, we will assume that *all polytopes Δ_i are simplices*. At the base of the abelian cover, the map $(f_1,\dots,f_r)\colon(\mathbb{C}^*)^\ell\to\mathbb{C}^r$

splits as a product and, by the assumption of Δ_i to be a simplex, the hypersurface $\{f_i = 0\}$ in $(\mathbb{C}^*)^{k_i}$ is a cover of the pair-of-pants $\{\bar{f}_i = 0\}$. The right column is entirely defined once we specify $\bar{\Phi} = \bar{\Phi}_1 \times \cdots \times \bar{\Phi}_r$ if we additionally require that the two squares adjacent to the right column be Cartesian. Note that Δ_i is a unimodular simplex with respect to L_i (by definition of L_i). In Theorem 1 we will use a universal notation Φ for any of the maps $\bar{\Phi}_i$.

As before, we denote by Y the codimension 1 skeleton of the normal fan to the standard simplex Δ^k. That is, the cones Y_J in Y are labeled by subsets $J \subseteq \{0,1,\ldots,k\}$ of size at least 2. On the other hand, the coamoeba of the $(k-1)$-pants has a well-known $(k-1)$-dimensional skeleton $S \subset \mathbb{T}^k$ (the boundary of a permutahedron), whose faces are labeled by cyclic partitions σ of $\{0,1,\ldots,k\}$ with at least 2 parts, see [RZ20] for details. The ober-tropical pair-of-pants $\mathscr{H}$ is the subcomplex of $Y \times S \subset \Delta^k \times \mathbb{T}^k$:

$$\mathscr{H} = \bigcup Y_J \times S_\sigma$$

such that J does not lie in a single part of σ.

Theorem 1 ([RZ20]). *Let $H = \{1 + y_1 + \cdots + y_k = 0\} \subset (\mathbb{C}^*)^k$ be the $(k-1)$-dimensional pair-of-pants. Then there is a homeomorphism of the pairs*

$$\Phi\colon ((\mathbb{C}^*)^k, H) \to ((\mathbb{C}^*)^k, \mathscr{H}),$$

where $\mathscr{H}$ is the ober-tropical pair-of-pants which is mapped by log *to Y with equidimensional fibers. The homeomorphism restricts well to the boundary under compactifying $(\mathbb{C}^*)^k$ to the product $\Delta \times \mathbb{T}^k$ using the moment map $\mu\colon (\mathbb{C}^*)^k \to \Delta^\circ$ (that maps to the interior of the simplex).*

The homeomorphism Φ in the theorem may be viewed as a deformation the log map so that the image of the pair-of-pants become the tropical hyperplane Y (the spine of the amoeba), rather then the amoeba itself. Now the fibration $X_\tau \to \mathbb{R}^n$ is induced after applying the homeomophism $\bar{\Phi}_i$ on the $(\mathbb{C}^*)^{k_i}$ factor (replacing V by a homeomorphic space $\tilde{V}$). The discriminant of the fibration is precisely $S_x = \bigcup_i S_{x,i}$, see (1), and this has codimension two in $\mathbb{R}^n$.

Finally, to be able to view the fibration $X_\tau \to \mathbb{R}^n$ as a compactification of the smooth fibration $X_0 = T^*B_0/\check{\Lambda} \to B_0$ in a neighbourhood of τ, one needs to replace the log map on the $(\mathbb{C}^*)^\ell$ factor by a suitable (other) moment map μ so that its image is the interior of a polytope rather than all of $\mathbb{R}^n$. A straight forward calculation then shows that the monodromy agrees with the local description in the neighborhood of $x \in B$, see, e.g [GS].

4 Gluing the torus fibration with parameters

Let B be an integral affine manifold with semi-simple polytopal singularities so that each Δ_i is a simplex.

Theorem 2 ([RZ2]). *There is a topological orbifold X which compactifies the torus bundle $X_0 = T^*B_0/\check{\Lambda}$ to a fibration $X \to B$ with n-dimensional fibers (singular over $D \subset B$). If all local cones Σ_τ are unimodular simplicial cones then X is a manifold.*

In fact one can vary the gluing data (a.k.a. B-field) to get a whole family of torus fibrations X_γ over B. The parameter space of gluings is a torsor over $H^1(B, \iota_*\check{\Lambda} \otimes U(1))$. One can make sense of the parameter space itself being $H^1(B, \iota_*\check{\Lambda} \otimes U(1))$ by carefully choosing preferred matching sections for local models - for these additional constructions, we refer to [RZ2]. By the universal coefficient theorem, $H^1(B, \iota_*\check{\Lambda}) \otimes_{\mathbb{Z}} U(1) \subseteq H^1(B, \iota_*\check{\Lambda} \otimes U(1))$ is the component of the identity. We will relate the resulting family

$$\mathscr{X} \to H^1(B, \iota_*\check{\Lambda}) \otimes_{\mathbb{Z}} U(1) \tag{2}$$

to the Gross-Siebert program in the next section.

For the remainder of this section, we are going to carry out the compactification procedure in dimension 3 (and thus for $r = 1$ at all strata of D as we already pointed out in Section 2) and we additionally assume that both Δ_τ and $\check{\Delta}_\tau$ are unimodular simplices. We follow the approach of Gross [G01], that is, we successively compactify the fibration over the star neighborhoods of vertices in D (the barycenters of faces in B) ordered by dimension of the corresponding face in B.

We begin with $X_0 = T^*B_0/\check{\Lambda}$. As mentioned before, there are two types of vertices in D: barycenters of 2-faces (negative vertices if more than bivalent) and barycenters of 1-faces (positive vertices if more than bivalent) in B. We will first compactify over the star neighborhoods (in the $\mathrm{bsd}B$) of the 2-face vertices. Let $x \in D$ be the barycenter of a 2-face Q of B. Then $\check{\Delta}$ is the unit interval and there are two possibilities for Δ_x: the standard 2-simplex or the unit interval.

Case 1: For Δ the unit interval, we have the $(2,2)$-case in [G01] which is the standard focus-focus compactification times a $\mathbb{C}^*$-factor.

Case 2: For Δ the standard 2-simplex, we find D has a trivalent vertex at x and this is referred to as the $(2,1)$-case in [G01]. Let Δ° be the interior of Δ and Y be the union of the 3 intervals in $\mathrm{bsd}\,\Delta$ that connect the barycenter of Δ with the barycenter of an edge of Δ respectively. In a neighborhood of x, the torus bundle $T^*B_0/\check{\Lambda}$ becomes a trivial $\mathbb{T}^2$-bundle once we take the quotient by the coinvariant (vanishing) circle $(\iota_*\check{\Lambda})_x \otimes U(1)$, hence this $\mathbb{T}^2$-bundle extends over the discriminant. Precisely, to glue in the cotangent torus bundle $T^*\Delta^\circ/L^* \cong \Delta^\circ \times \mathbb{T}^2 \cong Q \times (\check{\Lambda}/\check{L} \otimes U(1))$ over the simplex (here L^* is the dual lattice to L) we just need to identify the interior of the simplex Δ° with the cell Q. Let Q_1, Q_2, Q_3 be the 3 boundary intervals of Q which meet D.

Lemma 1. *There is a homeomorphism $\psi\colon (\Delta^\circ, Y) \to (Q, D \cap Q)$ which extends to a homeomorphism including the 3 boundary intervals of Δ and identifying these with the 3 boundary intervals Q_1, Q_2, Q_3 of Q.*

Thus we have a well-defined (trivial) $\mathbb{T}^2$-bundle (the quotient by the $\check{L}$-circle) over the star neighborhood of x in $\mathrm{bsd}B$. Our next step is to compactify the circle bundle over the 5-dimensional manifold $((\Delta^\circ \times \mathbb{T}^2) \times \mathbb{R}) \setminus (\mathscr{H} \times \{0\})$ to a fibration

over $\Delta^\circ \times \mathbb{T}^2 \times \mathbb{R}$. Here, $\mathscr{H} \subset (\mathbb{C}^*)^2 \cong \Delta^\circ \times \mathbb{T}^2$ denotes the ober-tropical pair-of-pants from [RZ20], a 2-dimensional submanifold of $\Delta^\circ \times \mathbb{T}^2 \times \mathbb{R}$. We can either glue in the local model X_Q from the previous section, or use the following proposition, leading to a homeomorphic result:

Lemma 2 (cf. [G01], Proposition 2.5). *Let U be the complement of an oriented connected submanifold S of codimension 3 in a manifold $\bar{U}$ and let $\pi\colon X \to U$ be a principal S^1-bundle with the Chern class $c_1 = \pm\kappa$ in $H^3_S(\bar{U},\mathbb{Z}) \cong H^0(S,\mathbb{Z})$ for some $\kappa > 0$. Then there is a unique compactification to an orbifold $\bar{X} = X \cup S$ such that $\bar{\pi}\colon \bar{X} \to \bar{U}$ is a proper map and $\bar{X}$ is a manifold if $\kappa = 1$.*

Our local description of the monodromy implies that κ coincides with the lattice length of $\check{\Delta}_x$ which we assumed to be one for this article. As pointed out by [G01] already, changing the orientation of the S^1-action, changes the sign of c_1. The label "negative" for the vertex x of D doesn't stem from the sign of c_1 but the Euler number of the local model X_Q which is -1 (and the same as the Euler number of the fiber over x).

Now comes the important step of **unwiggling** the ober-tropical fiber tori for extending the compactification over the rest of D. Figure 4 shows the $\mathbb{T}^2$-fibers over different points in Δ° indicated by little squares. The red locus inside each square is the intersection of $\mathscr{H}$ with the corresponding $\mathbb{T}^2$-fiber, the "ober-tropical fibers" over points of Y. As we move away from $x \in D$ we deform the ober-tropical fibers so that they become more and more straight circles. Outside the second barycentric star of x (the shaded region) they are true linear circles in $\mathbb{T}^2$ and are ready to be glued with the neighboring model. The S^1-fibration over $\Delta^\circ \times \mathbb{T}^2 \times \mathbb{R}$ collapses precisely over the red circles in the $\mathbb{T}^2$-fibers over $Y \times \{0\} \subset \Delta^\circ \times \mathbb{R}$.

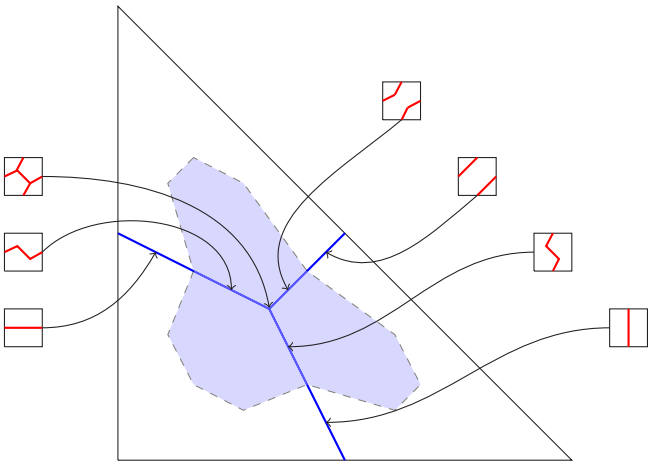


Fig. 4 Fading off the wiggling of red circles along $Y \subset \Delta^\circ$.

The main point of unwiggling is to achieve the following property: close to the boundary of the cell Q the constructed space X may not only be thought not only

as an S^1-fibration over $Q\times\mathbb{R}\times\mathbb{T}^2$ but also as a $\mathbb{T}^2$-fibration over $Q\times\mathbb{R}\times S^1$ via taking the quotient by the circle in the base $\mathbb{T}^2$ that is the homotopy class of the respective red ober-tropical fiber circle. From this perspective, the fibers over $D\times S^1$ are pinched tori (homeomorphic to I_1-degenerate elliptic curves). This helps us to do the last step, namely compactify the fibration over the vertices of D which are barycenters of the 1-dimensional strata in B.

Let us finally discuss the compactification over a vertex $x\in D$ that is the barycenter of a one-cell in B. At this barycenter, Δ is the unit interval. If $\check\Delta$ is also a unit interval then we are back to the focus-focus $(2,2)$-case which is straightforward to compactify, so assume $\check\Delta$ is a standard 2-simplex, this is the $(1,2)$-case in [G01].

We state a more general result that relates back to Figure 3. Let Σ be a cone in $\mathbb{R}^k\times\mathbb{R}$ over a lattice simplex $\check\Delta\subset\mathbb{R}^k\times\{1\}$, and let $\Sigma^\vee$ be the dual cone. The projection $\mathbb{R}^k\times\mathbb{R}\to\mathbb{R}$ gives a linear map of cones $w\colon\Sigma\to\mathbb{R}_{\geq 0}$ which, in turn, defines a map of affine toric varieties $z^w\colon U_\Sigma=\operatorname{Spec}\mathbb{C}[\Sigma^\vee_{\mathbb{Z}}]\to\mathbb{C}$. Let $\mu_0\colon U_\Sigma\to\mathbb{R}^k$ be the moment map with respect to the $\mathbb{T}^k$-action on U_Σ which fixes z^w.

We consider the $\mathbb{T}^k$-torus fibration $\pi\colon U_\Sigma\to\mathbb{R}^k\times\mathbb{C}$ given by $z\mapsto(\mu_0,z^w)$. Recall that $\check Y\subset\mathbb{R}^k$ stands for the codimension 1 skeleton of the normal fan to $\check\Delta$. Away from $\check Y\times\{0\}$, the map π is a $\mathbb{T}^k$-bundle with the monodromy prescribed by the pair $(\Delta=[0,1],\check\Delta)$. Over the strata of $\check Y$, the fibers of π become lower-dimensional tori (reflecting the dimension of the stratum) with the fiber over $\{0\}$ being just a single point. We denote by $\pi_0: U_\Sigma\setminus\{0\}\to(\mathbb{R}^k\times\mathbb{C})\setminus\{(0,0)\}$ the restriction of the fibration π to the complement of the origin.

Lemma 3 (cf. [G01], Proposition 2.9, for the 3-dimensional case). *Let $X\to(\mathbb{R}^k\times\mathbb{C})\setminus\{(0,0)\}$ be a torus fibration homeomorphic to π_0. There is a unique one point compactification to an orbifold $\bar X=X\cup\{pt\}$ such that $\bar\pi_0\colon\bar X\to\mathbb{R}^k\times\mathbb{C}$ is a proper map. Consequently, the fibration $\bar\pi_0\colon\bar X\to\mathbb{R}^k\times\mathbb{C}$ is homeomorphic to $\pi\colon U_\Sigma\to\mathbb{R}^k\times\mathbb{C}$ and $\bar X$ is a manifold if $\check\Delta$ is a unimodular simplex.*

There is a generalization of this statement when the pair $(\mathbb{C},\{0\})$ is replaced by a pair (U,S) of S being a submanifold in U of codimension 2. This, in particular, covers Lemma 1 as a special case $k=1$. The relevant case for us is $k=2$, $U=\mathbb{R}\times(\mathbb{R}/\mathbb{Z})$ and S is the point $(0,0)\in\mathbb{R}\times(\mathbb{R}/\mathbb{Z})$. We identify $S^1=\mathbb{R}/\mathbb{Z}$ and refer to 0 as the corresponding point in S^1 in the following.

First, we note that similar to the $(2,1)$-vertex, the torus bundle $T^*B_0/\check\Lambda$ in a neighborhood W of x becomes a trivial S^1-bundle once we take the quotient by the coinvariant (vanishing) $\mathbb{T}^2$-subbundle $(\iota_*\check\Lambda)_x\otimes S^1$, thus it extends over D. Second, we may view the torus bundle $T^*(B_0\cap W)/\check\Lambda$ as a $\mathbb{T}^2$-bundle over $(W\setminus D)\times S^1$.

Lemma 4. *The $\mathbb{T}^2$-bundle over $(W\setminus D)\times S^1$ extends to a singular $\mathbb{T}^2$-fibration $\pi_0\colon X\to(W\times S^1)\setminus(x\times\{0\})$ by adding the I_1-fibers over $(\check Y\setminus x)\times S^1$. The resulting fibration X agrees with those coming from the neighboring $(2,1)$-vertices of D after the unwiggling. Moreover, the fibration $X\to(W\times S^1)\setminus(x\times\{0\})$ is homeomorphic to π_0, so satisfies the hypothesis of Lemma 3.*

Applying Lemma 3, the space X compactifies to $\bar X\to W\times S^1$ by adding the point $x\times\{0\}$. This completes the compactification process.

Finally we briefly comment on the parameter space $H^1(B,\iota_*\check{\Lambda}\otimes U(1))$ of glu-ings. Already in building the smooth part $T^*B_0/\check{\Lambda}$ one can twist by a Čech cocycle representing an element in $H^1(B_0,\check{\Lambda}\otimes U(1))$. Furthermore when gluing in the local models around the vertices the twisting can be made when identifying the $\mathbb{T}^3$-fibers of $T^*B_0/\check{\Lambda}$ with the S^1-bundle over $\mathbb{T}^2$ (for the $(2,1)$-vertices) or $\mathbb{T}^2$-bundles over S^1 (for the $(1,2)$-vertices). Lastly, when identifying the models between $(2,1)$ and $(1,2)$-vertices there is only $\mathbb{T}^2$-freedom of twistings which corresponds to the sheaf $\iota_*\check{\Lambda}$ dropping the rank along the edges of D.

5 Canonical Calabi-Yau families and their Kato-Nakayama spaces

Recall from [GS06, Theorem 5.2, Theorem 5.4] and [GS10, Remark 5.3] that, given an integral affine manifold with simple singularities B and a compatible polyhedral decomposition $\mathscr{P}$ with multivalued strictly convex piecewise affine function φ with integral slopes, there is an associated algebraic family of toric log Calabi-Yau spaces[1]

$$X_0(B,\mathscr{P},\varphi)\to S:=\operatorname{Spec}\mathbb{C}[H^1(B,\iota_*\check{\Lambda})^*] \tag{3}$$

which is semi-universal by [RS, Theorem C.6]. Analytification and application of the Kato-Nakayama functor associates to the log morphism (3) a continuous surjection of topological spaces

$$\mathscr{X}''\to\mathscr{S}'':=\left(H^1(B,\iota_*\check{\Lambda})\otimes_{\mathbb{Z}}\mathbb{C}^*\right)\times U(1)$$

which is a fiber bundle by [NO10, Theorem 5.1]. This family has also been studied with regards to its real locus in [AS]. We restrict the family to

$$\mathscr{X}'\to\mathscr{S}':=\left(H^1(B,\iota_*\check{\Lambda})\otimes_{\mathbb{Z}}U(1)\right)\times U(1). \tag{4}$$

Let $c_1(\varphi)$ denote the class of φ in $H^1(B,\iota_*\check{\Lambda})$. The inclusion $\mathbb{Z}c_1(\varphi)\subseteq H^1(B,\iota_*\check{\Lambda})$ induces a map of real Lie groups

$$\phi_\varphi\colon U(1)\to H^1(B,\iota_*\check{\Lambda})\otimes_{\mathbb{Z}}U(1)$$

where we have identified $(\mathbb{Z}c_1(\varphi))\otimes U(1)=U(1)$.

As explained in [RS, §4.1], there is an equivariant $U(1)$ action on the family (4) and for the base space, by [RS, (4.14)], it is given by

$$U(1)\times\mathscr{S}'\to\mathscr{S}',\quad \lambda.(s,t)=\left(\phi_\varphi(\lambda)\cdot s,\lambda^{-1}\cdot t\right).$$

Consequently, the family (4) is a base change of the restricted family $\mathscr{X}\to\mathscr{S}:=H^1(B,\iota_*\check{\Lambda})\otimes_{\mathbb{Z}}U(1)$ under the base change homomorphism

[1] We have implicitly picked a splitting of the surjection $H^1(B,\iota_*\check{\Lambda})\to H^1(B,\iota_*\check{\Lambda})/H^1(B,\iota_*\check{\Lambda})_{\text{tors}}$.

$$\mathrm{id}\times\phi_\varphi\colon \left(H^1(B,\iota_*\check{\Lambda})\otimes_{\mathbb{Z}} U(1)\right)\times U(1)\to H^1(B,\iota_*\check{\Lambda})\otimes_{\mathbb{Z}} U(1).$$

The relevant topological information is therefore already contained in $\mathscr{X}\to\mathscr{S}$. In [RS, Section 2.1], a moment map $X_0(B,\mathscr{P},\varphi)\to B$ was given under the assumption that $X_0(B,\mathscr{P},\varphi)$ is projective over S. Since we restricted to the $U(1)$-part of the gluing torus when taking (4), a moment map exists even without the projectivity condition. Composing with the log forget morphism yields a fibration

$$\pi\colon \mathscr{X}\to B$$

whose discriminant has codimension one (being a union of amoebae in real hyperplanes). In our upcoming work [RZ2], we are going to prove the following result which may be viewed as deforming π near the discriminant so that the new discriminant has codimension two. In other words, up to this deformation, the Gross-Siebert fibration π agrees with the compactification of the Strominger-Yau-Zaslow fibration constructed in the previous sections.

Theorem 3 ([RZ2]). *The topological family $\mathscr{X}\to\mathscr{S}$ is homeomorphic to the family of compactified torus bundles given in* (2) *(over the identity on $\mathscr{S}$).*

The homomorphism in the theorem commutes with the respective torus fibration maps to B away from a tubular neighbourhood of the discriminant $D\subset B$.

Acknowledgements We are indebted to Mark Gross and Bernd Siebert for sharing their ideas and unpublished notes on the subject, parts of which will enter [RZ2]. Our gratitude for hospitality goes to Mittag-Leffler Institute, Oberwolfach MFO, University of Miami, MATRIX Institute, JGU Mainz and Kansas State University.

References

AAK. Abouzaid, M., Auroux, D. Katzarkov, L.: "Lagrangian fibrations on blowups of toric varieties and mirror symmetry for hypersurfaces", Publ.math.IHES **123**, 199–282 (2016), https://doi.org/10.1007/s10240-016-0081-9

AGIS. Abouzaid, M., Ganatra, S., Iritani, H., Sheridan, N.: "The Gamma and Strominger-Yau-Zaslow conjectures: a tropical approach to periods", https://arxiv.org/abs/1809.02177

AS. Hülya Argüz, H., Siebert, B.: "On the real locus in the Kato-Nakayama space of logarithmic spaces with a view toward toric degenerations", https://arxiv.org/abs/1610.07195

Au07. Auroux, D.: "Mirror symmetry and T-duality in the complement of an anticanonical divisor", J. Gökova Geom. Topol. 1 (2007), 51–91.

Au09. Auroux, D.: "Special Lagrangian fibrations, wall-crossing, and mirror symmetry", Surveys in differential geometry. Vol. XIII. Geometry, analysis, and algebraic geometry: forty years of theJournal of Differential Geometry, Surv. Differ. Geom., vol. 13, Int. Press, Somerville, MA, 2009, 1–47.

CBM09. Castaño Bernard, R., Matessi, D.: "Lagrangian 3-torus fibrations", J. Differential Geom. **81** (2009), 483–573.

Ch. Chan, K.: "The Strominger-Yau-Zaslow conjecture and its impact", Selected Expository Works of Shing-Tung Yau with Commentary. Vol. II", 1183–1208, Adv. Lect. Math. (ALM) 29, Int. Press, Somerville, MA, 2014.

CLL. Chan, K., Lau, S.-C., Leung, N. C.: "SYZ mirror symmetry for toric Calabi-Yau manifolds", J. Differential Geom.90, no. 2, (2012), 177–250.

CL. Chan, K., Leung, N. C.: "Mirror symmetry for toric Fano manifolds via SYZ transformations", Adv. Math. 223 no. 3, (2010), 797–839.

EM19. Evans, J., Mauri, M.: "Constructing local models for Lagrangian torus fibrations", https://arxiv.org/abs/1905.09229.

FLTZ12. Fang, B., Liu, C-C M., Treumann, D., Zaslow, E.: "T-duality and homological mirror symmetry for toric varieties", Adv. Math.229, no. 3, (2012), 1875–1911.

G00. Gross, M.: "Examples of special Lagrangian fibrations", Symplectic geometry and mirror symmetry (Seoul, 2000), World Sci. Publ., River Edge, NJ, 2001, 81–109.

G01. Gross, M.: "Topological mirror symmetry", Inv. Math., volume 144 (2001), 75–137.

G09. Gross, M.: "The Strominger-Yau-Zaslow conjecture: from torus fibrations to degenerations", Algebraic geometry, Seattle 2005. Part 1, 149–192, Proc. Sympos. Pure Math. 80, Part 1, Amer. Math. Soc., Providence, RI, 2009.

G12. Gross, M.: "Mirror symmetry and the Strominger–Yau–Zaslow conjecture", Current developments in mathematics 2012, Int. Press, Somerville, MA, 2013, pp. 133–191.

GW00. Gross, M., Wilson, P. M. H.: "Large Complex Structure Limits of K3 Surfaces", J. Differential Geom., Volume 55, Number 3 (2000), 475–546.

GS11. Gross, M., Siebert, B.: "From real affine geometry to complex geometry", Annals of Math. **174**, 2011, 1301–1428.

GS06. Gross, M., Siebert, B.: "Mirror symmetry via logarithmic degeneration data I", J. Differential Geom. **72**, 2006, 169–338.

GS10. Gross, M., Siebert, B.: "Mirror symmetry via logarithmic degeneration data II", J. Algebraic Geom. **19**, 2010, 679–780.

GTZ. Gross, M., Tosatti, V., Zhang, Y.: "Collapsing of abelian fibred Calabi-Yau manifolds", Duke Math. J., Volume 162, Number 3, (2013), 517–551.

HZ05. Haase, C., Zharkov, I.: "Integral affine structures on spheres: complete intersections", Int. Math. Res. Not. 2005, no. 51, 3153–3167.

H19. Hicks, J.: "Tropical Lagrangians and Homological Mirror Symmetry", https://arxiv.org/abs/1904.06005

J03. Joyce, D.: "Singularities of special Lagrangian fibrations and the SYZ conjecture", Comm. Anal.Geom. 11(5), 2003, 859–907.

KS00. Kontsevich, M., Soibelman Y., "Homological mirror symmetry and torus fibrations", In: "Symplectic geometry and mirror symmetry", Seoul, 2000, 203–263.

KS06. Kontsevich, M., Soibelman Y., "Affine structures and non-archimedean analytic spaces", In: Etingof P., Retakh V., Singer I.M. (eds), The Unity of Mathematics. Progress in Mathematics, vol 244. Birkhäuser Boston.

Leung. Leung, N.C.: "Mirror Symmetry Without Corrections", Communications in Analysis and Geometry, 13(2), 2001.

LYZ. J. Loftin, J. Yau, S.-T., Zaslow, E.: "Affine manifolds, SYZ geometry and the "Y" vertex", J. Differential Geom.71, no. 1, (2005), 129–158.

MR. Mak, C.Y., Ruddat, H.: "Tropically constructed Lagrangians in mirror quintic threefolds", https://arxiv.org/abs/1904.11780.

Mat. Matessi, D.: "Lagrangian pairs of pants", https://arxiv.org/abs/1802.02993.

Mi04. Mikhalkin, G.: "Decomposition into pairs-of-pants for complex algebraic hypersurfaces", Topology Vol. 43 (2004), Issue 5, 1035-1065.

Mi19. Mikhalkin, G.: "Examples of tropical-to-Lagrangian correspondence", European Journal of Mathematics 5, 2019, 1033–1066.

NO10. Nakayama, C., Ogus, A.: "Relative rounding in toric and logarithmic geometry", Geom. & Topol. **14**(4), 2010, 2189–2241.

NXY. Nicaise, J., Xu, C., Yu, T.Y.: "The non-archimedean SYZ fibration", Compos. Math. 155, Issue 5, 2019, 953–972.

Pa07. Parker, B.: "Exploded fibrations", Proceedings of 13th Gökova Geometry-Topology Conference, 52–90.

P18. Prince, T.: "Lagrangian torus fibration models of Fano threefolds" , https://arxiv.org/abs/1801.02997.

Ru20. Ruddat, H.: "A homology theory for tropical cycles on integral affine manifolds and a perfect pairing", https://arxiv.org/abs/2002.12290.

RS. Ruddat, H., Siebert, B.: "Period integrals from wall structures via tropical cycles, canonical coordinates in mirror symmetry and analyticity of toric degenerations", Publ.math.IHES (2020), https://doi.org/10.1007/s10240-020-00116-y.

RZ20. Ruddat, H., Zharkov, I.: "Tailoring a pair-of-pants", https://arxiv.org/abs/2001.08267.

RZ2. Ruddat, H., Zharkov, I.: "Topological Strominger-Yau-Zaslow fibrations", in preparation.

Rua. Ruan, W.-D.: "Lagrangian torus fibrations and mirror symmetry of Calabi–Yau manifolds", Symplectic geometry and mirror symmetry (Seoul, 2000), World Sci. Publ., River Edge, NJ, 2001, 385–427.

SS18. Sheridan, N., Smith, I.: "Lagrangian cobordism and tropical curves", https://arxiv.org/abs/1805.07924.

SYZ96. Strominger A., Yau S.-T., Zaslow, E.: "Mirror symmetry is T-duality", Nuclear Phys. B 479 (1996), no. 1-2, 243-259.

T06. Thomas, R. P.: "The Geometry of Mirror Symmetry", Encyclopedia of Mathematical Physics. Elsevier, (2006), 439–448.

Zh00. Zharkov, I.: "Torus fibrations of Calabi-Yau hypersurfaces in toric varieties", Duke Math. J., 101:2, 2000, 237–258.

Part II

Other Contributed Articles

Chapter 10

Topology of Manifolds: Interactions Between High And Low Dimensions

Graphical neighborhoods of spatial graphs

Stefan Friedl and Gerrit Herrmann

Abstract We give a definition of a graphical neighborhood of a spatial graph which generalizes the tubular neighborhood of a link in S^3. Furthermore we prove existence and uniqueness of graphical tubular neighborhoods.

1 Introduction

In this paper we give a precise definition of the notion of a spatial graph. In our opinion the goals of any good definition in knot theory and related subjects should be two-fold:

1. the definition should be flexible enough to encompass all "physical" objects that one has in mind,
2. the definition should be rigid enough to allow for a reasonable theory.

At least to our taste the definitions of a spatial graph used in the literature are often vague or fall short of (1) or (2). For example, a spatial graph is often defined as "an embedded (topological) graph in S^3". Since a topological graph (i.e. a finite 1-dimensional CW-complex) is in general not a manifold it is not entirely clear what the word "embedded" should really mean in this context.

We start out with a precise definition of a "spatial graph". We hope that the reader will be convinced that it satisfies (1). Afterwards we will attempt to show "spatial graphs" in our sense satisfy (2). More precisely, we will show that spatial graphs admit a graphical neighborhood, which is unique in an appropriate sense, which makes

Stefan Friedl
Fakultät für Mathematik, Universität Regensburg, 93040 Regensburg, Germany
e-mail: sfriedl@gmail.com

Gerrit Herrmann
Fakultät für Mathematik, Universität Regensburg, 93040 Regensburg, Germany
e-mail: mi@gerhit.de

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_38

it possible to sensibly study spatial graphs. These results had been announced, without proofs, in [2].

Before we turn to the definition of a spatial graph we recall the notion of an abstract graph.

Definition 1. An *abstract graph* G is a triple (V,E,φ) where V is a finite non-empty set, E is a finite set and φ is a map

$$\varphi\colon E \to \{\text{subsets of } V \text{ with one or two elements}\}.$$

The elements of V are called *the vertices of* G and the elements of E are called the *edges of* G. Furthermore, given $e \in E$ the elements of $\varphi(e) \subset V$ are called the *endpoints of* e.

We turn to topology.

Definition 2. An *arc* in S^3 is a subset E of S^3 for which there exists a map $\varphi\colon [0,1] \to S^3$ with the following properties:

1. the map φ is smooth, i.e. all derivatives are defined on the open interval $(0,1)$ and they extend to continuous maps on the closed interval $[0,1]$ that we also call derivatives,
2. the first derivative $\varphi'(t)$ is non-zero for all $t \in [0,1]$,
3. the restriction of φ to $(0,1)$ is injective,
4. $\varphi((0,1)) \cap \varphi(\{0,1\}) = \varnothing$ and
5. $\varphi((0,1)) = E$.

Given an arc E as above we refer to φ and $\varphi(1)$ as the *endpoints of* E. (Note that the endpoints of E do *not* lie in E.)

Fig. 1 Illustration of arcs with one or two endpoints.

Definition 3. A *spatial graph* G is a pair (V,E) with the following properties:

1. V is a finite non-empty subset of S^3.
2. E is a subset of S^3 with the following properties:
   a. E is disjoint from V,
   b. E has finitely many components,
   c. each component of E is an arc and the endpoints of each arc lie in V.

We refer to the points in V as the *vertices of* G and we refer to the components of E as the *edges of* G. Furthermore, given a spatial graph $G = (V,E)$ we write $|G| = V \cup E \subset S^3$.

Note that for a spatial graph (V,E) as above the corresponding triple

$$(V, \pi_0(E), \varphi(\text{arc}) := \text{endpoints of the arc})$$

is an abstract graph. Also note that a spatial graph admits an obvious CW-structure, i.e. the underlying topological space is indeed a topological graph in the above sense.

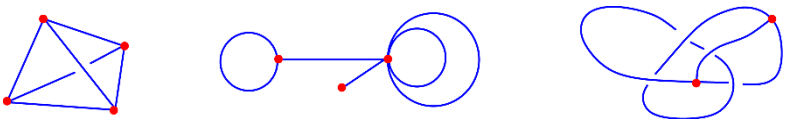


Fig. 2 Examples of spatial graphs.

As mentioned above, a good definition in topology should be flexible enough to capture the examples one has in mind, but it should also be rigid enough to allow for a sensible theory. One can easily convince oneself that the Figure 2 gives three examples of spatial graphs, so our definition seems to be broad enough to capture all "reasonable" examples. On the other hand, one of the key tools in the study of knots and links is the tubular neighborhood. The goal of the remainder of this paper is to show that our notion of a spatial graph is rigid enough to define an analogue of the tubular neighborhood of a knot or link.

More precisely, in Section 3 we introduce the notion of a "graphical neighborhood" of a spatial graph. The following three theorems summarize the key properties of graphical neighborhoods.

Theorem 1. *Every spatial graph admits a graphical neighborhood.*

Theorem 2. *Let $G = (V,E)$ be a spatial graph and let N be a graphical neighborhood for G.*

1. *N contains $|G| = V \cup E$ in the interior $N^o = N \setminus \partial N$ of N,*
2. *$|G|$ is a deformation retract of N,*
3. *∂N is a deformation retract of $N \setminus |G|$,*
4. *the exterior $E_G = S^3 \setminus N^o$ is a compact 3-dimensional manifold that is a deformation retract of $S^3 \setminus |G|$.*

Since every compact 3-dimensional manifold admits a finite CW-structure we obtain the following corollary to Theorem 2 (4).

Corollary 1. *Given a spatial graph G the fundamental group $\pi_1(S^3 \setminus G)$ is finitely presented and all homology groups $H_*(S^3 \setminus G)$ are finitely generated.*

The following theorem concludes our list of three theorems dealing with the key properties of graphical neighborhoods.

Theorem 3. *Any two graphical neighborhoods of a given spatial graph are equivalent* (*see Section 5 for the precise statement*).

The following corollary is an immediate consequence of Theorem 3.

Corollary 2. *Let G be a spatial graph and let N be a graphical neighborhood for G. The diffeomorphism type of the exterior $E_G := S^3 \setminus N^o$ does not depend on the choice of a graphical neighborhood N.*

Definition 4. We say that two spatial graphs $G = (V,E)$ and $G' = (V',E')$ are *equivalent* if there exists an orientation-preserving homeomorphism $\Psi\colon S^3 \to S^3$ with $\Psi(V) = V'$, $\Psi(E) = E'$ and which restricts to a diffeomorphism $S^3 \setminus V \to S^3 \setminus V'$.

The following lemma relates graphical neighborhoods of spatial graphs.

Lemma 1. *Let $G = (V,E)$ and $G' = (V',E')$ be two spatial graphs.*

1. *Let $\Psi\colon S^3 \to S^3$ be an orientation-preserving homeomorphism with $\Psi(V) = V'$ and $\Psi(E) = E'$ and which restricts to a diffeomorphism $S^3 \setminus V \to S^3 \setminus V'$. If N is a graphical neighborhood for G, then $\Psi(N)$ is a a graphical neighborhood of G'.*
2. *If G and G' are equivalent, then the exteriors of G and G' are diffeomorphic.*

Proof. The lemma follows immediately from the definitions and the following basic fact: if $f\colon \overline{B}^3 \to \mathbb{R}^3$ is a homeomorphism onto its image such that the restriction of f to $\overline{B}^3 \setminus \{0\}$ is a diffeomorphism onto its image, then $f(\overline{B}^3)$ is a submanifold of $\mathbb{R}^3$ that is diffeomorphic to the closed 3-ball, even though $f\colon \overline{B}^3 \to f(\overline{B}^3)$ is not necessarily a diffeomorphism. □

We conclude this introduction with a few remarks:

Remark 1. 1. It is also interesting to consider unions $G \sqcup L$ where G is a spatial graph and L is a link. A graphical neighborhood in this setting is defined as the union $Z \sqcup W$ where Z is a graphical neighborhood for Z and W is a tubular neighborhood for the submanifold L. All of the previous results also hold in that more general context.

2. The theory of graphical neighborhoods works basically the same for spatial graphs in any closed orientable 3-manifold. For simplicity's sake we only deal with the most important case, namely the case of spatial graphs in S^3.

3. The reader might be surprised to note how much effort we spend in our proofs on ensuring that maps are actually smooth and not just continuous. Even though in 3-dimensional topology we have Moise's Theorem which says that any two 3-manifolds that are homeomorphic are also diffeomorphic, this does not imply that analogous statements hold if one wants to keep more control over subsets. For example, consider the two three spatial graphs G, G' and G'' that are shown in Figure 3. They are equivalent in our sense, but there is no self-diffeomorphism h of S^3 that turns any of the spatial graphs into any of the other two spatial graphs.

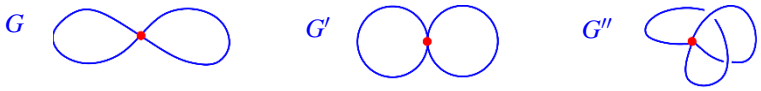


Fig. 3

4. Some authors on spatial graphs write that they work in the PL-category and use the notion of a regular neighborhood, that is for example discussed in [7, Chapter 3], [4, p. 7f] or [3, Chapter III.B]. A regular neighborhood is a much more general concept than a graphical neighborhood. The regular neighborhood of a spatial graph is unique in an appropriate sense (see [7, Theorem 3.24] or alternatively [3, Theorem II.16n]) and by [7, Corollary 3.30] the analogue of Theorem 2 (2) holds. With some effort one can use [7, Corollary 3.18] to show that (3), and thus also (4), are satisfied. But it takes some dedication to understand what a "regular neighborhood" is really supposed to be and at first glance it is not entirely clear that the various definitions given in [7, 3, 4] are actually consistent. At least to the authors it seems like working in the smooth category and working with our graphical neighborhoods is esthetically more pleasing and "closer to reality". On the other hand, for implementing algorithms it seems more reasonable to work in the PL-category.

The paper is organized as follows. In Section 2 we recall the existence and uniqueness of tubular neighborhoods of 1-dimensional submanifolds of 3-dimensional manifolds. In Section 3 we prove Theorem 1 and in Section 4 we provide a proof for Theorem 2. Finally in Section 5 we deal with the hardest part of the paper, namely we prove Theorem 3.

2 Tubular neighborhoods

Before we get started with the technical details we would like to introduce some conventions, definitions and notations:

1. By a manifold we mean a topological manifold equipped with a smooth structure. Every manifold is assumed to be compact and orientable unless we say otherwise. Throughout this paper we try to follow the definitions and conventions of [9].
2. Given a homotopy $F\colon X\times[0,1]\to Y$ and $t\in[0,1]$ we denote by $F_t\colon X\to Y$ the map that is given by $F_t(x)=F(x,t)$.
3. Let M be a manifold. A *diffeotopy of* M is a smooth map $F\colon M\times[0,1]\to M$ such that each $F_t\colon M\to M$ is a diffeomorphism.
4. Given a topological space X and a subset A we denote by A^o its interior and we denote by $\overline{A}$ its closure.
5. As usual we make the identification $S^3=\mathbb{R}^3\cup\{\infty\}$.
6. Given $r\in\mathbb{R}_{\geq 0}$ we denote by $B^3_r\subset\mathbb{R}^3$ the open ball of radius r around the origin. We write $S^2_r=\partial\overline{B^3_r}$.

7. Given $0 < s < t$ we identify $S^2 \times [s,t]$ with $\{z \in \mathbb{R}^3 \,|\, s \leq \|z\| \leq t\} = \overline{B^3_t} \setminus B^3_s$ in the obvious way.

In the next section we will introduce the concept of a graphical neighborhood of a spatial graph. It will build on the notion of a tubular neighborhood of a 1-dimensional submanifold.

Definition 5. 1. Let M be a 3-manifold. A *proper submanifold* of M is a compact submanifold C of M with $\partial C = C \cap \partial M$ and that meets ∂M transversally.
2. A *tubular neighborhood* for a proper 1-dimensional submanifold C of M is an embedding $F\colon C \times \overline{B}^2 \to M$ with the following properties:
   a. we have $F(\partial C \times \overline{B}^2) = F(C \times \overline{B}^2) \cap \partial M$,
   b. the image $F(C \times \overline{B}^2)$ is a submanifold of M with corners (see [9, p. 30] and [1, Chapter 86] for the definition of a submanifold with corners),
   c. there exists a collar neighborhood $\partial M \times [0,1]$ such that the tubular neighborhood of $C \cap (\partial M \times [0,1])$ is a product, i.e. we have

$$(\partial M \times [0,1]) \cap \Phi(C \times \overline{B}^2) = \Phi((\partial M \cap C) \times \overline{B}^2) \times [0,1].$$

Remark 2. Note that a tubular neighborhood N of a 1-dimensional submanifold C with non-empty boundary is a submanifold of M with non-empty corners, in particular N is strictly speaking not a smooth submanifold. Fortunately in practice this is not a problem. For example we are mostly interested in considering the exterior $E_C := M \setminus N^o$ where N^o is the interior of N. The exterior E_C is a smooth manifold with corner, but by "straightening of corners", see [9, Proposition 2.6.2] we can view $E_C = M \setminus N^o$ as a smooth manifold in a canonical way.

The following two theorems show the existence and uniqueness of tubular neighborhoods in our setting.

Theorem 4. *Every proper 1-dimensional submanifold C of every 3-manifold M admits a tubular neighborhood $F\colon C \times \overline{B}^2 \to M$.*

Proof. This theorem is basically a consequence of [9, Theorem 2.3.3]. We have the extra condition (2c) in the definition of a tubular neighborhood, which is not explicitly mentioned in [9], but using [9, Proposition 1.5.6] one can see that this condition can also be arranged. □

Theorem 5. *[9, Chapter 2.5] Let M be a 3-manifold and let C be a proper 1-dimensional submanifold of M. If $F, G\colon C \times \overline{B}^2 \to M$ are two tubular neighborhoods of C, then there exists a diffeotopy $\Phi\colon M \times [0,1] \to M$ rel C with the following properties:*

1. *$\Phi_t = \mathrm{id}_M$ for small t,*
2. *the restriction of Φ_1 to $F(C \times \overline{B}^2)$ defines a fiber-preserving diffeomorphism from $F(C \times \overline{B}^2)$ to $G(C \times \overline{B}^2)$.*

3 Definition and existence of graphical neighborhoods

Definition 6. Let $G = (V, E)$ be a spatial graph.

1. We say that an orientation-preserving map $\Theta\colon \overline{B^3_R} \to S^3$ is *transverse at* $v \in V$ if Φ is an embedding, if $\Phi = v$ and if for each $r \in (0, R]$ the image $\Theta(S^2_r)$ is a submanifold of $S^3 \setminus V$ that is transverse to the submanifold E of $S^3 \setminus V$.
2. A *small neighborhood of* V is a compact 3-dimensional submanifold X of S^3 with components $\{X_v\}_{v \in V}$ such that for each $v \in V$ there exists a map $\Theta\colon \overline{B^3_R} \to S^3$ that is transverse to v with $\Theta(\overline{B^3_r}) = X_v$.

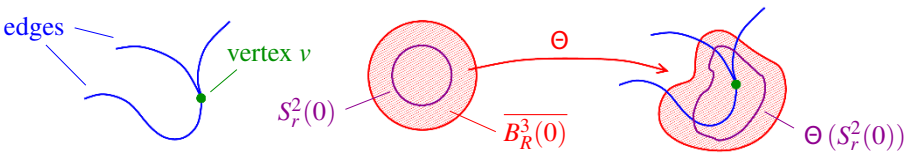


Fig. 4 Illustration of a small neighborhood.

We point out that $E \cap (S^3 \setminus X^o)$ is a proper submanifold of $S^3 \setminus X^o$. Now we can define a graphical neighborhood of a spatial graph.

Definition 7. Let $G = (V, E)$ be a spatial graph. A *graphical neighborhood of* (V, E) is a subset N of S^3 that can be written as a union $N = X \cup Y$ where X is a small neighborhood of V and Y is a tubular neighborhood of $E \cap (S^3 \setminus X^o)$ in $S^3 \setminus X^o$.

Remark 3. As remarked above, after "straightening of corners", we can view $S^3 \setminus N^o$ as a smooth manifold in a canonical way.

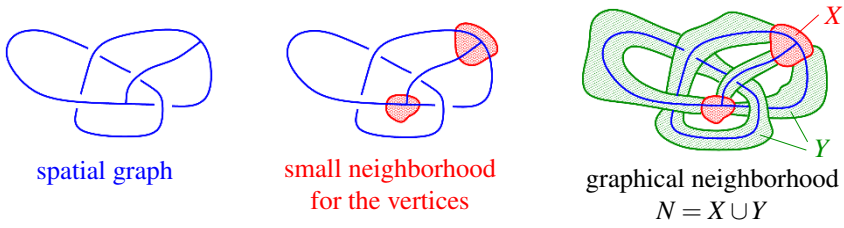


Fig. 5

In the following we will see that every spatial graph admits a graphical neighborhood and that graphical neighborhoods have properties that are very similar to the properties of tubular neighborhoods.

We now prove that every spatial graph admits a graphical neighborhood. This statement should be viewed as the analogue of Theorem 4.

Theorem 1. *Every spatial graph admits a graphical neighborhood.*

Proof. Let $G=(V,E)$ be a spatial graph. First we show that V admits a small neighborhood. Since we can always shrink small neighborhoods it suffices to show that each vertex admits a small neighborhood. Let v be a vertex. To simplify the notation we might as well assume that $v=0\in\mathbb{R}^3\cup\{\infty\}$. By definition of a spatial graph there exist smooth injective maps $\varphi_i\colon[0,\frac{1}{4}]\to S^3$, $i=1,\dots,k$, with the following properties:

1. For each $i\in\{1,\dots,k\}$ we have $\varphi_i=0$ and $\varphi_i(\frac{1}{4})\neq 0$,
2. for each $i\in\{1,\dots,k\}$ we have $\varphi_i'(t)\neq 0$ for all $t\in[0,\frac{1}{4}]$,
3. there exists an $s>0$ such that $\overline{B^3_s}\cap E=\overline{B^3_s}\cap\Big(\bigcup_{i=1}^{k}\varphi_i([0,\frac{1}{4}])\Big)$.

Let $i\in\{1,\dots,k\}$. Since the φ_i are smooth and since $\varphi_i=0$ we see that we can write $\varphi_i(t)=t\cdot\varphi_i'+\nu(t)$ where $\lim_{t\to 0}\frac{\|\nu(t)\|}{t}=0$. It follows from this observation and the fact that $\varphi_i'\neq 0$ that there exists an $s_i\in(0,\frac{1}{4})$ such that $\varphi_i(t)\cdot\varphi_i'(t)>0$ for all $t\in(0,s_i)$. We set $R:=\frac{1}{2}\cdot\min\{s,s_1,\dots,s_k\}$. It is now straightforward to verify that $\Phi=\mathrm{id}_{\overline{B^3_R}}$ has the desired properties.

Now let X be a small neighborhood for V. By the Tubular Neighborhood Theorem 4 there exists a tubular neighborhood Y of the proper submanifold $E\cap(S^3\setminus X^o)$ of $S^3\setminus X^o$. Then $X\cup Y$ is a graphical neighborhood of G. □

4 Properties of Graphical Neighborhoods

In this section we want to provide the proof for Theorem 2. Whereas statements (1) and (4) of Theorem 2 are basically obvious, the proof of the remaining two statements requires a little effort. We will need the following proposition.

Proposition 1. *Let $G=(V,E)$ be a spatial graph and let X be a small neighborhood for V. Let $v\in V$. We denote by X_v the corresponding component of X.*

1. *There exist points $P_1,\dots,P_k\in S^2$ and a homeomorphism $\Theta\colon\overline{B^3}\to X_v$ such that*

$$\Theta\Big(\bigcup_{i=1}^{k}\{r\cdot P_i\in\overline{B^3}\,|\,r\in[0,1]\}\Big)\;=\;E\cap X_v$$

   and such that each $\Theta(S^2_t)$ is transverse to E.
2. *Given any neighborhood U of v there exist points $P_1,\dots,P_k\in S^2$, some $\eta>0$ with $\Theta(\overline{B^3_\eta})\subset U$ and an orientation-preserving diffeomorphism $\Theta\colon\overline{B^3}\to X_v$ such that*

$$\Theta^{-1}(E)\cap(S^2\times[\eta,1])\;=\;\bigcup_{i=1}^{k}\{P_k\}\times[\eta,1]$$

and such that each $\Theta(S^2_t)$ is transverse to E.

Definition 8. Let $a < b$ be real numbers.

1. A *string* in $S^2 \times [a,b]$ is a connected 1-dimensional submanifold with one boundary point on $S^2 \times \{a\}$ and one boundary point on $S^2 \times \{b\}$.
2. A string is called *linear* if it is of the form $\{x\} \times [a,b]$ for some $x \in S^2$.
3. A *collection of strings* is defined as a finite set of disjoint strings.
4. We call a collection of strings E in $S^2 \times [a,b]$ *unknotted* if there exists a diffeomorphism $\Phi\colon S^2 \times [a,b] \to S^2 \times [a,b]$ with $\Phi_{S^2 \times \{a\}} = \mathrm{id}$ such that for every $t \in [a,b]$ the submanifold $\Phi(E)$ is transverse to $S^2 \times \{t\}$.

Lemma 2. *Let E be a collection of strings in $S^2 \times [a,d]$. We suppose that E is transverse to $S^2 \times \{t\}$ for all $t \in [a,d]$. Let $c \in [a,d]$. We write $b = \frac{a+c}{2}$. There exists a level-preserving diffeomorphism $\phi\colon S^2 \times [a,d] \to S^2 \times [a,d]$ with $\phi|_{S^2 \times [a,b]} = \mathrm{id}$ such that every component of $\phi(E) \cap (S^2 \times [c,d])$ is a linear string.*

Proof. We enumerate the components of E by $E_1, \ldots, E_n$ and write $v_i = E_i \cap S^2 \times \{a\}$. For notational simplicity we assume the case that $[a,d] = [0,2]$. We pick parametrizations $\alpha_i\colon [0,2] \to E_i$ for each i. Since E_i intersects each $S^2 \times \{t\}$ transversally we can reparametrize α_i to a smooth map $\tilde{\alpha}_i\colon [0,2] \to S^2 \times [0,2]$ which is level preserving, i.e. such that for any $t \in [0,2]$ we have $\tilde{\alpha}_i(t) \in S^2 \times \{t\}$. In other words, we can write $\tilde{\alpha}_i(t) = (\beta_i(t), t)$ for some smooth map $\beta_i\colon [0,2] \to S^2$. Thus we obtain a map:

$$\begin{aligned} h\colon V \times [0,2] &\to S^2 \times [0,2] \\ (v_i, t) &\mapsto (\beta_i(t), t). \end{aligned}$$

By the diffeotopy extension theorem [9, Theorem 2.4.2], we obtain a level preserving diffeomorphism $\phi\colon S^2 \times [0,2] \to S^2 \times [0,2]$ extending the map h. A quick look at the proof of the diffeotopy extension theorem [9, Theorem 2.4.2] shows that ϕ can be chosen to be the identity on $S^2 \times \{0\}$. If $c = 0$, then ϕ^{-1} is the desired diffeomorphism. If $c \neq 0$, then again by notational convenience we assume $c = 1$. We take a smooth function $f\colon \mathbb{R} \to \mathbb{R}$ which is f is monotonously increasing, with $f(t) = 0$ for $t \leq \frac{1}{2}$ and with $f(t) = 1$ for $t \geq 1$. Since ϕ is level-preserving, there is a smooth map $\varphi\colon S^2 \times [0,2] \to S^2$ such that $\phi(x,t) = (\varphi(x,t), t)$. Note that for every $t \in [0,2]$ the map $\varphi(\,\cdot\,, t)$ is a diffeomorphism. The map $\widetilde{\phi}(x,t) := (\varphi(x, t \cdot f(t)), t)$ is a diffeomorphism with inverse $(\varphi(x, t \cdot f(t))^{-1}, t)$. Moreover, for $t \in [0, \frac{1}{2}]$ we have $f(t) = 0$ and hence $\widetilde{\phi} = \mathrm{id}$ and for $t \in [1,2]$ we have $f(t) = 1$ and hence $\widetilde{\phi}^{-1}(\,\cdot\,, t) = \phi^{-1}(\,\cdot\,, t)$. Therefore in $S^2 \times [1,2]$ the map $\widetilde{\phi}$ maps linear strings to E. This shows that $\widetilde{\phi}^{-1}$ is the desired diffeomorphism. □

Proof (Proof of Proposition 1). Let $G = (V,E)$ be a spatial graph and let X be a small neighborhood for V. Given $v \in V$ and $X_v \subset X$ we can and will pick an orientation-preserving diffeomorphism $\Theta\colon \overline{B^3_R} \to X_v$ with $\Theta = v$ and such that for each $r \in (0,R]$ the image $\Theta_i(S^2_r)$ is a submanifold of $S^3 \setminus V$ that is transverse to the submanifold E of $S^3 \setminus V$.

