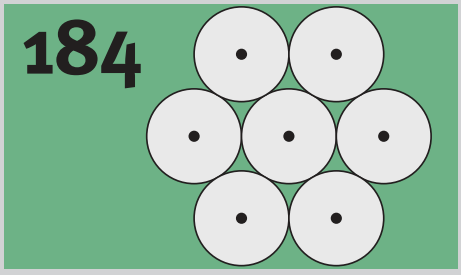
Inhoud | Contents

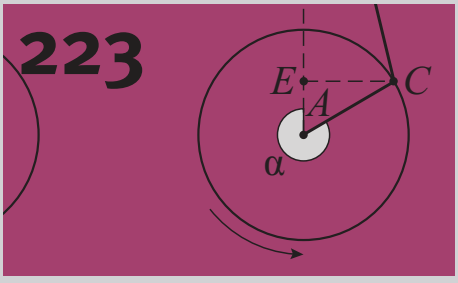
Artikelen | Features

Rubrieken | Departments

Uitgelicht



Een recente doorbraak in het bewijzen van de stelling over de meest efficiënte bolstapeling, wordt beschreven door David de Laat en Frank Vallentin.



Het examen vwo wiskunde B wordt besproken door Wim Caspers.

Colofon

Het Nieuw Archief voor Wiskunde is een uitgave van het Koninklijk Wiskundig Genootschap en verschijnt vier maal per jaar. Het tijdschrift richt zich op eenieder die zich beroepsmatig met wiskunde bezighoudt, als academisch of industrieel onderzoeker, student, leraar, journalist of beleidsmaker. Het stelt zich ten doel te berichten over ontwikkelingen in de wiskunde in het algemeen en in de Nederlandse wiskunde in het bijzonder.

ISSN 0028-9825

Adres
Nieuw Archief voor Wiskunde
Centrum Wiskunde & Informatica
Postbus 94079
1090 GB Amsterdam
tel. 020-5924288
redactie@nieuwarchief.nl
www.nieuwarchief.nl

Hoofdredactie
Jan van Neerven (j.m.a.m.vanneerven@tudelft.nl)

Eindredactie/Bladmanager
Fred Drissen (fred@nieuwarchief.nl)

Redactie
Hans Cuypers (TU/e), Jan Draisma (TU/e), Geertje Hek (UvA), Jan Hogendijk (UU), Teun Koetsier (VU), Sjoerd Rienstra (TU/e), Hans Sterk (TU/e), Mark Timmer (UT)

Problemenrubriek
Gabriele Dalla Torre (UL), Christophe Debry (KU Leuven), Jinbi Jin (UL), Marco Streng (UL), Wouter Zomervrucht (UL)

Bureauredactie
Lætitia Diemer

Illustraties
Ryu Tajiri, Amsterdam

Vormgeving
Susanne Laws (omslag, begeleiding binnenwerk)
Kitty Molenaar (ontwerp)

Druk
Drukkerij Ten Brink, Meppel

Advertenties
advertenties@nieuwarchief.nl

Koninklijk Wiskundig Genootschap
www.wiskgenoot.nl

Platform Wiskunde Nederland
www.platformwiskunde.nl

European Mathematical Society
www.euro-math-soc.eu

International Mathematical Union
www.mathunion.org

Koninklijk Wiskundig Genootschap

Het Koninklijk Wiskundig Genootschap (KWG) is een landelijke vereniging van beoefenaren van de wiskunde. In 1778 opgericht onder de zinspreuk: 'Een onvermoeide arbeid komt alles te boven' is het 's werelds oudste nationale wiskundegenootschap. Leden ontvangen het Nieuw Archief voor Wiskunde als onderdeel van hun lidmaatschap.

Bestuur
Erik van den Ban (UU, voorzitter), Joke Blom (CWI, penningmeester), Sonja Cox (UvA, secretaris), Fetsje Bijma (VU), Theo van den Bogaart (HU), Geurt Jongbloed (TUD), André Ran (VU), Herman te Riele (CWI), Mark Veraar (TUD)

Secretariaat
Joke Blom, CWI
Postbus 94079, 1090 GB Amsterdam
wiskgenoot@wiskgenoot.nl
www.wiskgenoot.nl

Ledenadministratie KWG / Abonnementen
Postbus 1064, 7940 KB Meppel
admin@wiskgenoot.nl, tel. 085-0160252

Contributie
De contributie voor gewone leden bedraagt in 2016 € 90 per jaar. Voor gepensioneerden en werklozen € 45. Voor studenten, aio's en oio's van een Nederlandse universiteit of hbo-opleiding € 35. Studenten en pas afgestudeerden kunnen een jaar lang gratis lid worden. Voor leden van de wiskundige vakverenigingen VvS+OR, NVvW en NVORWO geldt het gereduceerd tarief van € 70. Leden die het Nieuw Archief voor Wiskunde niet willen ontvangen, betalen slechts € 50 contributie. Losse nummers zijn te bestellen voor € 20.
Instituten in Nederland en buitenland kunnen zich abonneren op het Nieuw Archief voor Wiskunde voor € 100 (Nederland), € 110 (Europa) of € 112,50 (buiten Europa).

Members of foreign societies
Members of *SIAM*, the *Société Mathématique de France*, the *Gesellschaft für Angewandte Mathematik und Mechanik*, the *Deutsche Mathematiker-Vereinigung*, and of the *American*, *Australian*, *Belgian*, *Indian*, *London*, *Spanish*, and *South-African Mathematical Societies* living outside of the Netherlands pay € 70 instead of the full membership fee of € 90 a year. Conversely, these foreign societies, as well as the VvS+OR and the NVORWO, have reduced membership fees for KWG members. A postage charge of € 10 is added for members living abroad but in Europe, and a charge of € 12,50 for members living outside of Europe.

Exchange subscriptions
Koninklijk Wiskundig Genootschap Library
Centrum Wiskunde & Informatica
P.O. Box 94079, NL-1090 GB Amsterdam
KWG-exchange@cwi.nl

Informatie voor auteurs

Kopij
Het Nieuw Archief voor Wiskunde is een gerediger tijdschrift. Kopij dient per e-mail te worden aangeboden aan de hoofdredactie. Uitgebreide informatie en auteursinstructies kunt u vinden op www.nieuwarchief.nl.

Volgende nummers

nummer	*maand verschijnen*	*deadline kopij*
17(4)	december 2016	verstreken
18(1)	maart 2017	01–11-2016
18(2)	juni 2017	01–02-2017
18(3)	september 2017	01-052017

Instituutsleden

De publicaties van het Koninklijk Wiskundig Genootschap worden mede mogelijk gemaakt door de bijdragen van de volgende instituten.

Centrum Wiskunde & Informatica

Rijksuniversiteit Groningen

Universiteit van Amsterdam

Universiteit Leiden

Vrije Universiteit

Universiteit Utrecht

Technische Universiteit Eindhoven

Eurandom

Technische Universiteit Delft

Universiteit Twente

Radboud Universiteit Nijmegen

Freudenthal Instituut

Op het omslag

Het vermoeden van Kepler zegt dat de meest voor de hand liggende stapeling van bollen in de ruimte ook de meest efficiënte is. Lees het artikel van Freek Wiedijk, Herman Geuvers en Josef Urban op pagina 177 van dit themanummer over 'Wiskunde en computers'.

Wiskunde en computers

Dit themanummer van het *Nieuw Archief voor Wiskunde* staat in het teken van het gebruik van computers in de wiskunde. Natuurlijk worden computers al decennia gebruikt in de wiskunde om berekeningen en simulaties uit te voeren. Sterker nog, numerieke methoden en scientific computing vormen een centraal onderdeel van de toegepaste wiskunde. Maar dit themanummer gaat over een andere combinatie van computers en wiskunde, namelijk het gebruik van computers voor het controleren en vinden van bewijzen van wiskundige stellingen.

Het beroemdste voorbeeld van een computer-ondersteund bewijs is het bewijs van de vierkleurenstelling dat Appel en Haken in 1976 gaven. Veertig jaar later is er nog steeds geen alternatief bewijs bekend dat geen gebruik maakt van een computer. Nu kun je van alles vinden van computers in de wiskunde, bijvoorbeeld dat de geproduceerde bewijzen onbetrouwbaar zijn omdat ze niet met de hand (en mensenhersenen) te controleren zijn, of dat zo'n bewijs juist een gebrek aan begrip blootlegt. Sommige wiskundigen vinden computer-ondersteunde bewijzen daarom zelfs controversieel. Deels is dit een kwestie van smaak, deels is dit te verklaren vanuit de gedachte 'onbekend maakt onbemind'. Een voorbeeld daarvan is de scepsis van veel wiskundigen over de betrouwbaarheid van interval-aritmetiek gebaseerd op 'floating point'-getallen, terwijl de IEEE-standaard voor het rekenen daarmee juist een van de best gecontroleerde standaarden ter wereld is. Dit themanummer biedt een inkijkje in de mogelijkheden en beperkingen van het gebruik van computers bij het bewijzen van stellingen. Daarbij wordt ook duidelijk hoe computer-ondersteunde bewijzen en bewijsverificatie gecombineerd kunnen worden met pen-en-papier- en krijtjes-wiskunde.

Dit nummer bevat vier artikelen over het gebruik van interactieve bewijsassistenten in de wiskunde. Mahboubi geeft een overzicht van het gebruik van computers in het controleren van wiskundige bewijzen, met de nadruk op de recente formalisering van het bewijs van de oneven-orde-stelling van Feit–Thompson, een onderdeel van de classificatie van eindige enkelvoudige groepen. Rijke en Spitters geven een overzicht van de homotopie-type-theorie, een nieuwe grondslag voor de wiskunde en een uitbreiding van de verzamelingenleer. Deze theorie sluit beter aan bij de praktijk van de wiskunde en is al succesvol gebruikt voor het construeren en de verificatie van abstracte bewijzen. Wiedijk, Geuvers en Urban beschrijven de verificatie van Hales' bewijs van het Kepler-vermoeden over de meest efficiënte bolstapeling. Dit bewijs is gevonden met behulp van een computer, en later is het hele bewijs ook met behulp van een computer gecontroleerd.

Zeer recent is de stelling over de meest efficiënte bolstapeling ook bewezen voor dimensies 8 en 24. De Laat en Vallentin beschrijven deze doorbraak. Ook in dit geval spelen computers een cruciale rol, niet voor het controleren van een groot aantal gevallen, maar voor het vinden van een goede kandidaat voor een 'magische' functie. Met behulp van numerieke wiskunde werden goede criteria voor een kandidaatfunctie gevonden. Daarna werd met behulp van traditionele wiskunde zo'n functie geconstrueerd. Een verslag van deze zoektocht is te vinden in het interview met Cohn, Kumar, Miller en Viazovska. Cohen geeft een overzicht van het gebruik van computers in het wiskundeonderwijs. De Carli Silva, de Oliveria Filho en Sato beschrijven hoe met behulp van *flag algebras* vragen uit de grafentheorie kunnen worden gereduceerd tot optimalisatieproblemen die efficiënt met de computer opgelost kunnen worden.

Kalies en Vandervorst beschrijven Conley-theorie, een methode om met behulp van algebra, topologie en combinatoriek computationele algoritmes te maken waarmee stellingen over het globale gedrag van dynamische systemen kunnen worden bewezen. Van den Berg, Groothedde en Sheombarsing beschrijven een algemene methodologie voor computerbewijzen in de dynamische systemen (simulatie + contractie = bewijs). Ten slotte beschrijft Collins het gebruik van wiskundig rigoureus gevalideerde numerieke methoden voor het bewijzen van stellingen over zogenaamde hybride systemen (met zowel een discrete als een continue factor).

Jan Bouwe van den Berg, *Vrije Universiteit Amsterdam*
Bas Spitters, *Aarhus University*
Frank Vallentin, *Universität zu Köln*
gastredacteuren

Agenda

| Upcoming Events

september 2016

12–16 september 2016
❑ **Extremes and Risks in Higher Dimensions**
Workshop mede georganiseerd door Juan-Juan Cai (TUD) en Pasquale Cirillo (TUD).
plaats *Lorentz Center@Oort*
info *lorentzcenter.nl*

16 september 2016
❑ **Finale Nederlandse Wiskunde Olympiade**
38 onderbouwleerlingen, 44 vierdeklassers en 40 vijfdeklassers zijn door naar de finale van de Nederlandse Wiskunde Olympiade 2016.
plaats *TU Eindhoven*
info *wiskundeolympiade.nl*

22 september 2016
❑ **Fundamentals of Networks**
Satellietsessie georganiseerd door Remco van der Hofstad (TU/e), Frank den Hollander (UL) en Diego Garlaschelli (UL), als onderdeel van de Conference on Complex Systems 2016.
plaats *Beurs van Berlage, Amsterdam*
info *www.eurandom.nl*

23 september 2016
❑ **25ste wiskundetoernooi**
Wiskundetoernooi voor teams van middelbare scholieren. Een team bestaat uit 4 of 5 leerlingen uit 5–6 vwo en een begeleider. Thema is dit jaar 'Geodesie en Cartografie'.
plaats *Radboud Universiteit Nijmegen*
info *ru.nl/wiskundetoernooi*

25–30 september 2016
❑ **WIN-E2: Women in Numbers Europe 2**
Workshop voor vrouwen die actief zijn in de getaltheorie.
plaats *Lorentz Center@Snellius, Leiden*
info *lorentzcenter.nl*

30 september 2016
❑ **Mini-Workshop on Symplectic Geometry**
Mini-workshop voor onderzoekers, PhD-studenten en masterstudenten, werkzaam in symplectische meetkunde of aanverwante gebieden.
plaats *VU Amsterdam*
info *www.gqt.nl*

oktober 2016

1 oktober 2016
❑ **Amsterdam Science Park Open dag**
Jaarlijkse open dag van Amsterdam Science Park dé gelegenheid om een kijkje te nemen achter de schermen.
plaats *Amsterdam Science Park*
info *amsterdamsciencepark.nl*

1 oktober 2016
❑ **Junior Wiskunde Olympiade**
Jaarlijkse landelijke wiskundewedstrijd voor leerlingen uit de onderbouw van havo en vwo.
plaats *Vrije Universiteit, Amsterdam*
info *few.vu.nl*

5 oktober 2016
❑ **Dag van de Leraar**
Jaarlijkse internationale Dag van de Leraar waarop tal van activiteiten voor leraren worden georganiseerd, waaronder de verkiezing van de Leraren van het Jaar.
plaats *overal*
info *www.dagvandeleraar.nl*

5–7 oktober 2016
❑ **Eenenveertigste Woudschotenconferentie**
Jaarlijkse conferentie georganiseerd door de Nederlands/Vlaamse Werkgemeenschap Scientific Computing (WSC).
plaats *Conferentiecentrum Woudschoten, Zeist*
info *wsc.project.cwi.nl/conferentieN.php*

10–14 oktober 2016
❑ **Transformations in Statistical Mechanics: Pathologies and Remedies**
Workshop georganiseerd door Henk Broer (RUG), Frank den Hollander (LU), Frank Redig (TUD) en Wioletta Ruszel (TUD).
plaats *Lorentz Center@Snellius*
info *lorentzcenter.nl*

24–26 oktober 2016
❑ **Workshop Data Driven Operation Management**
Workshop van het STAR-cluster.
plaats *Eurandom, Eindhoven*
info *www.eurandom.nl*

Suggesties voor deze agenda zijn welkom.
agenda@nieuwarchief.nl

24–28 oktober 2016
❑ **Logics, Decisions, and Interactions**
Workshop mede georganiseerd door Elias Tsakas (UM) en Alessandra Palmigiano (TUD).
plaats *Lorentz Center@Snellius*
info *lorentzcenter.nl*

31 oktober – 4 november 2016
❑ **HighLight: High-Security Lightweight Cryptography**
Workshop mede georganiseerd door Lejla Batina, Joan Daemen, Peter Schwabe (RU) en Tanja Lange (TU/e).
plaats *Lorentz Center@Snellius*
info *lorentzcenter.nl*

november 2016

5 november 2016
❑ **Jaarvergadering NVvW**
Jaarlijkse vergadering en studiedag van de Nederlandse Vereniging van Wiskundeleraren.
plaats *Ichthus College, Veenendaal*
info *nvvw.nl*

7–9 november 2016
❑ **YEQT X workshop**
Tiende editie van de Young European Queueing Theorists (YEQT) workshop.
plaats *Eurandom, Eindhoven*
info *www.eurandom.nl*

9 november 2016
❑ **ICAB conference 2016**
Jaarlijkse conferentie voor iedereen die werkt in het academisch bètaonderwijs, met als thema 'Diversity in teaching'.
plaats *TU Delft*
info *icab.nl*

10 november 2016
❑ **Science Night**
Inspiratieavond voor docenten onderbouw van het voortgezet onderwijs.
plaats *NEMO, Amsterdam*
info *nemosciencemuseum.nl*

18 november 2016
❑ **Voorronde A-lympiade**
Voorronde van de jaarlijkse competitie voor leerlingen uit de bovenbouw havo/vwo, die wiskunde A in hun pakket hebben.
plaats *eigen school*
info *www.fi.uu.nl/alympiade*

18 november 2016
❑ **Wiskunde B-dag 2016**
Een eendaagse opdracht voor op school voor teams uit 5 havo en 5/6 vwo met wiskunde B in hun profiel.
plaats *deelnemende scholen*
info *uu.nl/wisbdag*

december 2016

9 december 2016
❑ **L. E. J. Brouwer, fifty years later**
Symposium georganiseerd door het KWG ter ere van de 50ste sterfdag van L. E. J. Brouwer.
plaats *Science Park, Amsterdam*
info *wiskgenoot.nl*

12–16 december 2016
❑ **Guided Tour through Random Media**
Workshop georganiseerd door Remco van der Hofstad (TU/e), Aernout van Enter (RUG) en Marek Biskup (UCLA Los Angeles).
plaats *Eurandom, Eindhoven*
info *www.eurandom.nl*

januari 2017

14 januari 2017
❑ **Wintersymposium KWG 2017**
Jaarlijks symposium van het KWG met dit jaar als thema 'Raakvlakken tussen wiskunde en informatica'.
plaats *Academiegebouw, Utrecht*
info *wiskgenoot.nl*

19–20 januari 2017
❑ **35ste Panama-conferentie**
Jaarlijkse conferentie die zich richt op alle aspecten van leren en onderwijzen van rekenen-wiskunde. Dit jaar is het thema 'Goed en fout in het reken-wiskundeonderwijs'.
plaats *Congrescentrum Koningshof, Veldhoven*
info *www.fi.uu.nl/panama*

23–25 januari 2017
❑ **16th Winter School Mathematical Finance**
Winterschool over financiële wiskunde met onder andere minicursussen van Damir Filipović (Lausanne) en Jan Kallsen (Kiel).
plaats *Congrescentrum De Werelt, Lunteren*
info *staff.science.uva.nl/~spreij/winterschool/winterschool.html*

23–27 januari 2017
❑ **Studiegroep Wiskunde met de Industrie**
Jaarlijkse studiegroep waarin wiskunde en industrie samen problemen oplossen, dit jaar georganiseerd door UvA en CWI.
plaats *Amsterdam Science Park*
info *swi-wiskunde.nl*

februari 2017

3–4 februari 2017
❑ **Nationale Wiskunde Dagen**
23ste editie van de tweedaagse conferentie voor wiskundeleraren.
plaats *Noordwijkerhout*
info *uu.nl/nationale-wiskunde-dagen*

april 2017

11–12 april 2017
❑ **NMC 2017**
53ste Nederlands Mathematisch Congres, jaarlijkse bijeenkomst van wiskundig Nederland.
plaats *Utrecht*
info *wiskgenoot.nl*