1. We pick a strictly decreasing sequence $R = a_1, a_2, a_3, \dots$ of real numbers with $\lim_{i\to\infty} a_i = 0$ and we iteratively apply Lemma 2 (with $c = a$) to corresponding strings in $S^2 \times [a_{i+1}, a_i] = \overline{B^3_{a_i}} \setminus B^3_{a_{i+1}}$. We combine the resulting diffeomorphisms to obtain the desired homeomorphism.
2. Let U be neighborhood U of v. We pick $a < c < R$ such that $\Theta(S^2 \times [a,c]) \subset U$. We apply Lemma 2 and obtain a map $\phi\colon S^2 \times [a,R] \to S^2 \times [a,R]$. Since ϕ is the identity in a neighborhood of $S^2 \times \{a\}$ we see that ϕ extends to a smooth map $\phi\colon \overline{B^3_R} \to \overline{B^3_R}$ that is the identity on $\overline{B^3_a}$. The map $\Theta \circ \phi\colon \overline{B^3_R} \to X_v$ has the desired properties. □

Now we can finally prove Theorem 2.

Theorem 2. *Let G be a spatial graph and let N be a graphical neighborhood for G.*

1. *N contains $|G|$ in the interior $N^o = N \setminus \partial N$ of N,*
2. *$|G|$ is a deformation retract of N,*
3. *∂N is a deformation retract of $N \setminus |G|$,*
4. *the exterior $E_G = S^3 \setminus N^o$ is a compact 3-dimensional manifold that is a deformation retract of $S^3 \setminus |G|$.*

Proof. Let $N = X \cup Y$ be a graphical neighborhood. Statement (1) is immediate. Statement (4) is a consequence of of statement (3). Statement (2) and (3) can be proved easily using Proposition 1 (1), using the fact that Y is a product and using the following elementary claim.

Claim.

1. a. There exists a deformation retraction from $[0,1] \times \overline{B^2}$ to $([0,1] \times \{0\}) \cup (\{0,1\} \times \overline{B^2})$,
   b. there exists a deformation retraction from $[0,1] \times (\overline{B^2} \setminus \{0\})$ to $[0,1] \times S^1$.
2. Let $P_1, \dots, P_k \in S^2$ with $k \geq 1$. We write $Y := \bigcup_{i=1}^{k} \{r \cdot P_i \in \overline{B^3} \mid r \in [0,1]\}$.
   a. There exists a deformation retraction from $\overline{B^3}$ to Y.
   b. There exists a deformation retraction from $\overline{B^3} \setminus Y$ to $S^2 \setminus \{P_1, \dots, P_k\}$.

The proof of the claim is left to the reader. □

5 Uniqueness of graphical neighborhoods

Finally we define what is means for two graphical neighborhoods of a given spatial graph to be equivalent.

Definition 9. Let G be a spatial graph and let N and N' be two graphical neighborhoods of G. We say N and N' are *equivalent* if there exists a map $\Phi\colon S^3 \times [0,1] \to S^3$ with the following properties:

1. each $\Phi_t\colon S^3 \to S^3$ is a diffeomorphism,
2. each Φ_t is the identity on the vertex set and it preserves each edge setwise,
3. $\Phi_0 = \mathrm{id}$,
4. we have $\Phi_1(N) = N'$.

The goal of this section is to prove Theorem 3 from the introduction, i.e. we want to show that any two graphical neighborhoods of a given spatial graph are equivalent.

We turn to our first technical lemma of this section.

Lemma 3. *We consider the manifold $S^2 \times [a,b]$. Suppose we are given points $P_1, \ldots, P_n$ in S^2 with $n \geq 1$. Let C be a proper submanifold of $S^2 \times (a,b)$ which is diffeomorphic to S^2, which is transverse to all the strings $P_i \times [a,b]$ and which meets each string $P_i \times [a,b]$ exactly once. Then there exist a diffeomorphism*

$$\Psi\colon S^2 \times [a,b] \;\to\; S^2 \times [a,b]$$

and some $c \in (a,b)$ with the following properties:

1. *Ψ is the identity near the boundary,*
2. *Ψ preserves each string $\{P_i\} \times [a,b]$ setwise,*
3. *$\Psi(S^2 \times \{c\}) = C$,*
4. *for each $t \in [a,b]$ the image $\Psi(S^2 \times \{t\})$ is transverse to each string $\{P_i\} \times [a,b]$.*

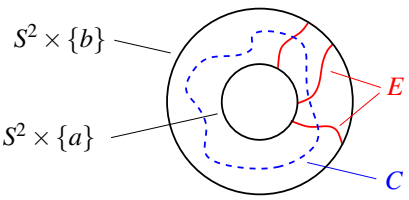


Fig. 6

In the proof of Lemma 3 we need the following well-known purely group theoretic statement.

Lemma 4. *Let $\varphi\colon A \to B$ be a homomorphism between free groups of the same finite rank. If $\varphi_*\colon H_1(A) \to H_1(B)$ is an epimorphism, then φ is a monomorphism.*

For the reader's convenience we include the proof of the lemma.

Proof. Let k be the rank of A and B. Note that $\Gamma := \varphi(A)$ is a subgroup of the free group B, thus it is also free group. The rank of Γ is evidently $\leq k$. The epimorphism $\varphi_*\colon H_1(A) \to H_1(B)$ factors through the inclusion induced map $H_1(\varphi(A)) \to H_1(B)$. Therefore we see that the rank of $\varphi(A)$ is $\geq k$. Thus $\varphi(A)$ is in fact a free group of

rank k. The map $A \to \varphi(A)$ is evidently an epimorphism. Since free groups are Hopfian [6, p. 109] we see that $A \to \varphi(A)$ is also a monomorphism. But this implies that φ itself is a monomorphism. □

We will also make use of the following theorem.

Theorem 6. *Let F be a surface (possibly with boundary) that is not diffeomorphic to S^2. Let $D \subset F \times (0,1)$ be a properly embedded surface. If D is incompressible, i.e. if the inclusion induced map $\pi_1(D) \to \pi_1(F \times [0,1])$ is a monomorphism, then there exists an orientation-preserving diffeomorphism $\Psi\colon F \times [0,1] \to F \times [0,1]$ with $\Psi(F \times \{\frac{1}{2}\}) = D$, Ψ is the identity in a neighborhood of $F \times \{0,1\}$ and the restriction of Ψ to $\partial F \times [0,1]$ is diffeotopic to the identity.*

Proof. This follows, with minor effort, from [8, Proposition 3.1 and Corollary 3.2]. □

Proof (Proof of Lemma 3). For notational convenience we suppose that $a = -1$ and $b = 1$. Thus we consider the manifold $S^2 \times [-1,1]$. We denote by $p\colon S^2 \times [-1,1] \to S^2$ the obvious projection. Suppose we are given points $P_1, \dots, P_n \in S^2$ with $n \geq 1$. Let C be a submanifold of $S^2 \times (-1,1)$ which is diffeomorphic to S^2, which is transverse to each string $\{P_i\} \times [-1,1]$ and which meets each string $\{P_i\} \times [-1,1]$ exactly once.

By picking small enough closed disks D_i around the P_i we obtain tubular neighborhoods $D_i \times [-1,1]$ for the strings such that for each i the intersection $C \cap (D_i \times [-1,1])$ is a single disk and such that the projection $p\colon S^2 \times [-1,1] \to S^2$ restricts to a diffeomorphism $C \cap (D_i \times [-1,1]) \to D_i$. We write $\Sigma = S^2 \setminus \bigcup_{i=1}^{n} D_i^o$. Since p also restricts to a diffeomorphism $C \cap (\partial D_i \times [-1,1]) \to \partial D_i$ we see that the curve $C \cap (\partial D_i \times [-1,1])$ represents a generator for $H_1(\partial D_i \times [-1,1])$.

Claim. The surface $C' = C \cap (\Sigma \times [-1,1])$ is incompressible in $\Sigma \times [-1,1]$, i.e. the inclusion induced map $\pi_1(C') \to \pi_1(\Sigma \times [-1,1])$ is a monomorphism.

As we noted above, the intersection of the sphere C with each cylinder $D_i \times [-1,1]$ is a single disk. Thus we see that C' is a sphere with n open disks removed, i.e. C' is diffeomorphic to Σ. Thus we see that $\pi_1(C')$ and $\pi_1(\Sigma \times [-1,1])$ are free groups of the same rank. By Lemma 4 it suffices to show that $H_1(C') \to H_1(\Sigma \times [-1,1])$ is an epimorphism.

We consider the following commutative diagram of inclusion induced maps:

$$\begin{array}{ccccc}
H_1(\partial\Sigma) & \xrightarrow{\cong} & H_1((\partial\Sigma)\times[-1,1]) & \twoheadleftarrow & H_1(C' \cap ((\partial\Sigma)\times[-1,1])) \\
\downarrow & & \downarrow & & \downarrow \\
H_1(\Sigma) & \xrightarrow{\cong} & H_1(\Sigma \times [-1,1]) & \leftarrow & H_1(C').
\end{array}$$

The two horizontal maps on the left are evidently isomorphisms. Furthermore, since Σ is a sphere minus some open disks we see that the left vertical map is an epimorphism. Thus the middle vertical map is an epimorphism. Furthermore, since

$C \cap (\partial D_i \times [-1,1])$ represents a generator for $H_1(\partial D_i \times [-1,1])$ we see that the top right horizontal map is an epimorphism. Thus it follows that the horizontal map on the bottom right is also an epimorphism. This concludes the proof of the claim.

It follows from the fact that C' is properly embedded in $\Sigma \times (-1,1))$ and Theorem 6 that there exists a self-diffeomorphism Ψ of $\Sigma \times [-1,1]$, that is the identity in a neighborhood of $\Sigma \times \{\pm 1\}$, that sends $\Sigma \times \{0\}$ to C' and which has the property that the restriction to each annulus $\partial D_i \times [-1,1]$ is diffeotopic to the identity.

It remains to extend Ψ over the cylinders $D_i \times [-1,1]$ in a suitable way. Recall that the projection $p\colon S^2 \times [-1,1] \to S^2$ restricts for each i to a diffeomorphism $C \cap (D_i \times [-1,1]) \to D_i$. The existence of the desired extensions is thus an immediate consequence of Lemma 5 below. □

The following technical lemma concludes the previous proof.

Lemma 5. *Let $\varepsilon > 0$. We denote by $p\colon \overline{B^2_{1+\varepsilon}} \times [-1,1] \to \overline{B^2_{1+\varepsilon}}$ the obvious projection. Let C be a properly embedded disk in $\overline{B^2_{1+\varepsilon}} \times (-1,1)$ such the restriction of p to $C \cap (\overline{B^2_1} \times [-1,1]) \to \overline{B^2_1}$ is a diffeomorphism. Furthermore suppose we are given an orientation-preserving self-diffeomorphism Ψ of $(S^1 \times [1,1+\varepsilon]) \times [-1,1]$ with the following properties:*

1. *Ψ is the identity near $(S^1 \times [1,1+\varepsilon]) \times \{\pm 1\}$,*
2. *the restriction of Ψ to $S^1 \times \{1\} \times [-1,1]$ is diffeotopic to the identity,*
3. *$\Psi((S^1 \times [1,1+\varepsilon]) \times \{0\}) = C \cap (S^1 \times [1,1+\varepsilon]) \times [-1,1]$.*

Then there exists a self-diffeomorphism Φ of $\overline{B^2_{1+\varepsilon}} \times [-1,1]$ with the following properties:

(0) it equals Ψ on $S^1 \times [1,1+\varepsilon] \times [-1,1]$,
(1) Φ is the identity in a neighborhood of $\overline{B^2_{1+\varepsilon}} \times \{\pm 1\}$,
(2) Φ preserves $\{0\} \times [-1,1]$ setwise,
(3) we have $\Phi\big(\overline{B^2_{1+\varepsilon}} \times \{0\}\big) = C$.

Proof. We write $\Theta = \Psi|_{S^1 \times [-1,1]}$. This is an orientation-preserving self-diffeomorphism of the annulus $S^1 \times [-1,1]$ that is the identity near $S^1 \times \{\pm 1\}$. By hypothesis there exists a diffeotopy $S^1 \times [-1,1] \times [0,1]$ from Θ to the identity and the diffeotopy can be chosen to be the identity near $S^1 \times \{\pm 1\} \times [0,1]$. We use this diffeotopy to extend Ψ to $S^1 \times [1, \frac{1}{2}] \times [-1,1]$. Finally we extend Ψ via the identity to $\overline{B^2_1} \times [-1,1]$. Note that the restriction of the projection p to $\Psi'(C) \to \overline{B^2_{1+\varepsilon}}$ is still a diffeomorphism. We can postcompose Ψ' with a suitable self-diffeomorphism of $\overline{B^2_{1+\varepsilon}} \times [-1,1]$ of the form $(P,z) \mapsto (P, f(P,z))$ to obtain the desired self-diffeomorphism Φ. □

5.1 Proof of Theorem 3

The theorem will be proved in several steps. To formulate the first step we need to introduce a new definition.

Definition 10. Let $G = (V,E)$ be a spatial graph. Let X and X' be small neighborhoods of V.

1. Given $v \in V$ we say that X'_v is *covered* by X_v if the following properties hold:
   a. $X'_v \subset X_v$,
   b. there exists a diffeomorphism $\Theta\colon \overline{B^3_R} \to X_v$ that is transverse at v and an $R' < R$ such that $\Theta(\overline{B^3_{R'}}) = X'_v$,
   c. there exist $P_1,\dots,P_n \in S^2$ such that $\Theta^{-1}(E) \cap S^2 \times [R',R] = \bigcup_{i=1}^{n} \{P_i\} \times [R',R]$.
2. We say X' is *covered* by X if each X'_v is covered by X_v.

Remark 4. By rescaling we can always arrange that $R' = 1$ and $R = 2$.

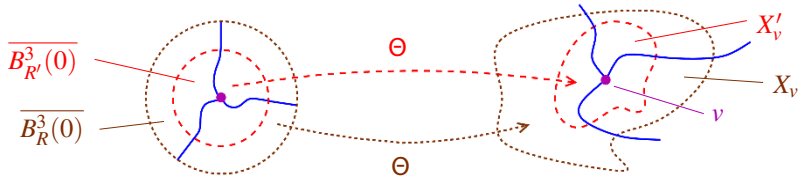


Fig. 7 The small neighborhood Θ'_i is covered by Θ_i.

Lemma 6 (Small neighborhood shrinking lemma). *Let $G = (V,E)$ be a spatial graph. Let X be X' two small neighborhoods of V. If $X' \subset X^o$, then X' is covered by X.*

Proof. Let X and X' be small neighborhoods of V with $X' \subset X^o$. Let $v \in V$. We pick a corresponding diffeomorphism $\Theta'_v\colon \overline{B^3_{R'}} \to X'_v$ that is transverse at v. By Proposition 1 (2) we can pick a diffeomorphism $\Theta_v\colon \overline{B^3_R} \to X_v$ that is transverse at v and which admits an $\eta > 0$ such that $\Theta_v(\overline{B^3_\eta}) \subset (X'_v)^o$ and such that $F := \Theta_v^{-1}(E) \cap S^2 \times [\eta,R]$ is linear. (Here we make the usual identification $\overline{B^3_R} \setminus B^3_\eta = S^2 \times [\eta,R]$.)

We write $C := \Theta_v^{-1}(\Theta'_v(S^2_{R'})) \subset \overline{B^3_R}$. Note that by the choice of η we have $C \subset S^2 \times [\eta,R]$. By Lemma 3 there exists a diffeomorphism

$$\Psi\colon S^2 \times [\eta,R] \;\to\; S^2 \times [\eta,R]$$

and some $c \in (\eta,R)$ with the following properties:

1. Ψ is the identity near the boundary,
2. Ψ preserves the linear strings $\Theta_v^{-1}(E) \cap (S^2 \times [\eta, R]$,
3. $\Psi(S^2 \times \{c\}) = C$,
4. for each $t \in [\eta, R]$ the image $\Psi(S^2 \times \{t\})$ is transverse to F.

We now consider the map $\Xi : \overline{B^3_R} \to \overline{B^3_R}$ that is given by the identity on $\overline{B^3_\eta}$ and that is given by Ψ on $S^2 \times [\eta, R] = \overline{B^3_R} \setminus B^3_\eta$. (Note that Ξ is smooth by the first property of Ψ.) Note that $\Theta_v \circ \Xi : \overline{B^3_R} \to X_v$ is transverse at v and that the image of $\overline{B^3_c}$ under this map is precisely X'_v. We have thus shown that X'_v is covered by X_v. □

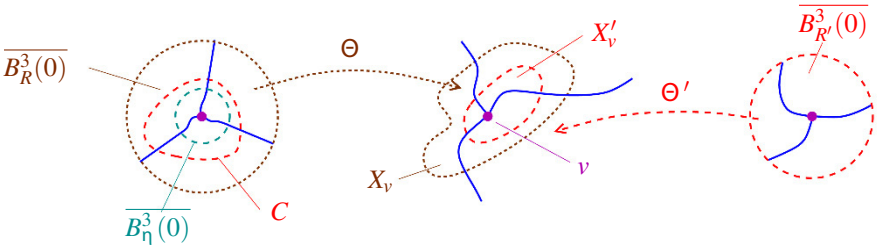


Fig. 8 Illustration of the proof of Lemma 6.

Lemma 7 (Graphical shrinking lemma). *Let $G = (V, E)$ be a spatial graph. Let $Z = X \cup Y$ be a graphical neighborhood for G. Suppose that X' is a small neighborhood for V. If X' is covered by X, then there is a graphical neighborhood Z' with decomposition $Z' = X' \cup Y'$ and which is equivalent to Z.*

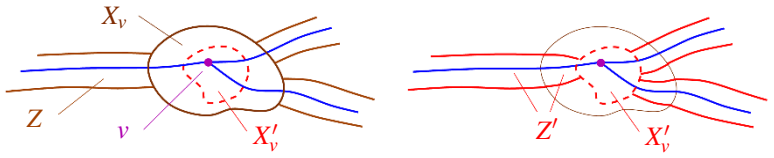


Fig. 9 Illustration of Lemma 7.

We will need the following rather technical lemma.

Lemma 8. *Let $P_1, \ldots, P_m$ be points in S^2. Let $\varepsilon \in (0,1)$ and let $f : S^2 \times [1, 1+\varepsilon] \to S^2 \times [1,2]$ be an embedding such that $f|_{S^2 \times \{1\}} = \mathrm{id}$ and such that f preserves $\{P_i\} \times [1,2]$ setwise. Then there exists a diffeomorphism $\Psi : S^2 \times [1,2] \to S^2 \times [1,2]$ with the following properties:*

1. Ψ equals f in a neighborhood of $S^2 \times \{1\}$,

2. *Ψ is the identity in a neighborhood of $S^2 \times \{2\}$,*
3. *for each $t \in [1,2]$ and for each $Q \in S^2$ the submanifolds $\Psi(S^2 \times \{t\})$ and $\{Q\} \times [1,2]$ are transverse,*
4. *for each i the map Ψ preserves $\{P_i\} \times [1,2]$ setwise.*

Proof. We denote by $p\colon S^2 \times [1,2] \to S^2$ the obvious projection. Since $f|_{S^2 \times \{1\}} = \mathrm{id}$ we can pick an $\mu > 0$ such that for all $Q \in S^2$ and $t \in [1, 1+\mu]$ the differential $Dp_{(Q,t)}$ is an isomorphism and such that for each i we have $f(\{P_i\} \times [1, 1+\mu]) \subset D_i \times [1,2]$. Note that for each $t \in [1, 1+\mu]$ the map $g_t := p \circ f_t\colon S^2 \to S^2$ is a diffeomorphism that is the identity on $\{P_1, \dots, P_m\}$.

Let $\sigma\colon [1, 1+\mu] \to [1, 1+\mu]$ be a smooth function which is equal to 1 on some interval $[1, \eta]$ with $\eta > 0$ and which has the property that there exists a $\nu > 0$ such that $\sigma(t) = t$ for all $t \in [1+\mu-\nu, 1+\mu]$. We consider the map

$$\begin{aligned} \alpha\colon S^2 \times [1, 1+\mu] &\to S^2 \times [1, 1+\mu] \\ (Q,t) &\mapsto f(g^{-1}_{\sigma(t)}(Q), t). \end{aligned}$$

Note that the map α has the property that $p(\alpha(Q,t)) = Q$ for all $Q \in S^2$ and all $t \in [1+\mu-\nu, 1+\mu]$. This means that there exists a smooth function $d\colon S^2 \times [1+\mu-\nu, 1+\mu] \to [1, 1+\mu]$ such that $\alpha(Q,t) = (Q, d(Q,t))$ for all $Q \in S^2$ and all $t \in [1+\mu-\nu, 1+\mu]$ and for some smooth function $d\colon S^2 \times [1+\mu-\nu, 1+\mu] \to [1, 1+\mu]$. Note that d has the property that for each $Q \in S^2$ the function $t \mapsto d(Q,t)$ has positive derivative. We pick an extension of d to a smooth function $d\colon S^2 \times [1+\mu-\nu, 2] \to [1,2]$ with the following properties:

1. for each $Q \in S^2$ the function $t \mapsto d(Q,t)$ has positive derivative,
2. there exists a neighborhood of $S^2 \times \{2\}$ such that the map d is just the projection.

Now we consider the map

$$\begin{aligned} \Psi\colon S^2 \times [1,2] &\to S^2 \times [1,2] \\ (Q,t) &\mapsto \begin{cases} \alpha(Q,t), & \text{if } t \in [1, 1+\mu], \\ (Q, d(Q,t)), & \text{if } t \in [1+\mu, 2]. \end{cases} \end{aligned}$$

One easily verifies that the above map Ψ has all the desired properties. □

Proof (Proof of the Graphical shrinking Lemma 7). Let $v \in V$.

1. Since X' is covered by X we can find for each $v \in V$ an orientation-preserving diffeomorphism $\Theta_v\colon \overline{B^3_2} \to X_v$ such that Θ_v is transverse at v and such that Θ_v restricts to a diffeomorphism $\overline{B^3_1} \to X'_v$ and such that there exists a finite subset $P_v \subset S^2$ with $\Theta_v(S^2 \times [\frac{1}{4}, 2]) \cap E = \Theta_v(P_v \times [\frac{1}{4}, 2])$.
2. By condition (3) on a tubular neighborhood we can find orientation-preserving embeddings $\Omega_v\colon S^2 \times [2,3] \to S^3 \setminus X^o$, $v \in V$ with the following properties:
   a. the images are disjoint,
   b. for each v we have $\Omega_v(S^2 \times \{2\}) = \partial X_v$,

c. for each v there exists a finite subset $Q_v \subset S^2$ with $E \cap \Omega_v(S^2 \times [2,3]) = \Omega_v(Q_v \times [2,3])$ and there exist disjoint closed disks $\{D_i\}_{i\in Q_v}$ in S^2 with $Z \cap \Omega_v(S^2 \times [2,3]) = \Omega_v\Big(\bigcup_{i\in Q_v} \{D_i\} \times [2,3]\Big)$.

Note that after possibly postcomposing Ω_v with the map

$$\begin{aligned} S^2 \times [2,3] &\mapsto S^2 \times [2,3] \\ (P,t) &\mapsto ((\Theta_v|_{S^2_2} \circ (\Omega_v|_{S^2_2})^{-1})(P), t) \end{aligned}$$

we can arrange that $\Theta_v|_{S^2_2} = \Omega_v|_{S^2_2}$. In particular we have $P_v = Q_v$.

3. Using the Whitney Approximation Theorem, as formulated in [5, Theorem 6.26], we can extend the embedding $\Omega_v \colon S^2 \times [2,3] \to S^3 \setminus X^o$ to a smooth map $\Omega_v \colon S^2 \times [2-\varepsilon, 3] \to S^3$ for a suitably small $\varepsilon > 0$. Furthermore we can arrange that $E \cap \Omega_v(S^2 \times [2-\varepsilon, 2]) = \Omega_v(Q_v \times [2-\varepsilon, 2])$. After possibly reducing the ε these maps are in fact embeddings. Note that Ω_v restricts to an embedding $S^2 \times [2-\varepsilon, 2] \to X_v$, in particular we obtain an embedding $f := (\Theta_v)^{-1} \circ \Omega_v \colon S^2 \times [2-\varepsilon, 2] \to S^2 \times [1,2]$ that is the identity on $S^2 \times \{2\}$ and that preserves $Q_v \times [2-\varepsilon, 2]$ setwise.

We pick $\Psi_v \colon S^2 \times [1,2] \to S^2 \times [1,2]$ as in Lemma 8. (Here we replace the endpoints $\{1,2\}$ by $\{2,1\}$ and we consider the map $f := (\Theta_v)^{-1} \circ \Omega_v \colon S^2 \times [2-\varepsilon, 2] \to S^2 \times [1,2]$.) We define

$$\begin{aligned} \Xi_v \colon \overline{B^3_3} &\to S^3 \\ (P,t) &\mapsto \begin{cases} \Omega_v(P,t), & \text{if } t \in [2,3], \\ \Theta_v(\Psi_v(P,t)), & \text{if } t \in [1,2], \\ \Theta_v(P,t), & \text{if } t \in [0,1]. \end{cases} \end{aligned}$$

We set

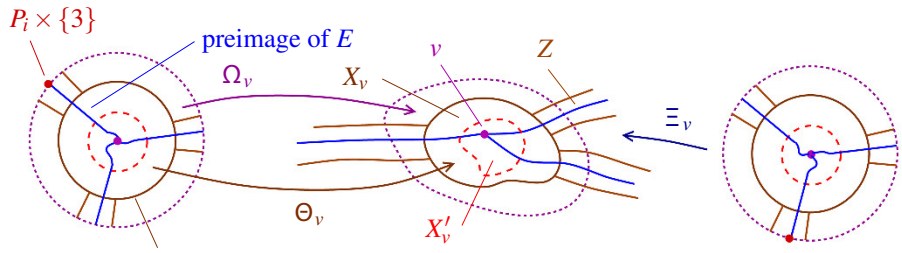


Fig. 10

$$Z' := Z \cup \bigcup_{v\in V} \bigcup_{i\in P_v} \Xi_v\big(D_i \times [1,2]\big).$$

It is fairly straightforward to show that Z' is indeed a graphical neighborhood for G.

It remains to show that the graphical neighborhoods Z and Z' are equivalent. We pick a smooth strictly monotonously function $f\colon [0,3] \to [0,3]$ with $f(t) = t$ for $t \in [0, \frac{1}{2}]$ and $t \in [\frac{5}{2}, 3]$ and with $f(2) = 1$. The map

$$\begin{aligned} \Phi\colon S^3 \times [0,1] &\to S^3 \\ (P,t) &\mapsto \begin{cases} \Xi_v(Q, s\cdot(1-t) + f(s)\cdot t), & \text{if } P = \Xi_v(Q,s) \text{ where } Q \in S^2, t \in [0,3] \\ P, & \text{otherwise} \end{cases} \end{aligned}$$

is a diffeotopy that fixes V pointwise and that fixes E setwise with $\Phi_0 = \mathrm{id}$ and with $\Phi_1(Z) = Z'$. $\square$

Lemma 9. *Let $P_1, \dots, P_m \in S^2$, let $\varepsilon > 0$ and let $\Psi\colon (S^2 \times [1, 1+\varepsilon)) \times [0,1] \to \mathbb{R}^3 \setminus B^3_1$ be a map with the following properties:*

1. *$\Psi_t = \mathrm{id}$ for small t,*
2. *each Ψ_t is an embedding,*
3. *each Ψ_t is the identity on each $\{P_i\} \times [1, 1+\varepsilon)$,*
4. *each Ψ_t preserves $S^2 \times \{1\}$ setwise.*

Then we can extend Ψ to a smooth map $(S^2 \times [\frac{1}{2}, 1+\varepsilon)) \times [0,1] \to \mathbb{R}^3 \setminus B^3_{\frac{1}{2}}$ with the following properties:

1. *$\Psi_0 = \mathrm{id}$,*
2. *each Ψ_t is an embedding,*
3. *for each i we have $\Psi_t(\{P_i\} \times [\frac{1}{2}, 1+\varepsilon)) \subset \{P_i\} \times [\frac{1}{2}, 1+\varepsilon)$,*
4. *there exists a $\nu > 0$ such that each Ψ_t is the identity on $S^2 \times [\frac{1}{2}, \frac{1}{2} + \nu]$.*

Proof. We start out with the following claim.

Claim. There exists a smooth map $\Psi'\colon S^2 \times [\frac{1}{2}, 1+\varepsilon) \times [0,1] \to \overline{B^3_{\frac{1}{2}}}$ with the following properties:

1. The map agrees with Ψ on $S^2 \times [1, 1+\varepsilon) \times [0,1]$,
2. for each i and each $t \in [0,1]$ we have $\Psi'_t(\{P_i\} \times [\frac{1}{2}, 1]) \subset \{P_i\} \times [\frac{1}{2}, 1+\varepsilon]$,
3. on $(S^2 \times [\frac{1}{2}, 1]) \times \{0\}$ the map Ψ' is the identity. (Strictly speaking it is the projection onto the factor in the parenthesis.)

First we extend Ψ to a map Ψ_1 on the following closed subset:

$$\begin{aligned} &\Psi_1 = (\Psi_1^x, \Psi_1^y)\colon (S^2 \times [1, 1+\varepsilon)) \times [0,1] \cup (S^2 \times [\tfrac{1}{2}, 1]) \times \{0\} \to (S^2 \times [\tfrac{1}{2}, 1]) \times [0,1] \\ &P \mapsto \Psi_1(P) = (\Psi_1^x(P), \Psi_1^y(P)) \end{aligned}$$

of $(S^2 \times [\frac{1}{2}, 1+\varepsilon]) \times [0,1]$ by defining it to be the identity on $(S^2 \times [\frac{1}{2}, 1]) \times \{0\}$. Since $\Psi_t = \mathrm{id}$ for small t we see that Ψ_1 is smooth on this closed subset. (Recall that by definition, [5, p. 45], a map $f\colon A \to N$ on an arbitrary subset A of a manifold is smooth if given any point $P \in A$ there exists an open neighborhood U of P and a

smooth map on U that agrees with f on $A \cap U$.) It follows from the Whitney Approximation Theorem, as formulated in [5, Theorem 6.26], that Ψ_1 can be extended to a smooth map on $(S^2 \times [\frac{1}{2}, 1+\varepsilon)) \times [0,1]$. We denote this extension again by Ψ_1. Next we pick disjoint open neighborhoods $U_1, \dots, U_m$ around $P_1, \dots, P_m$. In the following we make the identification $S^2 = \mathbb{R}^2 \cup \{\infty\}$ in such a way that $U_1, \dots, U_m \subset \mathbb{R}^2$. Thus we can consider the map

$$\Psi_2 \colon (S^2 \times [1, 1+\varepsilon)) \times [0,1] \cup S^2 \times [\tfrac{1}{2}, 1+\varepsilon) \times \{0\} \cup \bigcup_{i=1}^{m} U_i \times [\tfrac{1}{2}, 1+\varepsilon) \times [0,1] \to S^2 \times [\tfrac{1}{2}, 1+\varepsilon)$$

that is given by

$$(Q,s,t) \mapsto \begin{cases} (\Psi_1^x(Q,s,t) + P_i - \Psi_1^x(P_i,s,t), \Psi_1^y(Q,s,t)), & \text{if } Q \in U_i \text{ and } s \in [\frac{1}{2}, 1], \\ \Psi_1(Q,s,t), & \text{otherwise.} \end{cases}$$

One can easily see that Ψ_2 is smooth on the open subsets $U_i \times [\frac{1}{2}, 1+\varepsilon) \times [0,1]$. It follows in particular that the restriction of Ψ_2 to the closed subset

$$(S^2 \times [1, 1+\varepsilon)) \times [0,1] \cup \bigcup_{i=1}^{m} (P_i \times [\tfrac{1}{2}, 1]) \times [0,1] \cup (S^2 \times [\tfrac{1}{2}, 1]) \times \{0\}$$

is smooth. Thus, once again by the Whitney Approximation Theorem, as formulated in [5, Theorem 6.26], we can extend Ψ_2 to a smooth map $\Psi' \colon S^2 \times [\frac{1}{2}, 1+\varepsilon) \times [0,1]$ which now has all the desired properties. This concludes the proof of the claim.

It follows from the claim and the Whitney Approximation Theorem, as formulated in [5, Theorem 6.26], that Ψ' can be extended to a smooth map on $(S^2 \times [\frac{1}{2}, 1+\varepsilon)) \times [0,1]$. We denote this extension again by Ψ'.

Note that for a sufficiently small $\eta \in (0, \frac{1}{4})$ each map $\Psi'_t \colon S^2 \times [1-\eta, 1] \to \overline{B_1^3}$ is an embedding. Thus we can apply Lemma 8 (with the endpoints $\{1,2\}$ replaced by $\{\frac{1}{2}, 1\}$, with the interval $[1, 1+\varepsilon]$ replaced by $[1-\eta, 1]$ and with f replaced by Ψ') to obtain an extension of $\Psi' \colon (S^2 \times [1-\eta, 1]) \times [0,1] \to \overline{B_1^3}$ to a map on $(S^2 \times [\frac{1}{2}, 1]) \times [0,1] \to S^2 \times [\frac{1}{2}, 1]$ that has all the properties we expect on that domain. Together with our original map it defines the desired map $(S^2 \times [\frac{1}{2}, 1+\varepsilon)) \times [0,1] \to \mathbb{R}^3 \setminus B^3_{\frac{1}{2}}$. □

Now we can finally give the proof of Theorem 3.

Proof. Let G be a spatial graph and suppose that $Z = X \cup Y$ and $Z' = X' \cup Y'$ are two graphical neighborhoods of G. It follows easily from the proof of the existence of graphical neighborhoods that there exists a graphical neighborhood $Z'' = X'' \cup Y''$ with the following two properties:

1. we have $X'' \subset X^o$ and $X'' \subset (X')^o$,
2. for each $v \in V$ there exists a map $\Theta_v \colon \overline{B_2^3} \to X_v$ that is transverse at v and such that $\Theta_v(\overline{B_1^3}) = X''_v$ and such that the images $\Theta_v(\overline{B_2^3})$ are disjoint.

Since for graphical neighborhoods being equivalent is indeed an equivalence relation it suffices to prove the desired statement for X'' and X.

By the Shrinking Lemma 6 the small neighborhood X'' is covered by X. We apply the Graphical Shrinking Lemma 7 to X and X'' and we obtain a new graphical neighborhood $\widetilde{Z} = X'' \cup \widetilde{Y}$ which is equivalent to the graphical neighborhood $Z = X \cup Y$. Thus it remains to show that $Z'' = X'' \cup Y''$ is equivalent to $\widetilde{Z} = X'' \cup \widetilde{Y}'$. Now Y'' and $\widetilde{Y}$ are tubular neighborhoods of the proper submanifold $E \cap S^3 \setminus (X'')^o$ in $S^3 \setminus (X'')^o$. By uniqueness of tubular neighborhood, see Theorem 5, we obtain a diffeotopy Ψ of $S^3 \setminus (X'')^o$ rel $E \cap S^3 \setminus (X'')^o$ with $\Psi_0 = \mathrm{id}$ and with $\Psi(Y'') = \widetilde{Y}$.

We continue with the above maps $\Theta_v \colon \overline{B^3_2} \to X_v$. We apply Lemma 9 to the maps

$$\begin{aligned} S^2 \times [1, 1+\varepsilon] \times [0,1] &\to \mathbb{R}^3 \\ (P,s,t) &\mapsto \Theta_v^{-1}(\Psi_t(\Theta_v(P,s,t))) \end{aligned}$$

for a conveniently chosen $\varepsilon > 0$. We can use the resulting extensions given by Lemma 9 to extend Ψ over all of S^3. □

Acknowledgements We are very grateful to Erica Flapan for helpful comments. In particular our definition of a spatial graph arose from discussions with Erica Flapan. Both authors were supported by the SFB 1085 "higher invariants" funded by the DFG.

References

1. Friedl, S.: Algebraic Topology I-V. Lecture notes, University of Regensburg (2020) https://www.uni-regensburg.de/Fakultaeten/nat_Fak_I/friedl/papers/2020_friedl-algebraic-topology.pdf
2. Friedl, S. and Herrmann, G.: Spatial graphs, to appear in the Encyclopedia of Knot Theory to be published by CRC Press (2020).
3. Glaser L.: Geometrical combinatorial topology. Vol. I. Van Nostrand Reinhold Mathematics Studies, 27. Van Nostrand Reinhold Co. (1970).
4. Hempel J.: 3-manifolds. Annals of Mathematics Studies 86. Princeton, New Jersey: Princeton University Press and University of Tokyo Press. XII (1976).
5. Lee J.: Introduction to smooth manifold, Graduate Texts in Mathematics 218, second edition, Springer Verlag (2002).
6. Magnus W., Karrass A. and Solitar D.: Combinatorial group theory. Presentations of groups in terms of generators and relations. Reprint of the 1976 second edition. Dover Publications (2004).
7. Rourke C. P. and Sanderson B. J.: Introduction to piecewise-linear topology. Ergebnisse der Mathematik und ihrer Grenzgebiete 69 (1972).
8. Waldhausen F.: On irreducible 3-manifolds which are sufficiently large. Ann. of Math. (2) **87**, 56–88 (1968)
9. Wall C. T. C.: Differential topology. Cambridge Studies in Advanced Mathematics 156. Cambridge University Press (2016).

Open Problems in the Topology of Manifolds

Jonathan Bowden, Diarmuid Crowley, Jim Davis, Stefan Friedl, Carmen Rovi and Stephan Tillmann

Introduction

The problems in this list were collected at MATRIX, during the workshop on the Topology of Manifolds: Interactions between High and Low Dimensions, January 7th – 18th 2019. Several of the problems below were discussed in the problem sessions during the MATRIX workshop and *the organisers wish to thank all participants for their enthusiasm during the problem sessions and throughout the meeting.* A description of how the problem sessions were run can be found in the preface.

Below, we give a selection of eleven problems that were posed at the workshop. This selection illustrates the range and scope of the discussions at the meeting. We would like to thank all participants who contributed problems and further questions

Jonathan Bowden
Fakultät für Mathematik, Universität Regensburg, Germany
e-mail: jonathan.bowden@mathematik.uni-regensburg.edu

Diarmuid Crowley
School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia
e-mail: dcrowley@unimelb.edu.au

Jim Davis
Department of Mathematics, Indiana University, Bloomington, Indiana 47405, USA
e-mail: jfdavis@indiana.edu

Stefan Friedl
Fakultät für Mathematik, Universität Regensburg, Germany
e-mail: sfriedl@gmail.com

Carmen Rovi
Mathematisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg 69120 , Germany
e-mail: crovi@mathi.uni-heidelberg.de

Stephan Tillmann
School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia
e-mail: stephan.tillmann@sydney.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_39

that helped shape many of them. An evolving record of these and other problems and questions posed at the workshop can be found at the Manifold Atlas:

http://www.map.mpim-bonn.mpg.de/

We have attributed each problem in this list to the participant(s) who presented the problem at the workshop. The style in which problems were posed varied widely, and our selection reflects this. The first nine problems listed here are succinctly formulated and self-contained: references to the literature are minimal and references for each problem, where they exist, are at the end of the problem. The subject matter of the final two problems necessitated recalling more background and a somewhat more detailed referencing of the literature. The order in which we list the problems is chronological, rather than by subject matter.

Problem 1: A quotient of $\mathbf{S^2 \times S^2}$

presented by Jonathan Hillman

Let $C_4 = \langle \sigma \rangle$ act freely on $S^2 \times S^2$ with that action of σ defined by the equation $\sigma(x,y) = (y,-x)$ and let M be the quotient manifold.

The real projective plane $\mathbb{R}P^2 = S^2/\sim$ embeds in M via $[x] \mapsto [x,x]$ and its disk bundle neighborhood N in M is the tangent disk bundle of $\mathbb{R}P^2$. The complement of the open disk bundle neighborhood is the mapping cylinder of the double cover of lens spaces $L(4,1) \to L(8,1)$. Thus

$$M = N \cup \mathrm{MCyl}(\mathrm{L}(4,1) \to \mathrm{L}(8,1)).$$

This geometric analysis of M was given in [1] where it was shown that there are at most four closed topological manifolds in this homotopy type, half of which are stably smoothable.

The smooth manifold $M' = N \cup \mathrm{MCyl}(\mathrm{L}(4,1) \to \mathrm{L}(8,3))$ is homotopy equivalent to M.

Question. *Are M and M' homeomorphic? diffeomorphic?*

Reference

1. I. Hambleton and J. Hillman, *Quotients of $S^2 \times S^2$*, Preprint 2017. Available at arXiv1712.04572

Problem 2: Connected sum decompositions of high-dimensional manifolds

presented by Stefan Friedl

Let Cat be one of the categories Top, PL or Diff. A Cat-manifold M is called *irreducible* if, whenever we can write M as a connected sum of Cat-manifolds at least one of the summands is a homotopy sphere. The Kneser-Milnor theorem [2]

says that every compact Cat 3-manifold admits a connected sum decomposition into irreducible 3-manifolds, and this connected sum decomposition is unique up to permutation of the summands.

Stefan Friedl asked to what degree this statement holds in higher dimensions. During the two weeks of the workshop and during discussions afterwards Imre Bokor, Diarmuid Crowley, Stefan Friedl, Fabian Hebestreit, Daniel Kasprowski, Markus Land and Johnny Nicholson obtained a fairly comprehensive answer which appears in the proceedings [1]. Before we discuss the results, note that exotic spheres do not have a decomposition into irreducible manifolds. Thus it is reasonable to consider all questions "up to homotopy spheres".

In the following we summarize a few of the results.

1. It follows from standard algebraic topology and group theory that every Cat-manifold admits a connected sum decomposition into irreducible manifolds and a homotopy sphere.
2. The uniqueness statement (up to homotopy spheres) fails to hold in any of the dimensions ≥ 4 and any of the categories.
3. If one restricts attention to simply connected manifolds, then it is shown that in any dimension ≥ 17 uniqueness (up to homotopy spheres) fails to hold in any of the categories.
4. In contrast for many even dimensions $2k$, if one restricts attention to the case of $(k-1)$-connected smooth manifolds, uniqueness does hold.

References

1. I. Bokor, D. Crowley, S. Friedl, F. Hebestreit, M. Land, D. Kasprowski and J.Nicholson, *Connected sum decompositions of high-dimensional manifolds*, to appear in the MATRIX Annals (2019). Available at arXiv:1909.02628
2. J. Milnor, *A unique decomposition theorem for 3-manifolds,* Amer. J. Math. 84 (1962), 1–7.

Problem 3: An analogue of Casson-Gordon theory for trisections

presented by Stephan Tillmann

Heegaard splittings have long been used in the study of 3–manifolds. They were introduced in 1898 by Poul Heegaard, and provide a decomposition of each closed 3–manifold into two 1–handlebodies. A key concept introduced in the theory by Casson and Gordon [1] was the notion of *strong irreducibility*, with their main theorem stating that if a closed 3–manifold has a splitting that is not strongly irreducible, then either the splitting is reducible or the manifold contains an incompressible surface of positive genus. That is, one can either simplify the splitting, or one obtains topological information on the 3–manifold. Strongly irreducible Heegaard surfaces turn out to have many useful properties that one usually only associates with incompressible surfaces in 3–manifolds. Casson and Gordon also discovered a local condition, the *rectangle condition*, which guarantees that a Heegaard splitting is irreducible.

The challenge for the analogous theory of trisections of 4–manifold is to determine properties of trisections that have strong topological consequences and that can be determined by local information, for instance, from a trisection diagram.

Reference

1. A. Casson and C.McA. Gordon, *Reducing Heegaard splittings*, Topology Appl. 27 (1987), 275–283.

Problem 4: Aspherical manifolds whose fundamental group has nontrivial centre

presented by Fabian Hebestreit and Markus Land

Given a closed aspherical manifold M whose fundamental group has nontrivial centre, we can ask the following:

Question A. *Does there exist a finite cover of M with a principal S^1-action?*

Question B. *Is such an M null-cobordant?*

Motivation and background for these questions is found in [1, Section 7].

Reference

1. F. Hebestreit, M. Land, W. Lück and Oscar Randal-Williams, *A vanishing theorem for tautological classes of aspherical manifolds*, to appear in Geom. Topol.. Available at arXiv:1705.06232

Problem 5: Is the trisection genus additive under connected sum?

presented by Peter Lambert-Cole

Let M be a closed smooth 4-manifold. The "trisection genus" of M is the minimal genus of the central surface appearing in a trisection of M.

Question. *Is the trisection genus additive under connected sum?*

If so, then the following hold:

1. The trisection genus of M is a homeomorphism invariant.
2. The manifolds S^4, $\mathbb{C}P^2$, $S^2 \times S^2$, $\mathbb{C}P^2 \sharp \mathbb{C}P^2$ and $\mathbb{C}P^2 \sharp \overline{\mathbb{C}P^2}$ have a unique smooth structure.

An affirmative answer to the question is known for the class of all standard simply connected *PL* 4-manifolds [1].

Reference

1. J. Spreer and S. Tillmann, *The trisection genus of standard simply connected PL 4-manifolds*, 34th International Symposium on Computational Geometry, Art. No. 71, 13 pp., LIPIcs. Leibniz Int. Proc. Inform., 99, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

Problem 6: Compact aspherical 4-manifolds

presented by Jim Davis

Let M_0 and M_1 be a compact aspherical 4-manifolds with boundary. The Borel Conjecture in this setting states that a homotopy equivalence of pairs

$$f\colon (M_0, \partial M_0) \to (M_1, \partial M_1),$$

which is a homeomorphism on the boundary is homotopic, relative to the boundary, to a homeomorphism.

By topological surgery, the Borel Conjecture is valid when the fundamental group $\pi = \pi_1(M_0) \cong \pi_1(M_1)$ is good, for example, if π is elementary amenable. One now proceeds to the following three problems:

1. Decide which good π are the fundamental groups of compact aspherical 4-manifolds.
2. Determine the possible fundamental groups of the boundary components.
3. Determine the homeomorphism types of the boundary components.

These problems could be considered for compact aspherical 4-manifolds even when the fundamental group is not good, also in the smooth case.

Question. *Let M be a closed smooth aspherical 4-manifold. Is every smooth 4-manifold homotopy equivalent to M diffeomorphic to M?*

The question has not been answered for any M, not even the 4-torus.

Problem 7: Embedding integral homology 3-spheres into the 4-sphere

presented by Jonathan Hillman

Question. *Let Σ be an integral homology 3-sphere, not homeomorphic to S^3. Is there a locally flat embedding $\Sigma \hookrightarrow S^4$ such that one or both complementary regions are not simply-connected?*

This problem is motivated by the problem of classifying such embeddings up to isotopy. If a complement has non-trivial fundamental group, then a 'satellite' construction yields infinitely many isotopy classes of embeddings of Σ into S^4.

Problem 8: Stabilising number of knots and links

presented by Anthony Conway

Let W be a compact 4–manifold with boundary $\partial W \cong S^3$. We say that a properly embedded disk $(\Delta, \partial\Delta) \subset (W, \partial W)$ is *nullhomologous*, if its fundamental class $[\Delta, \partial\Delta] \in H_2(W, \partial W; \mathbb{Z})$ vanishes.

A link $L \subset S^3$ is *stably slice* if there exists $n \geq 0$ such that the components of L bound a collection of disjoint locally flat nullhomologous discs in the manifold $D^4 \# n(S^2 \times S^2)$. The *stabilising number* $\mathrm{sn}(L)$ of a stably slice link is the minimal such n.

Schneiderman proved that a link L is stably slice if and only if the following invariants vanish: the triple linking numbers $\mu_{ijk}(L)$, the mod 2 Sato-Levine invariants of L, and the Arf invariants of the components of L [2].

Question A. *Does the inequality* $\mathrm{sn}(L) \leq g_4^{\mathrm{top}}(L)$ *hold for stably slice links L of more than one component?*

This question is settled in the knot case: together with Matthias Nagel, we showed that $\mathrm{sn}(K) \leq g_4^{\mathrm{top}}(K)$ holds for stably slice knots [1]. We are currently unable to generalise this proof to links.

Remark. *The definition of the stabilising number also makes sense in the smooth category (one requires that the discs be smoothly embedded). Just as in the topological category, the inequality* $\mathrm{sn}^{\mathrm{smooth}}(K) \leq g_4^{\mathrm{smooth}}(K)$ *holds, and is unknown for links.*

This discussion of categories leads to the following question:

Question B. Is there a difference between the topological and smooth stabilising numbers of a knot? More precisely, is there a non-topological slice, Arf invariant zero knot K such that $0 < \mathrm{sn}^{\mathrm{top}}(K) < \mathrm{sn}^{\mathrm{smooth}}(K)$?

References

1. A. Conway and M. Nagel, *Stably slice disks of links*, J. Topol. 13 (2020), 1261–1301.
2. R. Schneiderman *Stable concordance of knots in 3-manifolds*, Algebr. Geom. Topol. 10 (2010), 373–432.

Problem 9: Unknotted surfaces in the 4-spheres

presented by Jim Davis

Kawauchi has published several accounts of the theorem below; however, none of them are satisfactory. The problem is to give a satisfactory proof.

Theorem. *Any two locally flat topological embeddings of a closed oriented surface in S^4 whose complement has infinite cyclic fundamental group are homeomorphic.*

There is a corresponding statement in the nonorientable case. Complex conjugation on $\mathbb{C}P^2$ has fixed set $\mathbb{R}P^2$ and orbit space S^4. Likewise for $\overline{\mathbb{C}P^2}$. The involution on $a\mathbb{C}P^2 \# b\overline{\mathbb{C}P^2}$ thus gives a locally flat embedding of $\#_{a+b}\mathbb{R}P^2$ in S^4.

Conjecture. *Any locally flat topological embedding of a closed nonorientable surface in S^4 whose complement has order 2 fundamental group is homeomorphic to one of the above embeddings.*

See also Massey [1] which determined the possible normal bundles.

Reference

1. W. S. Massey *Proof of a conjecture of Whitney*, Pacific J. Math. 31 (1969), 143–156.

Problem 10: Genus bounds for cancellations

presented by Diarmuid Crowley

This problem and the next are about the classification of compact $2q$-manifolds for $q \geq 2$. For simplicity, we assume that all manifolds are *connected*. We state these problems in the smooth category: there are obvious analogues for *PL*-manifolds and topological manifolds but we only discuss the topological case in dimension 4, which is of course an exceptional dimension and the *PL* case not at all.

For a natural number g, define $W_g := \#_g(S^q \times S^q)$ to be the g-fold connected sum of $S^q \times S^q$ with itself. If M_0 and M_1 are compact smooth $2q$-manifolds of the same Euler characteristic, then a *stable diffeomorphism* from M_0 to M_1 is a diffeomorphism

$$f\colon M_0 \sharp W_g \to M_1 \sharp W_g$$

for some $g \geq 0$. In this case we say write $M_0 \cong_{\mathrm{st}} M_1$ and we say that M_0 and M_1 are *stably diffeomorphic*. Of course, to define the connected sum operation, M_0, M_1 and W_g must be locally oriented. Since W_g admits an orientation reversing diffeomorphism for all g, if for $i = 0, 1$ the manifold M_i is orientable, then the diffeomorphism type of $M_i \sharp W_g$ does not depend on the orientation chosen for M_i.