Nieuws

| News

Deze rubriek is een kroniek van wiskundige activiteiten in Nederland. Toekomstige activiteiten worden aangekondigd en van voorbije activiteiten wordt verslag gedaan. Wilt u uw aankondiging of verslag in deze rubriek geplaatst zien? Stuur ons dan uw bijdrage van ± 350 woorden, zo mogelijk met illustratie. De redactie behoudt zich het recht voor berichten te weigeren of in te korten.

nieuws@nieuwarchief.nl

Driemaal brons op Internationale Wiskunde Olympiade

Bij de Internationale Wiskunde Olympiade in Hong Kong hebben drie Nederlandse leerlingen een bronzen medaille behaald. Levi van de Pol (14 jaar, Ichthus College, Veenendaal), Pim Spelier (16 jaar, Christelijk Gymnasium Sorghvliet, Den Haag) en Gabriel Visser (18 jaar, Stedelijk Gymnasium, Schiedam) wisten alle drie genoeg punten te behalen voor eremetaal. De andere drie teamleden, Wietze Koops (15 jaar, RSG Stad & Esch Lyceum, Meppel), Reinier Schmiermann (14 jaar, Stedelijk Gymnasium, 's-Hertogenbosch) en Erik van Cappellen (17 jaar, Johannes Fontanus College, Barneveld), losten allemaal minstens één van de zes pittige wiskundeopgaven op en behaalden hiermee een eervolle vermelding.

De Internationale Wiskunde Olympiade vond van 9 tot 16 juli plaats in Hong Kong. Voor de Nederlandse deelnemers was het de eerste wedstrijd op het hoogste internationale niveau. Coach Birgit van Dalen: "Dit team heeft hard gewerkt voor deze medailles en eervolle vermeldingen. De komende jaren kunnen we nog veel van deze talenten verwachten." *www.imo2016.org*

Ontwikkelingen rondom de rekentoets

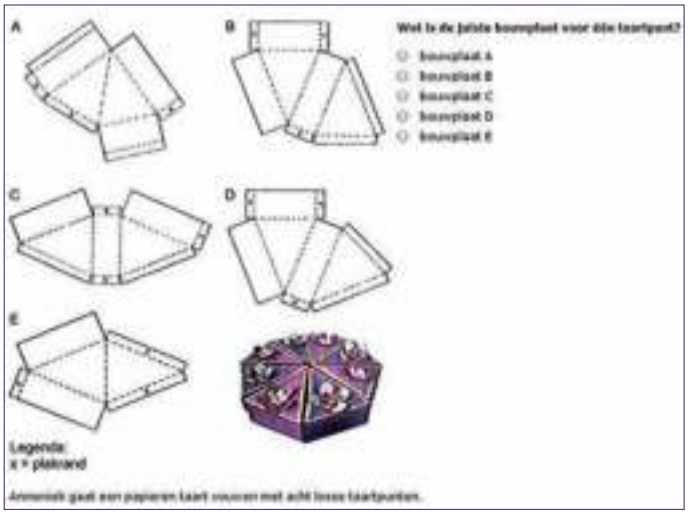
Dit jaar telde het resultaat van de rekentoets voor het eerst mee bij de diplomeringsbeslissing in het vwo: leerlingen moesten minimaal een 5 voor de toets halen om hun diploma in ontvangst te mogen nemen. Ze hebben hiertoe vier pogingen gehad, waarbij het hoogste cijfer telt.

De laatste gegevens van het ministerie geven aan dat op het vwo 98 procent van de leerlingen de benodigde 5 of hoger had. Aangezien de laatste poging nog niet in de statistieken is meegenomen, en sommige leerlingen wellicht ook los van de rekentoets gezakt zouden zijn, zal het percentage leerlingen dat gezakt is vanwege de rekentoets dus nog wel onder de 2 procent liggen. De vangnetconstructie, die het mogelijk maakte om het minimaal benodigde cijfer te verlagen naar een 4 als de resultaten tegen zouden vallen, is daarom niet in werking gesteld.

Het resultaat op de rekentoets telde op het vwo sinds dit jaar ook mee voor het al dan niet behalen van de kwalificatie cum laude. Bovendien is voor het vwo besloten de rekentoets per komend examenjaar deel te laten uitmaken van de kernvakkenregeling: leerlingen die een 5 voor de rekentoets scoren moeten dan een voldoende scoren voor de kernvakken Nederlands, Engels en wiskunde, terwijl leerlingen met een voldoende voor de rekentoets nog een 5 voor één van de kernvakken mogen halen. Een motie van de SP en CU om de toevoeging van de rekentoets aan de kernvakkenregeling tegen te houden is niet aangenomen.

Op de havo is de toets minder goed gemaakt: slechts 80 procent had de benodigde 5. Bij het vmbo was dit 62 procent, 74 procent en 93 procent, bij de bb-, kb- en gt-leerwegen, respectievelijk. Gelukkig voor de leerlingen van het vmbo en de havo was echter al eerder besloten dat hun resultaten voor de rekentoets niet mee zouden tellen voor de zak-slaagregeling. Op basis van de huidige gegevens is bovendien nu ook al besloten dat hetzelfde zal gelden voor het schooljaar 2016–2017. De huidige leerlingen in vmbo 3 en havo 4 'profiteren' dus van de tegenvallende resultaten van hun voorgangers. Volgend jaar zal besloten worden of er per 2017–2018 verandering in de situatie zal komen en de rekentoets ook voor het vmbo en/of de havo mee zal gaan tellen.

Verder is na lange tijd van geheimhouding nu een groot gedeelte van de dit jaar gebruikte opgaven openbaar gemaakt via https://opgavenetalage.facet.onl. Ongeveer 90 procent van de sommen van de niveaus 2F en 3F is openbaar gemaakt, terwijl zo'n 10 procent geheimgehouden wordt om in de toekomst nog te kunnen gebruiken als controleopgaven. *Mark Timmer*



Cryptografen winnen Internet Defense Prize 2016 voor 'NewHope'

De cryptografen Léo Ducas van het Centrum Wiskunde & Informatica en Peter Schwabe van de Radboud Universiteit hebben de internationale Internet Defense Prize 2016 gewonnen. Zij kregen de prijs samen met hun co-auteurs Erdem Alkim (Ege University, Turkije) en Thomas Pöppelmann (Infineon Technologies AG, Duitsland) voor hun artikel 'Post-Quantum Key Exchange – A New Hope'. De prijs werd uitgereikt op 10 augustus tijdens het 25ste USENIX Security Symposium in Austin, Texas. Facebook stelde de Internet Defense Prize in 2014 in, met USENIX als partner. De prijs bestaat uit 100.000 dollar.

"De informatiebeveiligingsindustrie houdt een race tegen de klok om sneller te innoveren dan de tegenstanders die consumenten en bedrijven schade willen toebrengen", schrijft Facebook. "De meeste security research richt zich echter vooral op aanvallende noviteiten op hackersgebied die weinig invloed hebben op het leven van de meeste mensen". Om ook de andere kant te stimuleren is de Internet Security Prize ingesteld om onderzoekers te belonen die een werkend prototype combineren met aanzienlijke bijdrages aan de veiligheid van het Internet — met name op het gebied van bescherming en verdediging, aldus Facebook.

Het winnende team stelde een verbeterd cryptosysteem voor, 'NewHope', dat is ontworpen om aanvallen van toekomstige kwantumcomputers te weerstaan. Zulke kwantumcomputers zouden een verwoestend effect hebben op de veiligheid van onze huidige protocollen — een komst die soms wordt aangeduid als een 'Cryptocalypse'. NewHope kan bijvoorbeeld worden geïntegreerd in TLS en HTTPS, twee security-protocollen die worden gebruikt door webbrowsers. Dit werd onlangs gedaan door Google, als een experiment voor post-kwantumbeveiliging. Hierover verscheen een artikel in *Wired*.

Facebook laat weten: "Voortbouwend op eerdere studies identificeerde dit nieuwe onderzoek een gepastere foutverdeling en een beter passend reconciliation-mechanisme, analyseerde het de robuustheid van het cryptosysteem tegen aanvallen van kwantumcomputers en identificeert het een verdedigingsmogelijkheid tegen mogelijke achterdeurtjes en alles-voor-de-prijs-van-één-aanvallen. Met deze maatregelen was het team in staat om de security-parameter met meer dan 100 procent te verhogen, de communicatieoverhead met meer dan de helft te verminderen en de rekensnelheid in portable C-implementatie en de huidige Intel CPU's aanzienlijk te verhogen — en dit alles met bescherming tegen timingaanvallen". *Annette Kik*

Beter kiessysteem voor presidentsverkiezingen

Het systeem van de Amerikaanse en Franse presidentsverkiezingen maakt ze kwetsbaar voor spoilers, kandidaten die één tegen één nooit zouden winnen, maar die wel de einduitslag veranderen. Nobelprijswinnaar Eric Maskin, expert in speltheorie, pleit voor het Condorcet-stemsysteem, waarbij iedere stemmer zijn eigen ranglijstje van kandidaten invult. Weliswaar is dan een eenduidige uitslag niet gegarandeerd, maar volgens Maskin is dit een theoretisch bezwaar dat in de praktijk zelden of nooit zal optreden. Dit zegt hij in een interview met Kennislink.