The *stable class* of a $2q$-manifold M is defined to be the set of diffeomorphism classes of $2q$-manifolds M' with same Euler characteristic as M and which are stably diffeomorphism to M:

$$\mathscr{S}^{\mathrm{st}}(M) = \{M' \mid \chi(M) = \chi(M') \text{ and } M \cong_{\mathrm{st}} M'\}/\text{diffeomorphism}$$

We say that *cancellation holds* for M if every manifold which is stably diffeomorphic to M is diffeomorphic to M; i.e. $|\mathscr{S}^{\mathrm{st}}(M)| = 1$. Our purpose here is to summarise some of what is known about when cancellation holds and to identify two basic problems about cancellation which remain open. For this we require a further definition, notation and discussion.

The *genus* of M, $g(M)$, is defined to be the largest natural number g such that

$$M \cong M' \sharp W_g$$

for some other compact smooth $2q$-manifold M'. Since we have assumed that $q \geq 2$, the fundamental group of M, which we denote by π, is unchanged by stabilisation

with W_g. So far, the majority of work on the cancellation problem has been to identify cases where cancellation holds via the fundamental group π, the genus g and the parity of q. For example, the following theorem of Hambleton and Kreck shows the power of cancellation as a classification technique in dimension 4.

Theorem A. (Topological cancellation for $q = 2$ and finite π; [2, Thm. B]) *Let M be a closed oriented topological 4-manifold with finite fundamental group and of genus at least 1. Then cancellation holds for M.*

Recall next that a finitely presented group π is polycyclic-by-finite if it has a finite index subgroup which has a subnormal series where each quotient is cyclic. The minimal number of infinite cyclic quotients is an invariant of π called the *Hirsch length* of π and is denoted $h(\pi)$. The results of the following theorem all use Kreck's theory of modified surgery: the first three are [3, Theorem 5] and the fourth is [2, Theorem 1.1].

Theorem B. (Cancellation results for $q \geq 3$; [3, Thm. 5] and [2, Thm. 1.1]) *Let M be a compact $2q$-manifold of genus g with polycyclic-by-finite fundamental group π and let N be stably diffeomorphic to M with the same Euler characteristic as M.*

1. *If q is odd and π is trivial then M and N are diffeomorphic;*
2. *If π is trivial and $g \geq 1$, then M and N are diffeomorphic;*
3. *If π is finite and $g \geq 2$, then M and N are diffeomorphic;*
4. *If $g \geq h(\pi) + 3$, then M and N are diffeomorphic.*

We now state two problems relating the genus of M to the cancellation problem. The first of these was explained to the author by Ian Hambleton and uses the following further terminology: let π be a finitely presented group and $\varepsilon \in \{\pm 1\}$. We say that g_0 is a *ε-genus cancellation bound* for π if cancellation holds for every $2q$-manifold M with $\varepsilon = (-1)^q$, $\pi_1(M) \cong \pi$ and genus $g(M) \geq g_0$. If such a g_0 exists, the *ε-cancellation genus of π* is defined to be minimum genus cancellation bound

$$\mathrm{cg}_\varepsilon(\pi) := \min\{g_0 \mid g_0 \text{ is an } \varepsilon\text{-genus cancellation bound for } \pi\}.$$

If there is no ε-genus cancellation bound for π, we set $\mathrm{cg}_\varepsilon(\pi) = \infty$.

Problem A. (Genus bounds for general groups) *Is there an example of a finitely presented group π which is not polycyclic-by-finite and for which $\mathrm{cg}_\varepsilon(\pi) < \infty$ for some ε?*

It perhaps remarkable that Problem A is still open, but in fact our knowledge of the cancellation bound for almost all groups π is minimal. By Theorem B(1), we have $\mathrm{cg}_-(\{e\}) = 0$ and there examples which combine with Theorem B(2) to give $\mathrm{cg}_+(\{e\}) = 1$ and indeed $\mathrm{cg}_+(\pi) \geq 1$ for all π. However, there are no known examples where $\mathrm{cg}_\varepsilon(\pi) \geq 2$; i.e. the following problem is still open.

Problem B. ((s the cancellation genus ever greater than one?) *Is there a finitely presented group π and $\varepsilon \in \{\pm 1\}$ such that $\mathrm{cg}_\varepsilon(\pi) \geq 2$? i.e. is there a pair of $2q$-manifolds M and N with $\pi_1(M) \cong \pi_1(M) \cong \pi$ such that we have $M \sharp W_g \cong N \sharp W_g$ for some g but $M \sharp W_1$ and $N \sharp W_1$ are not diffeomorphic?*

Remark. One source of manifolds in the stable class of M comes from the action of the L-group, $L_{2q+1}(\mathbb{Z}[\pi], w_1)$, by Wall realisation. To be precise about torsions, the torsion requirements correspond to the L-group denoted $L^E_{2q+1}(\pi)$ in [7, 17 D] and this L-group is defined as the group of units in the little-ℓ surgery monoid $l_{2q+1}(\mathbb{Z}[\pi], w)$; see [3, p. 773]. Hence the formations and lagrangians are based, but the formation is *not* required to be simple and there is an exact sequence

$$0 \to L^s_{2q+1}(\mathbb{Z}[\pi], w_1) \to L_{2q+1}(\mathbb{Z}[\pi], w_1) \xrightarrow{\tau} \mathrm{Wh}(\pi),$$

where $\mathrm{Wh}(\pi)$ is the Whitehead group of π and the image of τ is described precisely in [2, Lemma 6.2].

The Wall realisation procedure entails that if $\rho \in L_{2q+1}(\mathbb{Z}[\pi], w_1)$ is represented by a formation on a hyperbolic form of rank $2g_0$ and if $M' = \rho M$, then we have $M \sharp W_{g_0} \cong M' \sharp W_{g_0}$. Moreover, applying [2, First theorem of §1.3], it follows that if $g(M) \geq g_0$, then $M' \cong M$. Hence the cancellation problem is related to the algebraic problem of the determining the minimal rank of a formation representing a given $\rho \in L_{2q+1}(\mathbb{Z}[\pi], w_1)$. For example, if every element of $L_{2q+1}(\mathbb{Z}[\pi], w_1)$ is represented by a formation of rank $2g_0$ or less and $g(M) \geq g_0$, then $L_{2q+1}(\mathbb{Z}[\pi], w_1)$ acts trivially on $\mathscr{S}^{\mathrm{st}}(M)$.

References

1. D. Crowley and J. Sixt, *Stably diffeomorphic manifolds and $l_{2q+1}(\mathbb{Z}[\pi])$*, Forum Math. **23** (2011), 483–538.
2. I. Hambleton and M. Kreck, *Cancellation of hyperbolic forms and topological four-manifolds*, J. Reine Angew. Math. **443** (1993), 21–47.
3. M. Kreck, *Surgery and Duality*, Ann. of Math. **149** (1999), 707–754.
4. C. T. C. Wall, *Surgery on compact manifolds*, Second edition. Edited and with a foreword by A. A. Ranicki. Mathematical Surveys and Monographs, **69**. American Mathematical Society, Providence, RI, 1999.

Problem 11: The Q-form Conjecture

presented by Diarmuid Crowley

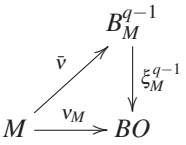
This problem follows on from the previous problem on genus bounds for cancellation. We use the same notation but now for simplicity *we assume that all manifolds are closed*, as well as connected. Recall that $\pi = \pi_1(M)$ and $w_1 = w_1(M)$ are the fundamental group and orientation character of M and that the L-group $L_{2q+1}(\mathbb{Z}[\pi], w_1)$ acts on the stable class of M via Wall realisation:

$$\mathscr{S}^{\mathrm{st}}(M) \times L_{2q+1}(\mathbb{Z}[\pi], w_1) \to \mathscr{S}^{\mathrm{st}}(M)$$

Given that the L-groups $L_{2q+1}(\mathbb{Z}[\pi], w_1)$ have been intensively studied, we focus on the quotient of the action above and suggest the following

Problem. *Determine* $\mathscr{S}^{\mathrm{st}}(M)/L_{2q+1}(\mathbb{Z}[\pi], w_1)$*, the set of orbits of the action of* $L_{2q+1}(\mathbb{Z}[\pi], w_1)$ *on the stable class.*

Below we present a conjectural solution to this problem, along with some evidence for the conjecture. To do this, we assume the reader is familiar with the setting of modified surgery; the details are found in [3, §1]. Let $\xi\colon B \to BO$ be a fibration over a connected space B. An m-dimensional *normal smoothing* in (B,ξ) is a pair $(M,\bar{\nu})$, where M is a compact m-manifold and $\bar{\nu}\colon M \to B$ is a lift of the stable normal bundle of M, $\bar{\nu}_M$, as in the following diagram:



If $\bar{\nu}$ is k-connected then $(M,\bar{\nu})$ is called a *normal* $(k-1)$*-smoothing* over (B,ξ) and if in addition ξ is k-coconnected, then the fibration ξ represents the normal $(k-1)$-type of M, which we denote by $\xi_M^{k-1}\colon B_M^{k-1} \to BO$. There is a well-defined notion of (B,ξ)-diffeomorphism, that is diffeomorphism preserving (B,ξ)-structures up to equivalence and also (B,ξ)-bordism of closed (B,ξ)-manifolds; the corresponding bordism group is denoted $\Omega_m(B;\xi)$. For an m-dimensional normal k-smoothing $(M,\bar{\nu})$ over (B,ξ), we let $[M,\bar{\nu}] \in \Omega_m(B;\xi)$ denote its bordism class and define

$$\begin{aligned}&\mathrm{NS}_{\xi}(M,\bar{\nu})\\ :=&\{(M',\bar{\nu}') \mid \chi(M') = \chi(M),\, [M'\bar{\nu}'] = [M,\bar{\nu}]\}/(B,\xi)\text{-diffeomorphism},\end{aligned}$$

to be the set of (B,ξ)-diffeomorphism classes of m-dimensional normal k-smoothings which are bordant to $(M,\bar{\nu})$ and have the same Euler characteristic as M.

For $m = 2q$ and $k = q-1$, a foundational result of Kreck [3, Corollary 3] states that if $(M_0,\bar{\nu}_0)$ and $(M_1,\bar{\nu}_1) \in \mathrm{NS}_{\xi}(M,\bar{\nu})$ then M_0 and M_1 are stably diffeomorphic. Combined with [2, Lemma 2.3], we obtain for $(B,\xi) = (B_M^{q-1},\xi_M^{q-1})$ that the forgetful map

$$\mathrm{F}\colon \mathrm{NS}_{\xi_M^{q-1}}(M,\bar{\nu}) \to \mathscr{S}^{\mathrm{st}}(M), \quad (M',\bar{\nu}') \mapsto M',$$

is onto. Moreover $\mathrm{aut}(\xi_M^{q-1})$, the group of fibre homotopy classes of fibre homotopy automorphisms of ξ_M^{q-1}, acts by post-composition $\mathrm{NS}_{\xi_M^{q-1}}(M,\bar{\nu})$ and by [4, Theorem 7.5], the universal properties of the Moore-Postnikov factorisation $\nu_M = \xi_M^{q-1} \circ \bar{\nu}$ ensure that the induced map

$$\mathrm{F}_{\mathrm{aut}(\xi_M^{q-1})}\colon \mathrm{NS}_{\xi_M^{q-1}}(M,\bar{\nu})/\mathrm{aut}(\xi_M^{q-1}) \to \mathscr{S}^{\mathrm{st}}(M) \tag{1}$$

is a bijection. Hence it makes sense to study $\mathrm{NS}_{\xi_M^{q-1}}(M,\bar{\nu})$ together with the action of $\mathrm{aut}(\xi_M^{q-1})$, in order to learn about $\mathscr{S}^{\mathrm{st}}(M)$.

We next define the key new invariant we shall use to formulate our conjectures and this is the *extended quadratic form* of $(M,\bar{\nu})$. Given $\xi : B \to BO$, we let $\pi = \pi_1(B)$ be the fundamental group of B and w_1 the orientation character of ξ. We fix a base-point in B and a local orientation of ξ at the base-point. We all assume that all normal smoothings $(M,\bar{\nu})$ over (B,ξ) are base-point preserving and that $\bar{\nu}_* : \pi_1(M) \to \pi_1(B)$ an isomorphism, which we use to identify $\pi_1(M) = \pi$. The local orientation of ξ gives M a local orientation and hence defines a fundamental class $[M] \in H_{2q}(M;\mathbb{Z}_{w_1})$ and also the equivariant intersection form $\lambda_{(M,\bar{\nu})} : H_q(M;\mathbb{Z}[\pi]) \times H_q(M;\mathbb{Z}[\pi]) \to \mathbb{Z}[\pi]$.

For every positive integer n, Ranicki [6, §10], defines a *quadratic form parameter* over the twisted group ring $(\mathbb{Z}[\pi], w_1)$, $Q_n(\xi)$, which is associated to the stable spherical fibration underlying the stable bundle ξ. In general, if we fix a ring with involution Λ, then a quadratic form parameter over Λ is a triple $Q = (Q, \mathrm{h}, \mathrm{p})$, written

$$Q = (Q \xrightarrow{\mathrm{h}} \Lambda \xrightarrow{\mathrm{p}} Q).$$

Here Q is an abelian group together with a *quadratic* action of Λ and h and p are equivariant homomorphisms with respect to the conjugation of $\mathbb{Z}[\pi]$ on itself, which satisfy certain equations. We refer the reader to [6, §10] for the details and point out that a similar but more general notion of quadratic form parameter can be found in the work of Baues [1]. We also mention that there is an exact sequence of abelian groups (see [6, p. 37])

$$Q_{(-1)^n}(\mathbb{Z}[\pi]) \to Q_\xi(n) \to H_n(B;\mathbb{Z}[\pi]) \to 0,$$

where $Q_{(-1)^n}(\mathbb{Z}[\pi])$ is the classical Q-group appearing in Wall's quadratic form [7, Theorem 5.2] and where the homomorphism $Q_\xi(n) \to H_n(B;\mathbb{Z}[\pi])$ is equal to the quotient map $Q_\xi(n) \to Q_\xi(n)/\mathrm{Im(p)}$.

An *extended quadratic form* over a form parameter Q, briefly a Q-form, is a triple

$$(H,\lambda,\mu),$$

where H is a Λ-module, $\lambda : H \times H \to \Lambda$ is a sesqui-linear form and $\mu : H \to Q$ is a quadratic refinement of λ which means in part that for all $x,y \in H$ we have

$$\mu(x+y) = \mu(x) + \mu(y) + \mathrm{p}(\lambda(x,y)) \quad \text{and} \quad \lambda(x,x) = \mathrm{h}(\mu(x)).$$

The *linearisation* of (H,λ,μ) is the Λ-module homomorphism

$$S(\mu) : H \to Q/\mathrm{Im(p)}, \quad x \mapsto [\mu(x)].$$

If (H,λ,μ) and (H',λ',μ') are Q-forms then an isometry between them is an λ-module isomorphism preserving the sesquilinear forms and their quadratic refinements and we write

$$\mathrm{Hom}_\Lambda(Q)$$

for the set of isometry classes of Q-forms on finitely generated Λ-modules.

The theory of [6, §10] ensures that a normal $(q-1)$-smoothing $\bar{\nu}\colon M \to B$ over a stable bundle $\xi\colon B \to BO$ defines a $Q_\xi(q)$-form

$$\mu(M,\bar{\nu}) := \big(H_q(M;\mathbb{Z}[\pi]), \lambda_{(M,\bar{\nu})}, \mu(\bar{\nu})\big),$$

where $(H_q(M;\mathbb{Z}[\pi]), \lambda_{(M,\bar{\nu})})$ is the equivariant intersection form of $(M,\bar{\nu})$ and the map $\mu(\bar{\nu})\colon H_q(M;\mathbb{Z}[\pi]) \to Q_\xi(q)$ is a quadratic refinement of $\lambda_{(M,\bar{\nu})}$, which has linearisation

$$S(\mu(\bar{\nu})) = \bar{\nu}_*\colon H_q(M;\mathbb{Z}[\pi]) \to H_q(B;\mathbb{Z}[\pi]).$$

It follows from the definitions that if $f\colon M_0 \to M_1$ is a (B,ξ)-diffeomorphism between $2q$-dimensional $(q-1)$-smoothings $(M_0,\bar{\nu}_0)$ and $(M_1,\bar{\nu}_1)$ over (B,ξ), then the induced homomorphism $f_*\colon H_q(M_0;\mathbb{Z}[\pi]) \to H_q(M_1;\mathbb{Z}[\pi])$ is an isometry of $Q_\xi(q)$-forms. It follows that there is a well-defined map

$$\mathrm{NS}_\xi(M,\bar{\nu}) \to \mathrm{Hom}_{(\mathbb{Z}[\pi],w_1)}(Q_\xi(q)), \quad (M,\bar{\nu}) \mapsto \mu(M,\bar{\nu}).$$

Now Wall realisation also defines an action of $L_{2q+1}(\mathbb{Z}[\pi],w_1)$ on $\mathrm{NS}(M,\bar{\nu})$ and it is elementary to check that the isometry class of the extended quadratic forms is invariant under this action. Hence the map above descends to define the map

$$\mu\colon \mathrm{NS}_\xi(M,\bar{\nu})/L_{2q+1}(\mathbb{Z}[\pi],w_1) \to \mathrm{Hom}_{(\mathbb{Z}[\pi],w_1)}(Q_\xi(q)). \tag{2}$$

At last, we can state the first version of the Q-form Conjecture.

Conjecture A. (The Q-form Conjecture for normal smoothings) If $q \geq 3$, then the map μ of (2) is injective; i.e. if $q \geq 3$ and $(M_0,\bar{\nu}_0)$ and $(M_1,\bar{\nu}_1)$ are $2q$-dimensional (B,ξ)-bordant normal $(q-1)$-smoothings with equal Euler characteristic and isometric $Q_\xi(q)$-forms, then $(M_0,\bar{\nu}_0)$ and $(M_1,\bar{\nu}_1)$ differ by the action of $L_{2q+1}(\mathbb{Z}[\pi],w_1)$.

Given the bijection of (1), Conjecture A allows us to formulate a conjectural determination of the stable class of M, at least for $q \geq 3$. For this, note that $\mathrm{aut}(\xi_M^{q-1})$ acts on $Q_\xi(q)$ by automorphisms and hence on $\mathrm{Hom}_{(\mathbb{Z}[\pi],w_1)}(Q_\xi(q))$ by post-composition. Thus we obtain the map

$$\mu_{/\mathrm{aut}(\xi)}\colon \mathrm{NS}_\xi(M,\bar{\nu})/\big(L_{2q+1}(\mathbb{Z}[\pi],w_1) \times \mathrm{aut}(\xi)\big) \to \mathrm{Hom}_{(\mathbb{Z}[\pi],w_1)}(Q_\xi(q))/\mathrm{aut}(\xi),$$

which is a bijection if Conjecture A holds. Since the bijection of (1) is equivariant with respect to the action of $L_{2q+1}(\mathbb{Z}[\pi],w_1)$, when $\xi = \xi_M^{q-1}$ is a representative of the normal $(q-1)$-type of M, the map $\mu_{/\mathrm{aut}(\xi)}$ induces another map, also denoted $\mu_{/\mathrm{aut}(\xi)}$,

$$\mu_{/\mathrm{aut}(\xi)}\colon \mathscr{S}^{\mathrm{st}}(M)/L_{2q+1}(\mathbb{Z}[\pi],w_1) \to \mathrm{Hom}_{(\mathbb{Z}[\pi],w_1)}(Q_{\xi_M^{q-1}}(q))/\mathrm{aut}(\xi_M^{q-1}). \tag{3}$$

Conjecture B. (The Q-form Conjecture for the stable class) If $q \geq 3$, then the map $\mu_{/\mathrm{aut}(\xi)}$ of (3) is injective; i.e. if $q \geq 3$ and we have $M_0, M_1 \in \mathscr{S}^{\mathrm{st}}(M)$ then

$M_0 \cong \rho M_1$ for some $\rho \in L_{2q+1}(\mathbb{Z}[\pi], w_1)$ if and only if for $i = 0, 1$, there are normal $(q-1)$-smoothings $\bar{\nu}_i \colon M_i \to B_M^{q-1}$ such that $\mu(M_0, \bar{\nu}_0)$ and $\mu(M_1, \bar{\nu}_1)$ are isometric $Q_{\xi_M^{q-1}}(q)$-forms.

We conclude by briefly discussing Conjectures A and B. Notice that since the map μ of Conjecture A is $\mathrm{aut}(\xi_M^{q-1})$-equivariant, Conjecture A implies Conjecture B. Both conjectures are inspired by the classification of the ℓ-monoids in [2] and to the best of our knowledge, both conjectures are consistent with the extensive literature on classifying $2q$-manifolds for $q \geq 3$. In addition, Conjecture A (hence Conjecture B) has been proven by Nagy in the case where q is even, $\pi = \{e\}$ and $H_q(B; \mathbb{Z})$ is torsion free [5].

At times, it has been tempting to propose Conjectures A and B as *hypotheses*; i.e. as sign posts for organising work on the classification of the stable class, as opposed to statements believed to be true. However, the resilience of these statements to date encourages their proposal as conjectures in the usual sense. This is also consistent with history of the exploration of the stable class, where the "unreasonable effectiveness" of the (equivariant) intersection form has often been observed.

References

1. H. J. Baues, *Quadratic functors and metastable homotopy*, J. Pure & App. Alg. **91** (1994) 49–107.
2. D. Crowley and J. Sixt, *Stably diffeomorphic manifolds and $l_{2q+1}(\mathbb{Z}[\pi])$*, Forum Math. **23** (2011), 483–538.
3. M. Kreck, *Surgery and Duality*, Ann. of Math. **149** (1999), 707–754.
4. M. Kreck, *An extension of the results of Browder, Novikov and Wall about surgery on compact manifolds*, preprint Mainz (1985).
   Availalbe at http://www.map.mpim-bonn.mpg.de/Template:Kreck1985
5. Cs. Nagy, *The classification of 8-dimensional E-manifolds*, PhD Thesis, University of Melbourne. In preparation 2020.
6. A. A. Ranicki, *Algebraic Poincaré cobordism*, Topology, geometry, and algebra: interactions and new directions (Stanford, CA, 1999), Contemp. Math., 279, Amer. Math. Soc., Providence, RI, (2001), 213–255.
7. C. T. C. Wall, *Surgery on compact manifolds*, Second edition. Edited and with a foreword by A. A. Ranicki. Mathematical Surveys and Monographs, **69**. American Mathematical Society, Providence, RI, 1999.

Chapter 11

Aperiodic Order meets Number Theory

Aperiodic order meets number theory: Origin and structure of the field

M. Baake, M. Coons, U. Grimm, J. A. G. Roberts and R. Yassawi

Abstract Aperiodic order is a relatively young area of mathematics with connections to many other fields, including discrete geometry, harmonic analysis, dynamical systems, algebra, combinatorics and, above all, number theory. In fact, number-theoretic methods and results are present in practically all of these connections. It was one aim of this workshop to review, strengthen and foster these connections.

Aperiodic structures and patterns have revolutionised parts of science, as evidenced by the 2011 Nobel Prize in Chemistry awarded to Dan Shechtman for the discovery of quasicrystals. Beautiful, yet profound, examples in mathematics have captured the attention of many, starting with the famous Penrose tiling from 1974. The deep connection between these topics emerged from the number-theoretic work of Yves Meyer (Abel Prize 2017). During this workshop, an international community of like-minded researchers came together to discuss recent results and develop research collaborations at the interface of aperiodic theory and number theory.

From the very beginnings of research on aperiodic order, intriguing links to number theory were observed; these have become increasingly apparent in recent years. Over the past decade, there has been a tremendous development in various directions, including spectra and transport theory of Schrödinger operators, reversing symmetry groups in dynamical systems and ergodic theory, and topological invariants in symbolic and algebraic dynamics. Additionally, the links between number

Michael Baake
Bielefeld University, Germany. e-mail: mbaake@math.uni-bielefeld.de

Michael Coons
University of Newcastle, Australia. e-mail: michael.coons@newcastle.edu.au

Uwe Grimm
The Open University, Milton Keynes, UK. e-mail: uwe.grimm@open.ac.uk

John A.G. Roberts
UNSW, Sydney, Australia. e-mail: jag.roberts@unsw.edu.au

Reem Yassawi
Université Claude Bernard Lyon 1, Villeurbanne, France. e-mail: ryassawi@gmail.com

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_40

theory, dynamical systems, and theoretical computer science are strengthening, and the lines between them are blurring.

Aperiodic order [36, 5, 6, 24] has several roots in mathematics — predating even the main impetus for the area: the discovery of quasicrystals. These include Harald Bohr's development of the theory of almost periodic functions [13], Robert Berger's proof of the undecidability of the tiling problem [11], and Yves Meyer's work on model sets [35], which was later developed further in [27, 37, 28, 38, 29], as well as early works on tilings and patterns including Roger Penrose's famous fivefold tiling of the plane [40]. From the very start, there have been connections to various areas of mathematics. Number theory features prominently, for instance in the early work of Peter Pleasants, as reviewed in [42]. Arguably the most obvious relation occurs for planar tilings with (non-crystallographic) rotational symmetry, which are closely related to rings of integers in cyclotomic fields.

At the same time, within number theory, the advent of modern (digital) computation has underscored the importance of understanding the relationship between base expansions and algebraic operations. This topic has a rich history, especially in Australian mathematics through the work of Loxton, Mahler and van der Poorten. It focussed mainly on results related to finite automata and their generalisations — structures of importance in aperiodic order.

Recent results and questions at the intersection of aperiodic order and number theory include the following.

- The study of weak model sets [7, 25] was partially motivated by the set of visible points of the integer lattice and the set of kth power-free integers [8, 43], and their connections to Sarnak's programme on the Möbius disjointness conjecture; see [39, 31] for recent developments. Under a rather natural extremality assumption, it is possible to establish pure point spectrum, and that such systems can be seen as natural generalisations of regular model sets.
- The diffraction theory of infinite point sets in Euclidean space with its corresponding inverse problem shows fundamental connections to almost periodicity [32, 48, 47] and Lyapunov exponents [34]. Likewise, there are similar structures in the theory of Schrödinger operators, with surprising applications to spectra on graphs [17].
- Rather than using diffraction, connections with Diophantine approximations can be used to investigate and quantify the nature of order in a cut and project set [21]. Quantities of interest here include the complexity function and the repetitivity function. Another interesting characteristic is the discrepancy, which describes the difference between the expected and actual number of appearances of a given patch in a large region.
- Constant-length substitutions are important objects for both number theory and aperiodic order. A generalisation of these are regular sequences, which are related to finitely generated semigroups of matrices. The growth properties of these sequences are related to questions in both areas, including spectral properties [1] as well as the finiteness conjecture for integer matrices [2, 30]. These results are

connected with the scaling structure of singular continuous measures, as recently analysed for the Thue–Morse measure; see [4] and reference therein.

- Logarithmic Mahler measures occur as the maximal Lyapunov exponents of matrix cocycles for binary constant-length substitutions [3]. In this way, Lehmer's problem for height-one polynomials having minimal Mahler measure becomes equivalent to a natural question from the spectral theory of binary constant-length substitutions. This supports another connection between Mahler measures and dynamics, beyond the well-known appearance of Mahler measures as entropies in algebraic dynamics [44].
- One of the questions that evolved from the early work on Wang tilings [11] and has attracted attention over the years is the question of the minimal set of tiles required to enforce quasiperiodicity. For a long time, it seemed that substitution-based structures were the way to go, until Kari [23] and Culik [16] came up with an ingenious way of assigning rational edge values to Wang tiles in a way that rules out periodicity by arithmetic constraints. This system continues to attract attention, see for instance [45, 22, 26], but the question whether it opens up a new approach to aperiodic structures remains to be explored. Closely related is the search for planar monotiles of hexagonal shape [41, 46].
- Constant-size substitutions and characteristic-p S-unit equations are also connected to the question of mixing in algebraic dynamics, as described recently in work by Derksen and Masser [18, 19]. Techniques in these articles should shed light on the nature of the symmetry groups of these algebraic dynamical systems, thus providing additional and powerful methods for the characterisation of (extended) symmetry groups in algebraic dynamics, as recognised in [9, 20, 15, 14], as well as the analysis of related systems [12].

References

1. Akiyama, S., Gähler, F., Lee, J.-Y.: Determining pure discrete spectrum for some self-affine tilings. Discr. Math. Theor. Comput. Sci. **14**, 305–316 (2014)
2. Baake, M., Coons, M.: A probability measure derived from Stern's diatomic sequence. Acta Arithm. **183**, 87–99 (2018). arXiv:1706.00187
3. Baake, M., Coons, M., Mañibo, N.: Binary constant-length substitutions and Mahler measures of Borwein polynomials. In: From Analysis to Visualization, eds. D.H. Bailey, N.S. Borwein, R.P. Brent, R.S. Burachik, J.H. Osborn, B. Sims and Q.J Zhu, PROMS 313, Springer, Cham (2020), pp. 303–322.
4. Baake, M., Gohlke, P., Kesseböhmer, M., Schindler, T.: Scaling properties of the Thue–Morse measure. Discr. Cont. Dynam. Syst. A **39**, 4157–4185 (2019).
5. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
6. Baake, M., Grimm, U. (eds.): Aperiodic Order. Vol. 2: Crystallography and Almost Periodicity. Cambridge University Press, Cambridge (2017)
7. Baake, M., Huck, C., Strungaru, N.: On weak model sets of extremal density. Indag. Math. **28**, 3–31 (2017)
8. Baake, M., Moody, R.V., Pleasants, P.A.B.: Diffraction from visible lattice points and k-th power free integers. Discr. Math. **221**, 3–42 (2000)

9. Baake, M., Roberts, J.A.G., Yassawi, R.: Reversing and extended symmetries of shift spaces. Discr. Cont. Dynam. Syst. A **38**, 835–866 (2018)
10. Baake, M., Scharlau, R., Zeiner, P.: Well-rounded sublattices of planar lattices. Acta Arithm. **166**, 301–334 (2014)
11. Berger, R.: The undecidability of the domino problem. Mem. Amer. Math. Soc. **66**, 1–72 (1966)
12. Berthé, V., Cecchi Bernales, P.: Balances and coboundaries in symbolic systems. Theor. Comput. Sci. **777**, 93–110 (2019)
13. Bohr, H.: Fastperiodische Funktionen. Springer, Berlin (1932)
14. Bustos, Á.: Extended symmetry groups of multidimensional subshifts with hierarchical structure, Discr. Cont. Dynam. Syst. A **40**, 5869–5895 (2020)
15. Cortez, M.I., Petite, S.: Realization of big centralizers of minimal aperiodic actions on the Cantor set, Discr. Cont. Dynam. Syst. A **40**, 2891–2901 (2020)
16. Culik, K.: An aperiodic set of 13 Wang tiles. Discr. Math. **160**, 245–251 (1996)
17. Damanik, D., Fillman, J., Sukhtaiev, S.: Localization for Anderson models on metric and discrete tree graphs. Math. Ann. **376**, 1337–1393 (2020)
18. Derksen, H., Masser, D.: Linear equations over multiplicative groups, recurrences, and mixing I. Proc. London Math. Soc. **104**, 1045–1083 (2012)
19. Derksen, H., Masser, D.: Linear equations over multiplicative groups, recurrences, and mixing II. Indag. Math. **26**, 113–136 (2015)
20. Fokkink, R., Yassawi, R.: Topological rigidity of linear cellular automaton shifts. Indag. Math. **29**, 1105–1113 (2018)
21. Haynes, A., Julien, A., Koivusalo, H., Walton, J.: Statistics of patterns in typical cut and project sets. Erg. Th. Dynam. Syst. **39**, 3365–3387 (2019)
22. Jeandel, E., Rao, M.: An aperiodic set of 11 Wang tiles. Preprint arXiv:1506.06492
23. Kari, J.: A small aperiodic set of Wang tiles. Discr. Math. **160**, 259–264 (1996)
24. Kellendonk, J., Lenz, D., Savinien J. (eds.): Mathematics of Aperiodic Order. Birkhäuser, Basel (2015)
25. Keller, G., Richard, C.: Periods and factors of weak model sets. Israel J. **229**, 85–132 (2019)
26. Labbé, S.: Substitutive structure of Jeandel–Rao aperiodic tilings. Discr. Comput. Geom., in press. arXiv:1808.07768
27. Lagarias, J.C.: Meyer's concept of quasicrystal and quasiregular sets. Commun. Math. Phys. **179**, 365–376 (1996)
28. Lagarias, J.C.: Geometric models for quasicrystals I. Delone sets of finite type. Discr. Comput. Geom. **21**, 161–191 (1999)
29. Lagarias, J.C., Pleasants, P.A.B.: Repetitive Delone sets and quasicrystals. Ergod. Th. & Dynam. Syst. **23**, 831–867 (2003)
30. Lagarias, J.C., Wang, Y.: The finiteness conjecture for the generalized spectral radius of a set of matrices. Lin. Alg. Appl. **214**, 17–42 (1995)
31. Lemańczyk, M., Müllner, C.: Automatic sequences are orthogonal to aperiodic multiplicative functions. Discr. Cont. Dynam. Syst. A **40**, 6877–6918 (2020)
32. Lenz, D., Strungaru, N.: On weakly almost periodic measures. Trans. Amer. Math. Soc. **371**, 6843–6881 (2019)
33. Loquias, M.J.C, Zeiner, P.: The coincidence problem for shifted lattices and crystallographic point packings. Acta Cryst. A **70**, 656–669 (2014)
34. Mañibo, N.: Lyapunov exponents for binary substitutions of constant length. J. Math. Phys. **58**, 113504:1–9 (2017)
35. Meyer, Y.: Algebraic Numbers and Harmonic Analysis. North-Holland, Amsterdam (1972)
36. Moody, R.V. (ed.): The Mathematics of Long-Range Aperiodic Order, NATO ASI Series C 489. Kluwer, Dordrecht (1997)
37. Moody, R.V.: Meyer sets and their duals. In [36], pp. 403–441 (1997)
38. Moody, R.V.: Model sets: A Survey. In: Axel, F., Dénoyer, F., Gazeau, J.P. (eds.), From Quasicrystals to More Complex Systems, pp. 145–166. EDP Sciences, Les Ulis, and Springer, Berlin (2000)

39. Müllner, C.: Automatic sequences fulfill the Sarnak conjecture. Duke Math. J. **166**, 3219–3290 (2017)
40. Penrose, R.: The role of aesthetics in pure and applied mathematical research. Bull. Inst. Math. Appl. **10**, 266–271 (1974)
41. Penrose, R.: Remarks on a tiling: Details of a $(1+\varepsilon+\varepsilon^2)$-aperiodic set. In: Moody R.V. (ed.) The Mathematics of Long-Range Aperiodic Order, pp. 467–497. Kluwer, Dordrecht (1997)
42. Pleasants, P.A.B.: Designer quasicrystals: Cut-and-project sets with pre-assigned properties. In: Baake, M., Moody, R.V. (eds.) Directions in Mathematical Quasicrystals, pp. 95–141. AMS, Providence, RI (2000)
43. Pleasants, P.A.B., Huck, C.: Entropy and diffraction of the k-free points in n-dimensional lattices. Discr. Comput. Geom. **50**, 39–68 (2013)
44. Schmidt, K.: Dynamical Systems of Algebraic Origin. Birkhäuser, Basel (1995)
45. Siefken, J.: A minimal subsystem of the Kari–Culik tilings. Ergodic Th. & Dynam. Syst. **37**, 1607–1634 (2017)
46. Socolar, J., Taylor, J.: An aperiodic hexagonal tile. J. Comb. Theory A **118**, 2207–2231 (2011)
47. Terauds, V.: The inverse problem of pure point diffraction — examples and open questions. J. Stat. Phys. **152**, 954–968 (2013)
48. Terauds, V., Strungaru, N.: Diffraction theory and almost periodic distributions. J. Stat. Phys. **164**, 1183–1216 (2016)

Delone sets on spirals

Shigeki Akiyama

Motivated by phyllotaxis in botany, the angular development of plants widely found in nature, we give a simple mathematical characterization of Delone sets on spirals.

Let X be a subset of $\mathbb{R}^2$ which is identified with the complex plane $\mathbb{C}$. Denote by $B(x,r)$ the open ball of radius r centered at x. We say X is *relatively dense* if there exists $r>0$ such that, for any $x\in\mathbb{C}$, $B(x,r)\cap X\neq\emptyset$ holds, X is *uniformly discrete* if there exists $r>0$ such that, for any $x\in\mathbb{C}$, we have $\mathrm{card}(B(x,r)\cap X)\leq 1$, and finally X is a *Delone set* if it is both relatively dense and uniformly discrete.

Set $\mathbf{e}(z)=e^{2\pi iz}$. Fix an *angle* $\alpha\in[0,1)$ and a strictly increasing function f from $\mathbb{R}_{\geq 0}$ to itself. We wish to characterize when the set

$$X_f=\left\{f(n)\mathbf{e}(n\alpha)\mid n\in\mathbb{N}\right\}$$

on a spiral curve $\left\{f(t)\mathbf{e}(t\alpha)\mid t\in\mathbb{R}_{\geq 0}\right\}$ forms a Delone set.

Let us collect necessary conditions. Clearly X_f is not relatively dense if the angle α is rational, since X_f is contained in a union of a finite number of lines passing through the origin. An easy discussion leads to the following result.

Lemma 1. *If X_f is relatively dense, then* $\limsup_{n\to\infty} f(n)/\sqrt{n}<\infty$. *If X_f is uniformly discrete, then* $\liminf_{n\to\infty} f(n)/\sqrt{n}>0$.

Hereafter, we assume that $f(n)=\sqrt{n}$ and that α is irrational, and study the set

$$X(\alpha)=\left\{\sqrt{n}\,\mathbf{e}(n\alpha)\mid n\in\mathbb{N}\right\}.$$

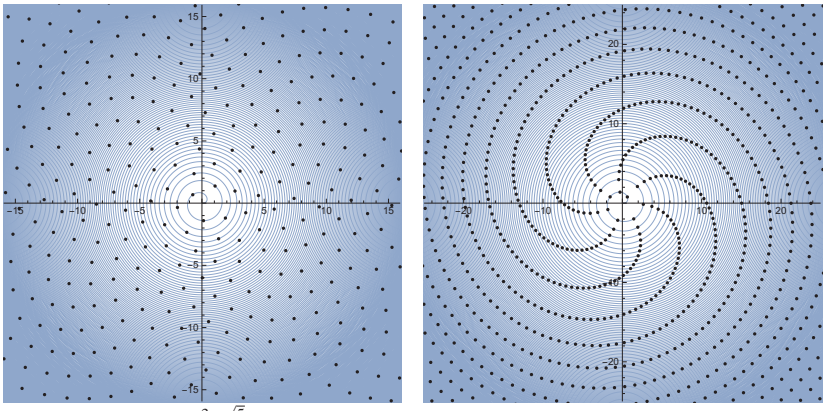
In other words, we are interested in the sequence of points on the Fermat spiral that progresses by a constant angle α (see Figure 1).

A real number α is *badly approximable* if there exists a positive constant C so that

$$q|q\alpha-p|\geq C$$

Shigeki Akiyama
University of Tsukuba, Japan. e-mail: akiyama@math.tsukuba.ac.jp

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_41



(a) $X(\frac{3-\sqrt{5}}{2})$ (b) $X(\pi-3)$

Fig. 1 Constant angular progressions on Fermat spiral

holds for all $(p,q)\in\mathbb{Z}\times\mathbb{N}$. It is well-known that α is badly approximable if and only if the partial quotients of the continued fraction expansion of α are bounded (compare [3, Theorem 23] and [1, Theorem 1.9]). In particular, if α is a real quadratic irrational, then α is badly approximable, due to Lagrange's theorem.

With the help of the three distance theorem on irrational rotation conjectured by Steinhaus and proved by Sós [5, 6] and then Świerczkowski [8], Surányi [7], Halton [2] and Slater [4], we can prove the following result.

Theorem 1. *The following four statements are equivalent.*

a) $X(\alpha)$ *is relatively dense,*
b) $X(\alpha)$ *is uniformly discrete,*
c) $X(\alpha)$ *is a Delone set,*
d) the angle α *is badly approximable.*

Searching for a possible higher-dimensional extension is an interesting problem.

References

1. Bugeaud, Y.: Approximation by Algebraic Numbers. Cambridge University Press, Cambridge (2004)
2. Halton, J.-H.: The distribution of the sequence $\{n\xi\}$ $(n=0,1,2,\cdots)$. Proc. Cambridge Philos. Soc. **61**, 665–670 (1965)
3. Khinchin, A.Ya.: Continued Fractions. The University of Chicago Press, Chicago, IL (1964)
4. Slater, N.B.: Gaps and steps for the sequence $n\theta$ mod 1. Proc. Cambridge Philos. Soc. **63**, 1115–1123 (1967)
5. Sós, V.T.: On the theory of diophantine approximations I. Acta Math. Acad. Sci. Hungar. **8**, 461–472 (1957)

6. Sós, V.T.: On the distribution mod 1 of the sequence $n\alpha$. Ann. Univ. Sci. Budapest Eötvös Sect. Math. **1**, 127–134 (1958)
7. Surányi, J.: Über die Anordnung der Vielfachen einer reellen Zahl mod 1. Ann. Univ. Sci. Budapest Eötvös Sect. Math. **1**, 107–111 (1958)
8. Świerczkowski, S.: On successive settings of an arc on the circumference of a circle. Fund. Math. **46**, 187–189 (1959)

Topological methods for symbolic discrepancy

Valérie Berthé

In this lecture, we discuss the notion of bounded symbolic discrepancy for infinite words and subshifts, both for letters and factors, from a topological dynamics viewpoint. We focus on three families of words, namely hypercubic words, words generated by substitutions, and dendric words. Symbolic discrepancy measures the difference between the numbers of occurrences of a given word v in some word of length n minus n times the frequency μ_v of v when it exists (in other words, μ_v is the measure of the cylinder $[v]$ for some invariant measure μ). Bounded discrepancy thus provides particularly strong convergence properties of ergodic sums toward frequencies. More precisely, let $u \in \mathscr{A}^{\mathbb{Z}}$ be a bi-infinite word and assume that each factor v in its language admits a frequency μ_v in u. The *discrepancy* $\Delta_v(u)$ of u with respect to v is defined as

$$\Delta_v(u) = \sup_{n\in\mathbb{N}} \big| |u_{-n}\cdots u_0 \cdots u_n|_v - (2n+1)\mu_v \big|.$$

This notion extends to any minimal subshift (X,T) in a straightforward way. If $\Delta_v(u)$ is finite, the cylinder $[v]$ is said to be a bounded remainder set, according to the terminology developed in classical discrepancy theory. Bounded discrepancy is closely related to the notion of balance in word combinatorics.

To illustrate the relevance of the topological approach for symbolic discrepancy, let us start with a first classical remark. Let (X,T) be a minimal and uniquely ergodic subshift and let μ stand for its invariant measure. Given a factor v in its language, define $f_v = \chi_{[v]} - \mu([v]) \in C(X,\mathbb{R})$, where $\chi_{[v]}$ stands for the characteristic function of the cylinder $[v]$. Then, according to the Gottschalk–Hedlund theorem, v has bounded discrepancy in (X,T) if and only if the map f_v is a coboundary. Bounded discrepancy thus implies that $\mu([v])$ is an additive topological eigenvalue of (X,T).

Here, we deduce that for hypercubic words produced by d-to-1 cut-and-project schemes (with irrationality assumptions that yield minimality), letters have bounded

Valérie Berthé
IRIF, CNRS UMR 8243, Université Paris Diderot – Paris 7, France. e-mail: berthe@irif.fr

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_42

discrepancy, whereas factors of length at least 2 do not have bounded discrepancy for $d \geq 3$. Indeed, frequencies of factors of length at least 2 do not belong to the group of additive eigenvalues.

In the substitutive case, we stress the role played by the existence of coboundaries taking rational values and show simple criteria when frequencies take rational values for exhibiting unbounded discrepancy. For more precise results, see [1].

The third family we consider here is the family of dendric words, and we present results from [3]. Given a subshift over a finite alphabet, one can associate with every word in the associated language a bipartite graph, called extension graph, in which one puts edges between left and right letter extensions of this factor in the language. If, for every word in this language, the extension graph is a tree, then the subshift is a dendric subshift. Dendric subshifts are therefore defined in terms of combinatorial properties of their language. This class of linear factor complexity subshifts encompasses Sturmian subshifts, Arnoux–Rauzy subshifts, as well as subshifts generated by regular interval exchanges. We study the dimension group of dendric subshifts, providing necessary and sufficient conditions for two dendric subshifts to be (strongly) orbit equivalent. More precisely, let (X,T) be a minimal dendric subshift on the alphabet $\mathscr{A} = \{1,\cdots,d\}$ and let $\mathscr{M}(X,T)$ stand for its set of invariant measures. Then, its dimension group with ordered unit is isomorphic to

$$\left(\mathbb{Z}^d, \left\{\boldsymbol{x} \in \mathbb{Z}^d \mid \langle \boldsymbol{x}, \boldsymbol{\mu} \rangle > 0 \text{ for all } \mu \in \mathscr{M}(X,T)\right\} \cup \left\{0\right\}, \vec{1}\right)$$

where $\boldsymbol{\mu}$ denotes the vector $\big(\mu([1]),\cdots,\mu([d])\big)$. We deduce that, as soon as dendric words are balanced on letters, they are balanced on factors. The proof relies on the following property of dendric subshift from [3]: let X be a minimal dendric subshift defined on the alphabet $\mathscr{A}$. Then, for any w in its language, the set of left return words to w is a basis of the free group over $\mathscr{A}$.

References

1. Berthé, V., Cecchi Bernales, P.: Balances and coboundaries in symbolic systems. Theor. Comput. Sci. **777**, 93–110 (2019). arXiv:1810.07453
2. Berthé, V., Cecchi Bernales, P., Durand, F., Leroy, J., Perrin, D., Petite, S.: Dimension groups of dendric subshifts, preprint arXiv:1911.07700
3. Berthé, V., De Felice, C., Dolce, F., Leroy, J., Perrin, D., Reutenauer, C., Rindone, G.: Acyclic, connected and tree sets. Monatsh. Math. **176**, 521–550 (2015)

Extended symmetry groups of multidimensional subshifts with hierarchical structure

Álvaro Bustos

In this contribution, we discuss the automorphism group, i.e., the centralizer of the shift action inside the group of self-homeomorphisms of a subshift, together with the extended symmetry group (the corresponding normalizer) of certain $\mathbb{Z}^d$ subshifts with a hierarchical structure, like bijective substitutive subshifts and the Robinson tiling. This group has been previously studied in the work of Baake, Roberts and Yassawi [1], among others.

Treating these subshifts as geometric objects, we introduce techniques to identify allowed extended symmetries from large-scale structures present in certain special points of the subshift, leading to strong restrictions on the group of extended symmetries. We prove that, in the aforementioned cases, $\mathrm{Sym}(X,\mathbb{Z}^d)$ (and thus $\mathrm{Aut}(X,\mathbb{Z}^d)$) is virtually-$\mathbb{Z}^d$, and we explicitly represent the non-trivial extended symmetries, associated with the quotient $\mathrm{Sym}(X,\mathbb{Z}^d)/\mathrm{Aut}(X,\mathbb{Z}^d)$, as a subset of rigid transformations of the coordinate axes. We also show how our techniques carry over to the study of the Robinson tiling, both in its minimal and non-minimal version. We emphasize the geometric nature of these techniques and how they reflect the capability of extended symmetries to capture such properties in a subshift.

Our discussion starts with the computation of the automorphism group for d-dimensional substitutive subshifts coming from bijective rectangular substitutions. By an application of desubstitution and some algebraic manipulations, we generalize Coven's theorem (see [2]) by showing the following.

Theorem 1. *For a non-trivial, primitive, bijective substitution θ on the alphabet $\mathscr{A} = \{0,1\}$, $\mathrm{Aut}(X_\theta,\mathbb{Z}^d)$ is generated by the shifts and the relabeling map (flip map) $\delta(x) := \bar{x}$, where $\bar{x}$ represents the sequence obtained from x by swapping all* 1*s with* 0*s and vice versa, and thus is isomorphic to $\mathbb{Z}^d \times (\mathbb{Z}/2\mathbb{Z})$.*

For more general alphabets, the same method of proof yields the following.

Álvaro Bustos
Departamento de Ingeniería Matemática, Universidad de Chile, Beauchef 851, Santiago, Chile.
e-mail: abustos@dim.uchile.cl

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_43

Corollary 1. *Let θ be a non-trivial, primitive, bijective substitution on an alphabet $\mathscr{A}$ with at least two symbols. For any $f \in \mathrm{Aut}(X_\theta, \mathbb{Z}^d)$, there exists a bijection $\tau\colon \mathscr{A} \to \mathscr{A}$ and a value $\boldsymbol{k} \in \mathbb{Z}^d$ such that $f = \sigma_{\boldsymbol{k}} \circ \tau_\infty$. Thus, $\mathrm{Aut}(X_\theta, \mathbb{Z}^d)$ is isomorphic to a subgroup of $\mathbb{Z}^d \times S_{|\mathscr{A}|}$, where S_n is the symmetric group in n elements.*

Next, we divert our attention towards *extended symmetries*, which are a generalization of shift automorphisms, in the sense that they are homeomorphisms $f\colon X \to X$ such that there exists a matrix $A_f \in \mathrm{GL}_d(\mathbb{Z})$ for which the following identity holds,

$$\forall \boldsymbol{n} \in \mathbb{Z}^d : f \circ \sigma_{\boldsymbol{n}} = \sigma_{A_f \boldsymbol{n}} \circ f.$$

The set of all such homeomorphisms is a group, $\mathrm{Sym}(X, \mathbb{Z}^d)$, which is a group extension of $\mathrm{Aut}(X, \mathbb{Z}^d)$ by some subgroup of $\mathrm{GL}_d(\mathbb{Z})$. Extended symmetries satisfy a variant of the Curtis–Hedlund–Lyndon theorem and thus are completely determined by a local mapping $F\colon \mathscr{A}^U \to \mathscr{A}$ (with $U \subset \mathbb{Z}^d$ finite) and the matrix A_f.

We devise a 'fracture method' in which we recognize special pairs of points from a subshift which match only on a half-space, in such a way that the discrepancy in the other half-space is preserved in the images under f by an application of the Curtis–Hedlund–Lyndon theorem. By showing limitations on the possible directions of these fractures, we can compute the extended symmetry group of several subshifts.

It is known that the automorphism group of the Robinson shift X_{Rob} and its minimal subshift M_{Rob} is isomorphic to $\mathbb{Z}^2$ (see e.g. [3]). We show the following.

Proposition 1. *For the Robinson shift, $\mathrm{Sym}(X_{\mathrm{Rob}}, \mathbb{Z}^2) \cong \mathbb{Z}^2 \rtimes D_4$, where D_4 is the dihedral group of order 8. The same holds for M_{Rob}.*

In the case of substitutive subshifts coming from bijective substitutions, the desubstitution technique allows us to apply a variant of the above fracture argument; the bijectiveness imposes a restriction on the possible directions on fracture, which lead to the following result.

Theorem 2. *For a d-dimensional, non-trivial, primitive, bijective substitution θ, the quotient group of all admissible lattice transformations of the subshift X_θ, $\mathrm{Sym}(X_\theta, \mathbb{Z}^d)/\mathrm{Aut}(X_\theta, \mathbb{Z}^d)$, is isomorphic to a subset of the hyperoctahedral group $Q_d \cong (\mathbb{Z}/2\mathbb{Z}) \wr S_d = (\mathbb{Z}/2\mathbb{Z})^d \rtimes S_d$, which is the symmetry group of the d-dimensional cube. Thus, the extended symmetry group $\mathrm{Sym}(X_\theta, \mathbb{Z}^d)$ is virtually-$\mathbb{Z}^d$.*

Acknowledgements The author was supported by ANID Doctoral Fellowship ANID–PFCHA/Doctorado Nacional/2017–21171061. A section of the discussed paper was written during a stay at Bielefeld, which was financed by the CRC 1283 at Bielefeld University. Additional financial support for this stay was received from University of Chile. The author thanks Michael Baake and Franz Gähler for their helpful comments and advice during this stay, and his advisor Michael Schraudner for his guidance.