Er zijn allerlei verkiezingssystemen, en geen ervan is in alle opzichten ideaal, maar volgens Maskin is de minst slechte van allemaal het Condorcet-systeem, waarbij elke kiezer een ranglijst van favorieten opgeeft. Als er veel kandidaten (of politieke partijen) zijn, hoef je ze niet allemaal in je ranglijst op te nemen; je geeft bijvoorbeeld alleen je 1ste, 2de en 3de favoriet aan. Maskin: "Je wilt dat kiezers die ranglijst samenstellen op basis van zinnige informatie. Dan is een ranglijst van twintig kandidaten voor de meesten niet realistisch." *kennislink.nl*

Beter rekenen in groep 8

Zwakke rekenaars uit groep 8 presteren beter als ze bij ingewikkelde sommen de berekeningen opschrijven. Dat ontdekte psycholoog Marije Fagginger Auer. De specialist in methodologie en statistiek ziet perspectief: "Na een training kiezen deze leerlingen vaker voor deze oplossingsstrategie."

Het rekenniveau van leerlingen in groep 8 is de afgelopen decennia op bepaalde onderdelen sterk gedaald, in het bijzonder bij vermenigvuldigen en delen met meercijferige getallen (bijvoorbeeld 23×56 en $544 \div 34$). Dit hangt samen met veranderingen in de strategieën die leerlingen gebruiken om opgaven op te lossen: leerlingen beantwoorden vaker opgaven zonder daarbij een berekening te noteren en ze maken daarbij veel fouten. Fagginger Auer: "We wilden meer inzicht krijgen in deze ontwikkelingen en in mogelijke oplossingen. Ons onderzoek laat zien hoe belangrijk het is om de oplossingsstrategieën van kinderen te bestuderen. Ook

Rectificatie

In het juninummer staat bij de foto van het Thomas Stieltjes Instituut op pagina 109 een verkeerde naam: op de achterste rij staat niet Chris Klaassen maar Michel Dekking.

toont het aan dat kinderen baat kunnen hebben bij het opschrijven van hun berekeningen, vooral de kwetsbare groep zwakkere rekenaars."

Fagginger Auer gaat haar in Leiden verworven expertise inzetten bij de Vereniging Hogescholen in Den Haag. *universiteitleiden.nl*

Tellende peuter is later beter in wiskunde

Het heeft nut om een peuter te leren tellen. Kinderen die op peuterleeftijd vertrouwd zijn met getallen, zijn op latere leeftijd beter in wiskunde. Dat blijkt uit onderzoek van de University of Missouri. Uit eerder onderzoek bleek al dat kinderen in groep 3 die goed zijn met getallen, later beter zijn in de wiskunde die ze kan helpen in hun loopbaan. Maar hoe ze aan die handigheid zijn gekomen, was tot nu toe nauwelijks onderzocht. Daarom volgden wetenschappers een groep kinderen vanaf drie jaar oud en legden ze de peuters allerlei oefeningen met getallen voor. Ze ontdekten dat twee vaardigheden bij peuters essentieel zijn om goed te worden met getallen: kennis van de telwoorden en wat ze betekenen en de mogelijkheid om hele simpele optelsommen te maken. *vakbladvroeg.nl*

Nieuwe app moet studenten leren wiskunde beter te begrijpen

Het Freudenthal Instituut start een miljoenenproject naar de inzet van nieuwe apps in de wiskundelessen op middelbare school en universiteit. "Het klinkt misschien vreemd," vertelt Arthur Bakker van het Freudenthal Instituut, "maar we gaan een stap terug doen om vooruit te komen. De resultaten van het invoeren van digitale hulpmiddelen onderwijs vallen immers tegen. Blijkbaar is te makkelijk gedacht dat plaatjes en filmpjes op de computer het wiskundeonderwijs zouden verbeteren." Niet dat Bakker alle computers en softwareprogramma's wil vervangen door pen en papier en het bord en krijtje, wel wil hij ervoor zorgen dat studenten weer aan het uitschrijven van sommen gaan. "We raken er steeds meer van overtuigd dat de handeling van het schrijven en tekenen een groot aandeel heeft in het leren begrijpen van wiskunde."

Bakker wist recent samen met Amerikaanse collega's het moment bij studenten vast te leggen van het ontstaan van wiskundig inzicht. Dit 'Aha-moment' kon hij destilleren uit de video-opnames en registratie van de oogbewegingen van studenten die zaten te ploeteren op sommen. "Vaak ontstaat het inzicht in een reflectie op een concrete handeling. Zoals het maken van een gebaar, het schetsen van een tekening maar ook het uitschrijven van een som."

Van Amerikaanse scholieren is bekend dat ze in wiskundelessen met grafische rekenmachine net iets minder goed wiskunde leren dan zonder. "Het vermoeden is dat het tempo van uitschrijven en het schetsen een rol speelt bij het ontstaan van het inzicht." Bakker gaat daarom voor Noordhoff een schriftherkenner programmeren die de tablet omtovert tot een schrijfblok voor wiskundeformules. In het project worden studenten met eye-trackers en camera's tijdens het maken van sommen in de gaten gehouden. Zo hoopt Bakker te ontdekken hoe ze hun 'Aha-momenten' krijgen.

Over drie jaar moet Bakkers schriftherkenner in de klas beschikbaar zijn. *dub.uu.nl*

Koninklijk Wiskundig Genootschap

❑ **L. E. J. Brouwerjaar, vijftig jaar later**

Het zal op 2 december precies vijftig jaar geleden zijn dat de beroemde Nederlandse wiskundige en filosoof L. E. J. (Bertus) Brouwer is overleden bij een tragisch verkeersongeluk. Naar aanleiding van zijn vijftigste sterfdag organiseert het KWG een aantal activiteiten.

Zo komt er een speciale uitgave van *Indagationes Mathematicae* gewijd aan het werk van Brouwer en de gevolgen daarvan voor de moderne topologie, logica en filosofie. De gastredacteuren voor deze speciale uitgave, die begin 2017 zal verschijnen, zijn Dirk van Dalen, Jan Willem Klop en Jan van Mill. Er hebben fantastische wiskundigen een bijdrage geleverd, waarvan sommigen ook zullen spreken op het Brouwer Symposium op 9 december in het Science Park Amsterdam (registratie is nog mogelijk, zie wiskgenoot.nl).

Verder heeft er, bij het verschijnen van dit *Nieuw Archief van de Wiskunde*, een Brouwer-middag plaatsgevonden in samenwerking met de Koninklijke Hollandsche Maatschappij der Wetenschappen. Deze middag vond plaats op 2 september in het Hodshonhuis in Haarlem en was gericht op een algemeen publiek. Jaap Korevaar, Alexander Rinnooy Kan en Dirk van Dalen hebben hun licht laten schijnen op 'L. E. J. Brouwer, Haarlems wiskundig wonderkind'. Tijdens deze bijeenkomst is ook het Brouwerarchief overgedragen aan het KWG, het zal worden ondergebracht bij het Noord Hollands Archief in Haarlem (waar tevens de rest van het KWG-archief zich bevindt).

Voor meer informatie zie wiskgenoot.nl, alwaar u zich ook kunt aanmelden voor het Brouwer Symposium op 9 december.

❑ **Oproep voor nominaties ASML Afstudeerprijs voor Wiskunde**

De jaarlijkse ASML Afstudeerprijs voor Wiskunde is verhoogd naar € 5000. Voordrachten voor de prijs kunnen tot uiterlijk 28 september 2016 gericht worden aan secretaris@khmw.nl.

Recent verschenen:

❑ **Indagationes Mathematicae** (www.elsevier.com/locate/indag)

Special Issue 'In memoriam J. G. van der Corput (1890–1975)', Part 2, J. Korevaar, R. Tijdeman en J. J. O. O. Wiegerinck, Volume 27, Issue 2, 2016.

❑ **Epsilon Uitgaven** (www.epsilon-uitgaven.nl)

85. Symmetrie in de deeltjesfysica, Walter D. van Suijlekom, € 17, 2016.

84. Wortels van de Wiskunde, William P. Berlinghoff en Fernando Q. Gouvêa, vertaling Desiree van den Bogaart, Jeanine Daems, € 25, 2016.