References

1. Baake, M., Roberts, J.A.G., Yassawi, R.: Reversing and extended symmetries of shift spaces. Discr. Cont. Dynam. Syst. A **38**, 835–866 (2018)
2. Coven, E.M.: Endomorphisms of substitution minimal sets. Z. Wahrscheinlichkeitsth. Verw. Geb. **20**, 129–133 (1971)
3. Donoso, S., Sun, W.: Dynamical cubes and a criteria for systems having product extensions. J. Mod. Dyn. **9**, 365–405 (2015)

Algebraic invariants for group actions on the Cantor set

María Isabel Cortez

The algebraic invariants[1] associated to the group actions on the Cantor set provide an interesting connection between the fields of dynamical systems and group theory. For instance, Giordano, Putnam and Skau have shown in [29] that the dimension group (see [24] for an introduction about dimension groups) of a minimal $\mathbb{Z}$-action on the Cantor set completely determines its strong orbit equivalence class. Furthermore, the topological full group of such a system, which is known from Juschenko and Monod [38] to be amenable, determines its flip-conjugacy class (see [6] and [30] for more details). On the other hand, the amenability of the topological full groups of minimal $\mathbb{Z}$-actions together with their properties shown in [41] by Matui make them the first known examples of infinite groups which are at the same time amenable, simple and finitely generated. Recently, another algebraic invariant, the group of automorphisms of actions on the Cantor set, has caught the eye of several researchers working in the field [13, 15, 16, 17, 14, 19, 20]. In [5], Boyle, Lind and Rudolph focused their attention on the group of automorphisms of subshifts of finite type, showing that these groups are always countable and residually finite. At the same time, they gave an example of a minimal $\mathbb{Z}$-action on the Cantor set whose group of automorphisms contains $\mathbb{Q}$, which implies that the automorphism group of a minimal action may be a non-residually finite group (recall that the $\mathbb{Z}$-subshifts of finite type are not minimal). This leads to the natural question about the relation between the algebraic properties of the group of automorphisms and the dynamics of the system. Indeed, the residually finite property of the group of automorphisms of the subshifts of finite type is a consequence of the existence of periodic points.

Facultad de Matemáticas, Pontificia Universidad Católica de Chile. e-mail: maria.cortez@mat.puc.cl

[1] By an *algebraic invariant* of the dynamical system (X, T, G) we understand any algebraic structure associated to the system which determines some dynamical properties of (X, T, G) or whose properties depend on the dynamics of (X, T, G).

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_44

1 Minimal Cantor systems

By a dynamical system we mean a continuous action $T\colon G\times X\to X$ of a countable group G on a compact metric space X (phase space). We denote this as (X,T,G), and for every $g\in G$, we call $T^g\colon X\to X$ the homeomorphism on X induced by the action of g on X. The dynamical system is *free* or *aperiodic* if $T^g(x)=x$ implies $g=1_G$ (the neutral element in G) for any $x\in X$. The *orbit* of $x\in X$ is the set $O_T(x)=\{T^g(x):x\in G\}$, and we say that the system (X,T,G) is *minimal* if for every $x\in X$ its orbit is dense in X. Minimality is also equivalent to the non-existence of non-trivial sub-dynamical systems of (X,T,G), i.e, the system is minimal if and only if the unique non-empty closed T-invariant set $Y\subseteq X$ is $Y=X$. As a consequence of Zorn's lemma, we get that every dynamical system (X,T,G) has a minimal sub-dynamical system (see for example [1, 3]). It is clear that, if (X,T,G) is aperiodic, the minimal sub-dynamical systems are also aperiodic.

A particular class of dynamical systems are the *Cantor systems*, which are defined as the systems (X,T,G) where X is a Cantor set. An example of a Cantor system is the full G-shift on the finite alphabet Σ. More precisely, given Σ^G, the set of all functions $x\colon G\to\Sigma$, the shift action σ of G on Σ^G is defined as $\sigma^g x(h)=x(g^{-1}h)$, for every $g,h\in G$ and $x\in\Sigma^G$. If we endow Σ with the discrete topology and Σ^G with the product topology, the space Σ^G becomes a Cantor set and every σ^g is a homeomorphism. Thus, (Σ^G,σ,G) is a Cantor dynamical system.

The full G-shift is neither aperiodic nor minimal. However, in [37], Hjorth and Molberg show that for every countable group G there exists an aperiodic Cantor system (X,T,G). Moreover, in [4] and [26], the authors show that this aperiodic Cantor system can be chosen as an aperiodic *G-subshift*, i.e., an aperiodic sub-dynamical system of a full G-shift.

2 Algebraic properties of the topological full group of Toeplitz subshifts

The *full group* of the dynamical system (X,T,G) is the subgroup $[G]$ of the group of homeomorphisms f on X such that for every $x\in X$ there exists $g\in G$ such that $f(x)=T^g(x)$. This is the topological version of the full group introduced by Dye [23] in the context of measure-theoretic dynamical systems. It was shown by Medynets in [43] that the full group of a Cantor aperiodic system is a complete invariant for topological orbit equivalence (see [27, 28, 29, 31] for the notion and results about topological orbit equivalence).

The *topological full group* of the dynamical system (X,T,G) is the subgroup $[[G]]$ of $[G]$ of all the homeomorphisms f on X such that for every $x\in X$ there exist a neighbourhood U of x and $g\in G$ such that $f|_U=T^g$ (see [30, 33] for definitions and results). It is straightforward to check that, when X is a connected space, $[[G]]$ is isomorphic to G. Conversely, when X is a Cantor set, the topological full group

depends not only on the group G, but on the dynamics of the system. Indeed, from [43], it is possible to deduce that, for aperiodic Cantor systems, the topological full group is a complete invariant for continuous orbit equivalence (see [9] and [40] for definitions and results about continuous orbit equivalence).

From a group theoretical point of view, Jushenko and Monod have shown in [38] that the topological full group of the minimal Cantor system $(X,T,\mathbb{Z})$ is amenable (see for example [7] for definitions and results about amenability of groups and [34, 35, 41, 42] for more algebraic properties of the topological full group). On the other hand, Elek and Monod exhibited in [25] an example of an aperiodic minimal Cantor system given by a $\mathbb{Z}^2$-action whose topological full group is not amenable. Thus the algebraic properties of the topological full groups of minimal Cantor systems (X,T,G), when the group G is not $\mathbb{Z}$, still remain unclear. In joint work with Medynets and Petite, we are investigating some of these algebraic properties for the class of the Toeplitz G-subshifts.

2.1 Toeplitz G-subshifts

Let Σ be a finite alphabet. An element $x \in \Sigma^G$ is *Toeplitz* if for every $g \in G$ there exists a finite index subgroup Γ of G such that $x(g) = x(\gamma g)$, for every $\gamma \in \Gamma$. A subshift $X \subseteq \Sigma^G$ is a *Toeplitz G-subshift* if there exists a Toeplitz element $x \in X$ such that $X = \overline{O_\sigma(x)}$; see [21] for a survey on Toeplitz $\mathbb{Z}$-subshifts and [8, 11, 12, 39] for results about Toeplitz G-subshifts. It is not difficult to show that the Toeplitz G-subshifts are Cantor minimal systems and that G admits an aperiodic Toeplitz G-subshift if and only if G is residually finite [12]. The aperiodic Toeplitz G-subshifts are characterized as the minimal almost one-to-one symbolic extensions of the G-odometers [12], which correspond to the minimal aperiodic equicontinuous actions of G on the Cantor set [9]. Furthermore, the G-odometers are among the only minimal aperiodic Cantor systems with a topological full group that can be described in an explicit way (see [18] for $\mathbb{Z}$-odometers and [9] for G-odometers when G is residually finite). This description allows to deduce that the topological full group of a G-odometer is amenable if and only if G is amenable (see [9]).

The existence of an almost one-to-one factor map from a Toeplitz G-subshift to a G-odometer makes it possible to define for those systems nice nested sequences of Kakutani–Rohlin partitions (see [22, 36] for definitions and results about Kakutani–Rohlin partitions for $\mathbb{Z}$-actions and [12, 11, 32] for Toeplitz G-subshifts), which provides a useful tool to study the properties of the topological full group of these subshifts in order to find examples of Toeplitz $\mathbb{Z}^2$-subshifts whose topological full groups are not amenable.

We are still working on the following general question: Which are the properties on a Toeplitz G-subshift that ensure that its topological full group is amenable?

3 Algebraic properties of the group of automorphisms of a group action on the Cantor set

Let (X,T,G) be a minimal aperiodic Cantor system. The *normalizer* group of (X,T,G), denoted $\mathrm{Norm}(X,T,G)$, is defined as the subgroup of all the homeomorphisms $h\colon X \to X$ such that there exists an isomorphism $\alpha_h\colon G \to G$ such that $h \circ T^g = T^{\alpha_h(g)} \circ h$, for every $g \in G$. The aperiodicity of the action implies the uniqueness of α_h for any element $h \in \mathrm{Norm}(X,T,G)$. Thus we can define the *automorphism group* of (X,T,G) as

$$\mathrm{Aut}(X,T,G) = \{h \in \mathrm{Norm}(X,T,G) : \alpha_h = id\}.$$

It is immediate that $\mathrm{Aut}(X,T,G)$ is a normal subgroup of $\mathrm{Norm}(X,T,G)$. For the case $G = \mathbb{Z}$, the quotient of the normalizer group by the group of automorphisms is either trivial or isomorphic to $\mathbb{Z}/2\mathbb{Z}$. Important progress has been made in the study of the group of automorphisms of minimal $\mathbb{Z}$-subshifts, establishing a connection between the complexity of the subshifts and the algebraic properties of the group of automorphisms; see [2, 15, 16, 17, 19].

In [10], we obtained results concerning the realization of groups as subgroups of the normalizer and the automorphism group of minimal aperiodic actions on the Cantor set as follows.

- Every countable group is the subgroup of the normalizer of some minimal aperiodic action of a countable Abelian free group on the Cantor set.
- Every residually finite group Γ can be realized as the subgroup of the automorphism group of a minimal $\mathbb{Z}$-action on the Cantor set [10, Prop. 7]. A key tool for the proof of this result is the characterization of residually finite groups as those groups G for which every full G-shift has a dense subset of points with finite orbit [7, Thm. 2.7.1].
- For any countable group G, the group of automorphisms of a minimal aperiodic G-action on the Cantor set is a subgroup of the group of automorphisms of a minimal $\mathbb{Z}$-action on the Cantor set.

References

1. Auslander, J.: Minimal Flows and their Extensions. North-Holland, Amsterdam (1988)
2. Baake, M., Roberts, J.A.G., Yassawi, R.: Reversing and extended symmetries of shift spaces. Discr. Cont. Dynam. Syst. **38**, 835–866 (2018)
3. Brin, M., Stuck, G.: Introduction to Dynamical Systems. Cambridge University Press, Cambridge (2002)
4. Aubrun, N., Barbieri, S., Thomassé, S.: Realization of aperiodic subshifts and densities. Groups Geom. Dyn., to appear
5. Boyle, M., Lind, D., Rudolph, D.: The automorphism group of a shift of finite type. Trans. Amer. Math. Soc. **306**, 71–114 (1988)

6. Boyle, M.: Topological Orbit Equivalence and Factor Maps in Symbolic Dynamics. PhD thesis, University of Washington, Seattle (1983)
7. Ceccherini-Silberstein, T., Coornaert, M.: Cellular Automata and Groups. Springer, Berlin (2010)
8. Cortez, M.I.: $\mathbb{Z}^d$ Toeplitz arrays. Discr. Cont. Dynam. Syst. **15**, 859–881 (2006)
9. Cortez, M.I., Medynets, C.: Orbit equivalence rigidity of equicontinuous systems. J. London Math. Soc. **94**, 545–556 (2016)
10. Cortez, M.I., Petite, S.: On the centralizers of aperiodic actions on the Cantor set. Preprint arXiv:1807.04654
11. Cortez, M.I., Petite, S.: Invariant measures and orbit equivalence for generalized Toeplitz subshifts. Groups Geom. Dynam. **8**, 1007–1045 (2014)
12. Cortez, M.I., Petite, S.: *G*-odometers and their almost one-to-one extensions. J. London Math. Soc. **78**, 1–20 (2008)
13. Coven, E., Quas, A., Yassawi, R.: Computing automorphism groups of shifts using atypical equivalence classes. Discr. Anal. **2016**, 611:1–28 (2016)
14. Cyr, V., Franks, J., Kra, B., Petite, S.: Distortion and the automorphism group of a shift. J. Mod. Dyn. **13**, 147–161 (2018)
15. Cyr, V., Kra, B.: The automorphism group of a shift of linear growth: beyond transitivity. Forum Math. Sigma **3**, e5:1–27 (2015)
16. Cyr, V., Kra, B.: The automorphism group of a shift of subquadratic growth. Proc. Amer. Math. Soc. **144**, 613–621 (2016)
17. Cyr, V., Kra, B.: The automorphism group of a minimal shift of stretched exponential growth. J. Mod. Dyn. **10**, 483–495 (2016)
18. de Cornulier, Y.: Groupes pleins-topologiques [d'après Matui, Juschenko, Monod,...]. Sém. Bourbaki, 65ème anné, no 1064 (2012/13)
19. Donoso, S., Durand, F., Maass, A., Petite, S.: On automorphism groups of low complexity subshifts. Ergodic Th. & Dynam. Syst. **36**, 64–95 (2016)
20. Donoso, S., Durand, F., Maass, A., Petite, S.: On automorphism groups of Toeplitz subshifts. Discr. Anal. **2017**, 11:1–19 (2017)
21. Downarowicz, T.: Survey of odometers and Toeplitz flows. In Kolyada, S., Manin, Y., Ward, T. (eds.), Algebraic and Topological Dynamics, pp. 7–37. AMS, Providence, RI (2005)
22. Durand, F., Host, B., Skau, C.: Substitutional dynamical systems, Bratteli diagrams and dimension groups. Ergodic Th. & Dynam. Syst. **19**, 953–993 (1999)
23. Dye, H.A.: On groups of measure preserving transformations. I. Amer. J. Math. **81**, 119–159 (1959)
24. Effros, E.G.: Dimensions and C*-Algebras. Conference Board of the Mathematical Sciences, Washington, D.C. (1981)
25. Elek, G., Monod, N.: On the topological full group of a minimal Cantor $\mathbb{Z}^2$-system. Proc. Amer. Math. Soc. **141**, 3549–3552 (2013)
26. Gao, S., Jackson, S., Seward, B.: A coloring property for countable groups. Math. Proc. Cambridge Philos. Soc. **147**, 579–592 (2009)
27. Giordano, T., Matui, H., Putnam, I.F., Skau, C.F.: The absorption theorem for affable equivalence relations. Ergodic Th. & Dynam. Syst. **28**, 1509–1531 (2008)
28. Giordano, T., Matui, H., Putnam, I.F., Skau, C.F.: Orbit equivalence for Cantor minimal $\mathbb{Z}^d$-systems. Invent. Math. **179**, 119–158 (2010)
29. Giordano, T., Putnam, I.F., Skau, C.F.: Topological orbit equivalence and C^*-crossed products. J. Reine Angew. Math. (Crelle) **469**, 51–111 (1995)
30. Giordano, T., Putnam, I.F., Skau, C.F.: Full groups of Cantor minimal systems. Israel J. Math. **111**, 285–320 (1999)
31. Giordano, T., Putnam, I.F., Skau, C.F.: Affable equivalence relations and orbit structure of Cantor dynamical systems. Ergodic Th. & Dynam. Syst. **24**, 441–475 (2004)
32. Gjerde, R., Johansen, O.: Bratteli–Vershik models for Cantor minimal systems: applications to Toeplitz flows. Ergodic Th. & Dynam. Syst. **20**, 1687–1710 (2000)
33. Glasner, E., Weiss, B.: Weak orbit equivalence of Cantor minimal systems. Intern. J. Math. **6**, 559–579 (1995)

34. Grigorchuk, R.I., Medynets, K.S.: On the algebraic properties of topological full groups. Sb. Math. **205**, 843–861 (2014)
35. Grigorchuk, R., Medynets, K.: Presentations of topological full groups by generators and relations. J. Algebra **500**, 46–68 (2018)
36. Herman, R., Putnam, I.F., Skau, C.F.: Ordered Bratteli diagrams, dimension groups and topological dynamics. Intern. J. Math. **3**, 827–864 (1992)
37. Hjorth, G., Molberg, M.: Free continuous actions on zero-dimensional spaces. Topology Appl. **153**, 1116–1131 (2006)
38. Juschenko, K., Monod, N.: Cantor systems, piecewise translations and simple amenable groups. Ann. Math. **178**, 775–787 (2013)
39. Krieger, F.: Sous-décalages de Toeplitz sur les groupes moyennables résiduellement finis. J. London Math. Soc. **75**, 447–462 (2007)
40. Li, X.: Continuous orbit equivalence rigidity. Ergodic Th. & Dynam. Syst. **38**, 1543–1563 (2018)
41. Matui, H.: Some remarks on topological full groups of Cantor minimal systems. Intern. J. Math. **17**, 231–251 (2006)
42. Matui, H.: Some remarks on topological full groups of Cantor minimal systems II. Ergodic Th. & Dynam. Syst., in press
43. Medynets, K.: Reconstruction of orbits of Cantor systems from full groups. Bull. London Math. Soc. **43**, 1104–1110 (2011)

Lyapunov exponents: recent applications of Fürstenberg's theorem in spectral theory

David Damanik

Abstract We discuss the phenomenon of Anderson localization and a new proof of it in one space dimension. This proof is due to V. Bucaj, D. Damanik, J. Fillman, V. Gerbuz, T. VandenBoom, F. Wang, Z. Zhang, and it is centered around the positivity of and large deviation estimates for the Lyapunov exponent — a strategy originally developed in non-random settings by J. Bourgain, M. Goldstein, W. Schlag.

1 The Anderson model

The Anderson model was proposed in 1958 by P. W. Anderson. Its main feature is that randomness can trap quantum states, a phenomenon called Anderson localization. Anderson received the Physics Nobel Prize in 1977 for this work. The model is given by a discrete Schrödinger operator on the d-dimensional standard lattice with potential values given by independent identically distributed random variables.

Concretely, given a probability measure ν on $\mathbb{R}$ whose topological support is compact and contains at least two points, we consider $\Omega = (\operatorname{supp}\nu)^{\mathbb{Z}^d}$ and $\mu = \nu^{\mathbb{Z}^d}$. For every $\omega \in \Omega$ and $n \in \mathbb{Z}^d$, we set $V_\omega(n) = \omega_n$. This defines, for $\omega \in \Omega$, a potential $V_\omega : \mathbb{Z}^d \to \mathbb{R}$, and a Schrödinger operator in $\ell^2(\mathbb{Z}^d)$:

$$[H_\omega \psi](n) = \sum_{|m-n|_1=1} \psi(m) + V_\omega(n)\psi(n).$$

The spectrum and the spectral type of H_ω are μ-almost surely independent of ω. The almost sure spectrum Σ is explicitly given by $\Sigma = [-2d, 2d] + \operatorname{supp}(\nu)$.

Anderson localization comes in two standard flavors: spectral localization and dynamical localization. Here, spectral localization refers to the statement that, μ-almost surely, H_ω has pure point spectrum in a suitable energy region $\Sigma_\ell \subseteq \Sigma$ with

David Damanik
Rice University, Houston, USA. e-mail: damanik@rice.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_45

exponentially decaying eigenvectors. Dynamical localization refers to the statement that quantum states remain trapped. Concretely, this means that $e^{-itH_\omega}\chi_{\Sigma_\ell}(H_\omega)\delta_0$ remains mostly in some fixed finite region for all times.

The size and shape of the localized energy region Σ_ℓ depends on the dimension d and the single-site measure ν:

- $d=1$: It is *known* that $\Sigma_\ell=\Sigma$.
- $d=2$: It is *conjectured* that $\Sigma_\ell=\Sigma$. Currently, the known result is the same as in the case $d\geq 3$.
- $d\geq 3$: It is *known* that Σ_ℓ contains nontrivial neighborhoods of $\partial\Sigma$. It is *conjectured* that $\Sigma_\ell\neq\Sigma$ if the diameter of $\operatorname{supp}\nu$ is not too large, and that there is a sharp transition from localization to diffusive transport.

We will focus on the case $d=1$. The first localization proof was given by Carmona, Klein and Martinelli in [4]. A different proof by Shubin, Vakilian and Wolff appears in [6]. Both of these proofs rely on multiscale analysis. A simpler and more direct proof was recently given by Bucaj et al. in [3]. The latter proof will be discussed in what follows.

This proof is centered around the positivity of and large deviation estimates for the Lyapunov exponent. Approaching localization proofs in this way is a strategy due to Bourgain, Goldstein in the case of quasi-periodic potentials [1] and to Bourgain, Schlag in the case of potentials generated by the doubling map [2].

This new proof also suggests how new results can be obtained. For example, Damanik, Fillman and Sukhtaiev implemented this approach in the setting of Anderson models on metric and discrete tree graphs and proved spectral and dynamical localization for these operators [5].

2 Main results

Here is the pair of theorems for the family $\{H_\omega\}_{\omega\in\Omega}$ of random Schrödinger operators, acting in $\ell^2(\mathbb{Z})$ via

$$[H_\omega\psi](n)=\psi(n+1)+\psi(n-1)+V_\omega(n)\psi(n),$$

where the potential V_ω is given by independent identically distributed random variables with a common distribution that has a compact support that contains at least two elements.

Theorem 1 (Spectral localization for the 1D Anderson model). *Almost surely, H_ω is spectrally localized, that is, it has pure point spectrum with exponentially decaying eigenfunctions.*

Theorem 2 (Exponential dynamical localization for the 1D Anderson model). *There is a constant $\gamma>0$ so that for almost every ω and every $\varepsilon>0$, there is a constant $C=C_{\omega,\gamma,\varepsilon}>0$ such that, for all $m,n\in\mathbb{Z}$,*

$$\sup_{t\in\mathbb{R}} |\langle \delta_n, e^{-itH_\omega}\delta_m\rangle| \le Ce^{\varepsilon|m|}e^{-\gamma|n-m|}.$$

3 Lyapunov exponents: positivity and large deviation estimates

The difference equation (or generalized eigenvalue equation) for the operator H_ω:

$$u(n+1)+u(n-1)+V_\omega(n)u(n) = Eu(n)$$

admits a two-dimensional solution space, as any two consecutive values of u determine all other values. Fixing $(u(0),u(-1))^\top$ as the point of reference, the linear map taking this vector to $(u(n),u(n-1))^\top$ is given by the so-called transfer matrix $M_n^E(\omega)$. Ergodicity of the standard shift transformation $T:\Omega\to\Omega$, $[T\omega]_n = \omega_{n+1}$ implies that for each E, there are $L(E)\ge 0$ and $\Omega_-^E, \Omega_+^E \subseteq \Omega$ with $\mu(\Omega_-^E) = \mu(\Omega_+^E) = 1$ such that

$$L(E) = \begin{cases} \lim_{n\to\infty} \frac{1}{n}\log\|M_n^E(\omega)\| & \text{for } \omega\in\Omega_+^E, \\ \lim_{n\to-\infty} \frac{1}{|n|}\log\|M_n^E(\omega)\| & \text{for } \omega\in\Omega_-^E. \end{cases}$$

The number $L(E)$ is called the *Lyapunov exponent*.

For E, let ν_E be the push-forward of ν under the map

$$x \mapsto \begin{pmatrix} E-x & -1 \\ 1 & 0 \end{pmatrix}$$

and let G_E be the smallest closed subgroup of $\mathrm{SL}(2,\mathbb{R})$ that contains $\operatorname{supp}\nu_E$.

By a result of Fürstenberg, a sufficient condition for $L(E)>0$ is

1. G_E is not compact,
2. G_E is strongly irreducible (which means that there is no finite non-empty invariant set of directions).

A modification of a result of Ishii shows that the condition

3. $\exists A,B\in G_E$ with $\mathrm{tr}(A)\neq 0$, $\mathrm{tr}(B)\neq 0$, $\det(AB-BA)\neq 0$,

implies 1 and 2. This condition is often easy to check.

For the Anderson model on $\mathbb{Z}$, let us verify 1. and 2. for arbitrary $E\in\mathbb{R}$. Since the support of the single-site distribution has cardinality at least two, it follows that ν_E also has at least two points in its support. Thus, G_E contains at least two distinct elements of the form

$$M_x = \begin{bmatrix} x & -1 \\ 1 & 0 \end{bmatrix},$$

say, M_a and M_b with $a\neq b$. Note that

$$A = M_a M_b^{-1} = \begin{bmatrix} 1 & a-b \\ 0 & 1 \end{bmatrix} \in G_E.$$

Taking powers of the matrix A, we see that G_{ν_E} is not compact, verifying 1.

Now, consider $V_1 := \operatorname{span}(\mathbf{e}_1)$, the projection of $\mathbf{e}_1 := (1,0)^\top$ to $\mathbb{RP}^1$. Then, one has $AV_1 = V_1$ and, for every $V \in \mathbb{RP}^1$, $A^n V$ converges to V_1. Thus, if there is a nonempty finite invariant set of directions $\mathscr{F} \subseteq \mathbb{RP}^1$, one must have $\mathscr{F} = \{V_1\}$. However, we also have

$$A' = M_a^{-1} M_b = \begin{bmatrix} 1 & 0 \\ a-b & 1 \end{bmatrix} \in G_E$$

and $A'V_1 \neq V_1$. This establishes 2.

So far, we have seen that the transfer matrices are almost surely asymptotically exponentially large. That is,

$$\lim_{|n|\to\infty} \frac{1}{|n|} \log \|M_n^E(\omega)\| = L(E) > 0$$

for every $E \in \mathbb{R}$ and μ-almost every ω.

It is natural to ask what can be said about the size of $\log \|M_n^E(\omega)\|$ before n is taken to infinity. Large deviation estimates address this issue.

Theorem 3 (Uniform LDT for the Lyapunov exponent). *For any $\varepsilon > 0$, there exist $C = C(\varepsilon) > 0$, $\eta = \eta(\varepsilon) > 0$ such that, for all $n \in \mathbb{Z}_+$ and all $E \in \Sigma$,*

$$\mu\left\{\omega \in \Omega : \left|\frac{1}{n} \log \left\|M_n^E(\omega)\right\| - L(E)\right| \geq \varepsilon\right\} \leq Ce^{-\eta n}.$$

4 Transfer matrices, Dirichlet determinants, and Green's functions

There are well known connections between the transfer matrices discussed above, the determinant of the restriction $H_{\omega,N}$ of H_ω to the finite interval $[0, N-1] \cap \mathbb{Z}$ with Dirichlet boundary conditions, and the Green function of $H_{\omega,N}$, defined by

$$G_{\omega,N}^E(m,n) = \langle \delta_m, (H_{\omega,N} - E)^{-1} \delta_n \rangle$$

for $0 \leq m, n \leq N-1$.

The following formula connects the transfer matrix and the determinants:

$$M_N^E(\omega) = \begin{pmatrix} \det(E - H_{\omega,N}) & -\det(E - H_{T\omega,N-1}) \\ \det(E - H_{\omega,N-1}) & -\det(E - H_{T\omega,N-2}) \end{pmatrix}, \quad N \geq 2.$$

This formula shows that if the transfer matrix has exponentially large norm, then at least one of the determinants must be exponentially large.

The following formula connects the determinants and the Green function:

$$G^E_{\omega,N}(m,n) = \frac{\det[H_{\omega,m} - E]\det[H_{T^{n+1}\omega,N-n-1} - E]}{\det[H_{\omega,N} - E]}.$$

If one has exponentially large determinants, this formula shows that the Green function must have exponential off-diagonal decay.

Finally, the following lemma connects the Green function and the solutions.

Lemma 1. *If u is a solution of the difference equation $H_\omega u = Eu$ and $E \notin \sigma(H_{\omega,N})$, then*

$$u(n) = -G^E_{\omega,N}(n,0)u(-1) - G^E_{\omega,N}(n,N-1)u(N)$$

for $0 \le n \le N-1$.

With this lemma and the results discussed earlier one can readily deduce for almost all ω's that a polynomially bounded solution must in fact decay exponentially, thus establishing spectral localization. A second look at the semi-uniform localization properties of the eigenvectors then allows one to establish dynamical localization as well.

References

1. Bourgain, J., Goldstein, M.: On nonperturbative localization with quasi-periodic potential. Ann. Math. **152**, 835–879 (2000)
2. Bourgain, J., Schlag, W.: Anderson localization for Schrödinger operators on $\mathbb{Z}$ with strongly mixing potentials. Commun. Math. Phys. **215**, 143–175 (2000)
3. Bucaj, V., Damanik, D., Fillman, J., Gerbuz, V., VandenBoom, T., Wang, F., Zhang, Z.: Localization for the one-dimensional Anderson model via positivity and large deviations for the Lyapunov exponent. Trans. Amer. Math. Soc., in press. arXiv:1706.06135
4. Carmona, R., Klein, A., Martinelli, F.: Anderson localization for Bernoulli and other singular potentials. Commun. Math. Phys. **108**, 41–66 (1987)
5. Damanik, D., Fillman, J., Sukhtaiev, S.: Localization for Anderson models on metric and discrete tree graphs. preprint arXiv:1902.07290
6. Shubin, C., Vakilian, R., Wolff, T.: Some harmonic analysis questions suggested by Anderson–Bernoulli models. Geom. Funct. Anal. **8**, 932–964 (1998)

Extended symmetries of Markov subgroups

Robbert Fokkink

A symmetry of a tesselation is an isometry of the plane, or space, preserving the tesselation. What symmetry groups can one get? This is a classical problem in geometry, leading to the wallpaper groups of the plane or crystallographic groups in higher dimensions. For dynamical systems with a $\mathbb{Z}^d$-action on X, the symmetries are the homeomorphisms that commute with the action. This is the centralizer of $\mathbb{Z}^d$ in $Homeo(X)$. The extended symmetries are given by the normalizer. Baake, Roberts and Yassawi [1] showed that the centralizer can be non-trivial for well-known systems such as the Thue–Morse shift or the Ledrappier shift. The latter is a standard example of a Markov subgroup and the topic of this talk is the extended symmetry group of arbitrary Markov subgroups in $\mathbb{Z}^2$ shifts.

A 2D Markov subgroup is described by a polynomial with two indeterminates $p(X,Y)$ with coefficients in $\mathbb{F}_2$. A fundamental result by Quas and Trow [3] gives precise conditions such that the symmetry group is 'algebraic': this occurs if $p(X,Y)$ has no collinear factors. Using results from algebraic geometry, one can deduce from this that the extended symmetry group is a finitely generated Abelian group if such a $p(X,Y)$ is squarefree. It is infinitely generated if it is not squarefree. This leaves the case of a polynomial with collinear factors. A prime example of this is $p(X,Y) = (1+X)(1+Y)$. It turns out that the elements of its extended symmetry group correspond to automorphisms of the Bernoulli shift $\{0,1\}^{\mathbb{Z}}$ that commute with the flip (the involution that flips 0 and 1). Which automorphisms have this property? This is not an easy question, and I left this as a homework exercise during the talk.

It seems likely that the extended symmetry group is non-amenable if $p(X,Y)$ has collinear factors. For some cases, other than $(1+X)(1+Y)$, this is not so dffi-cult to prove. The full result is topic of ongoing research with Dan Rust and Reem Yassawi [2]. I would like to thank the participants of the workshop for stimulating discussions and am looking forward to their solutions of the homework exercise.

Robbert Fokkink
TU Delft, The Netherlands. e-mail: R.J.Fokkink@tudelft.nl

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_46

References

1. Baake, M., Roberts, J.A.G., Yassawi, R.: Reversing and extended symmetries of shift spaces. Discr. Cont. Dynam. Syst. A **38**, 835–866 (2018)
2. Fokkink, R., Yassawi, R.: Topological rigidity of linear cellular automaton shifts Indag. Math. **27**, 1105–1113 (2018)
3. Quas, A., Trow, P.: Mappings of group shifts. Israel J. Math. **124**, 333–365 (2001)

Renormalisation for inflation tilings I: General theory

Franz Gähler

Inflation tilings are generated by iterating an inflation procedure ρ, which first expands a (partial) tiling linearly by a factor λ, and then divides each expanded tile (called supertile) according to a fixed rule into a set of original tiles. The relative positions of the tiles of type i within a supertile of type j are encoded in a set T_{ij} (we assume here finitely many tile types, up to translations). The associated inflation matrix M_ρ, which we assume to be primitive, has entries $\mathrm{card}(T_{ij})$ and a leading eigenvalue λ^d. The information contained in T_{ij} is also encoded in the matrix-valued function

$$B_{ij}(k) = \sum_{t \in T_{ij}} \mathrm{e}^{2\pi \mathrm{i} t \cdot k},$$

which is known as the Fourier matrix of the inflation. Note that the n^{th} power of ρ has the Fourier matrix $B^{(n)}(k) = B(k)B(\lambda k)\cdots B(\lambda^{n-1}k)$.

Suppose now an inflation tiling is decorated with point measures on the control points of its tiles, with weights which may depend on the tile type. The diffraction spectrum of the resulting measure is then given by the Fourier transform of its pair correlation measure. This is again a measure, which can be decomposed into pure-point (pp), absolutely continuous (ac), and a singular continuous (sc) parts. Here, we are mainly interested in the presence or absence of an ac part.

As first observed in [2], and further elaborated in [3], the self-similarity of inflation tilings results in exact renormalisation equations, which the pair correlation measure of the tiling must satisfy. These in turn lead to exact scaling relations for the diffraction measure, which must hold for each spectral component separately. For instance, the Radon–Nikodym density $v(k)$ of the ac part of the Fourier amplitude (a vector with one component per tile type) must satisfy the relation

$$v(\lambda k) = \lambda^{d/2} B^{-1}(k)\, v(k),$$

Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany. e-mail: gaehler@math.uni-bielefeld.de

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_47

provided $B(k)$ is invertible for almost all k. As $v(k)$ must be translation bounded, ac spectrum can exist only if the minimal Lyapunov exponent governing the asymptotic growth of $v(k)$,

$$\chi_{\min}(k) = \log \lambda^{d/2} + \liminf_{n\to\infty} \frac{1}{n} \log \left\| B(k)B(\lambda k)\cdots B(\lambda^{n-1}k) \right\|_{\mathrm{F}}^{-1},$$

vanishes for almost all k. Setting $\chi_{\min}(k) = \log \lambda^{d/2} - \chi^B(k)$, we need to investigate the behaviour of $\chi^B(k)$. Taking into account that the Frobenius norm is submultiplicative, we get the estimate

$$\chi^B(k) = \limsup_{n\to\infty} \frac{1}{n} \log \|B^{(n)}(k)\|_{\mathrm{F}} \leqslant \frac{1}{N} \mathbb{M}\big(\log \|B^{(N)}(k)\|_{\mathrm{F}}\big)$$

for any fixed N, where $\mathbb{M}(f)$ is the mean of the quasiperiodic function f. Moreover, to compute the mean of the quasiperiodic function $\log \|B^{(n)}(k)\|_{\mathrm{F}}$, or rather $\log \|B^{(n)}(k)\|_{\mathrm{F}}^2$, we can lift it to a section through a periodic function, and compute the mean as an integral over the unit cell,

$$\frac{1}{N} \mathbb{M}\big(\log \|B^{(N)}(.)\|_{\mathrm{F}}^2\big) = \frac{1}{N} \int_{\mathbb{T}^D} \log\Big(\sum_{i,j} \big| P_{ij}^{(N)}\big(\tilde{k}\big) \big|^2 \Big)\, \mathrm{d}\tilde{k},$$

where the $P_{ij}^{(N)}$ are trigonometric polynomials. In this way, for each N, an upper bound for $\chi^B(k)$ is obtained, which is readily computable for many examples. If that upper bound implies that $\chi^B(k) < c \cdot \log \lambda^{d/2}$ for some $c < 1$, the presence of ac spectrum can be ruled out.

This criterion has successfully been applied to many examples, among them several with non-Pisot inflation factors. These are known to have no non-trivial pp part in the spectrum, but the nature of the continous part has long remained unclear. Using our approach, it could be shown that the binary non-Pisot tiling [1], the (non-FLC) Frank–Robinson tiling [4], and the well-known Godrèche–Lançon–Billard tiling [5], all have singular diffraction spectrum. The same conclusion is obtained for several other examples with mixed pp and continuous spectrum. In fact, except for the few examples known to have an ac part in the spectrum, such as the Rudin–Shapiro tiling, in all examples studied the upper bound on $\chi^B(k)$ quickly drops below the threshold $\log \lambda^{d/2}$, showing that the diffraction spectrum is singular.

References

1. Baake, M., Frank, N., Grimm, U., Robinson, E.A.: Geometric properties of a binary non-Pisot inflation and absence of absolutely continuous diffraction. Studia Math. **247**, 4–27 (2019)
2. Baake, M., Gähler, F.: Pair correlations of aperiodic inflation rules via renormalisation: Some interesting examples. Topol. & Appl. **205**, 4–27 (2016)
3. Baake, M., Gähler, F., Mañibo, N.: Renormalisation of pair correlation measures for primitive inflation rules and absence of absolutely continuous diffraction, Commun. Math. Phys. 370(2) (2019) 591–635;

4. Baake, M., Grimm, U.: Renormalisation of pair correlations and their Fourier transforms for primitive block substitutions. In: Akiyama, S., Arnoux, P. (eds.), Substitution and Tiling Dynamics: Introduction to Self-inducing Structures, Lecture Notes in Mathematics 2273, Springer, Cham 2020 (in press); arXiv: 1906.10484
5. Godrèche, C., Lançon, F.: A simple example of a non-Pisot tiling with five-fold symmetry. J. Phys. I (France) **2**, 207–220 (1992)

Problems in number theory related to aperiodic order

Jeffrey C. Lagarias

This talk concerns properties of dilated floor functions $f_\alpha(x) = [\alpha x]$, where α takes a fixed real value. Such functions perform quantization of the linear function αx at length scale $\frac{1}{\alpha}$. For $\alpha > 1$, the set of values $[\alpha n]$ for positive integer n is called the Beatty sequence associated to α. It is an aperiodic sequence if α is irrational, and when extended to all integers it is a one-dimensional cut and project set; compare [1, Ex. 9.8]. This talk studies the composed functions $f_\alpha \circ f_\beta(x) = [\alpha[\beta x]]$. The set of parameter values (α, β) where the two floor functions commute were characterized in joint work [2] with Takumi Murayama and D. Harry Richman (2016). The solution set consists of three straight lines through the origin $(0,0)$ plus a countable set of 'exceptional' rational solutions $(\frac{1}{m}, \frac{1}{n})$ for positive integer m, n.

Ongoing joint work [3], [4] with D. Harry Richman (2019) determines the set S of all values (α, β) that satisfy $[\alpha[\beta x]] \geq [\beta[\alpha x]]$ for all real x. When α, β have opposite signs then $(\alpha, \beta) \in S$ if and only if $\alpha < 0$ and $\beta > 0$. For positive α, β, the solution set is a countable collection of half-lines and rectangular hyperbolas, passing through $(0,0)$, plus the vertical lines $\alpha = \frac{1}{m}, \beta > 0$ for positive integer m. The hyperbola solutions are associated to disjoint Beatty sequences when both α, β are irrational, but also include extra solutions with rational values. For negative α, β, the solution set consists of countably infinite families of lines and rectangular hyperbolas passing through $(0,0)$, plus vertical finite line segments at every rational $\alpha = -\frac{m}{n}$ (with $-\frac{1}{m} \leq \beta < 0$), plus a countable set of 'sporadic rational solutions'.

The existence of the sporadic rational solutions relates to the Diophantine Frobenius problem in two variables. The classification establishes that the set S is closed. It establishes various internal symmetries of the set S given by linear and bilinear changes of variables, for positive α, β (resp., negative α, β). The classification implies a pre-partial ordering on nonzero α where one says $\alpha \prec \beta$ if $(\alpha, \beta) \in S$. Namely, if $(\alpha, \beta) \in S$ and $(\beta, \gamma) \in S$ with $\alpha\beta\gamma \neq 0$, then $(\alpha, \gamma) \in S$.

Jeffrey C. Lagarias
University of Michigan, Ann Arbor, Michigan, USA. e-mail: lagarias@umich.edu

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_48

References

1. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
2. Lagarias, J.C., Murayama, T., Richman, D.H.: Dilated floor functions that commute. Amer. Math. Monthly **123**, 1033–1038 (2016)
3. Lagarias, J.C., Richman, D.H.: Dilated floor functions having nonnegative commutator. I. Positive and mixed sign dilations. Acta Arith. **187**, 271–299 (2019)
4. Lagarias, J.C., Richman, D.H.: Dilated floor functions having nonnegative commutator. II. Negative dilations. Acta Arithmetica **196**, 163–186 (2020).

Pure point spectrum and regular model sets in substitution tilings on $\mathbb{R}^d$

Jeong-Yup Lee

It has long been known that every regular model set has pure point spectrum, but the converse is not true in general. The relation between regular model sets and pure point spectrum is well studied in [3, 2, 13] in quite a general setting. When we restrict to substitution tilings, it has been shown in [8] that pure point spectrum and inter model set are equivalent. However the inter model set is a projected point set in a cut-and-project scheme (CPS) with an internal space which is constructed with an autocorrelation topology coming from pure point spectrum. It was not easy to extract information from the internal space.

In this joint work with Shigeki Akiyama, we show that the internal space can be a Euclidean space under some additional assumption. This result generalizes the remark [4], which shows the equivalence between regular model set and pure point spectrum in the case of one-dimensional substitution tilings, into d dimensions. From this result, we can think of 'Pisot conjecture' in more general setting of d-dimensional substitution tilings [1, 12]. Rigidity was introduced in [11] and primitive substitution tilings with finite local complexity (FLC) always show this type of rigidity. But the converse is not true as we can observe in an example in [5]. Under the rigidity assumption, pure point spectrum always gives FLC. So we do not asume FLC. Instead we assume rigidity.

We first show how to construct a CPS with Euclidean internal space. This construction was already introduced in [7] for other purposes. We make use of it in a more general setting. Under the assumption of pure point spectrum, we provide conditions under which the representive point set of a primitive repetitive substitution tiling is a model set. Using Keesling's argument [6, 9], the model set is in fact regular. In the process of the proof, we use the equivalence between pure point spectrum and algebraic coincidence which was introduced in [8].

Jeong-Yup Lee
Department of Mathematics Education, Catholic Kwandong University, Gangneung, Gangwon 210-701, Korea. e-mail: jylee@cku.ac.kr

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_49

References

1. Akiyama, S., Barge, M., Berthé, V., Lee, J.-Y., Siegel, A.: On the Pisot substitution conjecture. In: Kellendonk, J., Lenz, D., Savinien, J. (eds.) Mathematics of Aperiodic Order, pp. 33–72. Birkhäuser, Basel (2015)
2. Baake, M., Lenz, D., Moody, R.V.: Characterization of model sets by dynamical systems. Ergod. Th. & Dynam. Syst. **27** 341–382 (2007)
3. Baake, M., Moody, R.V.: Weighted Dirac combs with pure point diffraction. J. Reine Angew. Math. (Crelle) **573**, 61–94 (2004)
4. Barge, M., Kwapisz, J.: Geometric theory of unimodular Pisot substitutions. Amer. J. Math. **128**, 1219–1282 (2006)
5. Frank, N.P., Robinson, E.A.: Generalized β-expansions, substitution tilings, and local finiteness. Trans. Amer. Math. Soc. **360**, 1163–1177 (2008)
6. Keesling, J.: The boundaries of self-similar tiles in $\mathbb{R}^n$. Topology Appl. **94**, 195–205 (1999)
7. Lee, J.-Y., Akiyama, S., Nagai, Y.: Cut-and-project schemes for Pisot family substitution tilings. Symmetry **10**, 511:1–9 (2018)
8. Lee, J.-Y.: Substitution Delone sets with pure point spectrum are inter-model sets. J. Geom. Phys. **57**, 2263–2285 (2007)
9. Lee, J.-Y., Moody, R.V.: Lattice substitution systems and model sets. Discr. Comput. Geom. **25**, 173–201 (2001)
10. Lee, J.-Y., Moody, R.V.: Characterization of model multi-colour sets. Ann. H. Poincare **7**, 125–143 (2006)
11. Lee, J.-Y., Solomyak, B.: On substitution tilings and Delone sets without finite local complexity. Discr. Cont. Dynam. Syst. A **39**, 3149–3177 (2019)
12. Sing, B.: Pisot Substitutions and Beyond. PhD thesis, Bielefeld University (2006) https://pub.uni-bielefeld.de/download/2302336/2302339/diss.pdf
13. Strungaru, N.: Almost periodic pure point measures. In: Baake, M., Grimm, U. (eds.) Aperiodic order. Vol. 2: Crystallography and Almost Periodicity, pp. 271–342. Cambridge University Press, Cambridge (2017)

Automatic sequences are orthogonal to aperiodic multiplicative functions

Mariusz Lemańczyk

In 2010, P. Sarnak [7] formulated the following conjecture: For each zero entropy topological dynamical system (X,T), we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n\le N} f(T^n x)\mu(n)=0 \tag{1}$$

for all $f\in C(X)$ and $x\in X$. Sarnak's conjecture has been proved in many classes of zero entropy systems [1], including so-called automatic sequences (C. Müllner, [6]), that is when $X=X_\theta\subset A^{\mathbb{Z}}$ and $T=S$ (shift) is determined by a primitive substitution $\theta: A\to A^\lambda$ of constant length λ. One can ask, however, whether Eq. (1) holds when we replace μ by other arithmetic functions. Especially, we are interested in the class of multiplicative functions. My talk is to present the main strategies to prove the following result:

Theorem 1 (M. Lemańczyk, C. Müllner, 2018). *Each automatic sequence is orthogonal to an arbitrary bounded, aperiodic and multiplicative function, i.e. for each primitive substitution $\theta: A\to A^\lambda$, all $f\in C(X_\theta)$ and $x\in X$, we have*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n\le N} f(S^n x)u(n)=0$$

for each $u:\mathbb{N}\to\mathbb{C}$ as above.

Our main tool is the so-called DKBSZ criterion [3] which says that every bounded sequence (a_n) of complex numbers is orthogonal to *all* bounded multiplicative functions if

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n\le N} a_{pn}\overline{a}_{qn}=0 \tag{2}$$

Mariusz Lemańczyk
Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Ul. Chopina 12/18, 87-100 Toruń, Poland. e-mail: mlem@mat.umk.pl

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_50

for each pair of sufficiently large different primes p,q. Then, in the dynamical context, we consider $a_n = f(S^n x)$, and Eq. (2) leads us to study

$$\frac{1}{N}\sum_{n\le N} f(S^{pn}x)\overline{f(S^{qn}x)} = \int_X f\otimes\overline{f}\, d\left(\frac{1}{N}\sum_{n\le N}\delta_{(S^{pn}x,S^{qn}x)}\right),$$

which, by passing to a convergent subsequence of the empiric measures, yields the following: $\frac{1}{N_k}\sum_{n\le N_k}\delta_{(S^{pn}x,S^{qn}x)}\to\rho$ implies that the limit in Eq. (2) is equal to $\int_{X\times X} f\otimes\overline{f}\, d\rho$ and that ρ is a joining of S^p and S^q (remembering that primitivity implies that (X_θ,S^p) and (X_θ,S^q) are uniquely ergodic, say, ν denotes the unique S-invariant measure). Note that we cannot prove the theorem above for *all* bounded multiplicative functions, as periodic sequences are always automatic and can also be multiplicative.

In fact, more than that is true. By taking $u(1)=u(2)=1$, $u(2n)=u(n)$ and $u(2n+1)=(-1)^n$ we obtain an automatic sequence which is not periodic but represents a completely multiplicative function. To explain this phenomenon and the use of the DKBSZ criterion, we should remember that we do not expect the limit joinings ρ to be product measure $\nu\otimes\nu$. This is impossible as (X_θ,S) has the odometer (H_λ,R) as its factor. If we consider the odometer H_λ with its unique invariant measure (Haar measure) ν_{H_λ} the measure-theoretic systems $(H_\lambda,\nu_{H_\lambda},R^p)$ and $(H_\lambda,\nu_{H_\lambda},R^q)$ are isomorphic! Whence the only ergodic joinings between them are graphs of relevant isomorphisms. Hence the simplest possible (ergodic) joinings between S^p and S^q are relative products over graph joinings. However, if these are the *only* ergodic joinings, each pair $(x,x)\in X_\theta\times X_\theta$ is generic for such a relative product and the DKBSZ criterion works for all continuous functions $f\in C(X_\theta)$ provided that $f\perp L^2(H_\lambda,\nu_{H_\lambda})$. Therefore, we have two tasks:

- to show that in the case of primitive substitutions we have the minimal number of possible ergodic joinings between S^p and S^q and
- to describe the structure of continuous functions; more precisely, to show that there are continuous functions orthogonal to the L^2-space of the underlying odometer, in fact (surprisingly) that the conditional expectation of each continuous function with respect to the odometer factor remains continuous.

The first task is done by using some results from the 1980s of Host and Parreau [2] (and also of Lemańczyk and Mentzen [4]) on the measure-theoretic centralizer of substitutions of constant length by showing non-isomorphism of different prime powers and using Mentzen's theorem on factors (of substitutions) to conclude relative disjointness. The second task is fulfilled by developing a (new) theory of substitution joinings which culminates in showing that each substitution has a representation in which it is relatively bijective over its synchronizing part.

References

1. Ferenczi, S., Kułaga-Przymus, J., Lemańczyk, M.: Sarnak's Conjecture: What's New. In: Ferenczi, S., Kułaga-Przymus, J., Lemańczyk, M. (eds.) Ergodic Theory and Dynamical Systems in their Interactions with Arithmetics and Combinatorics, LNM 2213, pp. 163–235. Springer, Cham (2018).
2. Host, B., Parreau, F.: Homomorphismes entre systèmes dynamiques définis par substitutions. Ergodic Th. & Dynam. Syst. **9**, 469–477 (1989)
3. Kátai, I.: A remark on a theorem of H. Daboussi. Acta Math. Hungar. **47**, 223–225 (1986)
4. Lemańczyk, M., Mentzen, M.K.: On metric properties of substitutions. Compositio Math. **65**, 241–263 (1988)
5. Lemańczyk, M., Müllner, C.: Automatic sequences are orthogonal to aperiodic multiplicative functions. Preprint arXiv:1811.00594
6. Müllner, C.: Automatic sequences fulfill the Sarnak conjecture. Duke Math. J. **166**, 3219–3290 (2017)
7. Sarnak, P.: Three lectures on the Möbius function, randomness and dynamics http://publications.ias.edu/sarnak/

Similarity isometries of shifted lattices and point packings

Manuel Joseph C. Loquias

1 Some preliminaries

A *lattice* Γ (of rank and dimension d) is a discrete subset of $\mathbb{R}^d$ that is the $\mathbb{Z}$-span of d linearly independent vectors $v_1, \ldots, v_d \in \mathbb{R}^d$ over $\mathbb{R}$. The set $\{v_1, \ldots, v_d\}$ is called a basis for Γ, and $\Gamma = \mathbb{Z}v_1 \oplus \cdots \oplus \mathbb{Z}v_d$. As a group, Γ is isomorphic to the free Abelian group of rank d. Alternatively, a lattice Γ may be defined as a discrete co-compact subgroup of $\mathbb{R}^d$.

On the other hand, a (*crystallographic*) *point packing* Λ is a non-empty point set of $\mathbb{R}^d$ such that there exists a lattice Γ in $\mathbb{R}^d$ and a finite point set F such that

$$\Lambda = \Gamma + F = \{\ell + f \mid \ell \in \Gamma \text{ and } f \in F\}.$$

That is, a point packing is the union of a lattice Γ and a finite number of translated copies of Γ. We refer to Γ as a generating lattice for Λ and the shifted lattice $x+\Gamma$, where $x \in F$, as a component of Λ. Observe that a point packing need not be a lattice.

Point packings have appeared in the literature under different names and different contexts. For instance, they appeared as non-lattice periodic packings in relation to the sphere packing problem in [3]. Dolbilin et al. referred to point packings in [4] as ideal or perfect crystals, and gave minimal sufficient geometric conditions on a discrete subset of $\mathbb{R}^d$ to be an ideal crystal. The term multilattice has also been used to pertain to a point packing, and arithmetic classification of multilattices have been studied in [9, 6].

Point packings serve as a standard model for 'ideal crystals', that is, crystals having multiple atoms per primitive unit cell. Examples of point packings include the honeycomb lattice, diamond lattice (crystal structure of diamond, tin, silicon, and germanium), and hexagonal closed packing (crystal structure of quartz).

Manuel Joseph C. Loquias
Institute of Mathematics, College of Science, University of the Philippines Diliman, Philippines.
e-mail: mjcloquias@math.upd.edu.ph

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_51

As point sets, point packings are Meyer sets (relatively dense sets Λ such that $\Lambda - \Lambda$ is uniformly discrete) [2]. Recall that a periodic point set is a discrete set $\Lambda \subset \mathbb{R}^d$ for which $\mathrm{per}(\Lambda) := \{t \in \mathbb{R}^d \mid t + \Lambda = \Lambda\}$ is non-trivial. If $\mathrm{per}(\Lambda)$ forms a lattice in $\mathbb{R}^d$, then we say that Λ is crystallographic. Point packings are exactly the locally finite point sets that are crystallographic [2].

2 Symmetry groups of point packings

The symmetry group of a point packing is a crystallographic group [4]. In particular, denote an isometry of $\mathbb{R}^d$ by (v,R), where $(v,R)x = v + Rx$, with $v \in \mathbb{R}^d$ and $R \in \mathrm{O}(d,\mathbb{R})$. The crystallographic restriction for point packings [2] states that if Λ is a point packing with $\mathrm{per}(\Lambda) = \Gamma$, then R is a symmetry of the lattice Γ whenever (v,R) is a symmetry of Λ. Here, we discuss several additional results on the symmetry group of a given point packing Λ.