Wil Schilders
Platform Wiskunde Nederland
Amsterdam
bureau@platformwiskunde.nl

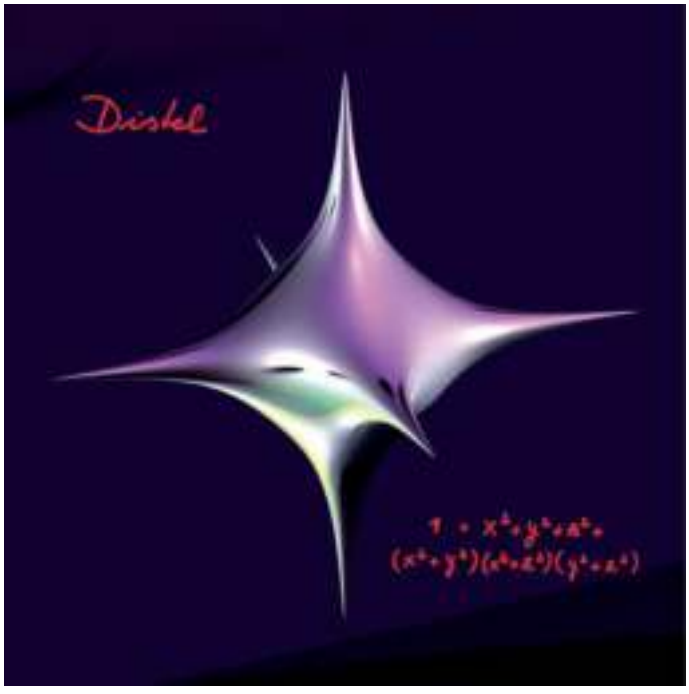
Nieuws Tentoonstelling

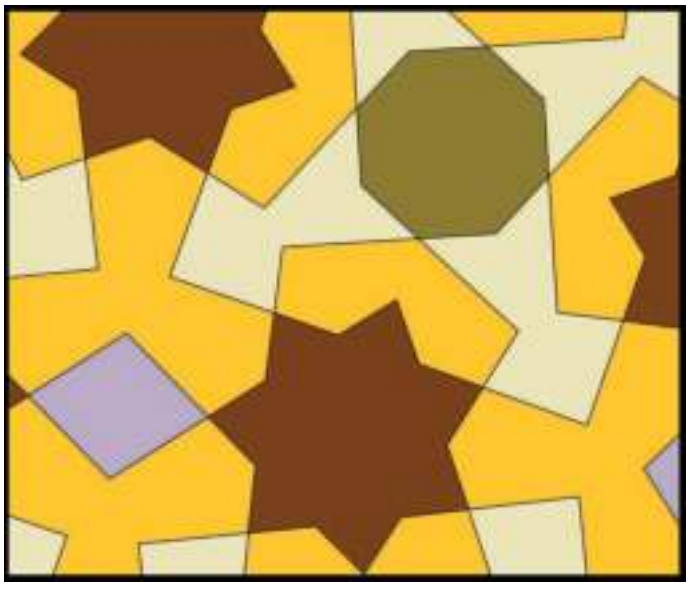
IMAGINARY in Nederland

Van september 2016 tot juli 2017 trekt IMAGINARY door Nederland. IMAGINARY is een tentoonstelling over wiskundige objecten, gecreëerd door het wiskundige onderzoeksinstituut Oberwolfach in Zuid-Duitsland. Het is ontstaan in 2008 tijdens het Duitse 'Jaar van de wiskunde' en is snel uitgegroeid tot een van de meest wijdverspreide reizende wiskunde-tentoonstellingen in de wereld, met meer dan één miljoen bezoekers in 29 landen en 120 steden.

IMAGINARY toont de schoonheid en kunstzinnigheid van wiskunde, en anderzijds een aantal verrassende toepassingen. De tentoonstelling bevat posters van prachtige figuren, zoals de distel en het zeepaardje hieronder, alsmede driedimensionale objecten met een wereldrecord aantal singulariteiten. Ook vlakvullingen, periodisch en a-periodisch, spelen een rol in de tentoonstelling

Kern van de tentoonstelling is het programma SURFER. Het presenteert visualisaties en hun theoretische achtergrond uit de algebraïsche meetkunde, singulariteitentheorie en differentiaalmeetkunde op een attractieve en begrijpelijke wijze. SURFER doet dit alles realtime, zodat bezoekers worden aangetrokken door de mogelijkheid om zelf bij te dragen aan de tentoonstelling. Het programma laat zien hoezeer wiskunde en kunst verweven kunnen zijn. Het stelt gebruikers in staat om begrip op te bouwen, en te experimenteren met de relatie tussen formule en vorm. Het gebruikersinterface van SURFER is simpel. Achter het programma gaat geen didactische theorie of kunstzinnige aanpak schuil, de intentie is om bezoekers te laten delen in de vreugde van creaties binnen de wiskunde en om het vak te bevrijden van zijn vaak droge en moeilijke imago.

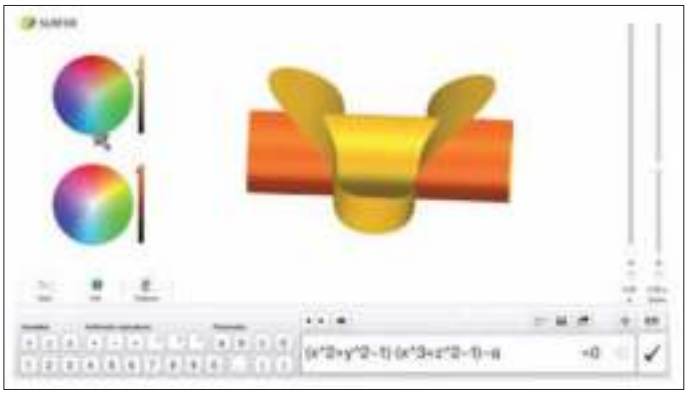




Heel erg bekend zijn de formules en bijbehorende figuren welke zijn geconstrueerd door het nog jonge talent, de Duitse Valentina Galata. Zij putte zich uit in het experimenteren met de SURFER-software, en was in staat om formules te creëren die leidden tot figuren van alledaagse voorwerpen, zoals een theepot (zie onder) of een kopje cappuccino.

De tentoonstelling is bestemd voor een groot publiek dat geïnteresseerd is in de schoonheid, kunstzinnige aspecten, potentie en toepassingen van wiskunde. Er worden ook rondleidingen georganiseerd, met name bedoeld voor middelbaar scholieren in de bovenbouw van vwo en havo.

De tentoonstelling is interactief, men kan de figuren van alle kanten bekijken, formules aanpassen en daarmee de figuren vervormen, kleuren aanpassen. Middels QR-codes kan men op de smartphone allerlei achtergrondinformatie krijgen.



Screenshot van het programma SURFER

Naast de posters is er ook een aantal gipsmodellen te zien, samen met moderne 3D-prints. Er is een tafel met Penrose-betegelingen om zelf in elkaar te puzzelen en diverse touchscreens bieden een scala aan leuke, mooie en uitdagende door wiskunde geïnspireerde toepassingen.

De IMAGINARY-tentoonstelling is prachtig, interessant, leerzaam! Een bezoek meer dan waard, bovendien gratis toegankelijk. Een impressie en allerlei interessante informatie, is te vinden op www.imaginarymaths.nl.

Locaties en data

Eindhoven	19 september – 7 oktober 2016
Enschede	24 oktober – 11 november 2016
Amsterdam	21 november – 11 december 2016
Utrecht	6–24 februari 2017
Leiden	6 maart–24 maart 2017
Groningen	22 mei – 9 juni 2017
Nijmegen	19 juni – 7 juli 2017

IMAGINARY wordt georganiseerd door Platform Wiskunde Nederland, in nauwe samenwerking met de Universiteit Utrecht en het Freudenthal Instituut. Sponsoren en lokale organisatoren zijn vermeld op de bovengenoemde website.

Sponsoring posters
Individuele sponsoring van posters, welke na de tentoonstelling eigendom worden, is mogelijk. Zie www.imaginarymaths.nl voor meer informatie.

$$(4\cdot(x+z+0.55)^2+4y^2+200z^6-0.1)\cdot((2.5x^2+2.5y^2+z^2)^2-(2.5x^2+2.5y^2))\cdot((40\cdot(x-0.65)^2+40y^2+30z^2+2)^2-9\cdot(30z^2+40x-0.65)^2)-8\cdot(0.7x-0.65))-0.4=0$$

$$(0.45x^2+0.45y^2+2\cdot(1.6z-1.39)^3\cdot(1+(1.35z-1.37)))\cdot(x^2+y^2+(z-0.85)^2-0.015)-0.000001=0$$

$$10y^2+10x^2-10-(20z+9)^5=0$$

$$(2x-0.8)^6+(2y+1.15)^6+(3z+1.31)^6-0.0001)\cdot((1.35x+0.9+1.35y)^6+(1.35y+1.2-1.35x)^6+(3z+1.31)^6-0.0001)=0$$

$$(0.7x+1+0.7y+y)^6+(0.7y+0.9-0.7x-x)^6+(3z+0.9)^6-0.0001=0$$

Egbert Rijke
Department of Philosophy
Carnegie Mellon University, Pittsburgh, USA
erijke@andrew.cmu.edu

Bas Spitters
Department of Computer Science
Aarhus University, Denmark
spitters@cs.au.dk

Homotopy type theory and the formalization of mathematics

Egbert Rijke is a PhD student at Carnegie Mellon University, working on homotopy type theory. Bas Spitters is associate professor at Aarhus university, working on type theory, topos theory and proof assistants. Both authors participated in the Univalent Foundations Program.

Formalization of mathematics, that is the complete reduction of mathematical arguments to the axioms, has become more and more feasible due to the development of better computer implementations and advances in languages and algorithms. In this article we focus on one such language, homotopy type theory, which Voevodsky has proposed as a new foundation for all of mathematics. It extends the usual set theoretic foundations with a computational meaning, while at the same time taking the more general notion of space as primitive.

Computer formalization of mathematics

Reasons for formalization

The present refereeing process for mathematical articles, based on social consensus by a small group of experts, does not guarantee that all proofs are flawless. Fields medalist Voevodsky courageously points to an error [23] which existed for a decade in one of his own fundamental papers. One reason for this is that checking proofs for correctness can be boring or hard, and may require a lot of effort and expertise. These issues are addressed by software called 'proof assistants', in which proofs can be programmed and their correctness is verified algorithmically.

Moreover, formalization of mathematics in proof assistants provides new insights and conveys new mathematical structures. It is also a source of pleasure, a computer game, for mathematicians with an appreciation of detail. Finally, it gives peace of mind when ones arguments have been checked very carefully with the aid of a computer.

Figure 1 Cover image of the HoTT book.

Great formalization feats include: a formal proof of the Kepler conjecture [10], the four color theorem [8] and the Feit–Thompson theorem [9] (the beginning of the classification of finite simple groups). Notably, the former two are huge proofs which crucially depend on computer computations which due to their sheer size cannot be checked as a whole by humans.

Importance for computer science

In computer science, ideally one would want the behavior of software to be specified and we would like a formal guarantee that it actually sticks to that specification. Such formal verification has become practical for large software products. Examples of formally verified software include a verified microkernel called seL4 [14], a verified compiler for the C language [17], and on top of the latter, parts of the SSL internet protocol [5]. Another example is a verified compiler [13] for the ML functional programming language.

Foundations: sets, categories, types

The *material* conception of the mathematical notion of set conceives sets as being build up iteratively from the empty set. For instance, the disjoint union of sets A, B may be encoded as

$$\{(\emptyset, a) \mid a \in A\} \cup \{(\{\emptyset\}, b) \mid b \in B\}$$

The natural numbers can be encoded as $0 = \emptyset$, $1 = \{0\}$, $2 = \{0, 1\}$, et cetera. In contrast, Lawvere's elementary theory of the category of sets avoids this encoding by providing a *structural* account using categorical methods [16]. Category theory considers not only objects, but also the relations between them. These are captured using arrows (abstract functions).

Many familiar objects of set theory, such as singletons, the disjoint union, the set of natural numbers or the power set, are introduced in the elementary theory of the category of sets by their *universal property*. The disjoint union will be specified as the least object which allows embeddings from A, B. The natural numbers are captured as the least object that has a 0 and a successor function S. Here we use 'least' informally, and indeed the formal way to express this is the universal property. For instance, the universal property of $\mathbb{N}$ is the familiar property that, for every set X, a base point $x_0 \in X$ and a map $f : X \to X$ *uniquely determines* a sequence $(x_n)_n$ of elements of X, satisfying $x_{n+1} = f(x_n)$.

In the following, we shall see that type theory shares many of the structural aspects of the theory of sets. In particular, many types are specified by a type theoretic analogue of the universal property: the induction principle.

Practical foundations

Presently, foundations for computer formalization roughly fall in two classes, both have a *structural* flavor. They are either based on higher order logic (HOL), or on dependent type theory. The former is a little simpler, but the latter is a more expressive language for mathematics and at the same time provides a functional programming language, like Haskell or OCaml.

Type systems are used in most modern programming languages to avoid programming errors, with richer type systems allowing one to capture more errors. For example, adding a function $\mathbb{R} \to \mathbb{R}$ to a 2×2-matrix would raise a type error. This is similar to the unit tests in physics which warn against adding meters to kilos. In an untyped language like set theory, such errors are more difficult to catch. For example, one could ask for the elements of the real number π, although this question has little meaning. This is the reason that even the systems which are based on set theory have a weak type system build on top of them.

Type theory shares much of the structural properties of category theory, including the specification of types by universal properties. A large class of types specified in this way are the *inductive types*. For example, the type of natural numbers is a type with $0 : \mathbb{N}$ and a successor function $\mathsf{S} : \mathbb{N} \to \mathbb{N}$ which adhere to an *induction principle*. We will describe the type theoretical induction principle of the natural numbers, but before we do so we need to explain some of the concepts of dependent type theory.

First of all, in dependent type theory, there is a notion of *type families* (or, dependent types). A type family B indexed by a type A, consists of a type $B(a)$ varying over a variable a of type A. As an example, there is a type $\mathsf{list}_A(n)$, depending on $n : \mathbb{N}$, which gives for each n the type of lists with n elements from a type A. We often write $B : A \to \mathsf{Type}$ if B is a type family indexed by a type A.

Like a section of a family of sets, a section b of a type family B indexed by A consists of a term $b(a) : B(a)$, depending on a variable $a : A$. The sections of the type family B are themselves terms of a type: the *dependent product type* $\Pi_{(a:A)} B(a)$. Note that there is a universal quantification involved in the dependent product type: a term f of $\Pi_{(a:A)} B(a)$ assigns to *every* $a : A$, a term $f(a) : B(a)$.

Now we can state the induction principle of the natural numbers. The induction principle tells us how to show $\Pi_{(n:\mathbb{N})} P(n)$, for an arbitrary type family $P : \mathbb{N} \to \mathsf{Type}$. In other words, it tells us how to prove a universal quantification over the natural numbers, just as the induction principle you might be familiar with.

Induction principle of the natural numbers: *Let* $P : \mathbb{N} \to \mathsf{Type}$ *be a family of types over the natural numbers. Given*

$$a_0 : P(0)$$
$$a_{\mathsf{S}} : \Pi_{(n:\mathbb{N})} P(n) \to P(n+1),$$

there is a section $f : \Pi_{(n:\mathbb{N})} P(n)$. *Moreover, we can* compute *with* f *as follows*:

$$f(0) :\equiv a_0$$
$$f(\mathsf{S}(n)) :\equiv a_{\mathsf{S}}(n, f(n)).$$

This allows us to compute, say, $f(2)$ recursively:

$$\begin{aligned} f(\mathsf{SS0}) &= a_{\mathsf{S}}(1, f(\mathsf{S0})) \\ &= a_{\mathsf{S}}(1, a_{\mathsf{S}}(0, f(0))) \\ &= a_{\mathsf{S}}(1, a_{\mathsf{S}}(0, a_0)). \end{aligned}$$

Computation is an important aspect of any formalization. Fortunately, it is safely supported by most modern proof assistants. Computation is crucial to formalize proofs efficiently. Already, Russell and Whitehead used hundreds of pages in their book *Principia Mathematica*, to arrive at their proof that $1 + 1 = 2$. With modern implementations such proofs are automatic. This is necessary as, even with this support, the formalization of the Feit–Thompson theorem, for example, still has 170.000 lines of code. Just by their sheer size, such formalization projects need to be treated as software projects. Without verified computation those projects would simply be unfeasible.

The alert reader might have observed that the induction principle of the natural numbers only asserts the *existence* of a section when certain conditions are met. Nothing is said about the uniqueness, whereas the universal property is about existence *and* uniqueness. Of course, being unique depends on the notion of equality. The notion of equality in dependent type theory is inductively defined, just as the natural numbers. Martin-Löf, who first wrote down their inductive definition, called these equality types *identity types*. The identity type $a =_A b$ is a family of types which depends on two variables a, b of type A, and every type is equipped with an identity type. The identity type is generated by a term $\mathsf{refl}(a) : a =_A a$, for every $a : A$.

Since the identity types are themselves type families, one might anticipate that it is possible to use the induction principle again, to show the uniqueness of the sections that are obtained by the induction principle. This is indeed the case, but we will not go into that here. We only note that this derivation requires function extensionality. Function extensionality is the principle that two sections $f, g : \Pi_{(a:A)} B(a)$ are equal precisely when they take equal values. Since we take identity types

as our notion of equality, f and g take equal values if there is a section of type $\Pi_{(a:A)} f(a) =_{B(a)} g(a)$. Function extensionality cannot be proved directly in dependent type theory. However, in homotopy type theory we can prove function extensionality from Voevodsky's univalence axiom which we will discuss below.

Martin-Löf's dependent type theory

Foundations based on dependent type theory are characteristically different from the set theoretic foundations, in that the set theoretic foundations have two layers (one for first-order logic, and one for the theory of sets), whereas in dependent type theory there is only one such layer. Predicates of first order logic are replaced by dependent types. Notably, we have a *dependent type* $x =_A y$ of equalities in A instead of an equality predicate.

Apart from identity types, we have types such as the empty type, a type with one element and a type $\mathbb{N}$ of natural numbers. Given two types A and B, we can form their cartesian product $A \times B$ and their disjoint union $A + B$. More generally, given a family B of types indexed by a type A, we can form their dependent product $\Pi_{(a:A)} B(a)$, and we can also form their dependent sum $\Sigma_{(a:A)} B(a)$.

There is a tight connection between logic and type theory, which is illustrated by the Curry–Howard correspondence, as shown in Table 1. Apart from these logical connectives, type theory also possesses types which are not directly linked to any logical connective, including a type $\mathbb{N}$ of natural numbers and a so-called *universe* $\mathcal{U}$. Famously, by the Russell paradox there cannot be a set of all sets. So, one usually distinguishes between small sets and large sets, sometimes called sets and classes. In type theory, a universe is a type which contains the 'small' types as its terms, and is closed under the mentioned constructions.

Logical connective	Type operation
truth	$\mathbf{1}$
falsehood	$\mathbf{0}$
conjunction	$A \times B$
disjunction	$A + B$
universal quantification	$\Pi_{(x:A)} P(x)$
implication	$A \to B$
existential quantification	$\Sigma_{(x:A)} P(x)$
equality	$x =_A y$

Table 1 Curry–Howard correspondence

Notably, there is no untyped elementhood predicate ($\in$) in type theory, as there is in set theory. A term always comes with a specified type. There are rules for how to derive a term of a given type, and for how to use them to derive other things. From one point of view, these rules are much like the rules for how to derive a proposition of a certain form. From another point of view, these rules are a version of the universal property of a particular type constructor. We have seen an instance of such rules in the case of the type $\mathbb{N}$, with its induction principle.

Homotopical models

Homotopy type theory refers to the use of homotopy theory to study models of dependent type theory or the use of ideas and constructions from homotopy theory to prove theorems in type theory. So, homotopy type theory is both a theory of homotopy types and a homotopical perspective of type theory:

(homotopy type) theory
=
homotopy (type theory).

Voevodsky proposes homotopy type theory together with his univalence axiom as a new foundation for mathematics, as we will now explain.

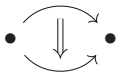
The set model

The collection of all sets forms a model of type theory. In this model, a type is interpreted as a set, and a family of types over a base type X is simply an indexed collection $(Y_i)_{i \in X}$ of sets. Universes in type theory can be modeled by so-called Grothendieck universes in set theory. In this way, types can be used to talk about sets in a structural way. In this model, any two terms of the same type can be equal in at most one way. This leads to a variant of the Curry–Howard correspondence where propositions are truncated. The (-1)-truncation $\|A\|$ of a type A identifies all its terms. Such truncated types are also referred to as mere propositions. Importantly, the (dependent) sum of propositions need not be a proposition, so we need to truncate it to soundly interpret first order logic. Having such propositional identity types seems natural, but models without this extra requirement are mathematically relevant too.

The groupoid model

In the groupoid model of type theory equality of terms is modeled by isomorphism. A groupoid is a category in which all morphisms are invertible. Groupoids are fundamental in homotopy theory. Given a topological space one defines the fundamental groupoid, the objects of which are the points of the space, and the arrows are the (homotopy equivalence classes of) paths between them. The identity arrow is the path that stands still, composition of arrows is composition of paths, the inverse of an arrow is the path in the opposite direction.