Suppose $\Lambda = \Gamma + \{x_0 = 0, x_1, \ldots, x_{m-1}\}$. It is easy to see that $\mathrm{per}(\Gamma) \subseteq \mathrm{per}(\Lambda)$. The reverse inclusion does not always hold. It has been shown in [8] that there exists a (maximal) generating lattice Γ' of Λ that contains Γ such that $\mathrm{per}(\Lambda) = \mathrm{per}(\Gamma')$. In fact, if $S := \{x_i \mid (x_i, \mathrm{id})\Lambda = \Lambda\}$, then $\Gamma' = \Gamma + S$ is a lattice that generates Λ with $\mathrm{per}(\Lambda) = \mathrm{per}(\Gamma')$; see [5].

Hence, without loss of generality, we may assume that Λ is a point packing with $\mathrm{per}(\Lambda) = \mathrm{per}(\Gamma)$. Then, the isometry (v,R) is a symmetry of Λ if and only if for each $i \in \{0,\ldots,m-1\}$, there exists $j \in \{0,\ldots,m-1\}$ such that $(v,R)(x_i + \Gamma) = x_j + \Gamma$; see [5]. In words, the symmetries of $\Lambda \subseteq \mathbb{R}^d$ are precisely the isometries of $\mathbb{R}^d$ that induce a permutation of the components of Λ.

3 Similarity isometries of shifted lattices and point packings

Two lattices Γ and Γ' of $\mathbb{R}^d$ are said to be commensurate if the intersection of Γ and Γ' is a sublattice (of finite index) of both Γ and Γ'. A linear isometry R of $\mathbb{R}^d$ is called a (*linear*) *similarity isometry* of Γ whenever Γ and $\alpha R\Gamma$ are commensurate for some $\alpha \in \mathbb{R}^+$. Equivalently, R is a similarity isometry of Γ whenever $\beta R\Gamma$ is a sublattice of Γ for some $\beta \in \mathbb{R}^+$. The lattice $\beta R\Gamma$ is referred to as a similar sublattice of Γ. Given a similarity isometry R of Γ, the set $\mathrm{scal}_\Gamma(R)$ of scaling factors of R is defined to be the set of all real numbers α for which Γ and $\alpha R\Gamma$ are commensurate. Meanwhile, the set $\mathrm{Scal}_\Gamma(R)$ is comprised of all real numbers β for which $\beta R\Gamma \subseteq \Gamma$. The set of similarity isometries of Γ forms a group which we denote by $\mathrm{OS}(\Gamma)$. We use the notation $\mathrm{SOS}(\Gamma)$ for the group of similarity rotations of Γ. Various studies have examined the existence of similar sublattices as well as the properties of similarity isometries for particular lattices. We now give some initial results on similarity isometries of a point packing [1]. To this end, we first consider affine similarity isometries of lattices which allow us to study similarity

isometries of shifted lattices. This line of attack is analogous to the one used to investigate coincidence isometries of point packings [7].

Let (v,R) be an isometry of $\mathbb{R}^d$ and Γ be a lattice in $\mathbb{R}^d$. Then, there exists $\alpha \in \mathbb{R}^+$ such that Γ and $(v,\alpha R)\Gamma$ are commensurate if and only if Γ and $\alpha R\Gamma$ are commensurate and $v \in \Gamma + \alpha R\Gamma$. On the other hand, there exists $\beta \in \mathbb{R}^+$ such that $(v,\beta R)\Gamma \subseteq \Gamma$ if and only if $\beta R\Gamma \subseteq \Gamma$ and $v \in \Gamma$. Observe that we obtain inequivalent statements when we extend the equivalent definitions of linear similarity isometries of lattices to affine similarity isometries of lattices. The same phenomenon occurs when we look at similarity isometries of shifted lattices.

Let $\mathrm{OS}_1(x+\Gamma)$ be the set of linear isometries R such that $x+\Gamma$ is commensurate with $\alpha R(x+\Gamma)$ for some $\alpha \in \mathbb{R}^+$, and $\mathrm{OS}_2(x+\Gamma)$ be the set of linear isometries R such that $\beta R(x+\Gamma) \subseteq x+\Gamma$ for some $\beta \in \mathbb{R}^+$. We obtain that

$$\mathrm{OS}_1(x+\Gamma) = \{R \in \mathrm{OS}(\Gamma) \mid \alpha Rx - x \in \Gamma + \alpha R\Gamma \text{ for some } \alpha \in \mathrm{scal}_\Gamma(R)\}, \text{ and}$$
$$\mathrm{OS}_2(x+\Gamma) = \{R \in \mathrm{OS}(\Gamma) \mid \beta Rx - x \in \Gamma \text{ for some } \beta \in \mathrm{Scal}_\Gamma(R)\}.$$

This implies that $\mathrm{OS}_2(x+\Gamma) \subseteq \mathrm{OS}_1(x+\Gamma)$. The reverse inclusion does not always hold. Nonetheless, a sufficient condition for $R \in \mathrm{OS}_1(x+\Gamma)$ to be in $\mathrm{OS}_2(x+\Gamma)$ is if there exists $M \in \mathbb{N}$ such that $Mkx - x \in \Gamma$, where $k = [\alpha R\Gamma : \Gamma \cap \alpha R\Gamma]$ with $\alpha \in \mathrm{scal}_\Gamma(R)$.

To illustrate, suppose Γ is the square lattice $\mathbb{Z}[i]$. Then $\mathrm{OS}_2(\sqrt{2}+\Gamma) = \varnothing$ while $\mathrm{OS}_1(\sqrt{2}+\Gamma) = \langle T_r \rangle$, where T_r is the reflection along the real axis. On the other hand,

$$\mathrm{SOS}_1(\tfrac{1}{1-i}+\Gamma) = \mathrm{SOS}_2(\tfrac{1}{1-i}+\Gamma) \subset \mathrm{SOS}(\Gamma).$$

From the above discussion, given a point packing $\Lambda = \Gamma + \{x_0 = 0, x_1, \ldots, x_{m-1}\}$, we will say that R is a similarity isometry of Λ if there exists $\beta \in \mathbb{R}^+$ such that $\beta R\Lambda = \bigcup_{i=0}^{m-1} \beta R(x_i+\Gamma) \subset \Lambda$.

References

1. Arias, J.C.H., Loquias, M.J.C.: Similarity isometries of point packings. Acta Cryst. A 76, 677–686 (2020)
2. Baake, M., Grimm, U.: Aperodic Order. Vol. 1. A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
3. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups. 3rd ed. Springer, New York (1999)
4. Dolbilin, N.P., Lagarias, J.C., Senechal, M.: Multiregular point systems. Discr. Comput. Geom. **20**, 477–498 (1998)
5. Estillore, N.V.D., Felix, R.P., Loquias, M.J.C.: Symmetries and colorings of point packings. In preparation.
6. Indelicato, G.: An algorithm for the arithmetic classification of multilattices. Acta Cryst. A **69**, 63–74 (2013)
7. Loquias, M.J.C., Zeiner, P.: The coincidence problem for shifted lattices and crystallographic point packings. Acta Cryst. A **70**, 656–669 (2014)
8. Parry, G.P.: On essential and non-essential descriptions of multilattices. Math. Mech. Solids **9**, 411–418 (2004)

9. Pitteri, M., Zanzotto, G.: Continuum Models for Phase Transitions and Twinning in Crystals. Chapman & Hall/CRC, Boca Raton, Florida (2003)

Renormalisation for inflation tilings II: Connections to number theory

Neil Mañibo

In the study of spectral properties of a d-dimensional aperiodic tiling which arises from an inflation rule ρ on a finite set of prototiles, one recovers a system of renormalisation relations for measures which make up the diffraction measure $\widehat{\gamma}$; see [2, 3] for general notions and [5, 4, 6] for the rigorous treatment of certain classes. One can show that each component of the Lebesgue decomposition of $\widehat{\gamma}$ satisfies these relations independently. In particular, the Radon–Nikodym density $h(k)$ representing the absolutely continuous component $\widehat{\gamma}_{\mathsf{ac}}$ exhibits a certain scaling behaviour, which is encoded in the Fourier matrix $B(k)$. Here, $k \in \mathbb{R}^d$ and the dimension of $B(k)$ is given by the number of prototiles.

When $d = 1$, the exponential growth of $h(k)$ along orbits of the dilation map $k \mapsto \lambda k$ is determined by the Lyapunov exponent $\chi^B(k)$ of the matrix cocycle induced by $B(k)$, where λ is the inflation multiplier of ρ. If, for a set of real parameters k of full measure, this exponent is bounded from above by $\log\sqrt{\lambda} - \varepsilon$ for some $\varepsilon > 0$, the diffraction $\widehat{\gamma}$ is singular, i.e., $\widehat{\gamma}_{\mathsf{ac}} = 0$.

What is described below is based on joint work with Michael Baake, Michael Coons, Franz Gähler, and Uwe Grimm. In general, one can view $B(k)$ as a section of a function $\widetilde{B}$ on $\mathbb{T}^r$, where r is the algebraic degree of λ. This allows one to obtain a sequence of upper bounds for $\chi^B(k)$ via the mean of $\log|\widetilde{B}(.)|$, which are normalised logarithmic Mahler measures of multivariate polynomials.

Proposition 1 ([7]). *Let ρ be a one-dimensional primitive inflation with inflation multiplier λ of algebraic degree r. Assuming that $B(k)$ is invertible for some $k \in \mathbb{R}$, there exists a sequence of multivariate polynomials $P_N \in \mathbb{Z}[x_1, \ldots, x_r]$ such that, for each $N \in \mathbb{N}$,*

$$\chi^B(k) \leqslant \frac{1}{N}\mathfrak{m}(P_N)$$

for a.e. $k \in \mathbb{R}$, where $\mathfrak{m}(P)$ is the logarithmic Mahler measure of P. □

Neil Mañibo
Faculty of Mathematics, Bielefeld University, Germany. e-mail: cmanibo@math.uni-bielefeld.de

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_52

Whenever λ is an integer or a Pisot number, the Lyapunov exponent $\chi^B(k)$ exists as a limit and is constant for a.e. $k \in \mathbb{R}$. In some cases, this a.e. value is given as a logarithmic Mahler measure. Finding suitable bounds for $\chi^B(k)$ then reduces to bounding logarithmic Mahler measures of certain polynomials.

A Borwein polynomial is a polynomial whose coefficients lie in $\{-1,0,1\}$. For an inflation derived from a binary substitution of constant length, one can show that $\chi^B(k) = \mathfrak{m}(P)$, for a.e. $k \in \mathbb{R}$, where P is a Borwein polynomial. Indeed, for this class of inflations, $\mathfrak{m}(P)$ is always strictly bounded from above by $\log\sqrt{\lambda}$, thus implying the singularity of $\widehat{\gamma}$ [6]. In fact, the correspondence also goes the other way as follows.

Proposition 2 ([1]). *Let P be a Borwein polynomial. Then, there exists at least one binary constant-length substitution ρ such that*

$$\mathfrak{m}(P) = \chi^B(k)$$

for a.e. $k \in \mathbb{R}$. □

A famous problem regarding logarithmic Mahler measures of polynomials in $\mathbb{Z}[x]$ is Lehmer's problem, which asks whether there is a universal non-zero constant c that serves as a lower bound for all non-zero logarithmic Mahler measures $\mathfrak{m}(P)$, with $P \in \mathbb{Z}[x]$. This question has been answered affirmatively for certain subclasses, but the general case remains open.

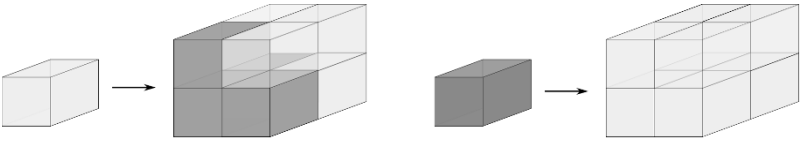
In view of Lehmer's problem, Borwein polynomials form an important class due to a result by Pathiaux which states that a polynomial $P \in \mathbb{Z}[x]$ with $\mathfrak{m}(P) < \log(2)$ must divide a Borwein polynomial, i.e., $Q = PR$ for some Borwein Q.

From his previous calculations, Boyd noted that the other divisor R can be chosen to have relatively small degree with respect to P and so that $\mathfrak{m}(R) = 0$. So far, there is still no proof of the existence of such a mollifier polynomial R; see [1] and references therein for details. If such an R always exists, the correspondence given in Proposition 2 gives the following dynamical version of Lehmer's problem.

Conjecture 1 ([1]). For all binary substitutions ρ of constant length, whose associated Lyapunov exponent $\chi^B(k)$ is a.e. non-zero, there exists a non-zero constant c such that $\chi^B(k) \geqslant c$ for a.e. $k \in \mathbb{R}$.

One should note that Proposition 2 extends to higher dimensions, where one can always realise a logarithmic Mahler measure of a multivariate Borwein polynomial as a Lyapunov exponent of a binary block inflation.

As an example, the logarithmic Mahler measure of $1+x+y+z$ is realised as the Lyapunov exponent associated to the following binary block substitution in $\mathbb{R}^3$.



It is interesting to note that $\mathfrak{m}(1+x+y+z) = \frac{7}{2\pi^2}\zeta(3)$, where $\zeta(s)$ is Riemann's zeta function.

References

1. Baake, M., Coons, M., Mañibo, N.: Binary constant length-substitutions and Mahler measures of Borwein polynomials. In: Bailey, D. et al. (eds.) From Analysis to Visualization, JBCC 2017. Springer Proceedings in Mathematics and Statistics vol. 313, pp. 303–322 (2020)
2. Baake, M., Frank, N.P., Grimm, U., Robinson, E.A.: Geometric properties of a binary non-Pisot inflation and absence of absolutely continuous diffraction. Studia Math. **247**, 109–154 (2019)
3. Baake, M., Gähler, F., Mañibo, N.: Renormalisation of pair correlation measures for primitive rules and absence of absolutely continuous diffraction. Commun. Math. Phys. **370**, 591–635 (2019)
4. Baake, M., Grimm, U.: Renormalisation of pair correlations and their Fourier transforms for primitive block substitutions. In: Akiyama, S., Arnoux, P. (eds.) Tiling and Discrete Geometry. Springer, Berlin, in press
5. Baake, M., Grimm, U., Mañibo, N.: Spectral analysis of a family of binary inflation rules. Lett. Math. Phys. **108**, 1783–1805 (2018)
6. Mañibo, N.: Lyapunov exponents for binary constant-length substitutions. J. Math. Phys. **58**, 113504:1–9 (2017)
7. Mañibo, N.: Lyapunov Exponents in the Spectral Theory of Primitive Inflation Systems. PhD thesis, Bielefeld University (2019)

The Penrose and the Taylor–Socolar tilings, and first steps to beyond

Robert V. Moody

The Penrose hexagonal tilings form a family of aperiodic tilings comprised of arrowed double-hexagon tiles based on the standard periodic tiling of the plane by equilateral triangles [7, 8]. Each Penrose tile consists of a hexagon whose edges are arrowed, and within it a smaller hexagon of $1/3$ the area with orthogonal orientation whose edges are also arrowed. The matching of the arrowing forces aperiodicity, and remarkably, in spite of the rich symmetry of the usual hexagonal tilings, none of the Penrose hexagonal tilings has any non-trivial symmetry at all.

As observed in [5], each Penrose hexagonal tiling can be described algebraically by a pair of inverse sequences based on the nesting of equilateral triangles. This description is almost always (in a measure theoretical sense) unambiguous, and the exceptions (singular cases) are seen algebraically to be directly due to the symmetries of the underlying triangular lattice. In fact, from this perspective, one can see the Penrose tilings as being a symmetry-breaking construction.

This talk, which is based on joint work with Jeong-Yup Lee, is concerned with these symmetries—how they appear both geometrically and algebraically, and how they are broken by the Penrose tilings themselves. We also explore the relationship of these singularities to the singular Taylor–Socolar tilings. The Penrose tilings have long been known to be closely connected with another aperiodic tiling based on the usual hexagonal tiling of the plane, this one due to Joan Taylor and Joshua Socolar [10, 9]. Still, the connection between the two tilings is subtle [1]; see also [2, Sec. 6.4]. Here we show the connection directly, seeing it as the outcome of a $3:1$ mapping of the inverse sequences which pass from the Penrose tilings to Taylor–Socolar tilings. Under this mapping, each Penrose tiling gives rise to a unique Taylor–Socolar tiling, although only one third of its tiles actually appear directly from the Penrose hexagons (as the smaller interior hexagons of the Penrose hexagons). In the reverse direction, each Taylor–Socolar tiling gives rise to three different Penrose tilings.

The original version of this chapter has been revised. Chapter title was incorrect in online TOC and has been updated. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-62497-2_66

Robert V. Moody
University of Victoria, Canada. e-mail: rvmoody@mac.com

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_53

A key feature of both tilings is the underlying geometry of what is called a Coxeter Euclidean kaleidoscope [3]—the infinite configuration of hyperplanes that define the reflections of a Euclidean Coxeter group. It seems to us that putting these two remarkable tilings in the setting of kaleidoscopes might lead to a deeper understanding of how they actually arise. We conclude the talk with a brief introduction to the full classification of all the Euclidean kaleidoscopes that range through the famous A, B, C, D, E, F, G series, of which the Penrose and Taylor–Socolar tilings belong to the Euclidean G_2 kaleidoscope. In spite of the complications of higher dimensions, the geometry of these kaleidoscopes is quite articulately described, and in particular there are good descriptions of both their Voronoi and Delaunay cells [6]. We make the suggestion that each kaleidoscope may give rise to families of aperiodic tilings in ways similar to those from which the Penrose and Taylor–Socolar tilings can be derived. We give a short initial foray into this by creating a new square substitution tiling which arises by the same sort of ideas from the Euclidean B_2 kaleidoscope.

References

1. Baake, M., Gähler, F., Grimm, U.: Hexagonal inflation tilings and planar monotiles. Symmetry **4**, 581–602 (2012)
2. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
3. Coxeter, H.S.M.: Regular polytopes. Dover Publications, 3rd edition (1973)
4. Lee, J.-Y., Moody, R.V.: Taylor–Socolar hexagonal tilings. Symmetry **5**, 1–46 (2013)
5. Lee, J.-Y., Moody, R.V.: On the Penrose and Taylor–Socolar hexagonal tilings. Acta Cryst. A **73**, 246–256 (2017)
6. Moody, R.V., Patera, J.: Voronoi domains and dual cells in the generalized kaleidoscope with applications to root and weight lattices. Can. J. Math. **47**, 573–605 (1995)
7. Penrose, R.: Remarks on a tiling: Details of a $(1+\varepsilon+\varepsilon^2)$-aperiodic set. In: Moody R.V. (ed.) The Mathematics of Long-Range Aperiodic Order, pp. 467–497. Kluwer, Dordrecht (1997)
8. Penrose, R.: Twistor Newsletter. Reprinted in: Roger Penrose: Collected Works, Volume 6: 1997-2003. Oxford University Press, New York (2010). Available at http://people.maths.ox.ac.uk/lmason/Tn/. See specifically: http://people.maths.ox.ac.uk/lmason/Tn/41/TN41-08.pdf, http://people.maths.ox.ac.uk/lmason/Tn/42/TN42-09.pdf, http://people.maths.ox.ac.uk/lmason/Tn/43/TN43-11.pdf
9. Socolar, J., Taylor, J.: An aperiodic hexagonal tile. J. Comb. Theory A **118**, 2207–2231 (2011)
10. Taylor, J.: Aperiodicity of a functional monotile. Preprint (2010). http://www.math.uni-bielefeld.de/sfb701/files/preprints/sfb10015.pdf

Scaling properties of the Thue–Morse measure: A summary

Tanja Schindler

This is an extended abstract of the paper 'Scaling properties of the Thue–Morse measure' by Baake, Gohlke, Kesseböhmer and Schindler [1].

The Thue–Morse diffraction measure for the balanced-weight case is given by the infinite Riesz product

$$\mu_{\mathrm{TM}} = \prod_{\ell=0}^{\infty} \bigl(1 - \cos(2\pi 2^{\ell} k)\bigr), \tag{1}$$

with convergence in the vague topology, see [2, Sec. 10.1] and references therein. As such, μ_{TM} is a translation-bounded, positive measure on $\mathbb{R}$ that is purely singular continuous and 1-periodic. Clearly, $\mu_{\mathrm{TM}} = \nu * \delta_{\mathbb{Z}}$, with $\nu = \mu_{\mathrm{TM}}|_{[0,1)}$ being a probability measure on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, the latter represented by $[0,1)$ with addition modulo 1. In this case, ν is the weak limit of Radon–Nikodym densities of finite products as the right-hand side of (1).

If we denote by $B(x,r)$ the ball around x with radius r (either with respect to the Euclidean or the subshift metric), one way to quantify how concentrated the measure ν is at a given point $x \in \mathbb{T}$ is to determine its *local dimension*, given by

$$\dim_{\nu}(x) = \lim_{r \to 0} \frac{\log \nu(B(x,r))}{\log(r)},$$

provided that the limit exists. Due to their highly irregular structure, we cannot hope to pin down the level sets of $\dim_{\nu}$ explicitly. However, the corresponding *Hausdorff dimension*,

$$f(\alpha) = \dim_{\mathrm{H}}\{x \in \mathbb{T} : \dim_{\nu}(x) = \alpha\},$$

yields a properly behaved function of α. The analysis of the *dimension spectrum* $f(\alpha)$ is one of the open questions considered in [7]. This problem turns out to be

Tanja Schindler
Research School of Finance, Actuarial Studies and Statistics, The Australian National University, 26C Kingsley St, Acton ACT, 2601, Australia. e-mail: tanja.schindler@anu.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_54

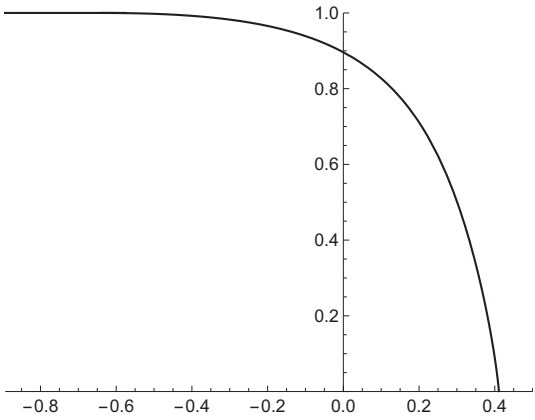


Fig. 1 The graph of the Birkhoff spectrum b from Eq. (5).

intimately related to pointwise scaling properties of the approximants in Eq. (1). More precisely, we consider

$$\beta(x) := \lim_{n\to\infty} \frac{1}{n\log(2)} \log \prod_{\ell=0}^{n-1} \bigl(1-\cos(2^{\ell+1}\pi x)\bigr),$$

for all $x \in \mathbb{T}$ for which the limit exists. The limit is known for Lebesgue-a.e. $x \in \mathbb{T}$, in which case it equals -1, and for some particular examples of non-typical points; see [4, 3].

There is a natural way to interpret β in terms of the Birkhoff average of some function $\psi\colon \mathbb{T} \longrightarrow [-\infty, \log(2)]$,

$$\psi(x) = \log\bigl(1-\cos(2\pi x)\bigr), \quad \beta(x) = \lim_{n\to\infty} \frac{\psi_n(x)}{n\log(2)}, \tag{2}$$

where $\psi_n(x) = \sum_{\ell=0}^{n-1} \psi(2^\ell x)$. With this, we are interested in the *Birkhoff spectrum* $b(\alpha) = \dim_{\mathrm{H}} \mathscr{B}(\alpha)$ with

$$\mathscr{B}(\alpha) = \Bigl\{x \in \mathbb{T} : \lim_{n\to\infty} \frac{\psi_n(x)}{n} = \alpha\Bigr\} = \Bigl\{x \in \mathbb{T} : \beta(x) = \frac{\alpha}{\log(2)}\Bigr\}. \tag{3}$$

It is one of the strengths of the thermodynamic formalism to connect such locally defined functions to the Legendre transform of a globally defined quantity. An adequate choice for the latter in our situation is the *topological pressure* of the function $t\psi$, $t \in \mathbb{R}$, defined by

$$p(t) := \mathscr{P}(t\psi) := \lim_{n\to\infty} \frac{1}{n} \log \sum_{J\in I_n} \sup_{x\in J} \exp\bigl(t\psi_n(x)\bigr), \tag{4}$$

where, for each $n \in \mathbb{N}$, I_n forms a partition of $[0,1]$ into intervals of length 2^{-n}.

Indeed, the relation between $f(\alpha)$ and $b(\alpha)$ given in [1, Thm. 1.1] is analogous to known results for Hölder continuous potentials [6, Cor. 1]: If p^* denotes the Legendre transform of p, one obtains

$$b(\alpha) = \max\Big\{\frac{-p^*(\alpha)}{\log(2)}, 0\Big\} \text{ and } f(\alpha) = b\big(\log(2)(1-\alpha)\big), \tag{5}$$

with the graph of $b(\alpha)$ given in Figure 1, confirming some numerical and scaling-based results of [5].

References

1. Baake, M., Gohlke, P., Kesseböhmer, M., Schindler, T.: Scaling properties of the Thue–Morse measure. Discr. Cont. Dynam. Syst. A, in press. arXiv:1810.06949
2. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
3. Baake, M., Grimm, U., Nilsson, J.: Scaling of the Thue–Morse diffraction measure. Acta Phys. Pol. A **126**, 431–434 (2014)
4. Cheng, Z., Savit, R., Merlin, R.: Structure and electronic properties of Thue–Morse lattices. Phys. Rev. B **37**, 4375–4382 (1982)
5. Godrèche, C., Luck, J.M.: Multifractal analysis in reciprocal space and the nature of the Fourier transform of self-similar structures. J. Phys. A: Math. Gen. **23**, 3769–3797 (1990)
6. Pesin, Y., Weiss, H.: The multifractal analysis of Birkhoff averages and large deviations. In: Broer, H.W., Krauskopf, B., Vegter, G. (eds.) Global Analysis of Dynamical Systems, pp. 419–431. IoP Publishing, Bristol and Philadelphia (2001)
7. Queffélec, M.: Questions around the Thue–Morse sequence. Unif. Distrib. Th. **13**, 1–25 (2018)

Weak model sets

Nicolae Strungaru

1 Square-free integers

In this talk, which is based on joint work with M. Baake and C. Huck, we will review the properties of weak model sets of extremal density. These results have been proved independently in [3] and [6], and we recommend these for more details.

Let us start by recalling the example of the set S of square-free integers,

$$S := \{n \in \mathbb{Z} : \forall p \in \mathbb{P}, p^2 \nmid n\},$$

where $\mathbb{P}$ denotes the set of primes.

Theorem 1 ([4]). *The autocorrelation measure γ of S, with respect to the natural van Hove sequence $\left(A_m = [-m,m]\right)_{m\in\mathbb{N}}$, exists. The correponding diffraction measure, $\widehat{\gamma}$, is a pure point measure.*

Note that, with respect to other van Hove sequences, S can have mixed diffraction spectrum. We will see below why the choice of the natural van Hove sequence is important, and leads to a connection to a cut and project scheme (CPS).

We assume below that the reader is familiar with the cut and project formalism and with regular model sets. For a review of this, we recommend the monograph [1] for $G = \mathbb{R}^d$, and [8] for general G.

To describe S, consider the following CPS,

$$\begin{array}{ccccc}
\mathbb{R} & \xleftarrow{\pi_G} & \mathbb{R}\times H & \xrightarrow{\pi_H} & H := \prod_{p\in\mathbb{P}} \mathbb{Z}/p^2\mathbb{Z} \\
\cup & & \cup & & \cup \text{ dense} \\
L & \xleftarrow{1-1} & \mathscr{L} := \{(n,\tau(n)) : n \in \mathbb{Z}\} & \longrightarrow & L^\star
\end{array}$$

Nicolae Strungaru
MacEwan University, Edmonton, Alberta, Canada. e-mail: StrungaruN@macewan.ca

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_55

where $\tau\colon\ \mathbb{Z} \longrightarrow H$ is defined by $\tau(n) = \left(n \bmod p^2\right)_{p\in\mathbb{P}}$. Consider the set

$$W := \prod_{p\in\mathbb{P}} \left((\mathbb{Z}/p^2\mathbb{Z}) \setminus \{0\}\right),$$

which acts as a window for the above CPS.

Theorem 2 ([4, 2]). *For the square-free integers, the following properties hold.*

1. *The window W is compact, with $W = \partial W$ and $S = ⅄(W)$.*
2. *The natural autocorrelation and diffraction measures of S are given by*

$$\gamma = \omega_{1_W * \widetilde{1_W}}, \quad \widehat{\gamma} = \omega_{|\check{1}_W|^2}. \tag{1}$$

While (1) is the usual formula of the diffraction of regular model sets, the window W has no interior and a boundary of positive measure. Thus, the standard proofs for the diffraction of regular model sets do not apply.

2 Weak model sets of maximal density

Definition 1. Let $(G,H,\mathscr{L})$ be a CPS. If $W \subseteq H$ is compact, we say that $⅄(W)$ is a *weak model set*. We say that the weak model set $⅄(W)$ has *maximal density* with respect to $\mathscr{A} = (A_n)_{n\in\mathbb{N}}$ if

$$\operatorname{dens}_{\mathscr{A}}(⅄(W)) := \lim_n \frac{\operatorname{card}(⅄(W)\cap A_n)}{\theta_G(A_n)} = \operatorname{dens}(\mathscr{L})\,\theta_H(W),$$

where θ_G and θ_H are the Haar measures of the groups G and H, respectively.

Note that the right-hand side is always an upper bound for the left-hand side [5]. Regular model sets have maximal density. Square-free integers, as well as visible lattice points, have maximal density with respect to the natural van Hove sequence. The next result shows that generic positions of compact windows define weak model sets of maximal density.

Proposition 1 ([7]). *Let $(G,H,\mathscr{L})$ be a CPS, let $W \subseteq H$ be compact, and let $\mathscr{A}$ be a tempered van Hove sequence. Then, for generic $(x,y)+\mathscr{L} \in (G\times H)/\mathscr{L}$, the weak model set $-x+⅄(y+W)$ has maximal density with respect to $\mathscr{A}$.*

For weak model sets of maximal density, we have the following result; see [3, 6] for details.

Theorem 3 ([3, 6]). *Let $(G,H,\mathscr{L})$ be a CPS, and $⅄(W)$ a weak model set of maximal density with respect to $\mathscr{A} = (A_n)_{n\in\mathbb{N}}$. Then, the following properties hold.*

1. *With respect to $\mathscr{A}$, the set $⅄(W)$ has autocorrelation and diffraction*

$$\gamma = \operatorname{dens}(\mathscr{L})\omega_{1_W * \widetilde{1_W}} \quad \text{and} \quad \widehat{\gamma} = (\operatorname{dens}(\mathscr{L}))^2\,\omega_{|\check{1}_W|^2}.$$

2. *For each $\chi \in \widehat{G}$, the Fourier–Bohr coefficient a_χ exists, with*

$$a_\chi := \lim_n \frac{1}{\theta_G(A_n)} \sum_{x \in \curlywedge(W) \cap A_n} \overline{\chi(x)} = \operatorname{dens}(\mathscr{L}) \int_W \overline{\chi^\star(t)}\, \mathrm{d}t.$$

3. *There exists an ergodic measure ν for the dynamical system $\big(\mathbb{X}(\curlywedge(W)), G\big)$ such that $\curlywedge(W)$ is generic for ν.*

The measure ν can be identified as the unique invariant measure with maximal density for generic configurations.

References

1. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
2. Baake, M., Huck, C.: Ergodic properties of visible lattice points. Proc. Steklov Inst. Math. **288**, 184-208 (2015)
3. Baake, M., Huck, C., Strungaru, N.: On weak model sets of extremal density. Indag. Math. **28**, 3–31 (2017)
4. Baake, M., Moody, R.V., Pleasants, P.A.B.: Diffraction from visible lattice points and kth power free integers. Discr. Math. **221**, 3-42 (2000)
5. Huck, C., Richard, C.: On pattern entropy of weak model sets. Discr. Comput. Geom. **54**, 741-757 (2015)
6. Keller, G., Richard, C.: Dynamics on the graph of the torus parametrisation. Ergod. Th. & Dynam. Syst. **38**, 1048–1085 (2018)
7. Moody, R.V.: Uniform distribution in model sets. Can. Math. Bull. **45**, 123-130 (2002)
8. Richard, C., Strungaru, N.: Pure point diffraction and Poisson summation. Ann. H. Poincaré **18**, 3903-3931 (2017)

Doubly sparse measures on locally compact Abelian groups

Venta Terauds

In this work, joint with Michael Baake and Nicolae Strungaru [3], we are interested in *doubly sparse* measures on a locally compact Abelian group (LCAG) G. By a doubly sparse measure, we mean a Fourier-transformable Radon measure μ such that both $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\widehat{\mu})$ are locally finite point sets in G and $\widehat{G}$, respectively. In particular, both μ and $\widehat{\mu}$ must then be pure point measures.

This work has its origins in the study of crystals and quasicrystals: a physical structure, represented by a point measure in $\mathbb{R}^d$, is considered to have long range order when its diffraction is also a pure point measure. In the simplest case, we have a periodic structure represented by the Dirac comb of a lattice, in which case its diffraction is also periodic: for a general lattice $\Gamma \subseteq \mathbb{R}^d$, we have from the Poisson summation formula (PSF) that $\widehat{\delta_\Gamma} = \mathrm{dens}(\Gamma) \cdot \delta_{\Gamma^*}$ and hence the diffraction

$$\widehat{\gamma_\Gamma} = \mathrm{dens}(\Gamma)^2 \cdot \delta_{\Gamma^*};$$

see [1] for general background.

A model set is a point set gained from a cut and project scheme (CPS) by projecting onto a group, G, from a lattice in a higher-dimensional superspace, $G \times H$, via a sufficiently nice, relatively compact window in H. Such sets form natural mathematical models of quasicrystals, being non-periodic with pure point diffraction. We apply some recent results of Strungaru [7], who characterised measures that may be written as model combs in a CPS, and Richard and Strungaru [6], who proved the PSF for measures supported on a model set, using the PSF of the underlying lattice.

Meyer sets possess a strong form of finite local complexity and may always be constructed as relatively dense subsets of model sets. In fact, a Fourier-transformable measure μ supported inside a Meyer set with pure point Fourier transform $\widehat{\mu}$ can be written as a model comb in a CPS, with its coefficients determined by a continuous function of compact support on the internal space, H. Using this, we show that, un-

Venta Terauds
Discipline of Mathematics, University of Tasmania, Private Bag 37, Hobart, TAS 7001, Australia.
e-mail: venta.terauds@utas.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_56

der relatively mild conditions of sparseness on the support of $\widehat{\mu}$, both μ and $\widehat{\mu}$ (and hence the diffraction of μ) are supported on finitely many translates of a lattice, and thus have a periodic structure.

If a measure μ has uniformly discrete support and is positive definite, with pure point Fourier transform $\widehat{\mu}$, then $\widehat{\mu}$ may again be written as a model comb in a CPS, however with coefficients determined by a continuous function vanishing at infinity on the internal space, H. In this case, with similarly mild conditions of sparseness on the support of $\widehat{\mu}$, we show that μ is the limit of a sequence of measures with periodic structure.

Our results can be seen as a generalisation of many of those of Lev and Olevskii [4, 5] from the Euclidean to the general LCAG setting. To conclude, we consider some consequences of our results for measures supported on $\mathbb{R}^d$. In particular, we show the following. If a measure μ, supported inside a model set in a fully Euclidean CPS, is such that the support of the pure point part of $\widehat{\mu}$, that is, $\mathrm{supp}(\widehat{\mu}_{\mathsf{pp}})$, is locally finite then, in fact, $\widehat{\mu}_{\mathsf{pp}} = 0$; see [3] for details.

References

1. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013).
2. Baake, M., Grimm, U. (eds.): Aperiodic Order. Vol. 2: Crystallography and Almost Periodicity. Cambridge University Press, Cambridge (2017).
3. Baake, M., Strungaru, N., Terauds, V.: On pure point measures with sparse support and sparse Fourier–Bohr support, arXiv:1908.00579 (2019).
4. Lev, N., Olevskii, A.: Quasicrystals and Poisson's summation formula. Invent. Math. **200**, 585–606 (2015).
5. Lev, N., Olevskii, A.: Fourier quasicrystals and discreteness of the diffraction spectrum. Adv. Math. **315**, 1–26 (2017).
6. Richard, C., Strungaru, N.: Pure point diffraction and Poisson summation. Ann. H. Poincaré **18**, 3903–3931 (2017).
7. Strungaru, N.: Almost periodic measures and Meyer sets. In [2], pp. 271–342.

The mean-median map

Franco Vivaldi

Consider a finite multiset $\xi = [x_1, \ldots, x_n]$ of real numbers. The *arithmetic mean* $\langle \xi \rangle$ and the *median* $\mathscr{M}(\xi)$ of ξ are defined, respectively, as

$$\langle \xi \rangle = \frac{1}{|\xi|} \sum_{x \in \xi} x \qquad \text{and} \qquad \mathscr{M}(\xi) = \begin{cases} x_{j_{\frac{n+1}{2}}} & n \text{ odd}, \\ \frac{1}{2}\left(x_{j_{\frac{n}{2}}} + x_{j_{\frac{n}{2}+1}} \right) & n \text{ even}, \end{cases}$$

where $x_{j_1} \leqslant x_{j_2} \leqslant \cdots \leqslant x_{j_n}$, for some permutation $k \mapsto j_k$ of indices.

We enlarge ξ by adjoining to it a new real number x_{n+1} determined by the requirement that the arithmetic mean of the enlarged multiset be equal to the median of the original multiset:

$$x_{n+1} = (n+1)\mathscr{M}(\xi) - n\langle \xi \rangle.$$

This rule is known as the *mean-median map* (MMM), which was introduced in [5], and subsequently studied in [2, 1, 3].

Since the MMM commutes with affine transformations [5], the simplest non-trivial case — three distinct initial numbers — may be studied in full generality by considering the initial multiset $[0, x, 1]$, with $x \in [\frac{1}{2}, \frac{2}{3}]$, exploiting symmetries [2]. Here one finds already substantial difficulties, which are synthesised in the following conjectures.

Conjecture 1 (Strong terminating conjecture [5]). The MMM sequence of any initial multiset is eventually constant.

For the system $[0, x, 1]$, we let the *transit time* $\tau(x)$ be the time at which the MMM sequence becomes constant (letting $\tau(x) = \infty$ if this does not happen). If the MMM sequence $(x_n)_{n=1}^{\infty}$ converges at x — with finite or infinite transit time — we have

Franco Vivaldi
School of Mathematical Sciences, Queen Mary, University of London, London E1 4NS, UK.
e-mail: f.vivaldi@qmul.ac.uk

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_57

a real function $x \mapsto m(x)$, called the *limit function*, which gives the limit of this sequence. This function has an intricate, distinctive structure.

Conjecture 2 (Continuity conjecture [2]). The function $x \mapsto m(x)$ is continuous.

In [2], both conjectures were proved to hold in a neighbourhood of $x = \frac{1}{2}$, where m turns out to be affine. Using a computer-assisted proof, this result was then substantially extended in [1], where the limit function was constructed in small neighbourhoods of all rational numbers with denominator at most 18 lying in the interval $\left[\frac{1}{2}, \frac{2}{3}\right]$. The authors also identified 17 rational numbers at which m is non-differentiable.

In this joint work with Jonathan Hoseana [4], motivated by the above investigations, we study the mean-median map as a dynamical system on the space of finite multisets $[Y_1(x), \ldots, Y_n(x)]$ of piecewise-affine continuous functions with rational coefficients, the MMM map being defined pointwise. We study the limit function in the vicinity of its local minima. The latter occur at a distinctive family of rational points, the so-called *X-points*, which are transversal intersections of the functions Y_k. We prove the existence of local symmetries (homologies) around X-points, which result in affine functional equations for the limit function. We establish the general form of the limit function near an X-point, and show that the X-points form a hierarchical structure, whereby each X-point typically generates an *auxiliary sequence* of like points; such sequences form the scaffolding of the intricate structure of local minima of the limit function.

We then show that there is a one-parameter family of dynamical systems over $\mathbb{Q}$ — the *reduced system* — which, after suitable scaling, represent the dynamics near any X-point with given transit time. This simplification results from the fact that the reduced dynamics is largely unaffected by the earlier history of the X-point.

By exploiting the dynamics of the reduced system, we have established the strong terminating conjecture for $[0, x, 1]$ in neighbourhoods of 2791 rational numbers in the interval $\left[\frac{1}{2}, \frac{2}{3}\right]$, thereby extending the results of [1] by two orders of magnitude. This large data collection makes it clear that the domains over which the limit function is regular do not account for the whole Lebesgue measure, suggesting the existence of a drastically different, yet unknown, dynamical behaviour.

For a quantitative assessment of this phenomenon, we have computed a lower bound for the total variation of the limit function, sampled over a set of some 202,000 Farey points. Our data suggest the following conjecture.

Conjecture 3. The Hausdorff dimension of the graph of the limit function of the system $[0, x, 1]$ is greater than 1.

References

1. Cellarosi, F., Munday, S.: On two conjectures for M&m sequences. J. Diff. Eqs. Appl. **22**, 428–440 (2016)

2. Chamberland, M., Martelli, M.: The mean-median map. J. Diff. Eqs. Appl. **13**, 577–583 (2007)
3. Hoseana, J.: The mean-median map. MSc thesis, Queen Mary, University of London (2015)
4. Hoseana, J., Vivaldi, F.: Geometrical properties of the Mean-Median map. J. Comp. Dyn, **7**, 83–121 (2020)
5. Shultz, H., Shiflett, R.: M&m sequences. College Math. J. **36**, 191–198 (2005)

Similar sublattices and submodules

Peter Zeiner

In this talk, we want to give an overview of similar sublattices and submodules and the recent progress in this area. A milestone are some general existence results for similar sublattices of rational lattices [6] by Conway, Rains and Sloane. Since then, detailed explicit results have been achieved for a large collection of lattices and $\mathbb{Z}$-modules in dimensions $d \leq 4$, including planar modules related to cyclotomic integers [1], root lattices such as the A_4-lattice [2] and related modules in 4 dimensions, where one makes use of certain quaternion algebras [3], but also for less symmetric lattices in the plane [4]. Another important result is the establishment of a close connection between similar submodules and coincidence site modules (CSMs) [7, 8, 10].

A similar sublattice (SSL) of a lattice Γ is a sublattice of full rank that is similar to Γ; see [6, 5]. By a $\mathbb{Z}$-module, we mean a $\mathbb{Z}$-module which is (properly) embedded in $\mathbb{R}^d$, that is, a module $M \subset \mathbb{R}^d$ such that there is a $\mathbb{Z}$-basis $\{b_1, \ldots, b_n\}$ of M whose $\mathbb{R}$-span is $\mathbb{R}^d$; see [5]. Likewise, a similar submodule (SSM) of M is a submodule of full rank that is similar to M. In particular, every similar submodule is of the form αRM, where $\alpha \in \mathbb{R}^* := \mathbb{R} \setminus \{0\}$ and $R \in \mathrm{O}(d, \mathbb{R})$. The key objects are the group of similarity isometries

$$\mathrm{OS}(M) = \{R \in \mathrm{O}(d, \mathbb{R}) \mid \exists \alpha \in \mathbb{R}^+ \text{ such that } \alpha RM \subseteq M\}$$

and the sets of scaling factors

$$\begin{aligned} \mathrm{Scal}_M(R) &:= \{\alpha \in \mathbb{R} \mid \alpha RM \subseteq M\} \quad \text{and} \\ \mathrm{scal}_M(R) &:= \{\alpha \in \mathbb{R} \mid \alpha RM \text{ is commesurate to } M\}, \end{aligned}$$

which are non-trivial if and only if R is a similarity isometry.

Peter Zeiner
Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor, Malaysia.
e-mail: pzeiner@xmu.edu.my

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_58

If E is the identity operation, then we have $\mathrm{Scal}_{\varGamma}(E) = \mathbb{Z}$ for a lattice $\varGamma$ and $\mathrm{scal}_{\varGamma}(E) \cup \{0\} = \mathbb{Q}$ is the corresponding field of fractions. In general, the set of 'trivial' scaling factors $\mathrm{Scal}_M(E)$ is an order in a real number field, whose rank satisfies certain restrictions [10, 5]. The set $\mathrm{scal}_M(E) \cup \{0\}$ is again the corresponding field of fractions.

The family of these sets, $\{\mathrm{scal}_M(R) : R \in \mathrm{OS}(M)\}$, has a natural group structure, which allows one to define a homomorphism [7, 8, 10, 5]

$$\begin{aligned} \phi : \mathrm{OS}(M) &\to \mathbb{R}/(\mathrm{scal}_M(E)), \\ R &\mapsto \mathrm{scal}_M(R). \end{aligned}$$

The kernel of this homomorphism is $\mathrm{OC}(M)$, the so-called group of coincidence isometries of M. The latter is defined as the group of all $R \in \mathrm{O}(d,\mathbb{R})$ such that M and RM are commensurate [10, 5]. This establishes a connection between similar submodules and coincidence site modules. In particular, $\mathrm{OS}(M)/\mathrm{OC}(M)$ is an Abelian group. If $M = \varGamma$ is a lattice, all elements of $\mathrm{OS}(M)/\mathrm{OC}(M)$ have a finite order which is a divisor of d. For instance, if M is the square lattice, then $\mathrm{OS}(M)/\mathrm{OC}(M)$ is an infinite 2-group [7]. In case of modules in general, this is not true any more, and $\mathrm{OS}(M)/\mathrm{OC}(M)$ may have factors isomorphic to $\mathbb{Z}$ [5].

Finding $\mathrm{OS}(M)$ and $\mathrm{Scal}_M(R)$ for all $R \in \mathrm{OS}(M)$ are typically the first steps if one wants to count the SSMs of a given module. If $b(n)$ denotes the number of SSMs of a given index n, then $b(n)$ is a supermultiplicative arithmetic function, that is, $b(mn) \geq b(m)b(n)$ whenever m and n are coprime. If the modules under consideration are related to number fields of class number one, this counting function is typically multiplicative and it makes sense to consider generating functions of Dirichlet series type

$$\varPhi(s) = \sum_{n \in \mathbb{N}} \frac{b(n)}{n^s}.$$

For many cases, these generating functions have been calculated explicitly, among others for certain planar modules of N-fold symmetry [1], and some root lattices and related modules up to dimension 4, see e.g. [3, 2]. These generating functions can be used to determine the asymptotic behaviour of the number of SSMs via Delange's theorem [9]. In particular, one can calculate the asymptotic behaviour of the summatory function $\sum_{n \leq x} b(n)$ by using some information on the poles of $\varPhi(s)$. For explicit calculations on the examples mentioned above (and many more), see [1, 3, 2, 5, 4] and references therein.

Acknowledgements This work has been supported by XMUM under grant no. XMUMRF/2019-C3/IMAT/0009.

References

1. Baake, M., Grimm, U.: Bravais colourings of planar modules with N-fold symmetry. Z. Krist. **219**, 72–80 (2004)
2. Baake, M., Heuer, M., Moody, R.V.: Similar sublattices of the root lattice A_4. J. Algebra **320**, 1391–1408 (2008)

3. Baake, M., Moody, R.V.: Similarity submodules and root systems in four dimensions. Canad. J. Math. **51**, 1258–1276 (1999)
4. Baake, M., Scharlau, R., Zeiner, P.: Similar sublattices of planar lattices. Canad. J. Math. **63**, 1220–1237 (2011)
5. Baake, M., Zeiner, P.: Geometric enumeration problems for lattices and embedded $\mathbb{Z}$-modules. In: Baake, M., Grimm, U. (eds.) Aperiodic Order. Vol. 2: Crystallography and Almost Periodicity, pp. 73–172. Cambridge University Press. Cambridge (2017)
6. Conway, J.H., Rains, E.M., Sloane, N.J.A.: On the existence of similar sublattices. Can. J. Math. **51**, 1300–1306 (1999)
7. Glied, S., Baake, M.: Similarity versus coincidence rotations of lattices. Z. Krist. **223**, 770–772 (2008)
8. Glied, S.: Similarity and coincidence isometries for modules. Can. Math. Bull. **55**, 98–107 (2011)
9. Tenenbaum, G.: Introduction to Analytic and Probabilistic Number Theory. Cambridge University Press, Cambridge (1995)
10. Zeiner, P.: Coincidence Site Lattices and Coincidence Site Modules. Habilitation thesis, Bielefeld University (2015)

Chapter 12

Ergodic Theory, Diophantine Approximation and Related Topics

A diffraction abstraction

Michael Coons

Abstract For some time now, I have been trying to understand the complexity of integer sequences from a variety of different viewpoints and, at least at some level, trying to reconcile these viewpoints. However vague that sounds—and it certainly is vague to me—in this short note, I hope to explain this sentiment.

1 Introduction

My interest in the complexity[1] of integer sequences is rooted in some classical results from the first part of the twentieth century concerning power series. These start with a result of Fatou [12], that a power series $F(z) \in \mathbb{C}[[z]]$ whose coefficients take only finitely many values is either rational or transcendental over $\mathbb{C}(z)$. Szegő [16] generalised Fatou's result to give, under the same assumptions, that $F(z)$ is either rational or has the unit circle as a natural boundary. Completing this picture in a certain sense, Carlson [9] then showed that if $F(z) \in \mathbb{Z}[[z]]$ converges in the unit disc, the same conclusion holds—either $F(z)$ is rational or it has the unit circle as a natural boundary. I use the word 'completing' as Carlson's theorem cannot be extended without adding more restrictive assumptions—there are irrational integer power series with a smaller radius of convergence that are meromorphic, such as the algebraic function

$$\frac{1}{\sqrt{1-4z}} = \sum_{n\geqslant 0} \binom{2n}{n} z^n.$$

Michael Coons
School of Mathematical and Physical Sciences, University of Newcastle, Australia
e-mail: michael.coons@newcastle.edu.au

[1] I mean complexity in the standard dictionary definition: the state or quality of being intricate or complicated.

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_59

Of course, one would like to know more about the behaviour of these power series as z approaches the unit circle. Towards addressing this, a beautiful result of Duffin and Schaeffer [11] states that a power series that is bounded in a sector of the unit disc and has coefficients from a finite set is necessarily a rational function. As well, this result cannot be extended to full generality—there are integer power series converging in the unit disc that are bounded in certain sectors, such as the series

$$\sum_{n\geqslant 0}(1-z)^n z^{n!},$$

which is bounded in the sector $\arg(z)\in[-\pi/4,\pi/4]$. One of the 'takeaways' for me from these results is the importance of asymptotics in relation to the complexity of integer sequences.

It is worth pointing out that these results occurred during an historically interesting time for integer sequences. Up to the year 1909, problems of probability were classified as either 'discontinuous' or 'continuous' (also called 'geometric'). Towards filling this gap, in that year, Borel [8] introduced what he called countable probabilities (probabités dénombrables). In this new type of problem, one asks probabilistic questions about countable sets. As a—now common—canonical example, Borel considered properties of the frequency of digits in the digital expansions of real numbers. A central concept in Borel's approach is that of normality. A real number x is called simply normal to the base k (or k-simply normal) if each of $0,1,\ldots,k-1$ occurs in the base-k expansion of x with equal frequency $1/k$. This number x is then called normal to the base k (or k-normal) provided it is k^m-simply normal for all positive integers m, and the number x is just called normal if this is true for all integers $k\geqslant 2$. Borel's use of the word 'normal' is well-justified; he showed, in that 1909 paper, that almost all real numbers, with respect to Lebesgue measure, are normal. The question he left was to determine if the decimal expansion of $\sqrt{2}$ is normal. It is now customary to attribute the following broader question to Borel: *Is the base expansion of an irrational algebraic real number normal?* It is not at all an exaggeration to say that nothing substantial is known now, 110 years later.

The strength of Borel's approach, as well as the difficulty, rests upon considering large blocks of a bounded integer sequence, the sequence of digits of a base expansion of a real number. But what if we relax this a bit and consider only the two-point correlations? This brings us squarely into the realm of diffraction.

In classical Fraunhofer (far-field) diffraction, monochromatic light waves from a (far) point source come into contact with an object, are scattered (diffracted) and then meet a (far) screen. The image left on the screen is (essentially) the Fourier transform of the object. The present situation concerns the specific case of a sequence of integers. For a bounded sequence w of integers, one arrives at the diagram

$$\omega:=\sum_{n\in\mathbb{Z}}w(n)\,\delta_n \xrightarrow{\circledast} \gamma_\omega:=\omega\circledast\omega=\sum_{m\in\mathbb{Z}}\eta(m)\,\delta_m \xrightarrow{\mathscr{F}} \widehat{\gamma_\omega}=\widehat{\omega\circledast\omega}$$

where ω is the (weighted) Dirac comb with weights w, $\circledast$ represents convolution, the values

$$\eta(m) := \lim_{N\to\infty} \frac{1}{2N+1} \sum_{i=-N}^{N} w(i)w(i+m)$$

are the autocorrelation coefficients and $\mathscr{F}$ is Fourier transformation. In the more general context, this diagram is commonly called a *Wiener diagram*, after the American applied mathematician Norbert Wiener, who pointed out the usefulness of the autocorrelation function for understanding X-ray diffraction patterns; see Senechal [15] and Patterson [14]. In fact, Wiener [17] instigated the use of diffraction methods on integer sequences. In his paper, "The spectrum of an array and its application to the study of the translational properties of a simple class of arithmetical functions, Part One," Wiener outlined a process whereby one uses the autocorrelation function to produce a spectral function that in a sense encodes some of the complexity of the underlying sequence. In modern day terms, he was showing, given a subset A of $\mathbb{Z}$, how to produce the diffraction measure $\widehat{\gamma}_\omega$ and then using the Lebesgue decomposition theorem to determine a sort of complexity for the set A. Recall that the Lebesgue decomposition theorem states that any regular Borel measure μ on $\mathbb{R}^d$ has a unique decomposition $\mu = \mu_{\mathrm{pp}} + \mu_{\mathrm{ac}} + \mu_{\mathrm{sc}}$ where μ_{pp}, μ_{ac} and μ_{sc} are mutually singular and also $|\mu| = |\mu_{\mathrm{pp}}| + |\mu_{\mathrm{ac}}| + |\mu_{\mathrm{sc}}|$. Here μ_{pp} is a pure point measure corresponding to the monotone step function part of Wiener's spectral function (the Bragg part), μ_{ac} is an absolutely continuous measure corresponding to the part of the spectral function that is the integral of its derivative, and μ_{sc} is a singular continuous measure corresponding to the continuous part of the spectral function which has almost everywhere a zero derivative. Wiener's purpose is exactly what I am aiming at, *"to extend the spectrum theory [...] to the harmonic analysis of functions only defined for a denumerable set of arguments—arrays, as we shall call them—and the application of this theory to the study of certain power series admitting the unit circle as an essential boundary."* [17]

In the remainder of this note, I will describe some examples of each (pure) type of measure with a number-theoretic flavour, then move on to an extended diffraction example before finishing our exposition with an example of a non-diffractive measure, which still gives some reasonable information, but for an unbounded sequence.

2 Three examples: diffraction measures of pure type

My current favourite three examples, illustrating each (pure) type of measure are the characteristic function on k-free integers, the Rudin–Shapiro sequence and the Thue–Morse sequence. Each of these sequences have power series generating functions having the unit circle as a natural boundary.

The k-free integers. Let $V_k \subset \mathbb{Z}$ be the set of k-free integers with fixed $k \geqslant 2$, that is, the elements of $\mathbb{Z}$ that are not divisible by a k-th power of any (rational) prime number. If one lets $w = \chi_k$ be the characteristic function on V_k and considers

$$\omega_k := \sum_{n\in\mathbb{Z}} \chi_k(n)\,\delta_n,$$

then a result of Baake, Moody and Pleasants [6] gives that the diffraction measure $\widehat{\gamma}_{\omega_k}$ is a pure point measure, which is explicitly computed in terms of elementary number-theoretic functions.

The Rudin–Shapiro sequence. In this example, one lets the sequence of weights w be the Rudin–Shapiro sequence $w_{\mathrm{RS}} : \mathbb{Z} \to \{\pm 1\}$ determined by the recurrences

$$w_{\mathrm{RS}}(4m+\ell) = \begin{cases} w_{\mathrm{RS}}(m), & \text{for } \ell \in \{0,1\}, \\ (-1)^{m+\ell} w_{\mathrm{RS}}(m), & \text{for } \ell \in \{2,3\}, \end{cases}$$

with initial conditions $w_{\mathrm{RS}}(0) = -w_{\mathrm{RS}}(-1) = 1$. Given this definition, using weights $w = w_{\mathrm{RS}}$, it turns out that the diffraction measure $\widehat{\gamma}_{\omega_{\mathrm{RS}}}$ is absolutely continuous with respect to Lebesgue measure; in fact, the two are equal. See Baake and Grimm [4, Section 10.2] for more details.

The Thue–Morse sequence. This example is a special case of a result of Kurt Mahler [13], who wrote "Part Two" of Wiener's above-mentioned paper [17].

Let $\{t(n)\}_{n\in\mathbb{Z}}$ be the Thue–Morse sequence defined on the alphabet $\{\pm 1\}$ by $t(0) = 1$, for $n \geqslant 1$ by the recurrences $t(2n) = t(n)$ and $t(2n+1) = -t(n)$ and extended to all of $\mathbb{Z}$ by the symmetric relation $t(-n) = t(n)$. The right half of this sequence, which starts

$$\{t(n)\}_{n\geqslant 0} = \{1,-1,-1,1,-1,1,1,-1,-1,1,1,-1,1,-1,-1,1,\ldots\},$$

is one of the most ubiquitous integer sequences and one of central importance in various areas within number theory, combinatorics, theoretical computer science and dynamical systems theory. In both theoretical computer science and dynamics one often views this sequence as the infinite iteration of the binary substitution (or morphism) ρ_{TM} defined on the two letter alphabet $\Sigma_2 := \{a,b\}$ by

$$\rho_{TM} : \begin{cases} a \mapsto ab \\ b \mapsto ba\,. \end{cases}$$

If one considers the Dirac comb

$$\omega_{\mathrm{TM}} = \sum_{n\in\mathbb{Z}} t(n)\,\delta_n,$$

then, as implied by the result of Mahler, the diffraction measure $\widehat{\gamma}_{\omega_{\mathrm{TM}}}$ is a purely singular continuous measure. Indeed, this was the first explicit example of such a measure, and appeared in Mahler's first published paper!

3 An extended diffraction example: the Thue–Morse sequence

Standing at the intersection of number theory, dynamics and theoretical computer science, the most widely interesting of these examples is that of the Thue–Morse sequence. Fortunately for us, it is also an example where much is known. In this section, I highlight a few of the known results concerning this sequence and provide some questions on further relationships.

Before continuing, we note that in this instance, the existence of the autocorrelation measure $\gamma_{\omega_{\mathrm{TM}}}$ is guaranteed by an application of Birkhoff's ergodic theorem; see Baake and Grimm [4, Section 10.1] for all details regarding measures associated to the Thue–Morse sequence. As well, the autocorrelation coefficients satisfy $\eta_{\mathrm{TM}}(-m) = \eta_{\mathrm{TM}}(m)$, so that one can determine the coefficients via the one-sided limit

$$\eta_{\mathrm{TM}}(m) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} t(i)t(i+m),$$

for all $m \in \mathbb{N}$. Using the recursions defining t, with some rearrangement, we arrive at the recursions

$$\eta_{\mathrm{TM}}(2m) = \eta_{\mathrm{TM}}(m) \quad \text{and} \quad \eta_{\mathrm{TM}}(2m+1) = -\frac{1}{2}\big(\eta_{\mathrm{TM}}(m) + \eta_{\mathrm{TM}}(m+1)\big),$$

for $m \geqslant 0$. Along with the fact that $\eta_{\mathrm{TM}}(0) = 1$, these recurrences specify a sequence $\{\eta_{\mathrm{TM}}(m)\}_{m\geqslant 0}$, which is 2-regular in the sense of Allouche and Shallit [2]. This presumably generalises, the moral result being that the autocorrelation coefficients of a k-automatic sequence should be k-regular. This added structure is useful and can be harnessed (as it is in the case for the Thue–Morse sequence) to help decide whether a given diffraction measure is continuous. See the monograph [1] for background and details on automatic sequences.

Since the Thue–Morse sequence is an automatic sequence, its generating function is a Mahler function. A Mahler function is a function $F(z) \in \mathbb{C}[[z]]$ for which there exist integers $d \geqslant 1$ and $k \geqslant 2$ and polynomials $p_0(z), \ldots, p_d(z)$ such that

$$p_0(z)F(z) + p_1(z)F(z^k) + \cdots + p_d(z)F(z^{k^d}) = 0.$$

That is to say, the function $F(z)$ behaves predictably under the map $z \mapsto z^k$. The generating function of the (one-sided) Thue–Morse sequence,

$$T_{\pm}(z) := \sum_{n\geqslant 0} t(n)z^n,$$

satisfies the Mahler-type functional equation

$$T_{\pm}(z) - (1-z)T_{\pm}(z^2) = 0.$$

The simplicity of this functional equation allows one to write $T_{\pm}(z)$ as the infinite product

$$T_{\pm}(z) = \prod_{j \geqslant 0} \left(1 - z^{2^j}\right).$$

It is evident by examining this product that as z radially approaches 1 from the origin, $T_{\pm}(z)$ is extremely flat. Indeed, de Bruijn [10] showed that as $z \to 1^-$, we have

$$T_{\pm}(z) = C_{\mathrm{TM}}(z) \cdot (1-z)^{1/2} \cdot 2^{-\log_2^2(1-z)/2} \cdot (1 + o(1)). \tag{1}$$

Here $\log_2^2(y) = (\log(y)/\log(2))^2$ is the square of the binary logarithm and $C_{\mathrm{TM}}(z)$ is a positive oscillatory term, which in $(0,1)$ is bounded away from 0 and infinity, is real-analytic, and satisfies $C_{\mathrm{TM}}(z) = C_{\mathrm{TM}}(z^2)$.

These asymptotics of $T_{\pm}(z)$ are reflected in the scaling behaviour of the distribution function of the Thue–Morse measure. Indeed, consider

$$\widehat{\gamma}_{\omega_{\mathrm{TM}}} = \mu_{\mathrm{TM}} * \delta_{\mathbb{Z}},$$

where

$$\mu_{\mathrm{TM}}(x) = \prod_{\ell \geqslant 0} \left(1 - \cos(2^{\ell+1}\pi x)\right),$$

and where this limit is taken in the vague topology. Setting $F_{\mathrm{TM}}(x) := \mu_{\mathrm{TM}}([0,x])$, Baake and Grimm [5] have shown that there are positive constants c_1 and c_2 such that for small x, we have

$$c_1 x^{2+\alpha} 2^{-\log_2^2(x)} \leqslant F_{\mathrm{TM}}(x) \leqslant c_2 x^{\alpha} 2^{-\log_2^2(x)}, \tag{2}$$

where $\alpha = -\log_2(\pi^2/2)$. Presumably inequality (2) holds with equal exponents of x on each side, but at the moment this remains an open question. A sort of heuristic for this is the validity of (1).

In this setting, one should view the function $T_{\pm}(z)$ as an exact "error term" in the following way. Let A_{01} be the set of nonnegative integers that have an odd number of ones in their binary expansion and denote by $T_{01}(z)$ the generating function of the characteristic function of A_{01}. Note that the set A_{01} has density $1/2$ in the nonnegative integers. With these definitions, we have

$$2\left(T_{01}(z) - \frac{1}{2} \cdot \frac{1}{1-z}\right) = T_{\pm}(z).$$

Now, de Bruijn's result gives,

$$\left(T_{01}(z) - \frac{1}{2} \cdot \frac{1}{1-z}\right)^2 = \tfrac{1}{4} \cdot C_{\mathrm{TM}}(z)^2 \cdot (1-z) \cdot 2^{-\log_2^2(1-z)} \cdot (1 + o(1)). \tag{3}$$

Equation (3) can be interpreted as a probabilistic (or statistical) statement about the set A_{01}.

The similarities between (2) and (3) are striking and though it is quite tempting, we refrain from making any direct conjectures, but ask the following question: *Is there a direct transformation (in general) between certain asymptotics of generat-*

ing functions and the asymptotic behaviour near zero of the associated distribution function of the diffraction measure?

4 A non-diffractive example: the Stern sequence

In the previous section, we considered an example of a bounded integer sequence, the Thue–Morse sequence. Because of this, one is able to use the setting of diffraction to consider complexity and to compare with asymptotics of the related generating function. In this section, I consider an example of an unbounded sequence, the Stern sequence. Due to this unboundedness, the traditional diffraction paradigm is not available.

Stern's sequence $\{s(n)\}_{n\geqslant 0}$, also called Stern's diatomic sequence, is defined by the initial conditions $s(0) = 0$ and $s(1) = 1$ and for $n \geqslant 1$ by the recurrences $s(2n) = s(n)$ and $s(2n+1) = s(n) + s(n+1)$. The sequence starts

$$\{s(n)\}_{n\geqslant 0} = \{0,1,1,2,1,3,2,3,1,4,3,5,2,5,3,4,1,5,4,7,3,8,5\ldots\}.$$

Stern's sequence has some interesting properties, maybe the most interesting of which is that the sequence $\{s(n)/s(n+1)\}_{n\geqslant 0}$ is an enumeration of the nonnegative rational numbers, without repeats, and already in reduced form! Like the Thue–Morse sequence, the generating function of the Stern sequence

$$S(z) := \sum_{n\geqslant 0} s(n+1)z^n$$

is a Mahler function given by an infinite product. In this case,

$$S(z) = \prod_{j\geqslant 0}\left(1 + z^{2^j} + z^{2\cdot 2^j}\right).$$

Due to the structure of this infinite product, one easily sees that the value $s(n+1)$ is the number of hyperbinary representations of n, that is, the number of ways of writing n as the sum of powers of two with each power being used at most twice.

As stated above, the unboundedness of the Stern sequence rules out the study of this sequence by means of traditional diffraction. Nonetheless, there is enough structure to form a measure associated to Stern's sequence. The following formulation follows my recent work [3] with Michael Baake. It was made possible due to the well-known relationship

$$\sum_{m=2^n}^{2^{n+1}-1} s(m) = 3^n, \tag{4}$$

for $n \geqslant 0$. This allowed us to define

$$\mu_n := 3^{-n} \sum_{m=0}^{2^n-1} s(2^n+m)\, \delta_{m/2^n}, \tag{5}$$

where δ_x denotes the unit Dirac measure at x. Here, we view $\{\mu_n\}_{n\geqslant 0}$ as a sequence of probability measures on the 1-torus—written as $\mathbb{T} = [0,1)$ with addition modulo 1—wherein we have re-interpreted the values of the Stern sequence in the interval $[2^n, 2^{n+1})$ as weights of a pure point probability measure on $\mathbb{T}$ with $\operatorname{supp}(\mu_n) = \left\{ \frac{m}{2^n} : 0 \leqslant m < 2^n \right\}$.

The main result of [3] is that the sequence $\{\mu_n\}_{n\geqslant 0}$ of probability measures on $\mathbb{T}$ converges weakly to a singular continuous probability measure μ_{S}, which we call the Stern measure. Moreover, one has $\mu_0 = \delta_0$ and $\mu_n = \mathop{\ast}_{m=1}^{n} \frac{1}{3}\left(\delta_0 + \delta_{2^{-m}} + \delta_{-2^{-m}}\right)$ for $n \geqslant 1$. The weak limit as $n \to \infty$ is given by the convergent infinite convolution product

$$\mu_{\mathrm{S}} = \mathop{\ast}_{m\geqslant 1} \tfrac{1}{3}\left(\delta_0 + \delta_{2^{-m}} + \delta_{-2^{-m}}\right).$$

Its Fourier transform $\widehat{\mu}_{\mathrm{S}}$ is given by

$$\widehat{\mu}_{\mathrm{S}}(k) = \prod_{m\geqslant 1} \tfrac{1}{3}\left(1 + 2\cos(2\pi k/2^m)\right) = \prod_{m\geqslant 1} \tfrac{1}{3}\left(1 + e^{2\pi k/2^m} + e^{-2\pi k/2^m}\right)$$

for $k \in \mathbb{Z}$. This infinite product is also well-defined on $\mathbb{R}$, where it converges compactly.

We also proved [3] that the distribution function $F_{\mathrm{S}}(x) := \mu_{\mathrm{S}}\left([0,x]\right)$ is strictly increasing and is Hölder continuous with exponent $\log_2(3/\tau)$, where $\tau := (1+\sqrt{5})/2$ is the golden mean. This implies that there is a positive constant c_3 such that

$$F_{\mathrm{S}}(x) \leqslant c_3 x^{\log_2(3/\tau)}.$$

Here, the comparison with known asymptotics is again striking. It follows from a result of mine with Bell [7], that as $z \to 1^-$,

$$S(z) = \frac{C_{\mathrm{S}}(z)}{(1-z)^{\log_2 3}} \cdot (1 + o(1)),$$

where, as in the case of the Thue–Morse sequence, $C_{\mathrm{S}}(z)$ is a positive oscillatory term, which in $(0,1)$ is bounded away from 0 and infinity, is real-analytic, and satisfies $C_{\mathrm{S}}(z) = C_{\mathrm{S}}(z^2)$. It is worth noting here, that while the constant 3 essentially comes from (4), the maximal values of the Stern sequence between 2^n and $2^{n+1}-1$ are proportional to τ^n, in fact, they are Fibonacci numbers. So here, the exponent in the scaling of the distribution function is the binary logarithm of the ratio of the average value $3/2$ and of the averaged maximum $\tau/2$.

5 Concluding remark

In this note (and the talks from whence it came), I discussed generating functions and measures associated to a few paradigmatic integer sequences. For the Thue–Morse sequence, I discussed the related diffraction measure and asked whether the asymptotics of the generating function near the unit circle are related to the scaling behaviour of the distribution function of the measure close to zero. Also, I used the example of the Stern sequence to define a measure (not a diffraction measure) for an unbounded integer sequence and again related properties of the distribution function of that measure to the asymptotics of the generating function of the Stern sequence near the unit circle. I find the similar structures of the asymptotics in these situations compelling and worthy of further study.

Acknowledgements Most of the results discussed in this work were joint with Michael Baake. I thank him for introducing me to the beautiful area of aperiodic order, a very enjoyable collaboration and, more locally, for his comments on this exposition.

References

1. J.-P. Allouche and J. Shallit, Automatic sequences, Cambridge University Press, Cambridge (2003).
2. J.-P. Allouche and J. Shallit, *The ring of k-regular sequences*, Theoret. Comput. Sci. **98** (1992), no. 2, 163–197.
3. M. Baake and M. Coons, *A natural probability measure derived from Stern's diatomic sequence*, Acta Arith. **183** (2018), no. 1, 87–99. arXiv:1706.00187
4. M. Baake and U. Grimm, Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013).
5. M. Baake and U. Grimm, *Scaling of diffraction intensities near the origin: some rigorous results*, J. Stat. Mech. (2019), 054003. arXiv:1905.04177
6. M. Baake, R. V. Moody, and P. A. B. Pleasants, *Diffraction from visible lattice points and kth power free integers*, Discrete Math. **221** (2000), no. 1–3, 3–42. arXiv:9906132
7. J. P. Bell and M. Coons, *Transcendence tests for Mahler functions*, Proc. Amer. Math. Soc. **145** (2017), no. 3, 1061–1070. arXiv:1511.07530
8. E. Borel, *Les probabilités dénombrables et leurs applications arithmétiques*, Palermo Rend. **27** (1909), 247–271.
9. F. Carlson, *Über ganzwertige Funktionen*, Math. Z. **11** (1921), no. 1–2, 1–23.
10. N. G. de Bruijn, *On Mahler's partition problem*, Nederl. Akad. Wetensch., Proc. **51** (1948), 659–669, Indagationes Math. **10**, 210–220 (1948).
11. R. J. Duffin and A. C. Schaeffer, *Power series with bounded coefficients*, Amer. J. Math. **67** (1945), 141–154.
12. P. Fatou, *Séries trigonométriques et séries de Taylor*, Acta Math. **30** (1906), no. 1, 335–400.
13. K. Mahler, *The spectrum of an array and its application to the study of the translation properties of a simple class of arithmetical functions. Part Two: On the translation properties of a simple class of arithmetical functions.*, J. Math. Phys., MIT **6** (1927), 158–163.
14. A. L. Patterson, *Experiences in crystallography–1924 to date*, in Fifty Years of X-Ray Diffraction, P. P. Ewald (Editor), Oosthoek, Utrecht, 1962, pp. 612–622.
15. M. Senechal, Quasicrystals and geometry, Cambridge University Press, Cambridge (1995).

16. G. Szegő, *Tschebyscheffsche Polynome und nichtfortsetzbare Potenzreihen*, Math. Ann. **87** (1922), no. 1–2, 90–111.
17. N. Wiener, *The spectrum of an array and its application to the study of the translation properties of a simple class of arithmetical functions. Part One: The spectrum of an array.*, J. Math. Phys., MIT **6** (1927), 145–157.

Chapter 13

Early Career Researchers Workshop on Geometric Analysis and PDEs

Extrinsic curvature flows and applications

Julian Scheuer

Abstract These notes arose from a mini lecture series the author gave at the Early Career Researchers Workshop on Geometric Analysis and PDEs, held in January 2020 at The Mathematical Research Institute MATRIX. We discussed some classical aspects of expanding curvature flows and obtained first applications. In these notes we will give a detailed account on what was covered during the lectures.

1 Introduction

Expanding curvature flows

This is an introduction to the theory of (expanding) extrinsic curvature flows, i.e. normal variations of hypersurfaces the speed of which are determined by the principal curvatures at each point. The flowing hypersurfaces are parametrized by a time-dependent family of embeddings

$$x\colon [0,T)\times\mathbb{S}^n\to\mathbb{R}^{n+1}$$

which satisfies

$$\dot{x}=\frac{1}{f(\kappa_1,\dots,\kappa_n)}\nu, \tag{1}$$

where

$$\kappa_1\le\cdots\le\kappa_n$$

Julian Scheuer
Columbia University New York/Cardiff University
e-mail: scheuerj@cardiff.ac.uk

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_60

are the principal curvatures at x, ν is the outward pointing unit normal and a dot denotes the partial time derivative.

Under a monotonicity assumption on f, this flow is a weakly parabolic system and we present proofs of the classical results due to Claus Gerhardt [2] and John Urbas [16]: Under certain assumptions on f and the initial embedding x_0 this flow exists for all times and after exponential blowdown converges to a round sphere. Furthermore we show that this flow can be used to prove so-called *Alexandrov-Fenchel inequalities*, which are inequalities between certain curvature functionals of a hypersurface. The approach is due to Pengfei Guan and Junfang Li [5]. Classical examples are the isoperimetric inequality and the Minkowski inequality

$$\int_M H \geq c_n |M|^{\frac{n-1}{n}},$$

which holds if M is mean-convex ($H > 0$) and starshaped. Here $|M|$ is the surface area of M. Equality holds precisely on every geodesic sphere. An appropriate rescaling of the flow (1) has nice monotonicity properties which, together with the convergence result, can be used to prove the inequalities. The approach we take slightly differs from the original works [2, 5]. Namely we use that the normal component of the rescaled flow actually moves by

$$\dot{x} = \left(\frac{1}{f} - \frac{u}{n}\right)\nu, \tag{2}$$

where u is the support function of the hypersurface. A priori estimates for (1) are directly deduced along this rescaling, which makes the estimates a little easier compared to Gerhardt's original arguments [2]. One interesting aspect of this particular rescaling is that (2) belongs to the class of so-called *locally constrained curvature flows*. The mean curvature type flow of this class,

$$\dot{x} = (n - uH)\nu,$$

was invented by Pengfei Guan and Junfang Li in [6] as a natural flow to prove the isoperimetric inequality in space forms: It preserves the enclosed volume and decreases surface area. A variety of such flows have appeared since then and they have been useful to obtain new geometric inequalities, cf. [7, 9, 13, 14, 17].

Outline

These notes are structured as follows. First we present some background on the curvature function f. It is known that the ordered principal curvatures are continuous in time, but if they have higher multiplicity they are in general not smooth. Hence at first sight the operator in (1) seems to lack regularity. However, this issue can be worked around by considering the function

$$F(A) = f \circ \mathrm{EV}(A),$$

where A is the Weingarten (or shape-) operator of the embedding and EV the eigenvalue map. Interestingly, even though EV is not smooth, if f is smooth and symmetric, F will be a smooth and natural map on the space of vector space endomorphisms. To people working with fully nonlinear curvature operators this is well known. We will give the precise setup to make this approach rigorous and state some important relations between derivatives of f and F, but skip most of the proofs in these notes. The material is taken from [12].

Afterwards we first fix some notation and conventions about hypersurface geometry and deduce the evolution equations for various geometric quantities. After these general considerations, we actually start with the a priori estimates for the inverse curvature flows and prove their convergence. We conclude by presenting the application to Alexandrov-Fenchel inequalities.

Up to some hard results from general parabolic PDE theory, i.e. short-time existence of fully nonlinear equations, Krylov-Safonov- and Schauder theory, the exposition should be mostly self-contained. However, on some occasions we will skip proofs for elementary statements.

2 Curvature functions

We quickly introduce the algebra of curvature functions using a new approach from [12]. Along a variation

$$\dot{x} = -f\nu$$

the function f is supposed to be a function of the principal curvatures of the flow hypersurfaces $M_t = x(t,M)$. As we deal with geometric flows, f has to be invariant under coordinate changes and thus we require it to be symmetric under all permutations. Hence we may assume the κ_i to be ordered,

$$\kappa_1 \leq \cdots \leq \kappa_n.$$

We assume that f is smooth. Along the curvature flows considered later, we derive estimates for the curvature and hence we would like to deduce a parabolic equation which is satisfied by the κ_i. However, those are in general not smooth functions, so we need to find another description of f, namely make it depend on the Weingarten operator A, the components of which are smooth.

This can be accomplished with the following idea: Suppose $\Gamma \subset \mathbb{R}^n$ is an open and symmetric set and

$$f \in C^\infty(\Gamma)$$

symmetric. It is a classical result [4] that f then is a function of the elementary symmetric polynomials

$$s_m(\kappa) := \sum_{1\leq i_1<\cdots<i_m\leq n} \prod_{j=1}^{m} \kappa_{i_j}, \tag{1}$$

or also of the power sums

$$p_m(\kappa) = \sum_{i=1}^{n} \kappa_i^m.$$

This means

$$f = \rho(s_1,\ldots,s_n) = \psi(p_1,\ldots,p_n)$$

for some smooth functions ρ and ψ. The crucial point is, that for the power sums it is very easy to make the transition from the dependence on the eigenvalues κ_i to dependence on the operator. This is formalized as follows:

Definition 1. Let V be an n-dimensional real vector space and $\mathscr{D}(V) \subset \mathscr{L}(V)$ be the set of real diagonalizable endomorphisms. Then we denote by EV the eigenvalue map, i.e.

$$\begin{aligned} \mathrm{EV}\colon \mathscr{D}(V) &\to \mathbb{R}^n/\mathscr{P}_n \\ A &\mapsto (\kappa_1,\ldots,\kappa_n), \end{aligned}$$

where $\kappa_1,\ldots,\kappa_n$ denote the eigenvalues of A and $\mathscr{P}_n$ is the permutation group of n elements.

For the power sums there is a very obvious candidate to serve as a function defined on linear maps, namely

$$P_k(A) = \mathrm{tr}(A^k).$$

Then there holds

$$P_k(A) = p_k(\mathrm{EV}(A)) \quad \forall A \in \mathscr{D}(V).$$

Now we can just insert the P_k into ψ, i.e. we define

$$F = \psi(P_1,\ldots,P_n).$$

Then $F \in C^\infty(\Omega)$ for some open set $\Omega \subset \mathscr{L}(V)$ and

$$F_{|\mathscr{D}_\Gamma(V)} = f \circ \mathrm{EV}_{|\mathscr{D}_\Gamma(V)},$$

where $\mathscr{D}_\Gamma(V)$ is the set of those real diagonalizable linear maps with eigenvalues in Γ. We obtain the following relations for the derivatives, see [12] for the details.

Denote by $F'(A)$ the gradient of F, i.e. by the relation

$$dF(A)B = \operatorname{tr}(F'(A) \circ B).$$

If A is real diagonalizable, then $F'(A)$ is real diagonalizable and if we denote by $F^i(A)$ its eigenvalues, then

$$F^i(A) = \frac{\partial f}{\partial \kappa_i}(\kappa),$$

where $\kappa = \mathrm{EV}(A)$. The second derivatives are related via

$$d^2F(A)(\eta,\eta) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial \kappa_i \partial \kappa_j} \eta_i^i \eta_j^j + \sum_{i \neq j}^{n} \frac{\frac{\partial f}{\partial \kappa_i} - \frac{\partial f}{\partial \kappa_j}}{\kappa_i - \kappa_j} \eta_j^i \eta_i^j,$$

where f is evaluated at the n-tuple (κ_i) of corresponding eigenvalues. The latter quotient is also well defined in case $\kappa_i = \kappa_j$ for some $i \neq j$. Here (η_j^i) is a matrix representation of some $\eta \in \mathscr{L}(V)$ with respect to a basis of eigenvectors of A.

Later we will require F to have certain properties, which we collect in the following definition.

Definition 2. The function F is called

(i) homogeneous of degree one, if Γ is a cone and

$$F(\lambda A) = \lambda F(A) \quad \forall \lambda > 0 \, \forall A \in \mathscr{D}_\Gamma(V),$$

(ii) strictly monotone, if

$$\mathrm{EV}(F'(A)) \in \Gamma_+ \quad \forall A \in \mathscr{D}_\Gamma(V),$$

(iii) concave, if

$$D^2F(A)(\eta,\eta) \leq 0$$

for all A and for all η which are jointly self-adjoint with A.

Here Γ_+ is the positive open cone on $\mathbb{R}^n$,

$$\Gamma_+ = \{\kappa \in \mathbb{R}^n : \kappa_i > 0 \quad \forall 1 \leq i \leq n\}.$$

Example 1. Important examples of functions f, such that F has the properties in the above definition, are the quotients

$$q_m = \frac{s_m}{s_{m-1}}$$

or the roots

$$\sigma_m = s_m^{\frac{1}{m}}.$$

In either case f has the mentioned properties in the cone

$$\Gamma_m = \{\kappa \in \mathbb{R}^n \colon s_k > 0 \quad \forall 1 \leq k \leq m\},$$

see for example [10]. Later we will use the quotients to deduce the Alexandrov-Fenchel inequalities.

3 Some hypersurface geometry

3.1 Conventions on Riemannian geometry

In this section we state the basic conventions concerning the elementary objects of Riemannian geometry. Let M be a smooth manifold of dimension n. For vector fields X, Y which are also derivations of $C^\infty(M)$, their Lie bracket is given by

$$[X,Y] = XY - YX$$

and for an endomorphism field A we denote by $\mathrm{tr}A \in C^\infty(M)$ its trace. Let g be a Riemannian metric on M with Levi-Civita connection ∇. The Riemannian curvature tensor is

$$\mathrm{Rm}(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,$$

and we also use Rm to denote the associated $(0,4)$-tensor,

$$\mathrm{Rm}(X,Y,Z,W) = g(\mathrm{Rm}(X,Y)Z,W).$$

The connection ∇ induces covariant derivatives of tensor fields T in the usual way via

$$\begin{aligned}
&\nabla T(X_1,\dots,X_l,Y^1,\dots,Y^k,X)\\
&= (\nabla_X T)(X_1,\dots,X_l,Y^1,\dots,Y^k)\\
&= X(T(X_1,\dots,X_l,Y^1,\dots,Y^k)) - T(\nabla_X X_1, X_2,\dots,X_l,Y^1,\dots,Y^k)\\
&\quad - \dots - T(X_1,\dots,X_l,Y^1,\dots,\nabla_X Y^k).
\end{aligned}$$

Let

$$x \colon M \to \mathbb{R}^{n+1}$$

be the smooth embedding of an n-dimensional manifold. The induced metric of $x(M)$ is given by the pullback of the ambient Euclidean metric $\langle \cdot, \cdot \rangle$,

$$g = x^* \langle \cdot, \cdot \rangle .$$

The second fundamental form h of the embedding x is given by the Gaussian formula

$$D_{x_*(X)} x_*(Y) = x_*(\nabla_X Y) - h(X,Y)\nu, \tag{1}$$

where D is the standard Euclidean connection. The Weingarten operator is defined via

$$g(A(X),Y) = h(X,Y)$$

and the Weingarten equation says that

$$D_{x_*(X)} \nu = x_*(A(X)). \tag{2}$$

Finally, we have the Gauss equation,

$$\mathrm{Rm}(W,X,Y,Z) = h(W,Z)h(X,Y) - h(W,Y)h(X,Z).$$

Remark 1. We will simplify the notation by using the following shortcuts occasionally:

(i) We will often omit x_*, i.e. when we insert a tangent vector field X into an ambient tensor field, we always understand X to be given by its pushforward.

(ii) When we deal with complicated evolution equations of tensors, we will occasionally use a local frame to express tensors with the help of their components, i.e. for a (k,l)-tensor field T, an expression like $T^{i_1 \dots i_k}_{j_1 \dots j_l}$ is understood to be

$$T^{i_1 \dots i_k}_{j_1 \dots j_l} = T(e_{j_1}, \dots, e_{j_l}, \varepsilon^{i_1}, \dots \varepsilon^{i_k}),$$

where (e_i) is a local frame and (ε^i) its dual coframe.

(iii) The coordinate expression for the m-th covariant derivative of a (k,l)-tensor field T is

$$\nabla^m T = \left(\nabla_{j_{l+m} \dots j_{l+1}} T^{i_1 \dots i_k}_{j_1 \dots j_l} \right),$$

where subscripts to ∇ represent the derivatives.

3.2 Hypersurfaces in polar coordinates

The punctured Euclidean space is isometric to

$$N = (0,\infty) \times \mathbb{S}^n, \quad \bar{g} = dr^2 + r^2 \sigma,$$

where σ is the round metric on $\mathbb{S}^n$ and $r = |x|$. We will deal with closed starshaped hypersurfaces, i.e. those which can be written as graphs over the fibre $\mathbb{S}^n$. We collect some useful formulae here.

Differentiating twice along M and using the Gaussian formula (1) gives

$$\frac{1}{2}\nabla^2|x|^2 = g - uh, \tag{3}$$

where u is the *support function*

$$u = \langle r\partial_r, \nu\rangle = \langle x, \nu\rangle .$$

The flow hypersurfaces we consider are graphs over $\mathbb{S}^n$, so let us recall some standard formulae, which can be found in [3, Sec. 1.5]. Let $M_0 = x(M) \subset \mathbb{R}^{n+1}$ be a graph over $\mathbb{S}^n$,

$$M_0 = \{(\rho(y), y) \colon y \in \mathbb{S}^n\} = \{(\rho(y(\xi)), y(\xi)) \colon \xi \in M\}.$$

Then the induced metric of M_0 is

$$g = d\rho \otimes d\rho + \rho^2\sigma.$$

We choose the normal ν to satisfy

$$\langle \nu, \partial_r\rangle > 0.$$

Let

$$\bar{h} = \rho\sigma$$

be the second fundamental form of the embedded slice $\{r = \rho\}$, then the second fundamental form of M_0 can be expressed with the help of the graph function,

$$uh = -\rho\nabla^2\rho + \rho\bar{h} = -\rho\nabla^2\rho + g - d\rho \otimes d\rho, \tag{4}$$

which is an easy exercise using the Gaussian formula and the Christoffel-symbols in polar coordinates. Also note that the principal curvatures $\bar{\kappa}$ of these slices are given by

$$\bar{\kappa} = \frac{1}{\rho}.$$

Formulae for hypersurface variations

As we consider time-dependent families of embedded hypersurfaces, we have to know how the previously discussed geometric quantities behave along variations

with arbitrary speed,

$$\dot{x} = -\mathscr{F}\nu,$$

where ν is the same normal as the one in the Gaussian formula (1).

Lemma 1. *Let $T > 0$, M^n a smooth orientable manifold and*

$$x\colon [0,T) \times M \to \mathbb{R}^{n+1}$$

be a normal variation with velocity $-\mathscr{F}$ of a smooth hypersurface $M_0 = x(0,M)$. Then the following evolution equations are satisfied.

(i) The induced metric g satisfies

$$\dot{g} = -2\mathscr{F}h. \tag{5}$$

(ii) The normal vector field satisfies

$$\frac{D}{dt}\nu = \operatorname{grad}\mathscr{F}, \tag{6}$$

where $\frac{D}{dt}$ is the covariant time derivative along the curve $x(\cdot,\xi)$ for fixed $\xi \in M$.

(iii) The Weingarten operator evolves by

$$\dot{A} = \nabla \operatorname{grad}\mathscr{F} + \mathscr{F}A^2. \tag{7}$$

Proof. Let X, Y be vector fields.

"(5)": Due to the Weingarten equation (2) we have

$$\dot{g}(X,Y) = \langle D_{\dot{x}}X, Y\rangle + \langle X, D_{\dot{x}}Y\rangle = -\mathscr{F}\langle D_X\nu, Y\rangle - \mathscr{F}\langle X, D_Y\nu\rangle = -2\mathscr{F}h(X,Y).$$

"(6)": We have

$$0 = \frac{\partial}{\partial t}\langle \nu, \nu\rangle = \left\langle \frac{D}{dt}\nu, \nu \right\rangle$$

and

$$\left\langle \frac{D}{dt}\nu, X \right\rangle = -\langle \nu, D_{\dot{x}}X\rangle = X\mathscr{F} = \langle \operatorname{grad}\mathscr{F}, X\rangle.$$

"(7)": Differentiate the Weingarten equation (2) with respect to time. The left hand side gives

$$D_{\dot{x}}D_X\nu = D_X D_{\dot{x}}\nu = \nabla_X \operatorname{grad}\mathscr{F} - h(X, \operatorname{grad}\mathscr{F})\nu,$$

where we have used (6). The right hand side gives

$$D_{\dot{x}}(A(X)) = D_{A(X)}\dot{x} + \dot{A}(X) = -h(X, \operatorname{grad} \mathscr{F})\nu - \mathscr{F}A^2(X) + \dot{A}(X).$$

Equate both sides to get the result.

4 Classical inverse curvature flows

We prove the classical result of Claus Gerhardt [2] and John Urbas [16], that the inverse curvature flow

$$\dot{x} = \frac{1}{F}\nu$$

in the Euclidean space $\mathbb{R}^{n+1}$, starting from starshaped and F-admissable[1] initial data converges to a round sphere after rescaling. Here is the result in detail.

Theorem 1 ([2, 16]). *Let $n \geq 2$ and $x_0 \in C^\infty(\mathbb{S}^n, \mathbb{R}^{n+1})$ be the embedding of a starshaped F-admissable hypersurface, where $F \in C^\infty(\Gamma) \cap C^0(\bar{\Gamma})$ is a positive, strictly monotone, 1-homogeneous and concave curvature function on a symmetric, open and convex cone Γ which contains $(1, \dots, 1)$. Suppose that*

$$F_{|\Gamma} > 0, \quad F_{|\partial\Gamma} = 0, \quad F(1, \dots, 1) = n.$$

Then the parabolic Cauchy-problem

$$\begin{aligned} \dot{x} &= \frac{1}{F}\nu \\ x(0, \cdot) &= x_0 \end{aligned}$$

has a unique solution $x \in C^\infty([0, \infty) \times \mathbb{S}^n, \mathbb{R}^{n+1})$. The rescaled hypersurfaces

$$\tilde{x}(t, \cdot) = e^{-\frac{t}{n}} x(t, \cdot)$$

converge smoothly to the embedding of a round sphere.

We use an approach slightly different from the original papers, namely we work directly on the rescalings. Note that $\tilde{x}$ will solve

$$\dot{\tilde{x}} = \frac{1}{F(e^{\frac{t}{n}}A)}\tilde{\nu} - \frac{1}{n}\tilde{x}. \tag{1}$$

As the Weingarten operator scales reciprocally to the hypersurfaces,

$$\tilde{A} = e^{\frac{t}{n}}A$$

[1] At every point the Weingarten operator is in the domain of definition

is the Weingarten operator of the rescaled surfaces

$$\tilde{M}_t = \tilde{x}(t, \mathbb{S}^n).$$

For technical reasons we only want to work with normal velocities, so we introduce a time-dependent family $y(t,\cdot) \in C^\infty(\mathbb{S}^n, \mathbb{S}^n)$ of diffeomorphisms in order to kill the tangent part in (1). We calculate

$$\frac{d}{dt}\tilde{x}(t, y(t,\cdot)) = \frac{1}{F(\tilde{A})}\tilde{\nu} - \frac{1}{n}\langle \tilde{x}, \tilde{\nu}\rangle \tilde{\nu} - \frac{1}{n}\left\langle \tilde{x}, \tilde{\nabla}_j \tilde{x}\right\rangle \tilde{\nabla}_i \tilde{x}\tilde{g}^{ij} + \tilde{\nabla}_i \tilde{x}\dot{y}^i.$$

Thus, if we solve the ODE system

$$\dot{y}^i = \frac{1}{n}\left\langle \tilde{x}, \tilde{\nabla}_j \tilde{x}\right\rangle \tilde{g}^{ij},$$

we see that $z(t) = \tilde{x}(t, y(t,\cdot))$ solves

$$\dot{z} = \left(\frac{1}{F(\tilde{A})} - \frac{1}{n}\tilde{u}\right)\tilde{\nu}, \tag{2}$$

where

$$\tilde{u} = \langle z, \tilde{\nu}\rangle$$

is positive due to the starshapedness of $\tilde{M}_t$. This formal discussion justifies that we as well may focus on the long-time existence and regularity for the flow (2). In order to facilitate notation, we will switch back to a more convenient notation and prove the following theorem, from which Theorem 1 then follows.

Theorem 2. *Let x_0 and F satisfy the assumption of Theorem 1. Then there exists a unique solution $x \in C^\infty([0,\infty) \times \mathbb{S}^n, \mathbb{R}^{n+1})$ of*

$$\begin{aligned} \dot{x} &= \left(\frac{1}{F(A)} - \frac{u}{n}\right)\nu \\ x(0,\cdot) &= x_0. \end{aligned} \tag{3}$$

The embeddings $x(t,\cdot)$ converge smoothly to the embedding of a round sphere.

Short time existence

To prove that the system (3) has a unique solution at least for a short time, we reduce it to a scalar parabolic equation and a system of ODEs. As we assume the initial hypersurface to be graphical over $\mathbb{S}^n$, if we already had a smooth solution for a while, the radial function would satisfy

$$\dot{\rho} = \frac{\langle x, \dot{x} \rangle}{|x|} = \left(\frac{1}{F} - \frac{u}{n} \right) \frac{u}{\rho}, \tag{4}$$

as can be seen by differentiation of $\rho = |x|$. From (4), [3, Equ. (2.4.21)] and [3, Lemma 2.7.6] we see that $\rho = \rho(t, x^i)$ would be the solution to the fully nonlinear equation

$$\begin{aligned} \partial_t \rho &= G(\bar{\nabla}^2 \rho, \bar{\nabla} \rho, \rho, \cdot) \\ \rho(0, \cdot) &= \rho_0, \end{aligned} \tag{5}$$

where ρ_0 is the radial function of the initial surface $M_0 = x(0, \mathbb{S}^n)$ and $\bar{\nabla}$ is the Levi-Civita connection of the round metric σ on $\mathbb{S}^n$. Also note that here (x^i) are the spherical coordinates of x in the polar coordinate system of the punctured Euclidean space. The idea is to solve this Cauchy-problem, which then determines the radial functions $\rho = \rho(t, x^i)$ of the flow hypersurfaces. Then we solve the following ODE initial value problem on $\mathbb{S}^n$:

$$\begin{aligned} \dot{x}^i &= \left(\frac{1}{F(A)} - \frac{u}{n} \right) \nu^i \\ x^i(0) &= x_0^i, \end{aligned}$$

where we note that the right hand side is fully determined by the function ρ and its derivatives, which itself solely depend on (x^i). Then we plug everything together and define

$$x(t, \xi) = (\rho(t, x^i(t, \xi)), x^i(t, \xi)),$$

which solves (3). In particular we note that *the maximal time of existence for* (3) *is entirely determined by the maximal time of existence for* (5).

It would miss the aim of this course to provide the rigorous argument behind this approach. The proof of existence for (5) uses solvability of linear parabolic equations in Hölder spaces and the implicit function theorem. In particular the maximal time of existence is controlled from below by estimates on the initial data. See [3, Sec. 2.5] for some more details. We have:

Theorem 3. *There exists $T^* \leq \infty$ and a unique maximal solution*

$$x \in C^\infty([0, T^*) \times \mathbb{S}^n, \mathbb{R}^{n+1})$$

to (3). *If $T^* < \infty$, then at T^* some derivative of x must blow up.*

Evolution equations

In order to prove the immortality of the maximal solution to (3), by Theorem 3 it suffices to prove uniform estimates on all derivatives of x. As those are controlled by derivatives of ρ, everything is reduced to prove regularity estimates for ρ.

The proof of these proceed by establishing estimates up to C^2-level as well as a lower F-bound by maximum principle, followed by regularity estimates for fully nonlinear parabolic operators due to Krylov and Safonov, as well as a bootstrapping argument using Schauder theory. We need further evolution equations, which are specifically adapted to the flow (3). We define the operator

$$\mathscr{L} = \partial_t - \frac{1}{F^2}\operatorname{tr}(F'(A)\circ(\nabla^2)^\sharp) - \frac{1}{n}\left\langle \rho\partial_r, \nabla^{(\cdot)}\right\rangle.$$

Lemma 2. *Along the flow* (3) *the radial function* $\rho = \rho(t,\xi)$ *satisfies*

$$\mathscr{L}\rho = \frac{2}{F}\frac{u}{\rho} - \frac{\rho}{n} - \frac{1}{\rho F^2}\operatorname{tr}F'(A) + \frac{1}{\rho F^2}\operatorname{tr}(F'\circ\nabla\rho\otimes(\nabla\rho)^\sharp),$$

while the support function u *satisfies*

$$\mathscr{L}u = \frac{1}{F^2}\left(\operatorname{tr}(F'(A)\circ A^2) - \frac{F^2}{n}\right)u.$$

Proof. (i) Use (4) to deduce

$$\operatorname{tr}(F'\circ(\nabla^2\rho)^\sharp) = \frac{1}{\rho}\operatorname{tr}F' - \frac{u}{\rho}F - \frac{1}{\rho}\operatorname{tr}(F'\circ\nabla\rho\otimes(\nabla\rho)^\sharp)$$

and hence, also using (4),

$$\mathscr{L}\rho = \left(\frac{2}{F} - \frac{u}{n}\right)\frac{u}{\rho} - \frac{\rho}{n}\left\langle\partial_r, \nabla^{(\cdot)}\rho\right\rangle - \frac{1}{\rho F^2}\operatorname{tr}F' + \frac{1}{\rho F^2}\operatorname{tr}(F'\circ\nabla\rho\otimes(\nabla\rho)^\sharp).$$

There holds

$$\begin{aligned}
-\frac{1}{n}\frac{u^2}{\rho} - \frac{\rho}{n}\left\langle\partial_r, \nabla^{(\cdot)}\rho\right\rangle &= -\frac{\rho}{n}\langle\partial_r,\nu\rangle^2 - \frac{\rho}{n}\left\langle\partial_r, x_*\nabla^{(\cdot)}\rho\right\rangle\\
&= -\frac{\rho}{n}\left(\langle\partial_r,\nu\rangle^2 + |\nabla\rho|^2\right)\\
&= -\frac{\rho}{n}\left(\langle\partial_r,\nu\rangle^2 + \sum_{i=1}^n\langle\partial_r,\nabla_i x\rangle^2\right)\\
&= -\frac{\rho}{n},
\end{aligned}$$

if coordinates are chosen such that $(\nu,\nabla_i x)$ is an orthonormal basis.

(ii) The position field $r\partial_r$ is a conformal vector field, hence for all vector fields $\bar{X}$ on $\mathbb{R}^{n+1}$ we have

$$D_{\bar{X}}(r\partial_r) = \bar{X}.$$

Hence, for vector fields X on M,

$$\dot{u} = \langle \dot{x}, \nu \rangle + \langle \rho\partial_r, D_{\dot{x}}\nu \rangle = \frac{1}{F} - \frac{u}{n} + \left\langle \rho\partial_r, \frac{\nabla F}{F^2} \right\rangle + \left\langle \rho\partial_r, \frac{\nabla u}{n} \right\rangle,$$

$$Xu = \langle \rho\partial_r, A(X) \rangle$$

and

$$\nabla^2 u(X,Y) = Y(Xu) - (\nabla_Y X)u = h(X,Y) - h(X,A(Y))u + \langle \rho\partial_r, \nabla_Y A(X) \rangle.$$

The result follows from combining these equalities, also using the Codazzi equation to cancel the ∇F-terms and the homogeneity of F which implies

$$\operatorname{tr}(F'(A) \circ A) = F.$$

We also need specific curvature evolution equations to estimate the principal curvatures and F from below.

Lemma 3. *The Weingarten operator satisfies*

$$\mathscr{L}A = \frac{1}{F^2}(F' \circ A^2)A - \frac{2A^2}{F} + \frac{A}{n} - \frac{2}{F^3}\nabla F \otimes (\nabla F)^\sharp + \frac{1}{F^2} d^2F(\nabla_{(\cdot)}A, \nabla^{(\cdot)}A),$$

while the curvature function F satisfies

$$\mathscr{L}F = -\frac{1}{F^2}\left(\operatorname{tr}(F'(A) \circ A^2) - \frac{F^2}{n}\right) F - \frac{2}{F^3}(F' \circ \nabla F \otimes (\nabla F)^\sharp).$$

Proof. (i) From (7) we calculate

$$\begin{aligned} \dot{A} &= \nabla \operatorname{grad}\left(\frac{u}{n} - \frac{1}{F}\right) + \left(\frac{u}{n} - \frac{1}{F}\right) A^2 \\ &= \frac{(\nabla^2 u)^\sharp}{n} + \frac{(\nabla^2 F)^\sharp}{F^2} - \frac{2}{F^3}\nabla F \otimes (\nabla F)^\sharp + \left(\frac{u}{n} - \frac{1}{F}\right) A^2 \\ &= \frac{A}{n} + \frac{1}{n}\left\langle \rho\partial_r, \nabla^{(\cdot)} A \right\rangle + \frac{(\nabla^2 F)^\sharp}{F^2} - \frac{2}{F^3}\nabla F \otimes (\nabla F)^\sharp - \frac{1}{F} A^2. \end{aligned} \tag{6}$$

We have to analyze the term $\nabla^2 F$ and do this is a local coordinate frame. There hold

$$\nabla_i F = dF(A)\nabla_i A$$

and

$$\nabla_{ji} F = d^2F(A)(\nabla_i A, \nabla_j A) + dF(A)\nabla_{ji} A.$$

We have to swap indices in $\nabla_{ij} A = \nabla_{ij} h^k_l$.

$$\begin{aligned}\nabla_{ji} h^k_l &= \nabla_{jl} h^k_i \\ &= \nabla_{lj} h^k_i + R_{jla}{}^k h^a_i - R_{jli}{}^a h_{ka} \\ &= \nabla^k_l h_{ij} + R_{jla}{}^k h^a_i - R_{jli}{}^a h_{ka} \\ &= \nabla^k_l h_{ij} + (h^k_j h_{la} - h_{ja} h^k_l) h^a_i - (h^a_j h_{li} - h^a_l h_{ij}) h_{ka}.\end{aligned}$$

Applying $dF = dF(A) = (F^l_k)$ to this, while using the 1-homogeneity and that $dF(A)$ commutes with A, gives

$$\begin{aligned}F^l_k \nabla_{ji} h^k_l &= F^l_k \nabla^k_l h_{ij} + F^l_k (h^k_j h_{la} - h_{ja} h^k_l) h^a_i - F^l_k (h^a_j h_{li} - h^a_l h_{ij}) h_{ka} \\ &= F^l_k \nabla^k_l h_{ij} - F^l_k h_{ja} h^k_l h^a_i + F^l_k h^a_l h_{ij} h_{ka} \\ &= F^l_k \nabla^k_l h_{ij} - F h_{ja} h^a_i + F^l_k h^a_l h_{ka} h_{ij}.\end{aligned}$$

Application of the sharp-operator gives

$$(\nabla^2 F)^\sharp = d^2F(\nabla_{(\cdot)} A, \nabla^{(\cdot)} A) + \operatorname{tr}(F' \circ (\nabla^2 A)^\sharp) - FA^2 + (F' \circ A^2)A.$$

Inserting this into (6) gives the first equation.

(ii) To get the equation for F calculate in local coordinates

$$\dot F = F^l_k \dot h^k_l$$

and use

$$F^l_k F^i_j \nabla^j_i h^k_l = F^i_j \nabla^j_i F - F^i_j d^2F(A)(\nabla_i A, \nabla^j A).$$

A priori estimates

The following estimates control the flow up to C^2-level for the function

$$\rho\colon [0, T^*) \times \mathbb{S}^n \to \mathbb{R}.$$

The following proof contains some common of tricks on how to estimate solutions to parabolic equations. It should be interesting even outside the world of curvature flows.

Lemma 4. *There exists a constant $c > 0$, which only depends on the initial hypersurface, such that*

(i)

$$\min_{\mathbb{S}^n} \rho(0,\cdot) \leq \rho \leq \max_{\mathbb{S}^n} \rho(0,\cdot), \tag{7}$$

(ii)

$$c^{-1} \leq u \leq c,$$

(iii)

$$c^{-1} \leq F \leq c,$$

(iv)

$$|A|^2 \leq c.$$

It follows that there exists a compact set $K \subset \Gamma$, in which the principal curvatures range during the whole evolution.