A homotopy between two paths:



Hofmann and Streicher [11] showed that many of the ideas from the set theoretical model can be extended to the groupoid model. Importantly, the groupoid model also contains a universe of all small sets, namely the groupoid of all sets in which the arrows are the bijections between sets. Since the equality type is modeled by isomorphism in the groupoid model, we see that the groupoid model gives a precise sense in which isomorphic sets may be identified. This is one of the early instances of the univalence axiom.

The way in which equality is modeled in the groupoid model gives a fundamental distinction between the set model and the groupoid model: in the set model there are many distinguishable natural numbers objects (e.g. $\mathbb{N}$, $2\mathbb{N}$), while in the groupoid model all of them are equal, because they are all isomorphic. This fits with our usual phrase: *the* natural numbers.

The homotopy interpretation of type theory

Again, it turns out to be impossible to show that any type is a groupoid, because it is not provable that any two proofs of equality are themselves equal in at most one way. Sets (discrete spaces) and groupoids (spaces for which all higher homotopy groups are trivial) are just two levels of a hierarchy of more and more complex structures, indeed: homotopy types.

Just as every space gives rise to its fundamental groupoid, we can associate to every space its fundamental ∞-groupoid. Whereas in the fundamental groupoid of a space we identified all the homotopic paths, in the fundamental ∞-groupoid these homotopies between paths will be the arrows at the next level.

This suggests a homotopy interpretation of type theory, in which identities are interpreted as paths in a space, identities between identities as homotopies between paths, and so on. This idea has been made precise by Voevodsky [12, 21] who showed the Kan simplicial sets form a model of type theory. Intuitively, simplicial sets are collections of triangulations. We will say more about them shortly. Independently, Awodey and Warren [2] showed that a relaxed version of identity types can be interpreted in an arbitrary Quillen model category (an abstract setting for homotopy theory). The original identity types can be interpreted in *cloven* weak factorization systems [4].

In the homotopy interpretation of type theory, types are interpreted by spaces, and dependent types are interpreted by fibrations. Terms of a type are points in a space, and a term of a dependent type is a section of the corresponding fibration.

A proof of equality is a point in the path space, so we see that a proof of equality between two equalities gets interpreted as a homotopy between paths. Higher equalities are interpreted accordingly by higher homotopies.

The simplicial set model

Voevodsky constructed, moreover, a universe in simplicial sets which has a very nice property: the path space between two Kan simplicial sets in the universe is weakly equivalent to the Kan simplicial set of weak equivalences between them. This property goes under the name *univalence*: for any two types X and Y in a universe U, we have a canonical (i.e. specified) equivalence

$$(X =_U Y) \simeq (X \simeq Y).$$

Grothendieck's *homotopy hypothesis* states that ∞-groupoids are a model for homotopy types, topological spaces up to homotopy. The hypothesis itself is very delicate due to the difficulty of giving a precise definition of ∞-groupoid. ∞-groupoids in turn can be modeled by a so-called model structure on simplicial sets. Just like sets, we can take products and sums (disjoint unions, i.e. coproducts) of such ∞-groupoids. Moreover, we can even take dependent (co)products. This makes ∞-groupoids (Kan simplicial sets) into a model of type theory.

Motivated by Grothendieck's insight that ∞-groupoids are ubiquitous in mathematics, and by the foundational role of type theory, Voevodsky proposes homotopy type theory as a new 'univalent' foundation for all of mathematics. The connection between types and ∞-groupoids is further strengthened by a theorem of Van den Berg, Garner and Lumsdaine, that types in Martin-Löf type theory indeed exhibit the structure of an ∞-groupoid [3, 15].

In the simplicial set model, we can find the set model and the groupoid model, and indeed a whole hierarchy of n-groupoid models. The 0-groupoids, which are also known as the (homotopy) discrete spaces, are our ordinary sets [19].

Improving set theory

Zermelo–Fraenkel set theory is often considered to be 'the' foundation of mathematics. This was introduced in the beginning of the previous century. There have been several proposals to update this foundation to more modern mathematical insights. As already discussed, one such proposal is Lawvere's structural set theory. This arose from the development of Grothendieck's topos theory, a vast generalization of topology applicable to a wide range of fields including algebraic geometry. Lawvere and Tierney showed that toposes can be described by a (structural) set theory. A prime example of a topos is the topos of sets. Moreover, for a general topological space, the variable sets, the so-called sheaves, over this space form again a topos. Another class of toposes is connected to the theory of computation. In other words, toposes are everywhere.

Since there are many toposes, and therefore many interpretations of the *structural* axioms of set theory, one might consider the axioms of set theory similar to the axioms of a group (in fact, both are essentially algebraic theories). By adding axioms to structural set theory, one reduces the class of models, just as one reduces the class of groups by adding the axiom of commutativity.

In the case of structural set theory, we can add for example the axiom of choice. Not every topos satisfies this property, as the logic of the set theory a topos adheres to is not classical logic, but Heyting's intuitionistic logic. Nevertheless, we can restrict our attention to the toposes in which classical logic is satisfied, by adding the axiom of choice. In this sense, topos theory *subsumes* classical set theory.

Similarly, type theory is by its computational nature an intuitionistic theory. But just as in toposes, homotopy type theory subsumes classical reasoning, in the sense that the axiom of choice can be added. The axiom of choice is satisfied, for instance, in the model of Kan simplicial sets.

Universes play a key part in the univalent foundations. The univalence axiom characterizes the identity types on the universe. In fact, by characterizing the identity types on the universe, one describes indirectly a universal property of the universe. This universal property is a homotopical analogue of the notion of a subobject classifier in a topos. In a topos, the subobject classifier can be seen as the object of all propositions in the logic of that topos. In the case of the sheaves over a topological space, this subobject classifier just consists of the opens of the topological space. Hence, this is an important object in the theory of toposes. By the propositions-as-types [22] paradigm (i.e. the Curry–Howard correspondence), the universe plays the same role for general types. Another way to express this is that the univalence axiom connects univalent type theory with higher topos theory. Higher topos theory is a joint generalization of topos theory and the abstract theory of homotopy types [18}.

Consequences homotopy interpretation

Structuralism

Another application of the univalence axiom is the structure invariance principle between general (algebraic) structures. For instance, any two isomorphic groups may be identified. This idea is prominent in Bourbaki's vision of mathematics as the science of abstract structures, where structures are important only up to isomorphism [20]. Univalent type theory satisfies the principle of structuralism [1]: that isomorphic structures may be identified. In set theoretic foundations for mathematics, the principle of structuralism is simply false.

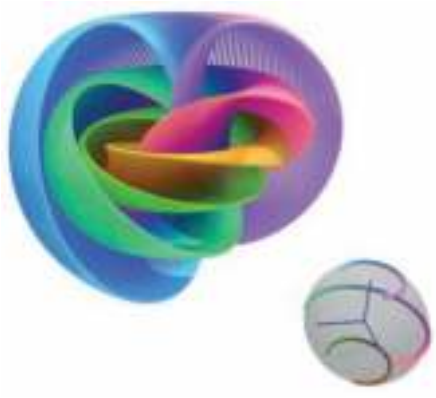


Illustration: Niles Johnson, Wikimedia

Figure 2 Hopf fibration.

Axiomatic reasoning

Since homotopy type theory can be interpreted in ∞-groupoids, we can actually use it to reason axiomatically about them. This may be compared to two ways of reasoning in geometry: analogous to traditional homotopy theory we have analytic reasoning using coordinates, and analogous to reasoning in homotopy type theory we have synthetic reasoning using axioms about points and lines and construction methods such as ruler and compass. Just like many theorems in geometry have beautiful synthetic proofs, likewise, many theorems in algebraic topology have elegant and very elementary proofs in homotopy type theory. One of the first examples of such reasoning, and indeed one of the first non-trivial applications of the univalence axiom, is Shulman's proof that the fundamental group of the circle is equivalent to the integers.

Many of the classical constructions of CW-complexes can be specified axiomatically in homotopy type theory by generalizing the construction of inductive types by specifying not only its points, but also (some) equalities. We call these constructions *higher* inductive types. The interval consists of two points and an equality (i.e., path) between them. A circle is a point with a loop, (using univalence one shows that in the free type with a point and a loop, the loop is distinguishable from the trivial path refl). The circle is also obtained as a special instance of suspensions of types, and by iterated suspensions one gets the n-sphere for any $n : \mathbb{N}$. Many other familiar spaces, including the torus, can be defined in this way.

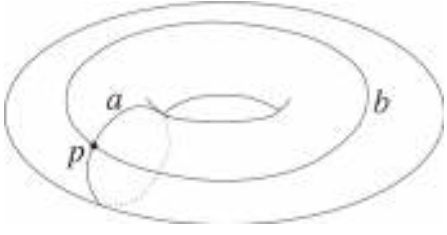


Illustration: Wikimedia

Figure 3 The fundamental group of the torus with base point p has two generators a and b.

Many theorems about fundamental groups of spheres have short new proofs in the language of homotopy type theory using reasoning principles from type theory. Also, many of the classic constructions are possible. For instance, once the spheres are defined, one may use the univalence axiom to construct the Hopf fibration.

One of the big accomplishments of homotopy type theory is Brunerie's proof [6] that the fourth homotopy group $\pi_4(\mathbb{S}^3)$ of the 3-sphere is $\mathbb{Z}/2\mathbb{Z}$. The first part of his proof shows that there is a natural number n for which $\pi_4(\mathbb{S}^3)$ is $\mathbb{Z}/n\mathbb{Z}$. He then proceeds by proving that this number is 2.

Cubical proof assistant

In principle one should be able to compute the number n, which was constructed in the first half of the proof, thus replacing tens of pages of advanced mathematics. However, since the univalence axiom is added as an axiom to Martin-Löf's dependent type theory, the computation may get stuck on applications of the univalence axiom. In other words, a *computational interpretation* of the univalence axiom has to be provided before the algorithms for computation can be extended to compute with applications of the univalence axiom.