Proof. (i) Define

$$\tilde{\rho}(t) = \max_{\mathbb{S}^n} \rho(t,\cdot).$$

Then $\tilde{\rho}$ is Lipschitz and hence differentiable almost everywhere. It can be shown that at points of differentiability there holds

$$\frac{d}{dt}\tilde{\rho} = \dot{\rho}(t,\xi_t),$$

where ξ_t is a point where the maximum is attained. This technical argument is due to Hamilton [8]. From (4) we get (note $d\rho = 0$)

$$\dot{\tilde{\rho}} = \frac{1}{F} - \frac{\tilde{\rho}}{n}.$$

Now we recall that F depends on the second fundamental form which is related to ρ via (3), which gives

$$A = \frac{\mathrm{id}}{\rho} - (\nabla^2 \rho)^\sharp \geq \frac{\mathrm{id}}{\rho}$$

at ξ_t. Hence at ξ_t we have

$$F(A) \geq \frac{n}{\tilde{\rho}},$$

and hence $\tilde{\rho}$ is non-increasing. The same argument at minimal points gives that the minimum of ρ is non-decreasing, which concludes the argument.

(ii) In order to bound u^{-1}, we first note that uF is bounded from above and below, which can be seen as follows. Define

$$w = \log u + \log F,$$

then w satisfies

$$\mathscr{L} w = \frac{1}{u^2 F^2} \operatorname{tr}(F' \circ \nabla u \otimes (\nabla u)^\sharp) - \frac{1}{F^4} \operatorname{tr}(F' \circ \nabla F \otimes (\nabla F)^\sharp).$$

At critical points of w there holds

$$\frac{\nabla u}{u} = -\frac{\nabla F}{F}.$$

Hence, as above, the functions $\max w$ and $\min w$ are non-increasing/decreasing. We can use the boundedness of uF to prove that u^{-1} is bounded as well. This and the subsequent estimates all boil down to finding appropriate test functions.

The evolution equation of u^{-1} has one bad positive term, which prevents us from estimating it directly. Namely there holds

$$\mathscr{L} u^{-1} \leq \frac{u^{-1}}{n} - \frac{2}{u^3 F^2} \operatorname{tr}(F' \circ \nabla u \otimes (\nabla u)^\sharp).$$

However, we already have one bounded quantity, ρ, and we can use it to build test functions. Define

$$w = \log u^{-1} + \lambda \rho, \quad \lambda > 0.$$

There holds, due to $Fu \geq c > 0$,

$$\mathscr{L} w \leq \frac{1}{n} + \lambda \frac{c}{\rho} u^2 - \frac{\lambda \rho}{n} < 0$$

at all critical points of w where w is large enough, provided λ is chosen large enough. Hence w is bounded. In turn u^{-1} is bounded. The upper bound for u simply follows from

$$u \leq \rho.$$

(iii) Follows directly from the bounds on u and those on uF.

(iv) We use

$$\dot{g} = 2\left(\frac{1}{F} - \frac{u}{n}\right)h$$

to deduce

$$\mathscr{L}h = \mathscr{L}h_i^k g_{kj} + 2\left(\frac{1}{F} - \frac{u}{n}\right)h_{kj}h_i^k$$

and hence

$$\mathscr{L}h = \frac{F' \circ A^2}{F^2}h - \frac{2u}{n}h(A,\cdot) + \frac{h}{n} - \frac{2}{F^3}\nabla F \otimes \nabla F + \frac{1}{F^2}d^2F(\nabla_{(\cdot)}A, \nabla_{(\cdot)}A).$$

The only angry looking term in the evolution of h is the first one. As it also appears in the evolution of u, we cancel it with this one. Suppose the function

$$z = u^{-1}\kappa_n$$

attains a maximal value at a point (t_0, ξ_0). Let $\eta \in T_{\xi_0}\mathbb{S}^n$ be an eigenvector corresponding to κ_n and extend η locally to a vector field such that $\nabla\eta(t_0, \xi_0) = 0$. Define

$$w = \frac{h(\eta,\eta)}{g(\eta,\eta)}u^{-1}.$$

Then, locally around (t_0, ξ_0) there holds

$$w \leq z, \quad w(t_0, \xi_0) = z(t_0, \xi_0).$$

Hence w also attains a local maximum at this point and it suffices to locally estimate w. At (t_0, ξ_0) there holds

$$\begin{aligned}\mathscr{L}w &\leq -\frac{2}{n}\frac{h(A(\eta),\eta)}{g(\eta,\eta)} + \frac{2}{n}w - 2\frac{h(\eta,\eta)^2}{g(\eta,\eta)^2}\left(\frac{1}{Fu} - \frac{1}{n}\right)\\ &= \frac{2}{n}w - \frac{2u}{F}w^2 - \frac{2}{n}\frac{h(A(\eta),\eta)g(\eta,\eta) - h(\eta,\eta)^2}{g(\eta,\eta)^2}\\ &= \frac{2}{n}w - \frac{2u}{F}w^2 - \frac{2}{n}\frac{|A(\eta)|^2|\eta|^2 - g(A(\eta),\eta)^2}{|\eta|^2},\end{aligned}$$

which is negative for large w due to Bunjakowski-Cauchy-Schwarz and where we used the concavity of F. Hence w is bounded and thus all eigenvalues of A are bounded from above. As F is also bounded from below, we deduce from the concavity of F that

$$0 < F \leq H,$$

[3, Lemma 2.2.20]. Hence

$$\kappa_1 > (1-n)\kappa_n \geq -c$$

and we obtain $|A|^2 \leq c$. If there existed a sequence $\kappa(t_n, \xi_n) \in \Gamma$ that leaves every compact set of Γ, any subsequential limit of this sequence would lie on $\partial\Gamma$, which is impossible due to

$$F(t_n, \xi_n) \geq c^{-1}, \quad F_{|\partial\Gamma} = 0.$$

As a corollary we obtain full spatial C^2-estimates for the radial function ρ.

Corollary 1. *There exists a constant c, which only depends on the initial hypersurface, such that*

$$|\rho(t,\cdot)|_{C^2(\mathbb{S}^n)} \leq c \quad \forall t \in [0, T^*).$$

Proof. The C^0-bound of ρ follows from (7). As we are dealing with graphs over $\mathbb{S}^n$ in the product space

$$\mathbb{R}^{n+1}\backslash\{0\} = (0,\infty) \times \mathbb{S}^n, \quad \langle\cdot,\cdot\rangle = dr^2 + r^2\sigma,$$

the normal $\nu(\rho(t,\xi))$ is given by

$$\nu = \frac{(1, -\rho^{-2}\sigma^{ik}\partial_i\rho)}{\sqrt{1+\rho^{-2}|d\rho|^2_\sigma}}$$

and hence the support function is

$$u = \rho\,\langle\partial_r, \nu\rangle = \frac{\rho}{\sqrt{1+\rho^{-2}|d\rho|^2_\sigma}}.$$

As ρ and u are uniformly bounded, so is $|d\rho|_\sigma$, which gives the C^1-estimate. C^2-estimates follow from curvature estimates and the representation of the second fundamental form in terms of the second derivatives of ρ, (3).

The key for higher order estimates is a regularity result due to Krylov [11]. We state a very accessible formulation of this result as it can be found in a note by Ben Andrews [1, Thm. 4].

Theorem 4. *Let $\Omega \subset \mathbb{R}^n$ be open and suppose $\rho \in C^4((0,T] \times \Omega)$ satisfies*

$$\partial_t \rho = G(D^2\rho, D\rho, \rho, \cdot),$$

where G is concave in the first variable. Then for any $\tau > 0$ and $\Omega' \Subset \Omega$ there holds

$$\sup_{s,t\in[\tau,T],p,q\in\Omega'}\left(\frac{|D^2\rho(p,t)-D^2\rho(q,t)|}{|p-q|^\alpha+|s-t|^{\frac{\alpha}{2}}}+\frac{|\partial_t\rho(p,t)-\partial_t\rho(q,t)|}{|p-q|^\alpha+|s-t|^{\frac{\alpha}{2}}}\right)$$
$$+\sup_{s,t\in[\tau,T],p\in\Omega'}\frac{|D\rho(p,t)-D\rho(p,s)|}{|s-t|^{\frac{(1+\alpha)}{2}}}\leq C,$$

where α depends on n and the ellipticity constants λ,Λ of F', and C depends on n, λ,Λ, bounds for $|D^2\rho|$ and $|\partial_t\rho|$, $d(\Omega',\partial\Omega)$, τ and the bounds on the other first and second derivatives of G.

As ρ satisfies the fully nonlinear equation

$$\partial_t\rho=G(\bar{\nabla}^2\rho,\bar{\nabla}\rho,\rho,\cdot)=\left(\frac{1}{F(A)}-\frac{u}{n}\right)\frac{\rho}{u},$$

cf. [3, Equ. (2.4.21)], let us quickly check the assumptions of this theorem are satisfied. We use (4), [3, Lemma 2.7.6] and Lemma 4 to obtain the uniform ellipticity of

$$\frac{\partial G}{\partial\rho_{ij}}=-\frac{1}{F^2}dF\frac{\partial A}{\partial\rho_{ij}}$$

and the convexity of G in the first variable,

$$\frac{\partial^2 G}{\partial\rho_{ij}\partial\rho_{kl}}=\frac{2}{F^3}dF\frac{\partial A}{\partial\rho_{ij}}dF\frac{\partial A}{\partial\rho_{kl}}-\frac{1}{F^2}d^2F\left(\frac{\partial A}{\partial\rho_{ij}},\frac{\partial A}{\partial\rho_{kl}}\right)$$

where we used that A depends on $\bar{\nabla}^2\rho$ linearly. Hence theorem 4 does not apply directly, but we see that $-\rho$ satisfies an equation with a concave operator, to which we can apply the theorem. Hence ρ lies in the parabolic Hölder space $H^{2+\alpha,\frac{2+\alpha}{2}}([0,T]\times\mathbb{S}^n)$ for every $T<T^*$ with estimates independent of T. A standard bootstrapping argument using parabolic Schauder estimates implies uniform C^k-estimates of ρ for every k. It follows:

Corollary 2. *There exists a constant c, depending only on initial data and k, such that*

$$|x(t,\cdot)|_{C^k(\mathbb{S}^n)}\leq c\quad\forall 0\leq t<T^*.$$

The solution to (3) *is immortal.*

Proof. We have already seen the argument for the uniform estimates. The argument for immortality of the solution goes as follows. Suppose $T^*<\infty$. From Theorem 3 we know that the maximal time of existence can be estimated from below in terms of estimates for the initial data. As we have uniform estimates up to T^*, we may move as close to T^* as required to exceed T^* once we start with

$$\tilde{M}_0 = M_t, \quad T^* - \varepsilon < t < T^*,$$

where ε is chosen such that the flow with initial data $\tilde{M}_0$ exists longer than ε. Due to uniqueness we can extend our original flow and thus have shown that $T^* = \infty$.

Convergence to a round sphere

To conclude the proof of theorem 2, we have to show convergence of the embeddings $x(t,\cdot)$ to the embedding of a round sphere. We use the strong maximum principle.

Theorem 5. *The solution x to* (3) *limits to the embedding of a round sphere as* $t \to \infty$.

Proof. We have shown that the radial function ρ satisfies a uniformly parabolic equation. Hence its oscillation

$$\operatorname{osc}\rho(t) = \max_{\mathbb{S}^n}\rho(t,x(t,\cdot)) - \min_{\mathbb{S}^n}\rho(t,x(t,\cdot)) = \max_{\mathbb{S}^n}x^0(t,\cdot) - \min_{\mathbb{S}^n}x^0(t,\cdot)$$

is strictly decreasing, unless it is zero. Suppose it would not converge to zero as $t \to \infty$. Then it converges to some other value

$$\operatorname{osc}\rho(t) \to c_0 > 0, \quad t \to \infty.$$

Due to our uniform estimates, a diagonal argument and Arzela-Ascoli, the sequence of flows

$$x_k(t,\xi) := x(t+k,\xi)$$

subsequentially converges to a limit flow x_∞ with corresponding radial function ρ_∞. There holds

$$\operatorname{osc}\rho_\infty(t) = \lim_{k\to\infty}\operatorname{osc}\rho_k(t) = c_0.$$

Hence the strong maximum principle holding for ρ_∞ is violated if $c_0 > 0$. Thus $\operatorname{osc}\rho(t) \to 0$ as $t \to 0$ and hence the flow converges to a sphere centered at the origin.

5 Alexandrov-Fenchel inequalities

We use Theorem 2 to prove the classical Alexandrov-Fenchel inequalities for star-shaped hypersurfaces with $\sigma_k > 0$, cf. [5]. These are inequalities between so-called

higher order volumes. To motivate the terminology, let us consider a convex body, i.e. a compact convex K set with non-empty interior and its ε-*parallel body*

$$K_\varepsilon = \{x \in \mathbb{R}^{n+1} : \operatorname{dist}(K,x) \leq \varepsilon\}.$$

A classical result is *Steiner's formula*, which provides a Taylor expansion of the volume of K_ε:

$$\operatorname{vol}(K_\varepsilon) = \sum_{k=0}^{n} \binom{n+1}{k} W_k(K)\varepsilon^k \quad \forall \varepsilon \geq 0,$$

where the $W_k(K)$ are called the quermassintegrals of K, cf. [15]. Locally, such an expansion even holds for non-convex domains. In the following we prove this and a useful representation formula. First we need a general variational formula, where S_k is the operator function associated to the elementary symmetric polynomial s_k, see (1). We also define

$$s_0 := 1, \quad s_{-1} := u.$$

We will use the following facts about the S_k without proof:

$$dS_k A = kS_k \quad \forall 0 \leq k \leq n,$$

$$\operatorname{tr}(dS_{k+1}) = (n-k)S_k \quad \forall 0 \leq k \leq n-1,$$

$$dS_k A^2 = S_1 S_k - (k+1)S_{k+1} \quad \forall 0 \leq k \leq n-1.$$

Furthermore dS_k is divergence free. Hence we can deduce:

Lemma 5. *Let x and $\mathscr{F}$ as in Lemma 1 with M compact. For every $1 \leq k \leq n$ there holds*

$$\partial_t \int_{M_t} S_{k-1} = -k \int_{M_t} \mathscr{F} S_k.$$

For $k = 0$ there holds

$$\partial_t \int_{M_t} \langle x, \nu \rangle = -(n+1) \int_{M_t} \mathscr{F}.$$

Proof. For $k = 0$ we have

$$\partial_t \int_{M_t} \langle x, \nu \rangle = -\int_{M_t} \mathscr{F} + \int_{M_t} (\langle x, \operatorname{grad} \mathscr{F} \rangle - \mathscr{F} H \langle x, \nu \rangle)$$
$$= -\int_{M_t} \mathscr{F} + \int_{M_t} \operatorname{div}_{M_t} (\mathscr{F} x^\top) - n \int_{M_t} \mathscr{F}.$$

For $k = 1$ we have

$$\partial_t \operatorname{Area}(M_t) = -\int_{M_t} \mathscr{F} H = -\int_{M_t} \mathscr{F} S_1,$$

while for $2 \le k \le n$ we calculate using (7):

$$\partial_t \int_{M_t} S_{k-1} = -\int_{M_t} S_{k-1} \mathscr{F} S_1 + \int_{M_t} \operatorname{tr}(dS_{k-1} \circ \nabla \operatorname{grad} \mathscr{F}) + \int_{M_t} \mathscr{F} \operatorname{tr}(dS_{k-1} \circ A^2)$$
$$= \int_{M_t} \mathscr{F} (dS_{k-1} A^2 - S_1 S_{k-1})$$
$$= -k \int_{M_t} \mathscr{F} S_k.$$

Now we can prove a local Steiner's formula for C^2-domains.

Lemma 6. *Let $\Omega \subset \mathbb{R}^{n+1}$ be a bounded domain with C^2-boundary and let $\bar{\Omega}_\varepsilon$ be the ε-parallel body. Then there exists some $\varepsilon_0 > 0$ such that for all $0 \le \varepsilon < \varepsilon_0$ we have the expansion*

$$\operatorname{vol}(\bar{\Omega}_\varepsilon) = \sum_{k=0}^{n} \binom{n+1}{k} W_k(\Omega) \varepsilon^k, \tag{1}$$

where $W_0(\Omega) = \operatorname{vol}(\Omega)$ and

$$W_k(\Omega) = \frac{1}{(n+1)\binom{n}{k-1}} \int_{\partial\Omega} s_{k-1}(\kappa_i), \quad 1 \le k \le n+1.$$

Proof. There holds

$$\operatorname{vol}(\bar{\Omega}_\varepsilon) = \frac{1}{n+1} \int_{\Omega_\varepsilon} \operatorname{div} x = \frac{1}{n+1} \int_{\partial \bar{\Omega}_\varepsilon} \langle x, \nu_\varepsilon \rangle$$

and hence

$$W_0(\Omega) = \operatorname{vol}(\Omega) = \frac{1}{n+1} \int_{\partial\Omega} s_{-1}.$$

The parallel hypersurfaces $\partial\bar{\Omega}_\varepsilon$, which are C^2-hypersurfaces for small ε, can be seen as the flow hypersurfaces of the flow

$$\partial_\varepsilon x = \nu_\varepsilon.$$

According to Lemma 5 we obtain

$$\partial_\varepsilon \operatorname{vol}(\Omega_\varepsilon) = \operatorname{Area}(\partial\Omega_\varepsilon)$$

and

$$\partial_\varepsilon^k \operatorname{vol}(\Omega_\varepsilon) = \partial_\varepsilon^{k-1} \int_{\partial\Omega_\varepsilon} 1 = (k-1)! \int_{\partial\Omega_\varepsilon} S_{k-1} \quad \forall 1 \leq k \leq n+1.$$

For $k > n+1$ there holds

$$\partial_\varepsilon^k \operatorname{vol}(\Omega_\varepsilon) = 0$$

due to Gauss-Bonnet. Defining the W_k according to the Taylor expansion in (1) we see that they must have the form

$$W_k(\Omega) = \frac{1}{k!\binom{n+1}{k}} \partial_\varepsilon^k \operatorname{vol}(\bar{\Omega}_\varepsilon)_{|\varepsilon=0} = \frac{1}{k\binom{n+1}{k}} \int_{\partial\Omega} S_{k-1},$$

which is the claimed formula.

Hence the $W_k(\Omega)$ are nothing but coefficients of higher order in the Taylor expansion of volume with respect to fattening of the boundary. The isoperimetric inequality provides an estimate between W_0 and W_1 and hence it is natural to ask whether such an estimate also holds between the other higher order volumes. While for convex bodies such estimates have long been known, see for example [15] for a broad overview, here we want to use Theorem 2 to prove them for starshaped hypersurface with a certain curvature condition. This approach is due to Pengfei Guan and Junfang Li [5].

Definition 3. A domain $\Omega \subset \mathbb{R}^{n+1}$ is called *k-convex*, if throughout $\partial\Omega$ the principal curvatures lie in the closure of the cone

$$\Gamma_k = \{\kappa \in \mathbb{R}^n \colon s_m(\kappa) > 0 \quad \forall m \leq k\}.$$

Ω is called *strictly k-convex*, if the principal curvatures lie in Γ_k.

Theorem 6. *Let $\Omega \subset \mathbb{R}^{n+1}$ be a starshaped and k-convex domain, then there holds*

$$\frac{W_{k+1}(\Omega)}{W_{k+1}(B)} \geq \left(\frac{W_k(\Omega)}{W_k(B)}\right)^{\frac{n-k}{n+1-k}},$$

where B is the unit ball in $\mathbb{R}^{n+1}$. Equality holds precisely if Ω is a ball.

Proof. We use Theorem 2 with

$$F = n\frac{\binom{n}{k-1}}{\binom{n}{k}}\frac{s_k}{s_{k-1}} = \frac{nk}{n-k+1}\frac{s_k}{s_{k-1}}$$

and start the flow with $M_0 = \partial\Omega$. Due to the k-convexity of Ω the assumptions of this theorem are satisfied. We calculate that along the flow

$$\dot{x} = \left(\frac{1}{F} - \frac{u}{n}\right)\nu$$

there holds

$$\partial_t W_k(\Omega_t) = \frac{k}{(n+1)\binom{n}{k-1}}\int_{M_t}\left(\frac{n-k+1}{nk}\frac{s_{k-1}}{s_k} - \frac{u}{n}\right)s_k.$$

As the dS_k are divergence free, we obtain after tracing (3) with respect to dS_k and integration:

$$(n-k+1)\int_{M_t} s_{k-1} = k\int_{M_t} us_k$$

and thus

$$\partial_t W_k(\Omega_t) = 0.$$

On the other hand we obtain

$$\begin{aligned}\frac{n(n+1)\binom{n}{k}}{k+1}\partial_t W_{k+1}(\Omega_t) &= \int_{M_t}\left(\frac{n-k+1}{k}\frac{s_{k-1}s_{k+1}}{s_k} - us_{k+1}\right)\\ &= \int_{M_t}\left(\frac{n-k+1}{k}\frac{s_{k-1}s_{k+1}}{s_k} - \frac{n-k}{k+1}s_k\right)\\ &\leq 0,\end{aligned}$$

by the Newton-Maclaurin inequalities. Hence W_{k+1} is decreasing along the flow. We know the flow converges to a sphere, on which the desired inequality holds with equality. Hence on Ω the inequality is valid. The equality case follows, since the Newton-Maclaurin inequalities hold with equality precisely in umbilical points. This concludes the proof.

Acknowledgements These lectures were held at the *Early Career Researchers Workshop on Geometric Analysis and PDEs* at the MATRIX Institute in Creswick, Australia. I would like to thank the Matrix Institute and the organizers Paul Bryan, Jiakun Liu, Mariel Saéz and Haotian Wu for the opportunity to give these lectures.

References

1. Ben Andrews, *Fully nonlinear parabolic equations in two space variables*, arxiv:0402235, 2004.
2. Claus Gerhardt, *Flow of nonconvex hypersurfaces into spheres*, J. Differ. Geom. **32** (1990), no. 1, 299–314.
3. ———, *Curvature problems*, Series in Geometry and Topology, vol. 39, International Press of Boston Inc., Sommerville, 2006.
4. Georges Glaeser, *Fonctions composées différentiables*, Ann. Math. **77** (1963), no. 1, 193–209.
5. Pengfei Guan and Junfang Li, *The quermassintegral inequalities for k-convex starshaped domains*, Adv. Math. **221** (2009), no. 5, 1725–1732.
6. ———, *A mean curvature type flow in space forms*, Int. Math. Res. Not. **2015** (2015), no. 13, 4716–4740.
7. Pengfei Guan, Junfang Li, and Mu Tao Wang, *A volume preserving flow and the isoperimetric problem in warped product spaces*, Trans. Am. Math. Soc. **372** (2019), 2777–2798.
8. Richard Hamilton, *Four-manifolds with positive curvature operator*, J. Differ. Geom. **24** (1986), no. 2, 153–197.
9. Yingxiang Hu, Haizhong Li, and Yong Wei, *Locally constrained curvature flows and geometric inequalities in hyperbolic space*, arxiv:2002.10643, 2020.
10. Gerhard Huisken and Carlo Sinestrari, *Convexity estimates for mean curvature flow and singularities of mean convex surfaces*, Acta Math. **183** (1999), no. 1, 45–70.
11. Nicolai Krylov, *Nonlinear elliptic and parabolic equations of the second order*, Mathematics and its applications, vol. 7, Springer, 1987.
12. Julian Scheuer, *Isotropic functions revisited*, Arch. Math. **110** (2018), no. 6, 591–604.
13. Julian Scheuer, Guofang Wang, and Chao Xia, *Alexandrov-Fenchel inequalities for convex hypersurfaces with free boundary in a ball*, to appear in J. Differ. Geom., arxiv:1811.05776, 2018.
14. Julian Scheuer and Chao Xia, *Locally constrained inverse curvature flows*, Trans. Am. Math. Soc. **372** (2019), no. 10, 6771–6803.
15. Rolf Schneider, *Convex bodies: The Brunn-Minkowski theory*, 2. ed., Encyclopedia of Mathematics and its Applications, no. 151, Cambridge University Press, 2014.
16. John Urbas, *On the expansion of starshaped hypersurfaces by symmetric functions of their principal curvatures*, Math. Z. **205** (1990), no. 1, 355–372.
17. Guofang Wang and Chao Xia, *Guan-Li type mean curvature flow for free boundary hypersurfaces in a ball*, arxiv:1910.07253, 2019.

Short time existence for higher order curvature flows with and without boundary conditions

Yuhan Wu

Abstract We prove short time existence for higher order curvature flows of plane curves with and without generalised Neumann boundary condition.

1 Introduction

We first introduce the ideal curve flow of plane curves with Neumann boundary condition. This ideal curve flow is a sixth order curvature flow which is the steepest descent gradient flow of the energy functional in L^2. The Neumann boundary condition here is that there are two parallel lines which has a distance between them, the two end points of the curve we study are on these two lines respectively and the ideal curves are orthogonal to the boundaries. We then give the definition of generalized $(2m+4)$th-order curvature flows of plane curves with Neumann boundary condition. Secondly, we introduce the closed curve diffusion flow of plane curves with constrained length.

1.1 The curvature flows of open curves with Neumann boundary condition

Let η_1, η_2 denote two parallel vertical lines in $\mathbb{R}^2$, with distance between them. The immersed curves $\gamma : [-1,1] \times [0,T) \to \mathbb{R}^2$ satisfying Neumann boundary condition.

$$\gamma(-1) \in \eta_1(\mathbb{R}), \gamma(1) \in \eta_2(\mathbb{R}).$$

Yuhan Wu
University of Wollongong, Australia, e-mail: yw120@uowmail.edu.au

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_61

Denote $\tau = \gamma_s$ is the unit tangent vector field and ν the unit normal vector along γ. The Neumann condition is equivalent to $\langle \nu(\pm 1,t), \nu_{\eta_{1,2}} \rangle = 0$, here $\nu_{\eta_{1,2}}$ is the unit normal vector field to $\eta_{1,2}$. See Figure 1.1.

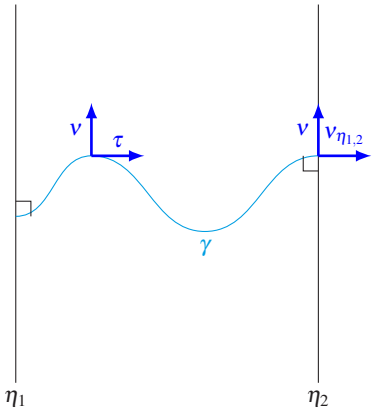


Figure 1.1

1.1.1 A sixth order flow of plane curves with Neumann boundary condition

We consider the energy functional

$$E(\gamma) = \frac{1}{2}\int_\gamma k_s^2 ds,$$

where k is the scalar curvature, ds the arclength element and k_s is the derivative of curvature with respect to arclength s. The corresponding gradient flow has normal speed given by F, that is

$$\partial_t \gamma = F\nu.$$

Under the evolution of the functional $E(\gamma)$ a straightforward calculation yields

$$\frac{d}{dt}\frac{1}{2}\int_\gamma k_s^2 ds = -\int_\gamma F \cdot \left(k_{s^4} + k^2 k_{ss} - \frac{1}{2}k k_s^2\right) ds,$$

where $k_{s^4} = k_{ssss}$. For the flow to be the steepest descent gradient flow of $E(\gamma)$ in L^2, we require

$$F = k_{s^4} + k^2 k_{ss} - \frac{1}{2}k k_s^2. \tag{1}$$

Let γ be a smooth curve satisfying $F = 0$, that is a stationary solution to the L^2-gradient flow of E. We call such curves *ideal*.

We define sixth order curvature flow with Neumann boundary condition as follows, see more details in [3].

Definition 1. [3] Let $\gamma : [-1,1] \times [0,T) \to \mathbb{R}^2$ be a family of smooth immersed curves. γ is said to move under sixth order curvature flow (1) with homogeneous Neumann boundary condition, if

$$\begin{cases} \frac{\partial}{\partial t}\gamma(s,t) = F\nu, & \forall (s,t) \in [-1,1] \times [0,T) \\ \gamma(\cdot,0) = \gamma_0, & \\ \langle \nu, \nu_{\eta_{1,2}} \rangle = k_s = k_{s^3} = 0, & \forall (s,t) \in \eta_{1,2}(\mathbb{R}) \times [0,T) \end{cases} \tag{2}$$

where $F = k_{s^4} + k_{ss}k^2 - \frac{1}{2}k_s^2 k$ denotes the normal speed of the curves, ν and $\nu_{\eta_{1,2}}$ are the unit normal fields to γ and $\eta_{1,2}$ respectively.

Here we give the long-time existence result as:

Theorem 1. *[3] Let γ_0 be a smooth embedded regular curve. Let $\gamma : [-1,1] \times [0,T) \to \mathbb{R}^n$ be a solution to* (2)*. If the initial curve γ_0 satisfies $\omega = 0$ and*

$$\|\kappa_s\|_2^2 \leq \frac{\pi^3}{7L_0^3},$$

here L_0 is the length of γ_0 and $\kappa = k(\cdot,0)$ is the curvature, then the flow exists for all time $T = \infty$ and $\gamma(\cdot,t)$ converges exponentially to a horizontal line segment γ_∞ in the C^∞ topology.

We use w to denote the winding number, defined here as

$$w := \frac{1}{2\pi}\int_\gamma k ds.$$

For closed curves, $w \in \mathbb{Z}$, in our setting, the winding number must be a multiple of $\frac{1}{2}$. For example in Figure 1.1.1,

$$\gamma_1 \text{ has } w[\gamma_1] = \tfrac{1}{2\pi}\textstyle\int_{\gamma_1} k ds = 1;$$
$$\gamma_2 \text{ has } w[\gamma_2] = \tfrac{1}{2\pi}\textstyle\int_{\gamma_2} k ds = \tfrac{1}{2}.$$

Lemma 1. *The hypothesis of theorem 1 implies that $\omega[\gamma] = \omega[\gamma_0] = 0$.*

1.1.2 Higher order flows of plane curves with Neumann boundary condition

We generalise the sixth order case where we considered the L^2-gradient flow for the energy

$$\frac{1}{2}\int_\gamma k_s^2 ds.$$

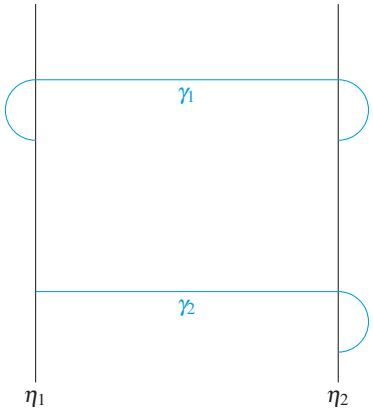


Figure 1.1.1

Our work is also the arbitrary even order generalisation of [11], where the fourth order curve diffusion and elastic flow of curves between parallel lines are investigated.

We consider the L^2-gradient flow for the energy

$$E(\gamma) = \frac{1}{2}\int_\gamma k_{s^m}^2 ds$$

with suitable associated generalised Neumann boundary conditions, here $m \in \mathbb{N} \cup \{0\}$.

Under a normal variation of the energy, straightforward calculations yield the normal flow speed

$$F = (-1)^{m+1} k_{s^{2m+2}} - \sum_{j=1}^{m} (-1)^j k k_{s^{m+j}} k_{s^{m-j}} - \frac{1}{2} k k_{s^m}^2. \tag{3}$$

And we set the Neumann boundary condition as:

$$< \nu, \nu_{\eta_{1,2}} > (\pm 1, t) = k_s(\pm 1, t) = ... = k_{s^{2m-1}}(\pm 1, t) = k_{s^{2m+1}}(\pm 1, t) = 0.$$

We define (2m+4)th order curvature flow with Neumann boundary condition in Definition 2, see more details in [5].

Definition 2. [5] Let $\gamma : [-1,1] \times [0,T] \to \mathbb{R}^2$ be a family of smooth immersed curves. γ is said to move under (2m+4)th-order curvature flow (3) with homogeneous Neumann boundary condition, if

$$\begin{cases} \frac{\partial}{\partial t}\gamma(s,t) = -F\nu, & \forall (s,t) \in [-1,1] \times [0,T) \\ \gamma(\cdot,0) = \gamma_0, & \\ < \nu, \nu_{\eta_{1,2}} >= k_s = ... = k_{s^{2m-1}} = k_{s^{2m+1}} = 0, & \forall (s,t) \in \eta_{1,2}(\mathbb{R}) \times [0,T) \end{cases} \tag{4}$$

where $F=(-1)^{m+1}k_{s^{2m+2}}+\sum_{j=1}^{m}(-1)^{j+1}kk_{s^{m+j}}k_{s^{m-j}}-\frac{1}{2}kk_{s^m}^2$ denotes normal speed of the curves, $m\in\mathbb{N}\cup\{0\}$, ν and $\nu_{\eta_{1,2}}$ are the unit normal fields to $\gamma(\pm1)$ and $\eta_{1,2}$ respectively.

We are also interested in one-parameter families of curves $\gamma(\cdot,t)$ satisfying the polyharmonic curvature flow

$$\frac{\partial}{\partial t}\gamma(s,t)=(-1)^{m+1}k_{s^{2m+2}}\nu, \tag{5}$$

here general $m\in\mathbb{N}\cup\{0\}$. Above ν is the smooth choice of unit normal such that the above flow is parabolic in the generalised sense.

Lemma 2. *While a solution to the flow* (5) *with generalised Neumann boundary conditions exists, we have*

$$\frac{d}{dt}L(t)=-\int_{\gamma}k_{s^{m+1}}^2\,ds,$$

where $L(t)$ denotes the length of the curve.

In view of this lemma and the separation of the supporting parallel lines $\eta_{1,2}$, the length $L(t)$ of the evolving curve $\gamma(\cdot,t)$ remains bounded above and below under the flow (5).

1.2 The length-constrained curve diffusion flow of closed curves

We consider one-parameter families of immersed closed curves $\gamma:\mathbb{S}^1\times[0,T)\to\mathbb{R}^2$. The energy functional

$$L(\gamma)=\int_{\gamma}|\gamma_u|\,du.$$

The curve diffusion flow is the steepest descent gradient flow for length in H^{-1}. We define the constrained curve diffusion flow here, see more details about this flow in [4].

Definition 3. [4] Let $\gamma:\mathbb{S}^1\times[0,T)\to\mathbb{R}^2$ be a $C^{4,\alpha}$-regular immersed curve. The length constrained curve diffusion flow

$$\begin{cases}\partial_t\gamma=-(k_{ss}-h(t))\nu,\ \forall(s,t)\in\mathbb{S}^1\times[0,T)\\ \qquad\gamma|_{t=0}=\gamma_0,\end{cases} \tag{6}$$

where ν denotes a unit normal vector field on γ.

To preserve length of the evolving curve $\gamma(\cdot,t)$, we take

$$h(t) = -\frac{\int k_s^2 ds}{2\pi w}.$$

Length-constrained curve diffusion flow fixes length and increases area. Regular curve diffusion flow fixes area and reduces length. We can say that the length-constrained curve diffusion flow is "dual" to curve diffusion flow.

The following theorem is the long time existence result for the length-constrained curve diffusion flow.

Theorem 2. *[4] Suppose γ_0: $\mathbb{S}^1 \to \mathbb{R}^2$ is a regular smooth immersed closed curve with $A[\gamma_0] > 0$ and $w[\gamma_0] = 1$. Then there exists a constant $K^* > 0$ such that if*

$$K_{osc}[\gamma_0] < K^*, I[\gamma_0] < \frac{4\pi^2}{4\pi^2 - K^*},$$

then the length-constrained curve diffusion flow γ with initial data γ_0 exists for all time and converges exponentially to a round circle with radius $\frac{L_0}{2\pi}$.

In Theorem 2, $A[\gamma]$ denotes the area, $K_{osc}[\gamma]$ is the oscillation and $I[\gamma]$ is the isoperimetric of the flow. From our calculation, we know that $K^* \simeq \frac{1}{9}$.

In the setting for this flow, the winding number must be an integer and is always 1 under the assumption in Theorem 2. For example in Figure 1.2,

$$\gamma_1 \text{ has } w[\gamma_1] = \tfrac{1}{2\pi}\int_{\gamma_1} k ds = 0;$$
$$\gamma_2 \text{ has } w[\gamma_2] = \tfrac{1}{2\pi}\int_{\gamma_2} k ds = 1;$$
$$\gamma_3 \text{ has } w[\gamma_3] = \tfrac{1}{2\pi}\int_{\gamma_3} k ds = 2.$$

Lemma 3. *The hypothesis of theorem 2 implies that $\omega[\gamma] = \omega[\gamma_0] = 1$.*

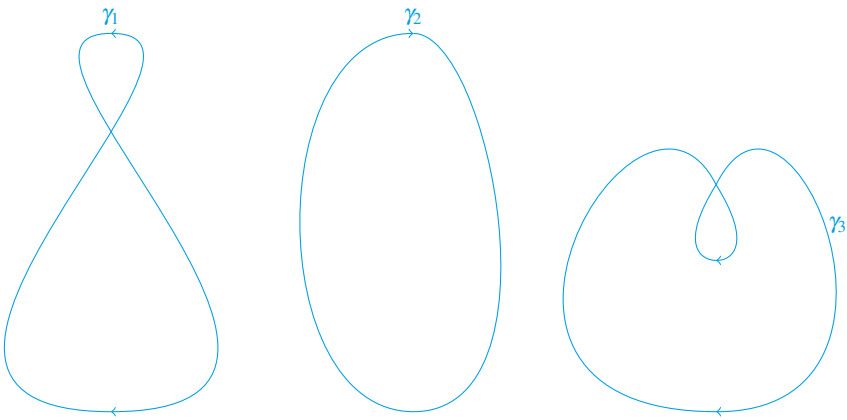


Figure 1.2

2 Short time existence for higher order curvature flows with Neumann boundary condition

Here we state the way to prove the short time existence for higher order curvature flows of plane curves with generalised Neumann boundary condition. The first step is to convert the weakly parabolic system (2) together with boundary conditions to a corresponding nonlinear scalar parabolic equation. This involves fixing a graphical parametrisation over a reference curve. The reference curve here is a straight line segment. The conversion process using generalised Gaussian coordinates in the case with boundary conditions is described for example in Section 2 of [7] (The case of higher codimension is covered in [9]). The second step is for the scalar parabolic equation with boundary conditions, we consider the corresponding linearized equation, for which existence of a unique (smooth) solution is well-known. By using the solution existence of the linearized problem together with the general result on the nonlinear evolutionary boundary value problems (for example, Theorem 4.4 in [6]) to see that the scalar graph equation has a unique solution at least for a short time. We then prove scalar graph equation is equivalent to the flow system (2), thus a solution to (2) exists for a short time. The solution to (2) is necessarily not unique due to the possibility of choosing different parametrisations, however the image curve is unique. This method also works for the generalised case (4).

2.1 A sixth order flow of plane curves with boundary condition

Let $l([-1,1])$ be a straight line segment which is perpendicular to boundaries η_1, η_2. Define the flux lines $\Phi = \Phi(u,\cdot)$ to $l([-1,1])$ are perpendicular to $l([-1,1])$ and tangential to η_1, η_2. Define a neighbourhood $\mathscr{U} \subset \mathbb{R}^2$ of $l([-1,1])$, $\mathscr{U}_\varepsilon := \{\Phi(u,x) : u \in [-1,1], |x| < \varepsilon\}$. In $\mathscr{U}$, let $\rho(p_0)$ denote tangential coordinate of p_0 on $l([-1,1])$, we can define a smooth normal vector field ξ with the following properties:

$$\langle \xi, \rho \rangle|_{l([-1,1])} = 0, \quad \xi|_{\eta_{1,2} \cap \mathscr{U}_\varepsilon} \in T\eta_{1,2}, \quad \|\xi\| = 1,$$

where $\eta_{1,2} \cap \mathscr{U}_\varepsilon = \{p \in \mathbb{R}^2 : p = \Phi(u,x), u \in \eta_{1,2}, x \in (-\varepsilon, \varepsilon)\}$.

Hence for any given point $p = \Phi(u,x)$, we can define $x(p)$ is the length of the flux line through p between p and intersection point $p_0 = \Phi(u,0)$ on $l([-1,1])$. We define $M = \{p \in \mathbb{R}^2 : p = \Phi(u, w(u,t)), u \in [-1,1]\}$, here $w(u,t) : [-1,1] \times [0,\sigma] \to \mathbb{R}$ and $\sigma \in [0,T)$.

We transform the problem (2) to a scalar initial-boundary-value graph problem as follows,

$$\begin{cases} \frac{\partial w}{\partial t}(u,t) = f(u,t), & \forall (u,t) \in [-1,1] \times [0,\sigma] \\ w(\cdot,0) = w_0, & \\ w_u = w_{u^3} = w_{u^5} = 0, & \forall (u,t) \in \eta_{1,2} \times [0,\sigma] \end{cases} \tag{7}$$

where $f(u,t) = v^{-6}w_{u^6} + g(w_u, w_{uu}, w_{u^3}, w_{u^4}, w_{u^5})$, $v(u,t) = |\gamma_u|$ and g is a function depending only on $w_u, w_{uu}, w_{u^3}, w_{u^4}, w_{u^5}$.

Next let $\bar{\gamma}(\phi(u,t),t) = (\phi(u,t), w(\phi(u,t),t))$ and define $\phi : [-1,1] \times [0,\sigma] \to [-1,1]$ by the following system of ordinary differential equation:

$$\begin{cases} \frac{d}{dt}\phi(u,t) = -(\bar{\gamma}_u)^{-1} \cdot \left(\frac{\partial}{\partial t}\bar{\gamma}\right)^T (\phi(u,t),t) \\ \phi(u,0) = u, \end{cases}$$

where $\alpha^T := \alpha - \langle \alpha, \bar{\nu} \rangle \cdot \bar{\nu}$ denotes the tangential component of a vector α, $\bar{\nu}(\phi(u,t),t) = v^{-1} \cdot (-w_\phi, 1)$ denotes normal vector field, $w_\phi(\phi(u,t),t) = \frac{\partial w}{\partial \phi}(\phi(u,t),t)$.

At least for a short time, ϕ is a diffeomorphism on $[-1,1]$, it's equivalent to that ϕ is tangential to the boundaries $\eta_{1,2}$, i.e. $u \in \eta_{1,2} \Longrightarrow \phi(u,t) \in \eta_{1,2}, \forall t \in [0,\sigma]$.

We prove the original problem (2) and the scalar graph problem (7) are equivalent under tangential diffeomorphism. See [8] for the proof.

Lemma 4. *The boundary conditions in* (7) *satisfy the compatibility condition,* $\forall (u,t) \in \eta_{1,2} \times [0,\sigma]$*, we have*

$$\partial_t^j w_u\Big|_{t=0} = \partial_t^j w_{u^3}\Big|_{t=0} = \partial_t^j w_{u^5}\Big|_{t=0} = 0, \quad j = 0,1,2,...,n.$$

Lemma 5. *The boundary conditions in* (7) *satisfy the normal boundary conditions.*

For the definition of the normal boundary condition, see [6].

Now we do the linearization at any $a \in \{w : [-1,1] \times [0,\sigma] \to \mathbb{R}\}$ for nonlinear problem (7). The linear problem of (7) can be written as:

$$\begin{cases} \frac{\partial w}{\partial t}(u,t) = f_a(a)w(u,t), \ \forall (u,t) \in [-1,1] \times [0,\sigma] \\ w(\cdot,0) = w_0, \\ w_u = w_{u^3} = w_{u^5} = 0, \quad \forall (u,t) \in \eta_{1,2} \times [0,\sigma] \end{cases} \tag{8}$$

where $f_a(a)w(u,t) = -v^{-6}(a) \cdot w_{u^6} + g_n w_{u^n}$, g_n are depends only on $a, a_u, ..., a_{u^{7-n}}$ and are all smooth in space and time, $n = 1,2,...,5$.

Proposition 1. *There is always a unique solution for the linear problem (8).*

The proof for Proposition 1 refers to classical results on linear parabolic boundary value problem (for example [2], Ch IV, 6.4).

Lemma 4, Lemma 5 and Proposition 1 allow us to use the result in [6], then our graph boundary value problem (7) has a unique solution. As (7) is equivalent to (2) under tangential diffeomorphism, thus we get the short time existence for (2):

Theorem 3. *There exists a smooth solution* $\gamma : [-1,1] \times [0,T) \to \mathbb{R}^2$ *of the flow system (2), unique up to parametrisation. This solution is in the class* $C^{6,1,\alpha}([-1,1] \times [0,T))$ *(with arbitrary* $0 < \alpha < 1$*).*

2.2 Higher order flows of plane curves with boundary conditions

The way to proof the short time existence for the $(2m+4)$th-order curvature flows of plane curves with generalised Neumann boundary condition (4) is similar to the method used for the sixth order curvature flow problem (2).

Firstly, we transform the given problem into an equivalent initial-boundary-value problem for a scalar function and using standard results of the parabolic theory. The scalar initial-boundary-value problem:

$$\begin{cases} \frac{\partial w}{\partial t}(u,t) = f(u,t), & \forall (u,t) \in [-1,1]\times[0,\sigma] \\ w(\cdot,0) = w_0, & \\ w_u = w_{u^3} = w_{u^5} = \ldots = w_{u^{2m+3}} = 0, & \forall (u,t) \in \eta_{1,2}\times[0,\sigma] \end{cases} \tag{9}$$

here $w : [-1,1]\times[0,\sigma] \to \mathbb{R}$.

Secondly, we prove that scalar nonlinear initial-boundary-value problem (9) has a unique solution for a short time. Next we show that equations (9) and (4) are equivalent, see the sixth order case for the proof. Thus, flow problem (4) has a unique solution for finite time up to reparameterization.

Theorem 4. *There exists a smooth solution* $\gamma : [-1,1]\times[0,T) \to \mathbb{R}^2$ *of the system (4), unique up to parametrisation. This solution is in the class* $C^{2m+4,1,\alpha}([-1,1]\times[0,T))$ *(with arbitrary* $0<\alpha<1$*).*

Directly, we can get the short time existence for flow (5) satisfying Neumann boundary condition and $k_s = \ldots = k_{s^{2m-1}} = k_{s^{2m+1}} = 0$ at the boundary and with smooth initial curve $\gamma(\cdot,0) = \gamma_0$ compatible with the boundary conditions, the solution is also unique up to parametrisation.

3 Short time existence for flow of closed planar curves without boundary

The framework of short time existence for flow of closed planar curves without boundary is that we first write the length-constrained curve diffusion flow as a graph over the initial curve for unknown function of time, we have the scalar nonlinear parabolic problem. Secondly, we prove there is a unique solution for the graph problem and the length-constrained curve diffusion flow is invariant under tangential diffeomorphisms. Then these is a unique solution for length-constrained flow with the unknown time function. Thirdly, we use the Schauder fixed point theorem to prove the unique solution exists for our original problem with specific $h(t)$.

The constrained curve diffusion flow (6) is introduced in Definition 3. Firstly we write $\gamma : \mathbb{S}^1\times[0,T) \to \mathbb{R}^2$ as a graph for unknown function of time $\bar{h}(t)$ over the initial curve γ_0, using $\nu(u,t)$, $\tau(u,t)$ to denote the tangential and normal vector fields of the curve $\gamma(u,t)$ respectively, then $\langle \nu,\tau\rangle(u,t) = 0$, $\nu(u,t) = rot_{\pi/2}\tau(u,t)$. Let $f : \mathbb{R}\times[0,T) \to \mathbb{R}$, $\nu_0(u) = \nu(u,0)$ write

$$\gamma(u,t) = \gamma_0(u) + f(u,t)\nu_0(u).$$

We write the scalar nonlinear parabolic problem as:

$$\begin{cases} (\partial_t f)(u,t) = Q(f),\ \forall (u,t) \in \mathbb{S}^1 \times [0,T) \\ f(\cdot,0) = 0, \end{cases} \tag{10}$$

where

$$Q(f) = -\frac{V_0^2(1-k_0 f)^2}{V^6} \cdot f_{u^4} + b\left(\bar{h}, f, f_u, f_{uu}, f_{u^3}\right),$$

here $V = |\gamma_u(u,t)|$, $V_0 = |\gamma_u(u,0)|$ and b is a function depending only on $\bar{h}(t), f, f_u, f_{uu}, f_{u^3}$. As $f \in C^{4,1,\alpha}\left(\mathbb{S}^1 \times [0,T)\right)$ which means that f is $C^{4,\alpha}$ in space and $C^{1,\alpha}$ in time, then $b\left(f, f_u, f_{uu}, f_{u^3}\right)$ is bounded and continuous in sapce and time.

Secondly, we linearize $Q(f)$ at $f_0 = f(\cdot,0) = 0$, then our linearized scalar graph problem at $f_0 = 0$ is

$$\begin{cases} (\partial_t f)(u,t) = -V_0^{-4} \cdot f_{u^4} + g_l \cdot f_{u^l},\ \forall (u,t) \in \mathbb{S}^1 \times [0,T) \\ f(\cdot,0) = 0. \end{cases}$$

As $f \in C^{4,1,\alpha}\left(\mathbb{S}^1 \times [0,T)\right)$, thus the leading coefficient $-V_0^{-4}$ and g_l, $l = 0,1,2,3$ are continuous at u,t and uniformly bounded. We also can see that the leading coefficient satisfies Legendre-Hadamard condition (See [1], there exists a positive constant $\lambda \in \mathbb{R}$ such that the leading coefficient satisfies $|-V_0^{-4}| \geq \lambda$). Thus we can refer to Main Theorem 5 in [1] and proof that there is a unique solution for the nonlinear scalar graph problem (10) when $\bar{h}(t)$ is an unknown function of time.

Proposition 2. *There exists a positive time $T > 0$ such that the problem* (10) *has a unique solution $f \in C^{4,1,\alpha}(\mathbb{S}^1 \times [0,T))$.*

Lemma 6. *Length-constrained curve diffusion flow is invariant under tangential diffeomorphisms.*

For the proof of Lemma 6, we refer to Lemma 2.11 in [10].

Before giving the fixed point argument, we calculate $\frac{d^2}{dt^2}h(0) \leq c(\gamma_0)$ first, we do the second derivative of $h(t) = -\frac{\int_\gamma k_s^2 ds}{2\pi w}$ with respect to time, the highest order term is $\int_\gamma k_{s^5}^2 ds$. So $\frac{d^2}{dt^2}h(0)$ is bounded if $\gamma_0 \in C^{7,\alpha}(\mathbb{S}^1)$.

Theorem 5. *(Schauder fixed point theorem) Let I be a compact, convex subset of a Banach space B and let J be a continuous map of I into itself. Then J has a fixed point.*

We get the short time existence for the flow problem (6) by applying Theorem 5 together with $\frac{d^2}{dt^2}h(0)$ is bounded. (For the proof see [10].)

Theorem 6. *Let $\gamma_0 : \mathbb{S}^1 \to \mathbb{R}^2$ be a $C^{7,\alpha}$-regular immersed curve. Then there exists a maximal $T \in (0,\infty]$ such that the constrained curve diffusion flow system* (6) *is uniquely solvable with γ of degree $C^{4,1,\alpha}\left(\mathbb{S}^1 \times [0,T)\right)$.*

References

1. C. Baker, The mean curvature flow of submanifolds of high codimension, PhD thesis, ANU (2010).
2. J.L. Lions and E. Magenes, Non-homogeneous boundary value problems and application, vol. I–II, Springer, Berlin (1972).
3. J. McCoy, G. Wheeler and Y. Wu, A sixth order flow of plane curves with boundary conditions, Tohoku Mathematical Journal, Vol. 72, (2020), No. 3, pp. 379-393.
4. J. McCoy, G. Wheeler and Y. Wu, Evolution of closed curves by length-constrained curve diffusion, Proc. Amer. Math. Soc. Vol. 147 (2019), pp. 3493-3506.
5. J. McCoy, G. Wheeler and Y. Wu, Higher order curvature flows of plane curves with generalised Neumann boundary conditions, arXiv:2001.06140.
6. M. Poppenberg, Nash-moser techniques for nonlinear boundary-value problems, Electronic Journal of Differential Equations, Vol. 2003(2003), No. 54, pp. 1-33.
7. A.Stahl, Regularity estimates for solutions to the mean curvature flow with a Neumann boundary condition, Calculus of Variations and Partial Differential Equations, Vol. 4(1996), pp. 385-407.
8. A.Stahl, Uber den mittleren Krümmungsfluss mit Neumannrandwerten auf glatten Hyperflächen, PhD thesis, Fachbereich Mathematik, Eberhard-Karls-Universität, Tüebingen, Germany, (1994).
9. A. Spener, Short time existence for the elastic flow of clamped curves, Mathematische Nachrichten, Vol. 291 (2018), pp. 2115-2116.
10. G. Wheeler, Fourth order geometric evolution equations, PhD thesis, UOW (2009).
11. G. Wheeler and V. Wheeler, Curve diffusion and straightening flows on parallel lines, arXiv:1703.10711.

Chapter 14

Harmonic Analysis and Dispersive PDEs: Problems and Progress

Hankel transforms and weak dispersion

Federico Cacciafesta and Luca Fanelli

Abstract This survey is concerned with a general strategy, based on Hankel transforms and special functions decompositions, to prove weak dispersive estimates for a class of PDE's. Inspired by [2], we show how to adapt the method to some scaling critical dispersive models, as the Dirac-Coulomb equation and the fractional Schrödinger and Dirac equation in Aharonov-Bohm field.

1 Introduction

Let $a \in \mathbb{R}$ and let us consider the Hamiltonians

$$H_0 := -\Delta, \qquad H_a := -\Delta + \frac{a}{|x|^2},$$

on $L^2(\mathbb{R}^n)$, with $n \geq 2$. It is well known that, under the condition

$$a \geq -\frac{(n-2)^2}{4}, \tag{1}$$

the Hamiltonian H_a can be realized as the Friedrichs' extension of the symmetric semi-bounded operator $-\Delta + a/|x|^2$, acting on the natural domain induced by the quadratic form

$$q[u] := \int_{\mathbb{R}^n} |\nabla u(x)|^2 \, dx + a \int_{\mathbb{R}^n} \frac{|u(x)|^2}{|x|^2} \, dx.$$

Federico Cacciafesta
Dipartimento di Matematica, Università degli studi di Padova, Via Trieste, 63, 35131 Padova PD, Italy. e-mail: cacciafe@math.unipd.it

Luca Fanelli
Universidad del País Vasco - Euskal Herriko Unibertsitatea & Ikerbasque, Bilbao, Spain
e-mail: luckyfim200479@gmail.com

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_62

In particular, by the Spectral Theorem we can define the Schrödinger flow e^{itH_a} on the domain of H_a, for any a satisfying (1). For $a \neq 0$, in dimension $n \geq 3$, we can consider H_a as a critical linear perturbation of H_0, due to the Hardy's inequality

$$\int_{\mathbb{R}^n} \frac{|u(x)|^2}{|x|^2}\,dx \leq \frac{4}{(n-2)^2}\int_{\mathbb{R}^n} |\nabla u(x)|^2\,dx \qquad (n \geq 3).$$

In addition, the Schrödinger equation

$$\partial_t u(t,\cdot) = -iH_a u(t,\cdot) \tag{2}$$

is invariant under the scaling

$$u_\lambda(t,x) := u\left(\frac{t}{\lambda^2}, \frac{x}{\lambda}\right).$$

In the recent years, a new interest has been devoted to the study of the dispersive properties of flows as e^{itH_a}, once it was realized that the somehow vintage business of special functions and Hankel transforms could play a role in the analysis of critical and scaling invariant models (see e.g. [2, 3, 6, 7, 10, 11, 12, 14, 17, 20]). In this topic, we will point our attention on dispersive models which usually arise in Quantum Mechanics and always enjoy the above mentioned property of criticality. The inspiration, and main motivation of the project, comes form the papers [2, 3], of which we now briefly review the main results. Let us first recall the spherical harmonics decomposition of $L^2(\mathbb{R}^2)$, which is a peculiar feature of the 2D-space. Given the complete orthonormal set $\{\phi_m\}_{m\in\mathbb{Z}}$ on $L^2(\mathbb{S}^1)$, with $\phi_m = \phi_m(\theta) = \frac{e^{im\theta}}{\sqrt{2\pi}}$, $\theta \in [0, 2\pi)$, one has the canonical isomorphism

$$L^2(\mathbb{R}^2) \cong \bigoplus_{m\in\mathbb{Z}} L^2(\mathbb{R}_+, rdr) \otimes [\phi_m] \tag{3}$$

where we are denoting with $[\phi_m]$ the one dimensional space spanned by ϕ_m and with $\|f\|^2_{L^2_{rdr}} = \int_0^\infty |f(r)|^2 rdr$. We denote by $L^2_{\geq d(\mathbb{R}^2)}$, the subspace of L^2 consisting of all functions that are orthogonal to all spherical harmonics of degree less than d. In [2], the authors managed to prove the following family of estimates

$$\||x|^{-1/2-2\alpha}(H_a^{1/4-\alpha})e^{itH_a}f\|_{L^2_t L^2_x} \leq C\|f\|_{L^2(\mathbb{R}^n)} \tag{4}$$

where we are denoting with

$$H_a = -\Delta + \frac{a}{|x|^2}$$

for $n \geq 2$, $\alpha \in (0, \frac{1}{4} + \frac{1}{2}\mu_d)$, with $\mu_d = \sqrt{(\lambda(n)+d)^2 + a}$, $d \geq 0$, $\lambda(n) = \frac{n-2}{2}$, and $f \in L^2_{\geq d(\mathbb{R}^n)}$. The strategy developed in [2] can be roughly summarized in the following steps.

1. Use *spherical harmonics decomposition* to reduce the equation to a radial problem;

2. Use *Hankel transform* to "diagonalize" the reduced problem and to define fractional powers of the operator $-\Delta + \frac{a}{|x|^2}$;
3. Prove the smoothing estimate on a fixed spherical space using Hankel transform properties and the explicit integral representation of the fractional powers;
4. Sum back: use triangle inequality and L^2-orthogonality of spherical harmonics to obtain the desired estimate for the original dynamics. To conclude, it will be crucial to show that the constant obtained in step (3) is a bounded function of the spherical parameter.

In [2, 3], as an appliction of of (4), the authors proved that the usual Strichartz estimates hold for the flow e^{itH_a}, when a satisfies (1). More precisely,

$$\left\| e^{itH_a} f \right\|_{L_t^p L_x^q} \leq C \|f\|_{L^2}, \tag{5}$$

for some $C > 0$ independent on f, provided (1) and

$$\frac{2}{p} = \frac{n}{2} - \frac{n}{q}, \qquad p \geq 2, \qquad (p,q,n) \neq (2, +\infty, 2).$$

At that time, it was a striking result, since it was completely unclear whether Strichartz estimates would have been true for critical perturbations of the free Hamiltonian. Moreover, it is known that the inverse-square potential represents a threshold, among homogeneous perturbations, for the validity of Strichartz estimates (see [13, 16]). In addition, it is now known that the usual time decay estimate

$$\sup_{x \in \mathbb{R}^n} \left| e^{itH_a} f(x) \right| \leq C |t|^{-\frac{n}{2}} \|f\|_1$$

fails, in general, as soon as $a < 0$ (see [10, 11, 12]), which possibly gives strength to the averaging property of Strichartz estimates. To complete the state of the art, for the critical value $a = -(n-2)^2/4$, in the recent papers [21, 25] the authors proved the validity of Strichartz estimates for the Schrödinger and wave equations, provided the admissible couple is not endpoint.

To prove (5) by (4) is a quite simple application of a $TT^\star$ argument, mixing free Strichartz estimates and (4) when $\alpha = \frac{1}{4}$, $d = 0$, which is

$$\| |x|^{-1} e^{itH_a} f \|_{L_t^2 L_x^2} \leq C \|f\|_{L^2(\mathbb{R}^n)}. \tag{6}$$

It is important to notice that $\alpha = \frac{1}{4}$ is in the range of estimate (4), for $d = 0$, thanks to (1), which implies $\mu_d > 0$ (see [2, 3] for details).

The aim of this survey is to describe under which extent we can hope to generalize estimates (4) to other dispersive models and which is the quantitative role played by the number a, interpreted as the bottom of the spectrum of the angular component of H_a. In the following, we will restrict our attention on fractional Schrödinger equations in Aharonov-Bohm fields and Dirac equations, both in Coulomb and Aharonov-Bohm fields.

2 Fractional Schrödinger in Aharonov-Bohm field

An interesting 2D-example of scaling critical, first order perturbation of the Laplace operator is given by the so called *Aharonov-Bohm field*: such a field is given by

$$A_B : \mathbb{R}^2 \setminus \{(0,0)\} \to \mathbb{R}^2, \qquad A_B(x) = \left(-\frac{x_2}{|x|^2}, \frac{x_1}{|x|^2} \right), \qquad x = (x_1, x_2) \tag{1}$$

so that we can define for each $\alpha \in \mathbb{R}$ the Hamiltonian

$$H_\alpha = \left(-i\nabla + \alpha \left(-\frac{x_2}{|x|^2}, \frac{x_1}{|x|^2} \right) \right)^2 \tag{2}$$

which is self-adjoint. By the Spectral Theorem we can thus define fractional powers of the hamiltonian H_α, and thus we can in particular consider for $a > 0$ the following Cauchy problems

$$\begin{cases} \partial_t u = i H_\alpha^{a/2} u \\ u(0,\cdot) = f(\cdot) \in L^2(\mathbb{R}^2) \end{cases} \tag{3}$$

which we will refer to as *fractional Schrödinger equation with Aharonov-Bohm field*. We note that the cases $a = 1$ and $a = 2$ correspond, respectively, to the Schrödinger and wave flows.

The main result in this case is given by the following Theorem, that is proved in [6].

Theorem 1 ([6]). *Let $a > 0$, $\alpha \in \mathbb{R}$ and*

$$0 < \varepsilon < \frac{1}{4} + \frac{1}{2} dist(\alpha, \mathbb{Z}).$$

Then for every $f \in L^2$ the following estimate holds

$$\| |x|^{-\frac{1}{2}-2\varepsilon} H^{\frac{a-1}{4}-\varepsilon} e^{itH^{a/2}} f \|_{L_t^2 L_x^2} \leq C \|f\|_{L^2} \tag{4}$$

with a constant C depending on α and ε.

In addition, in the endpoint case $\varepsilon = 0$ the following local estimate holds

$$\sup_{R>0} R^{-1/2} \| e^{itH^{a/2}} f \|_{L_t^2 L^2_{|x|<R}} \leq C \| H^{\frac{1-a}{4}} f \|_{L_x^2}. \tag{5}$$

Remark 1. It is worth noticing that (4) fails for $\alpha \in \mathbb{Z}$ and $\varepsilon = \frac{1}{4}$. Indeed, the dimension $d = 2$ is critical with respect to estimate (4), with $\varepsilon = \frac{1}{4}$, due to the fact that the weight $|x|^{-1}$ is too singular at the origin. Nevertheless, the presence of the field A_B, as it is well known, improves the angular ellipticity of H, if $\alpha \notin \mathbb{Z}$, and this usually permits to obtain better estimates than in the free case, as (4) shows. Roughly speaking, the higher the spherical frequency is, the better is the dispersive phenomenon we are measuring. The improvement arises since the introduction of

the external potential is cutting the 0-frequency from the spectrum of the spherical operator.

Remark 2. Notice that estimates above can be extended, by following the argument in [8], to deal with the Klein-Gordon flow $e^{it\sqrt{H^a+1}}$. Also, as an immediate corollary of the result above, it is possible to prove weighted Strichartz estimates for the dynamics (3) (simply by interpolating estimate (4) with the 2D Sobolev inequality)

2.1 The massless Dirac-Coulomb equation

The Cauchy problem for the 3D massless Dirac equation with a Coulomb potential reads as

$$\begin{cases} i\partial_t u + \mathscr{D}u + \dfrac{\nu}{|x|}u = 0, \quad u(t,x) : \mathbb{R}_t \times \mathbb{R}^3_x \to \mathbb{C}^4 \\ u(0,x) = f(x) \end{cases} \tag{6}$$

where we recall that the massless Dirac operator $\mathscr{D}$ is defined in terms of the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{7}$$

as

$$\mathscr{D} = -i\sum_{k=1}^{3} \alpha_k \partial_k = -i(\alpha \cdot \nabla)$$

where the 4×4 Dirac matrices are given by

$$\alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix}, \quad k = 1,2,3. \tag{8}$$

The charge ν is assumed to be in the interval $(-1,1)$, as the operator $\mathscr{D}_\nu = \mathscr{D} + \frac{\nu}{|x|}$ needs to be self-adjoint (see [9]).

The adaptation of the machinery to this setting is a bit more tricky, and requires to deal with some additional technical difficulties, which are mainly due to the rich algebraic structure of the Dirac operator, that are essentially the following:

- The Dirac operator does not preserve radiality, meaning that the standard spherical harmonics decomposition does not represent a "good" setting, which is instead given by the so called *partial wave decomposition*, that we now briefly introduce. First of all, we use spherical coordinates to write

  $$L^2(\mathbb{R}^3, \mathbb{C}^4) \cong L^2((0,\infty), r^2 dr) \otimes L^2(S^2, \mathbb{C}^4)$$

  with S^2 being the unit sphere. Then, we have the orthogonal decomposition on S^2:

$$L^2(S^2,\mathbb{C}^4)\cong\bigoplus_{k\in\mathbb{Z}\setminus\{0\}}\bigoplus_{m=-|k|+1}^{|k|}h_{m,k}$$

where the spaces $h_{m,k}:=\mathbb{C}\Phi^+_{m,k}+\mathbb{C}\Phi^-_{m,k}$ with

$$\Phi^+_{m,k}=\begin{pmatrix}\phi^+_{m,k}\\0\end{pmatrix},\qquad \Phi^+_{m,k}=\begin{pmatrix}0\\\phi^-_{m,k}\end{pmatrix}$$

and the functions $\phi^\pm_{m,k}$ can be explicitly written in terms of standard spherical harmonics as

$$\phi^\pm_{m,k}=\frac{1}{\sqrt{|2k\pm1|}}\begin{pmatrix}\sqrt{|k\mp(m-1)|}Y^{m-1}_{|k|-H(\mp k)}\\ \mp\mathrm{sgn}(k)\sqrt{|k\pm m)|}Y^m_{|k|-H(\mp k)}\end{pmatrix}$$

and H is the Heaviside function. The action of the Dirac-Coulomb operator leaves invariant these subspaces and this decomposition it is represented by the radial matrix

$$\mathscr{D}_{\nu,k}=\begin{pmatrix}\frac{\nu}{r} & -\frac{d}{dr}+\frac{1+k}{r}\\ \frac{d}{dr}-\frac{1-k}{r} & \frac{\nu}{r}\end{pmatrix}. \tag{9}$$

We mention the fact that a similar decomposition holds in any dimension $n\geq 2$. The standard reference for this and related problem is the book [24]

Remark 3. We have to stress the fact that the decomposition introduced in [24] (see in particular Subsection 4.6.5) is slightly different, as it also relies on the isomorphism $u\to r\tilde{u}$. This has the effect of rather "simplifying" the expression of the action of the radial Dirac operator given in (9) and, of course, affects the presence of a weight in the radial scalar product. This same approach is used also e.g. in [4]. Also, we stress the fact that our index m is shifted by $1/2$ with respect to the one in [24], in order to avoid half integers.

- The construction of the analogous of the Hankel transform is a more delicate problem: roughly speaking, we will need to define a two-dimensional operator which projects onto the positive and negative part of the continuous spectrum of the Dirac Coulomb operator (we recall that these generalized eigenfunctions are explicit and well known, see e.g. [19]). If we thus define, for a fixed admissible couple m,k, the "relativistic Hankel transform" to be

$$\mathscr{H}^\pm_{m,k}\Phi(r)=\langle\Psi^{\pm E}_{m,k}(r),\Phi(r)\rangle_{L^2(r^2dr)}$$

we obtain the properties we need, simply relying on the self-adjointess of the operator $\mathscr{D}_{\nu,k}$ with respect to the $L^2(r^2dr)$ scalar product

$$\mathscr{H}^\pm_{m,k}\mathscr{D}_{\nu,k}=\pm E\mathscr{H}^\pm_{m,k}. \tag{10}$$

With this, we mean that the transform $\mathscr{H}^\pm_{m,k}$ "diagonalizes" the equation.

- By relying on property (10) we can define fractional powers of the restricted Dirac-Coulomb operator. As a last (and technical) step then we will have to show that these fractional powers admit an integral kernel which can be explicitly written by solving suitably weighted interaction integrals of generalized eigenstates; this will allow to prove the estimate on a fixed partial wave subspace with a suitable constant depending on the spherical parameters which, again, will need to be bounded in order to allow the application of triangle inequality to sum back in the partial wave decomposition.

The following result is proved in [7] (we mention the fact that a similar result holds in 2D as well).

Theorem 2 ([7]). *Let K be a positive integer, and set*

$$h_{\geq K} = \bigoplus_{|k|\geq K} \bigoplus_{m=-|k|+1}^{|k|} h_{m,k}.$$

Then for any

$$1/2 < \varepsilon < \sqrt{K^2 - \nu^2} + 1/2$$

and any $f \in L^2((0,\infty), r^2 dr) \otimes h_{\geq K}$ there exists a constant $C = C(\nu, \varepsilon, K)$ such that the following estimate holds

$$\| |x|^{-\varepsilon} |\mathscr{D}_\nu|^{1/2-\varepsilon} e^{it(\mathscr{D} + \frac{\nu}{|x|})} f \|_{L^2_t L^2_x} \leq C \|f\|_{L^2_x}. \tag{11}$$

Remark 4. We need to point out a typo in formula (2.7) in [7]: in the definition of the space $\mathscr{H}^3_{\geq \bar{k}_3}$ the sum in j is in the range $j \geq |\bar{k}_3| - 1/2$. Also, we stress the fact that we are here providing a rather different (and somehow simplified) representation of the partial wave subspaces, and therefore of the spaces $h_{\geq K}$, with respect to [7]: in particular, we are here neglecting the sum in $j \in \frac{1}{2}\mathbb{N}$, that is "englobed" in the one in k (which is an integer) and, as mentioned, we have "shifted" the index m by $1/2$.

Remark 5. The analogous of estimate (5) seems to be more complicated to be proved in this contest. This is ultimately due to the much more complicated structure of the Hankel transform, which involves confluent hypergeometric functions: indeed, the key step is represented by the proof of a bound, uniform in R and l, of the form

$$\frac{1}{R} \int_0^R \chi_l(r)^2 r^{n-1} dr < C$$

for the radial components of the generalized eigenstates. This is well known in the case of Bessel functions (see [23]) but seems to be more complicated to be obtained in the Dirac-Coulomb case.

Remark 6. The restriction to the massless case is crucial in our result, as our strategy deeply relies on the scaling invariant structure of the equation, therefore leaving open the question whether the same result (or at least similar) holds in presence

of a mass. The application of the strategy presented in [8] to pass from the wave to the Klein-Gordon equation is not indeed completely straightforward, as here we are not simply "shifting" but we are "opening a gap" as, we recall, the spectrum of the free Dirac operator is unbounded both from above and below. Still, it seems to be possible to make things work in this contest, and this will be the object of future investigations. In any case, it is well known (see e.g. [19]) that the massive Dirac-Coulomb operator has eigevalues in the gap: therefore, in order to be able to obtain any kind of dispersive estimates, it will be necessary to project out of the point spectrum.

2.2 The massless Dirac equation in Aharonov-Bohm field

The two results above can be somehow merged to deal with the massless Dirac equation in Aharonov-Bohm field; in this case, the Hamiltonian reads as $\mathscr{D}_A = i\big(\sigma_1(\partial_x + A^1) + \sigma_2(\partial_y + A^2)\big)$, where the σ matrices are defined as in (7) and the magnetic field $A(x)$ is given by (1). In this case we claim that the following result holds (for simplicity we restrict to the case $\alpha \in (0,1)$ without losing in generality).

Theorem 3. *[5] Let $\alpha \in (0,1)$ and $A(x)$ given by* (1)*. Then for any*

$$1/2 < \varepsilon < 1 + |l + \alpha|. \tag{12}$$

and any $f \in L^2((0,\infty), rdr) \otimes \mathscr{H}_{\geq \bar{l}}$ there exists a constant $C = c(\alpha, \varepsilon, l)$ such that the following estimate holds

$$\Big\| \Omega^{-\varepsilon} \mathscr{D}_A^{1/2-\varepsilon} e^{it\mathscr{D}_A} f \Big\|_{L_t^2 L_x^2} \leq c \|f\|_{L_x^2}. \tag{13}$$

In addition, in the endpoint case $\gamma = 1/2$ the following estimate holds

$$\sup_{R>0} R^{-1/2} \|e^{it\mathscr{D}_A} f\|_{L_t^2 L^2_{|x|\leq R}} \lesssim \|f\|_{L_x^2}, \tag{14}$$

Remark 7. Notice that the range (12) is better than the one in the free case, as soon as $\alpha \notin \mathbb{Z}$. This fact can be interpreted as a *diamagnetic* behavior of this model, which is quite surprising for Dirac-type equations. Indeed, an original conjecture about universal paramagnetism in [18] was later disproved in [1]. The above model seems to go in the same direction of the example in [1], with the difference that it is critically singular at the origin.

Remark 8. It is important to notice that the local smoothing estimates obtained for the three models above do not allow to recover, following the perturbation arguments in [2, 3], the full set of Strichartz estimates for the corresponding flows. To make the argument work we would indeed need, for al the three models, an estimate of the form

$$\| |x|^{-1/2} u \|_{L_t^2 L_x^2} \leq C \|f\|_{L_x^2}$$

with u being a solution of any of the models above with initial condition f. But this, as it is seen, is exactly at the endpoint (and just outside) of our admissible ranges; in fact, we seem to have no hope to prove such an estimate even imposing further restrictions (say, radial initial data) as it does fail even for the free wave equation (see e.g. [22]). The loss of (fractional) derivatives in our local smoothing seems on the other hand to suggest that some "weak" (namely, with loss of derivatives) Strichartz estimates should hold in this setting.

Acknowledgements The authors are grateful to Kenji Nakanishi for providing several useful comments and suggestions. The first author is partially supported by the University of Padova STARS project "Linear and Nonlinear Problems for the Dirac Equation" (LANPDE).

References

1. Avron, J., and Simon, B., A counterexample to the paramagnetic conjecture, *Phys. Lett. A* **79** (1979/80), no. 1-2, 41-42.
2. Burq, N., Planchon, F., Stalker, J.G., and Tahvildar-Zadeh A., Strichartz estimates for the wave and Schrödinger equations with the inverse-square potential. *J. Funct. Anal.* 203 (2), 519-549 (2003).
3. Burq, N., Planchon, F., Stalker, J.G., and Tahvildar-Zadeh A., Strichartz estimates for the wave and Schrödinger equations with potentials of critical decay. *Indiana Univ. Math. J.* 53 (6), 1665-1680 (2004).
4. Cacciafesta, F., Global small solutions to the critical Dirac equation with potential. *Nonlinear Analysis* 74, pp. 6060-6073, (2011).
5. Cacciafesta F., and Fanelli L. Dispersive estimates for the Dirac equation in an Aharonov-Bohm field. *J. Differential equations* 263 7, 4382-4399, (2017).
6. Cacciafesta, F., and Fanelli, L., Weak dispersion and weighted Strichartz inequalities for fractional Schrödinger equations in Aharonov-Bohm magnetic fields. *Dynamics of PDE* Vol. 16 n.1, 95-103, (2019).
7. Cacciafesta, F., and Seré, E., Local smoothing estimates for the Dirac Coulomb equation in 2 and 3 dimensions. *J. Funct. Anal.* 271 no.8, 2339-2358 (2016)
8. D'Ancona, P., Kato smoothing and Strichartz estimates for wave equations with magnetic potentials. *Comm. Math. Phys.* 335, No. 1, 1-16 (2015)
9. Esteban, M., and Loss, M., Self-adjointness for Dirac operators via Hardy-Dirac inequalities. *J. Math. Phys.* 48, no. 11, 112107 (2007).
10. L. Fanelli, V. Felli, M. Fontelos, and A. Primo, Time decay of scaling critical electromagnetic Schrödinger flows, *Communications in Mathematical Physics* **324** (2013), 1033–1067.
11. L. Fanelli, V. Felli, M. Fontelos, and A. Primo, Time decay of scaling invariant electromagnetic Schrödinger equations on the plane, *Communications in Mathematical Physics* **337** (2015), 1515–1533.
12. L. Fanelli, V. Felli, M. Fontelos, and A. Primo, Frequency-dependent time decay of Schrödinger flows, *J. Spectral Theory* **8** (2018), 509–521.
13. Fanelli, L., and García, A., Counterexamples to Strichartz estimates for the magnetic Schrödinger equation, *Comm. Cont. Math.* **13** (2011) no. 2, 213–234.
14. Fanelli, L., Grillo, G., and Kovařík, H., Improved time-decay for a class of scaling critical electromagnetic Schrödinger flows, *J. Func. Anal.* **269** (2015), 3336–3346.

15. FANELLI, L., ZHANG, J., AND ZHENG, J, Strichartz estimates for 2D-scaling invariant electromagnetic waves, 2020 arXiv:2003.10356
16. GOLDBERG, M., VEGA, L., AND VISCIGLIA, N., Counterexamples of Strichartz inequalities for Schrödinger equations with repulsive potentials, *Int. Math Res Not.*, 2006 Vol. 2006: article ID 13927.
17. G. GRILLO AND H. KOVARIK, Weighted dispersive estimates for two-dimensional Schrödinger operators with Aharonov-Bohm magnetic field, *Journal of Differential Equations* **256** (2014), 3889–3911.
18. HOGREVE, H., SCHRADER, R., AND SEILER, R., A conjecture on the spinor functional determinant, *Nuclear Phys. B* **142**, no. 4 (1978), 525-534.
19. LANDAU, L.M., AND LIFSHITZ, L.D., Quantum mechanics - Relativistic quantum theory.
20. MIZUTANI, H., A note on smoothing effects for Schrödinger equations with inverse-square potentials, *Proc. Amer. Math. Soc.* **146** (2018), 295–307.
21. MIZUTANI, H., Remarks on endpoint Strichartz estimates for Schrödinger equations with the critical inverse-square potential, *J. Diff. Eq.* **263** (2017), 3832–3853.
22. OZAWA, T., AND ROGERS, K.M., Sharp Morawetz estimates. *J. Anal. Math.* 121, 163–175 (2013).
23. STRICHARTZ, R., Harmonic analysis as spectral theory of the Laplacians. *J. Func. Anal.* 87, 51-148 (1989).
24. THALLER, B., The Dirac Equation. Springer-Verlag, Texts and Monographs in Physics (1992).
25. ZHANG, J., AND ZHENG, J., Strichartz estimates and wave equation in a conic singular space, *Math. Ann.* **376** (2020), 525–581.

A priori bounds for the kinetic DNLS

Nobu Kishimoto and Yoshio Tsutsumi

Abstract In this note, we consider the kinetic derivative nonlinear Schrödinger equation (KDNLS), which arises as a model of propagation of a plasma taking the effect of the resonant interaction between the wave modulation and the ions into account. In contrast to the standard derivative NLS equation, KDNLS does not conserve the mass and the energy. Nevertheless, the dissipative structure of KDNLS enables us to show an a priori bound in the energy space and a lower bound of the L^2 norm for its solution, as we see in this note. Combined with the local well-posedness result, which we plan to show in a forthcoming paper, these bounds will give a global existence result in the energy space for small initial data.

1 Introduction

We consider the kinetic derivative NLS equation (KDNLS):

$$\partial_t u = \partial_x \Big[i u_x + \alpha |u|^2 u + \beta \mathscr{H}(|u|^2) u \Big], \qquad \alpha, \beta \in \mathbf{R}, \quad \beta < 0, \tag{1}$$

where the spatial domain is either $\mathbf{R}$ or $\mathbf{T} = \mathbf{R}/2\pi\mathbf{Z}$. We write $\mathscr{F}$ to denote the Fourier transform and use the notation: $u_x := \partial_x u$, $\mathscr{H} := \mathscr{F}^{-1}\big[-i\,\mathrm{sgn}(\xi)\big]\mathscr{F}$, $D := (-\partial_x^2)^{1/2} = \mathscr{F}^{-1}|\xi|\mathscr{F} = \partial_x \mathscr{H}$. The negative constant β represents the ratio of plasma pressure to magnetic pressure, which can be positive, negative or zero according to each physical situation. Equation (1) takes the resonant interaction be-

The original version of this chapter was revised: The correction to this chapter is available at https://doi.org/10.1007/978-3-030-62497-2_64

N. Kishimoto
Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, JAPAN
e-mail: nobu@kurims.kyoto-u.ac.jp

Y. Tsutsumi
Department of Mathematics, Kyoto University, Kyoto 606-8502, JAPAN
e-mail: tsutsumi@math.kyoto-u.ac.jp

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_63

tween the wave modulation and the ions into account, while it is ignored in the derivative NLS, i.e., in the case of $\beta = 0$. The word "kinetic" implies that the collective motion of ions in a plasma is modeled by the Vlasov equation and not by the fluid equation. If Maxwell's equations and the Euler equations are taken as a model system, then we have DNLS, i.e., (1) with $\beta = 0$. If Maxwell's equations and the Vlasov equation are taken as a model system, then we have KDNLS (1) (see Dysthe and Pécseli [1] and Mjølhus and Wyller [3, 4].

Due to the presence of the Hilbert transform, the mass and the energy corresponding to the standard DNLS ($\beta = 0$ in (1)) are not conserved under the flow when $\beta < 0$. However, the nonlinear term $\beta\partial_x(\mathscr{H}(|u|^2)u)$ has dissipative structure when $\beta < 0$. The aim of this note is to derive an a priori bound in the energy space H^1 by using this structure. The main result reads as follows:

Theorem 1. *Let u be a smooth solution to* (1) *on $[0,T]\times Z$, where Z is either* $\mathbf{R}$ *or* $\mathbf{T}$. *Then, it holds that*

$$\|u(t)\|_{L^2}^2 + |\beta|\int_0^t \|D^{1/2}(|u(\tau)|^2)\|_{L^2}^2\,d\tau \;=\; \|u(0)\|_{L^2}^2, \qquad t\in[0,T].$$

Moreover, there exist $C_*, C > 0$ depending only on α,β (and bounded when $\beta\to 0$) such that if $\|u(0)\|_{L^2} \le C_*^{-1}$, then

$$\|u(t)\|_{H^1}^2 + \frac{|\beta|}{4}\int_0^t \|D^{1/2}\partial_x(|u(\tau)|^2)\|_{L^2}^2\,d\tau \;\le\; 4\|u(0)\|_{H^1}^2 e^{C\|u(0)\|_{L^2}^2},$$

$$\|u(t)\|_{L^2}^2 \;\ge\; \|u(0)\|_{L^2}^2 \exp\Big[-C\|u(0)\|_{H^1}e^{C\|u(0)\|_{L^2}^2}|\beta|^{1/2}t^{1/2}\Big], \qquad t\in[0,T].$$

The proof of the theorem is based on the differential equalities (see Corollary 2) for the mass $\|u(t)\|_{L^2}^2$ and the energy functional

$$E[u] := \int\Big\{|u_x|^2 - \frac{3}{2}\Big(\alpha|u|^2 + \beta\mathscr{H}(|u|^2)\Big)\operatorname{Im}(\overline{u}u_x) + \frac{1}{2}\alpha^2|u|^6\Big\}\,dx.$$

In a forthcoming paper [2], we will consider the case $Z = \mathbf{T}$ and construct local-in-time solutions to the associated Cauchy problem for small initial data in Sobolev space $H^s(\mathbf{T})$ with $s > 1/2$. More precisely, we will prove the following result:

Theorem 2. *We assume $\alpha = 0$ and $\beta < 0$. Let $s \ge s_0 > 1/2$, then there exist $\eta = \eta(s_0,s) > 0$ and $T > 0$ such that for any $u_0 \in H^s(\mathbf{T})$ with $\|u_0\|_{H^{s_0}} \le |\beta|^{1/2}\eta$, the Cauchy problem of* (1) *with $u|_{t=0} = u_0$ has a unique solution $u \in C([0,T];H^s(\mathbf{T}))$ on $(0,T)\times\mathbf{T}$, which belongs to certain auxiliary spaces. Furthermore, the map $u_0 \mapsto u$ is continuous.*

The H^1 a priori bound in Theorem 1 and a standard approximation argument then show:

Corollary 1. *Let $\alpha = 0$ and $\beta < 0$. There exists $\eta > 0$ such that if $u_0 \in H^1(\mathbf{T})$ satisfies $\|u_0\|_{H^1} \le |\beta|^{1/2}\eta$, then the solution $u \in C([0,T];H^1(\mathbf{T}))$ of* (1) *on $(0,T)\times\mathbf{T}$ with $u|_{t=0} = u_0$ constructed in Theorem 2 can be extended to a global-in-time H^1 solution which is bounded and continuous in t.*

2 Energy conservation

In this section, we give a proof of Theorem 1. The next two lemmas follow from a direct calculation, so we omit their proofs.

Lemma 1. *Let u be a smooth solution to* (1)*. Then, it holds that*

$$\partial_t(|u|^2) = \partial_x\Big[-2\,\mathrm{Im}(\overline{u}u_x) + \frac{3}{2}\alpha|u|^4 + \beta|u|^2\mathscr{H}(|u|^2)\Big] + \beta|u|^2 D(|u|^2),$$

$$\begin{aligned}\partial_t\,\mathrm{Im}(\overline{u}u_x) = \partial_x\Big[\frac{1}{2}\partial_x^2(|u|^2) - 2|u_x|^2 + \Big(\alpha|u|^2 + \beta\mathscr{H}(|u|^2)\Big)\,\mathrm{Im}(\overline{u}u_x)\Big]\\ + 2\,\mathrm{Im}(\overline{u}u_x)\,\partial_x\Big(\alpha|u|^2 + \beta\mathscr{H}(|u|^2)\Big).\end{aligned}$$

Lemma 2. *Let u be a smooth solution to* (1)*. Then, it holds that*

$$\partial_t\int|u_x|^2\,dx \;=\; -3\int\big(\alpha|u|^2+\beta\mathscr{H}(|u|^2)\big)\partial_x(|u_x|^2)\,dx+\beta\big\|D^{1/2}\partial_x(|u|^2)\big\|_{L^2}^2,$$

$$\begin{aligned}&\partial_t\int|u|^2\,\mathrm{Im}(\overline{u}u_x)\,dx\\ &\quad=-2\int|u|^2\partial_x(|u_x|^2)\,dx+\int\Big[2\alpha\partial_x(|u|^4)+4\beta|u|^2D[|u|^2]\Big]\,\mathrm{Im}(\overline{u}u_x)\,dx,\end{aligned}$$

$$\begin{aligned}&\partial_t\int\mathscr{H}(|u|^2)\,\mathrm{Im}(\overline{u}u_x)\,dx\\ &=-2\int\mathscr{H}(|u|^2)\partial_x(|u_x|^2)\,dx-2\big\|D^{1/2}\,\mathrm{Im}(\overline{u}u_x)\big\|_{L^2}^2+\frac{1}{2}\big\|D^{1/2}\partial_x(|u|^2)\big\|_{L^2}^2\\ &\quad+\alpha\int\Big\{\frac{3}{2}D(|u|^4)-|u|^2D(|u|^2)+2\mathscr{H}(|u|^2)\partial_x(|u|^2)\Big\}\,\mathrm{Im}(\overline{u}u_x)\,dx\\ &\quad+\beta\int\Big\{D[|u|^2\mathscr{H}(|u|^2)]+\mathscr{H}\big(|u|^2D(|u|^2)\big)+\mathscr{H}(|u|^2)D(|u|^2)\Big\}\,\mathrm{Im}(\overline{u}u_x)\,dx,\end{aligned}$$

and that

$$\partial_t\int|u|^6\,dx \;=\; 6\int\partial_x(|u|^4)\,\mathrm{Im}(\overline{u}u_x)\,dx+2\beta\int|u|^6D(|u|^2)\,dx.$$

From these lemmas, we immediately obtain the following:

Corollary 2. *Let u be a smooth solution to* (1)*. Then, we have*

$$\partial_t\int|u|^2\,dt \;=\; \beta\big\|D^{1/2}(|u|^2)\big\|_{L^2}^2$$

and

$$\partial_t E[u] = \frac{1}{4}\beta \left\|D^{1/2}\partial_x(|u|^2)\right\|_{L^2}^2 + 3\beta \left\|D^{1/2}\operatorname{Im}(\overline{u}u_x)\right\|_{L^2}^2 + \alpha^2\beta \int |u|^6 D(|u|^2)\,dx$$
$$- \frac{3}{2}\alpha\beta \int \left\{\frac{3}{2}D(|u|^4) + 2\partial_x\left[|u|^2\mathscr{H}(|u|^2)\right] + |u|^2 D(|u|^2)\right\}\operatorname{Im}(\overline{u}u_x)\,dx$$
$$- \frac{3}{2}\beta^2 \int \left\{D[|u|^2\mathscr{H}(|u|^2)] + \frac{1}{2}\partial_x\left[\left(\mathscr{H}(|u|^2)\right)^2\right] + \mathscr{H}\left[|u|^2 D(|u|^2)\right]\right\}$$
$$\times \operatorname{Im}(\overline{u}u_x)\,dx.$$

We prepare some more lemmas:

Lemma 3. *There exists $C_0 > 0$ depending only on α, β (and bounded as $\beta \to 0$) such that*

$$\|u\|_{L^2} \le C_0^{-1} \quad \Longrightarrow \quad 2^{-1}\|u\|_{H^1}^2 \le E[u] + \|u\|_{L^2}^2 \le 2\|u\|_{H^1}^2.$$

Proof. This follows from the Gagliardo-Nirenberg inequality:

$$\|u\|_{L^6(Z)}^6 \lesssim \begin{cases} \|u\|_{L^2}^4\|u_x\|_{L^2}^2 & (Z = \mathbf{R}), \\ \|u\|_{L^2}^6 + \|u\|_{L^2}^4\|u_x\|_{L^2}^2 & (Z = \mathbf{T}). \end{cases} \qquad \square$$

Lemma 4. *The following estimates hold:*

$$\|D^{1/2}(|u|^2)\|_{L^2}^2 \lesssim \|u\|_{L^2}^2\|D^{1/2}\partial(|u|^2)\|_{L^2}, \tag{2}$$

$$\|\mathscr{F}^{-1}\left[|\mathscr{F}(|u|^2)|\right]\|_{L^\infty}\|D^{1/2}(|u|^2)\|_{L^2} \lesssim \|u\|_{L^2}^2\|D^{1/2}\partial(|u|^2)\|_{L^2}, \tag{3}$$

$$\|u\|_{L^\infty}\|\mathscr{F}^{-1}\left[|\mathscr{F}D^{1/2}(|u|^2)|\right]\|_{L^\infty} \lesssim \|u\|_{L^2}\|D^{1/2}\partial(|u|^2)\|_{L^2}. \tag{4}$$

Proof. We first derive (2). In the non-periodic case, by interpolation we have

$$\||u|^2\|_{L^2} \lesssim \||u|^2\|_{L^\infty}^{1/2}\||u|^2\|_{L^1}^{1/2} \lesssim \|\partial(|u|^2)\|_{L^2}^{1/4}\||u|^2\|_{L^2}^{1/4}\|u\|_{L^2}$$
$$\lesssim \|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/6}\||u|^2\|_{L^2}^{1/3}\|u\|_{L^2},$$

which implies

$$\||u|^2\|_{L^2} \lesssim \|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/4}\|u\|_{L^2}^{3/2}.$$

Using this, we have

$$\|D^{1/2}(|u|^2)\|_{L^2}^2 \lesssim \|D^{1/2}\partial(|u|^2)\|_{L^2}^{2/3}\||u|^2\|_{L^2}^{4/3} \lesssim \|D^{1/2}\partial(|u|^2)\|_{L^2}\|u\|_{L^2}^2.$$

The above argument also works in the periodic case if $|u|^2$ is replaced with $|u|^2 - \frac{1}{2\pi}\int_0^{2\pi}|u|^2\,dx$. Since we estimate $D^{1/2}(|u|^2)$, the same result (2) holds in the periodic case.

In what follows, $\chi = 0$ if $Z = \mathbf{R}$ and $\chi = 1$ if $Z = \mathbf{T}$. By interpolation inequalities, we have

$$\begin{aligned}\|u\|_{L^\infty}^2 = \||u|^2\|_{L^\infty} &\lesssim \chi\||u|^2\|_{L^2} + \|\partial(|u|^2)\|_{L^2}^{1/2}\||u|^2\|_{L^2}^{1/2}\\ &\lesssim \chi\|u\|_{L^\infty}\|u\|_{L^2} + \|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/3}\||u|^2\|_{L^2}^{2/3}\\ &\lesssim \chi\|u\|_{L^\infty}\|u\|_{L^2} + \|u\|_{L^\infty}^{2/3}\|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/3}\|u\|_{L^2}^{2/3},\end{aligned}$$

which implies that

$$\|u\|_{L^\infty}^2 \lesssim \chi\|u\|_{L^2}^2 + \|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/2}\|u\|_{L^2}. \tag{5}$$

The above argument also shows that

$$\|\mathscr{F}^{-1}\big[|\mathscr{F}(|u|^2)|\big]\|_{L^\infty} \lesssim \chi\|u\|_{L^2}^2 + \|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/2}\|u\|_{L^2}. \tag{6}$$

If $Z = \mathbf{R}$, (3) is obtained from (2) and (6). If $Z = \mathbf{T}$, it suffices to combine (2), (6) with the trivial estimate $\|D^{1/2}(|u|^2)\|_{L^2} \le \|D^{1/2}\partial(|u|^2)\|_{L^2}$.

The same argument for (3) but using (5) instead of (6) shows

$$\|u\|_{L^\infty}^2\|D^{1/2}(|u|^2)\|_{L^2} \lesssim \|u\|_{L^2}^2\|D^{1/2}\partial(|u|^2)\|_{L^2}.$$

Using this and interpolation, we see

$$\begin{aligned}\|u\|_{L^\infty}\|\mathscr{F}^{-1}\big[|\mathscr{F}D^{1/2}(|u|^2)|\big]\|_{L^\infty} &\lesssim \|u\|_{L^\infty}\|D^{1/2}(|u|^2)\|_{L^2}^{1/2}\|D^{1/2}\partial(|u|^2)\|_{L^2}^{1/2}\\ &\lesssim \|u\|_{L^2}\|D^{1/2}\partial(|u|^2)\|_{L^2},\end{aligned}$$

which shows (4). □

Proof (Proof of Theorem 1). The L^2 equality follows from the first equality in Corollary 2, so we focus on the H^1 a priori estimate and the exponential L^2 lower bound.

From (3), the integrals

$$\int D(|u|^4)\,\mathrm{Im}(\bar{u}u_x)\,dx, \qquad \int \partial_x\big[|u|^2\mathscr{H}(|u|^2)\big]\,\mathrm{Im}(\bar{u}u_x)\,dx,$$
$$\int D[|u|^2\mathscr{H}(|u|^2)]\,\mathrm{Im}(\bar{u}u_x)\,dx, \qquad \int \partial_x\big[\big(\mathscr{H}(|u|^2)\big)^2\big]\,\mathrm{Im}(\bar{u}u_x)\,dx$$

are bounded by

$$\begin{aligned}&\|\mathscr{F}^{-1}\big[|\mathscr{F}(|u|^2)|\big]\|_{L^\infty}\|D^{1/2}(|u|^2)\|_{L^2}\|D^{1/2}\,\mathrm{Im}(\bar{u}u_x)\|_{L^2}\\ &\quad\lesssim \|u\|_{L^2}^2\|D^{1/2}\partial(|u|^2)\|_{L^2}\|D^{1/2}\,\mathrm{Im}(\bar{u}u_x)\|_{L^2}\\ &\quad\lesssim \|u\|_{L^2}^2\|D^{1/2}\partial(|u|^2)\|_{L^2}^2 + \|u\|_{L^2}^2\|D^{1/2}\,\mathrm{Im}(\bar{u}u_x)\|_{L^2}^2.\end{aligned}$$

To estimate the integrals

$$\int |u|^2D(|u|^2)\,\mathrm{Im}(\bar{u}u_x)\,dx, \qquad \int \mathscr{H}\big[|u|^2D(|u|^2)\big]\,\mathrm{Im}(\bar{u}u_x)\,dx,$$

we denote the frequency variables for $|u|^2$, $D(|u|^2)$, and $\mathrm{Im}(\overline{u}u_x)$ by k_1, k_2, and k_3, respectively. Note that $k_1+k_2+k_3=0$. If $|k_3| \gtrsim |k_2|$, we can move half a derivative onto $\mathrm{Im}(\overline{u}u_x)$ and argue as before. If $|k_1| \sim |k_2| \gg |k_3|$, we apply (4) to estimate these integrals as

$$\begin{aligned}
&\|\mathscr{F}^{-1}\left[|\mathscr{F}D^{1/2}(|u|^2)|\right]\|_{L^\infty}\|D^{1/2}(|u|^2)\|_{L^2}\|u\|_{L^\infty}\|u_x\|_{L^2} \\
&\quad\lesssim \|u\|_{L^2}\|D^{1/2}\partial(|u|^2)\|_{L^2}\|D^{1/2}(|u|^2)\|_{L^2}\|u_x\|_{L^2} \\
&\quad\lesssim \|u\|_{L^2}^2\|D^{1/2}\partial(|u|^2)\|_{L^2}^2+\|D^{1/2}(|u|^2)\|_{L^2}^2\|u_x\|_{L^2}^2.
\end{aligned}$$

Finally, using (3) we have

$$\begin{aligned}
\left|\int |u|^6 D(|u|^2)\,dx\right| &\lesssim \|\mathscr{F}^{-1}\left[|\mathscr{F}(|u|^2)|\right]\|_{L^\infty}^2\|D^{1/2}(|u|^2)\|_{L^2}^2 \\
&\lesssim \|u\|_{L^2}^4\|D^{1/2}\partial(|u|^2)\|_{L^2}^2.
\end{aligned}$$

Combining these estimates and Corollary 2, we verify that

$$\begin{aligned}
\partial_t E[u(t)] \le &-\frac{|\beta|}{4}\left\|D^{1/2}\partial_x(|u|^2)\right\|_{L^2}^2-3|\beta|\left\|D^{1/2}\,\mathrm{Im}(\overline{u}u_x)\right\|_{L^2}^2 \\
&+C(|\alpha|+|\beta|)|\beta|\|D^{1/2}(|u|^2)\|_{L^2}^2\|u_x\|_{L^2}^2 \\
&+C(\alpha^2+|\alpha|+|\beta|)\left(\|u\|_{L^2}^2+\|u\|_{L^2}^4\right) \\
&\quad\times|\beta|\left(\left\|D^{1/2}\partial_x(|u|^2)\right\|_{L^2}^2+\left\|D^{1/2}\,\mathrm{Im}(\overline{u}u_x)\right\|_{L^2}^2\right).
\end{aligned}$$

Hence, there exist $C_1, C_2>0$ depending only on α,β (bounded as $\beta\to 0$) such that if $\|u\|_{L^2}\le C_1^{-1}$, then

$$\partial_t E[u] \;\le\; -\frac{|\beta|}{8}\left\|D^{1/2}\partial_x(|u|^2)\right\|_{L^2}^2+C_2|\beta|\|D^{1/2}(|u|^2)\|_{L^2}^2\|u_x\|_{L^2}^2$$

This inequality and Lemma 3, together with the L^2 equality, imply that if $\|u(0)\|_{L^2}\le \min\{C_0^{-1},C_1^{-1}\}$,

$$\begin{aligned}
\partial_t\left(E[u(t)]+\|u(t)\|_{L^2}^2\right) \le &-\frac{|\beta|}{8}\left\|D^{1/2}\partial_x(|u(t)|^2)\right\|_{L^2}^2 \\
&+2C_2|\beta|\|D^{1/2}(|u(t)|^2)\|_{L^2}^2\left(E[u(t)]+\|u(t)\|_{L^2}^2\right)
\end{aligned}$$

By the Gronwall inequality, we obtain the desired H^1 a priori bound:

$$\begin{aligned}
&\|u(t)\|_{H^1}^2 + \frac{|\beta|}{4}\int_0^t \left\|D^{1/2}\partial_x(|u(\tau)|^2)\right\|_{L^2}^2 d\tau \\
&\quad\le 2\Big(E[u(t)] + \|u(t)\|_{L^2}^2 + \frac{|\beta|}{8}\int_0^t \left\|D^{1/2}\partial_x(|u(\tau)|^2)\right\|_{L^2}^2 d\tau\Big) \\
&\quad\le 2\Big(E[u(0)] + \|u(0)\|_{L^2}^2\Big)\exp\Big[2C_2|\beta|\int_0^t \|D^{1/2}(|u(\tau)|^2)\|_{L^2}^2 d\tau\Big] \\
&\quad\le 4\|u(0)\|_{H^1}^2 \exp\Big[2C_2\|u(0)\|_{L^2}^2\Big].
\end{aligned}$$

For the lower bound of $\|u(t)\|_{L^2}^2$, we first note that $u(0) = 0$ implies $u(t) \equiv 0$ by the L^2 equality. We thus assume $u(0) \neq 0$, and consider the differential inequality for $\|u(t)\|_{L^2}^{-2}$. By Corollary 2 and the estimate (2), we see

$$\begin{aligned}
\partial_t\|u(t)\|_{L^2}^{-2} &= |\beta|\|u(t)\|_{L^2}^{-4}\left\|D^{1/2}(|u(t)|^2)\right\|_{L^2}^2 \\
&\le C|\beta|\|u(t)\|_{L^2}^{-2}\left\|D^{1/2}\partial_x(|u(t)|^2)\right\|_{L^2},
\end{aligned}$$

as long as $\|u(t)\|_{L^2} > 0$. Applying the H^1 a priori estimate shown above, we have

$$\begin{aligned}
\|u(t)\|_{L^2}^{-2} &\le \|u(0)\|_{L^2}^{-2}\exp\Big[C|\beta|\int_0^t \left\|D^{1/2}\partial_x(|u(\tau)|^2)\right\|_{L^2} d\tau\Big] \\
&\le \|u(0)\|_{L^2}^{-2}\exp\Big[C|\beta|^{1/2}t^{1/2}\Big(|\beta|\int_0^t \left\|D^{1/2}\partial_x(|u(\tau)|^2)\right\|_{L^2}^2 d\tau\Big)^{1/2}\Big] \\
&\le \|u(0)\|_{L^2}^{-2}\exp\Big[4C\|u(0)\|_{H^1}e^{C_2\|u(0)\|_{L^2}^2}|\beta|^{1/2}t^{1/2}\Big].
\end{aligned}$$

This completes the proof of Theorem 1. □

Acknowledgements The first author N.K is partially supported by JSPS KAKENHI Grant-in-Aid for Young Researchers (B) (16K17626). The second author Y.T is partially supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) (17H02853).

References

1. Dysthe, K.B., Pécseli, H.L.: Non-linear Langmuir wave modulation in collisionless plasma. Plasma Physics **19**, 931–943 (1977).
2. Kishimoto, N., Tsutsumi, Y.: Well-posedness of the Cauchy problem for the kinetic DNLS on **T**. In preparation.
3. Mjølhus, E., Wyller, J.: Alfvén solitons. Physica Scripta **33**, 442–451 (1986).
4. Mjølhus, E., Wyller, J.: Nonlinear Alfvén waves in a finite-beta plasma. J. Plasma Physics **40**, 299–318 (1988).

Correction to: A priori bounds for the kinetic DNLS

Nobu Kishimoto and Yoshio Tsutsumi

Correction to:
Chapter "A priori bounds for the kinetic DNLS" in:
D. R. Wood et al. (eds.), *2019–20 MATRIX Annals*,
MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_63

The original version of this chapter was revised in online version, the following correction has been incorporated: The author name "Federico Cacciafesta" has been changed to "Nobu Kishimoto" in Chapter 63. The chapter has been updated with the changes.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-030-62497-2_63

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_64

Correction to: Connected sum decompositions of high-dimensional manifolds

Imre Bokor, Diarmuid Crowley, Stefan Friedl, Fabian Hebestreit, Daniel Kasprowski, Markus Land, and Johnny Nicholson

Correction to:
Chapter "Connected sum decompositions of high-dimensional manifolds" in: D. R. Wood et al. (eds.), *2019–20 MATRIX Annals*, MATRIX Book Series 4, https://doi.org/10.1007/978-3-030-62497-2_1

In the original version of the book, the following belated correction has been incorporated: The author name has been changed from "Kasproswki" to "Kasprowski" in the Chapter 1. Both the Book and the chapter have been updated with the change.

The updated version of this chapter can be found at https://doi.org/10.1007/978-3-030-62497-2_1

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_65

Correction to: The Penrose and the Taylor–Socolar tilings, and first steps to beyond

Robert V. Moody

Correction to:
Chapter 53 in: D. R. Wood et al. (eds.),
***2019–20 MATRIX Annals*, MATRIX Book Series 4,**
https://doi.org/10.1007/978-3-030-62497-2_53

The original version of this chapter was revised in online version, An incorrect 53rd chapter title in A++ has been corrected. The correct title is "The Penrose and the Taylor–Socolar tilings, and first steps to beyond".

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-030-62497-2_53

D. R. Wood et al. (eds.), *2019-20 MATRIX Annals*, MATRIX Book Series 4,
https://doi.org/10.1007/978-3-030-62497-2_66