Fortunately, there is now a fully constructive model of the univalence axiom [7], which provides the desired computational interpretation. This is a model of cubical sets, instead of simplicial sets. In the cubical sets, spaces are built up from squares instead of triangles. This slight modification alters the algebraic structure of spaces slightly, making it possible to give a fully constructive proof that the univalence axiom is satisfied in this model, whereas Voevodsky's original proof makes use of non-computational reasoning principles in an essential way.

The cubical model of type theory turns out to provide a convenient language, cubical type theory, to reason about (homotopical) constructions. This type theory has been implemented so that one can now compute with univalence. The term for Brunerie's number n which we introduced above has been defined in cubical type theory. Unfortunately, the type checker is currently too slow to compute the answer. However, many possibilities for improvements exists, and we expect more such applications in the future.

Axiomatic physics

Schreiber and Shulman [24] use homotopy type theory to provide an axiomatic treatment of modern physics, more precisely of local higher gauge quantum field theory. To do so they greatly generalize Lawvere's axioms for synthetic differential geometry, which uses so-called cohesive toposes, by providing a modal type theory which can be modeled in cohesive *higher* toposes.

Open science

Voevodsky proposed the thematic year (2012–2013) about univalent foundations of mathematics at the Institute for Advanced Study to bring together researchers from a wide variety of backgrounds to develop mathematics based on his univalence axiom. He organized the year with Awodey and Coquand. During this year there was a pressure cooker full of ideas from researchers coming from different directions such as type theory and proof assistants, category theory and abstract homotopy theory. It was decided to write a book [20] as a means to collect the ideas and take a snapshot of the state of the field at the end of the year. As a proof of concept, the book develops a substantial amount of mathematics within the univalent foundations. This includes homotopy theory and algebraic topology (using synthetic reasoning), but also set theory, category theory and various constructions of the real numbers. Part of the book 'unformalizes' code that existed only in the computer before. This prominently includes both the 'Foundations' library by Voevodsky and the substantial work on using higher inductive types for synthetic reasoning. This was a new experience. Before, the usual procedure was to go from paper proofs in mathematics to computer formalization, but this time the formalization came first and the natural language presentation afterwards. In these formal proofs, the proof assistant is very helpful in developing the proofs, both for proofs by computation and in order to keep the bookkeeping manageable.

Photo: uf-ias-2012.wikispaces.com

Participants of the Univalent Foundations Program

The book is self-published under a permissive creative commons licence. One of the advantages of this process is that the book is still being updated regularly with additions and clarifications from readers. Often this is done by means of a 'pull-request', the common way of contributing to an open source programming project. The book was created in a true open source spirit and was developed using open source tools such as git version control. The LaTeX-sources are freely available to everyone. It was a unique experience to write a book with such a large number of authors. The efficient software tools were important in this process, but working collectively from one place where we could regularly meet was still crucial.

Many parts of the book have been formalized a number of times, and also in different proof assistants such as Coq, Agda and Lean. While some details of the semantics remain to be worked out, we can still be very confident that all these theorems are correct. Conversely, in the cases where the formalization came afterwards, usually a number of bugs were found in the book. These bugs have been corrected. It is worth stressing that the book has been proof read very carefully, similar bugs showed up when other parts of mathematics were formalized.

The book was included on the 2013 list of notable items published in computing by Computing Reviews of the Association of Computing Machinery (ACM).

References

1 Steve Awodey, Structuralism, invariance, and univalence, *Philosophia Mathematica* 22(1) (2013), 1–11.

2 Steve Awodey and Michael Warren, Homotopy theoretic models of identity types, *Mathematical Proceedings of the Cambridge Philosophical Society* 146(1) (2009), 45–55.

3 Benno van den Berg and Richard Garner, Types are weak ω-groupoids, *Proceedings of the London Mathematical Society* 102(2) (2011), 370–394.

4 Benno van den Berg and Richard Garner, Topological and simplicial models of identity types, *ACM transactions on computational logic (TOCL)* 13(1) (2012), 3.

5 Lennart Beringer, Adam Petcher, Katherine Q. Ye and Andrew W. Appel, Verified correctness and security of OpenSSL HMAC, in *USENIX Security* 15, 2015, pp. 207–221.

6 Guillaume Brunerie, *On the Homotopy Groups of Spheres in Homotopy Type Theory*, PhD thesis, Nice, 2016, arXiv:1606.05916.

7 Cyril Cohen, Thierry Coquand, Simon Huber, and Anders Mörtberg, Cubical type theory: a constructive interpretation of the univalence axiom, 2016, http://www.cse.chalmers.se/~coquand/cubicaltt.pdf.

8 G. Gonthier, Formal proof — the four-color theorem, *Notices of the AMS* 55(11) (2008), 1382–1393.

9 G. Gonthier, A. Asperti, J. Avigad, Y. Bertot, C. Cohen, F. Garillot, S. Le Roux, A. Mahboubi, R. O'Connor, S.O. Biha, I. Pasca, L. Rideau, A. Solovyev, E. Tassi and L. Théry, A machine-checked proof of the odd order theorem, in *Interactive Theorem Proving (ITP)*, Lecture Notes in Computer Science, Vol. 7998, Springer, 2013, pp. 163–179.

10 Thomas Hales, Mark Adams, Gertrud Bauer, Dat Tat Dang, John Harrison, Truong Le Hoang, Cezary Kaliszyk, Victor Magron, Sean McLaughlin, Thang Tat Nguyen, Truong Quang Nguyen, Tobias Nipkow, Steven Obua, Joseph Pleso, Jason Rute, Alexey Solovyev, An Hoai Thi Ta, Trung Nam Tran, Diep Thi Trieu, Josef Urban, Ky Khac Vu, Roland Zumkeller, A formal proof of the Kepler conjecture, 2015, arXiv:1501.02155.

11 Martin Hofmann and Thomas Streicher, The groupoid interpretation of type theory, in Sambin and Smith, eds., *Twenty-five Years of Constructive Type Theory*, Oxford Logic Guides, Vol. 36, OUP, 1998, pp. 83–111.

12 Chris Kapulkin and Peter Lumsdaine, The simplicial model of univalent foundations (after Voevodsky), 2012, arXiv:1211.2851.

13 Yong Kiam Tan, Magnus O. Myreen, Ramana Kumar, Anthony Fox, Scott Owens, and Michael Norrish, A new verified compiler backend for CakeML, in *ICFP16*, 2016.

14 Gerwin Klein, Kevin Elphinstone, Gernot Heiser, June Andronick, David Cock, Philip Derrin, Dhammika Elkaduwe, Kai Engelhardt, Rafal Kolanski, Michael Norrish, Thomas Sewell, Harvey Tuch and Simon Winwood, seL4: Formal verification of an OS kernel, in *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles*, ACM, 2009, pp. 207–220.

15 Peter LeFanu Lumsdaine, *Weak ω-Categories from Intensional Type Theory*, Springer, 2009, pp. 172–187.

16 Tom Leinster, Rethinking set theory, *American Mathematical Monthly* 121(5) (2014), 403–415.

17 X. Leroy, Formal certification of a compiler back-end or: programming a compiler with a proof assistant, in *POPL*, ACM, 2006, pp. 42–54.

18 J. Lurie, *Higher Topos Theory*, Vol. 170, Princeton, 2009.

19 Egbert Rijke and Bas Spitters, Sets in homotopy type theory, *MSCS* 25, Special Issue 05 (2015), 1172–1202.

20 The Univalent Foundations Program, *Homotopy Type Theory: Univalent Foundations for Mathematics*, Institute for Advanced Study, 2013, homotopytypetheory.org/book.

21 Vladimir Voevodsky, The equivalence axiom and univalent models of type theory, 2010, arXiv:1402.5556, to appear in *MSCS*.

22 Philip Wadler, Propositions as types, *Communications of the ACM* 58(12) (2015), 75–84.

23 https://www.math.ias.edu/vladimir/sites/math.ias.edu.vladimir/files/2014_08_ASC_lecture.pdf.

24 https://ncatlab.org/schreiber/show/Modern+Physics+formalized+in+Modal+Homotopy+Type+Theory.

Arjeh Cohen
Faculteit Wiskunde & Informatica
Technische Universiteit Eindhoven
a.m.cohen@tue.nl

Over digitaal wiskundig lesmateriaal

De digitalisering van wiskundig lesmateriaal is zover gevorderd dat het klassieke tekstboek overbodig wordt. Al jarenlang zijn online in LaTeX geschreven wiskundedictaten en tekstboeken te vinden, die op hun beurt langzamerhand vervangen worden door interactieve en adaptieve cursussen. In dit artikel laat Arjeh Cohen zien welke vooruitgang deze moderne middelen bieden. Het is geen compleet beeld van alle ontwikkelingen, maar een reeks observaties die gebaseerd zijn op eigen ervaringen.

Eind jaren zeventig werkte ik aan de Universiteit Twente. Ergens onderin het wiskundegebouw bevond zich een zaal die leeg was op een kring computerschermen met toetsenborden en stoelen na. De zaal was bedoeld voor experimentele wiskundepractica in computerondersteund onderwijs. Ik was geboeid door de poging zo wiskunde te instrueren, maar op die stoelen heb ik zelden studenten zien zitten. Het leek me ook erg optimistisch om te denken dat de trillende letters op het zwart/wit-scherm, de monotonie van de multiplechoicevragen, en de onwennigheid van studenten met computers overkoombare obstakels voor studiesucces zouden zijn.

Toch is de belangstelling, in het algemeen en bij mij, voor de mogelijkheden van de computer bij het aanleren van praktische wiskundevakken aan hbo en wo sindsdien gebleven. Nu, bijna vijftig jaar na de optimistische aanpak in de jaren zeventig, is wiskundeonderwijs met behulp van de computer de gewoonste zaak van de wereld geworden. Veel van het wiskundeonderwijs wordt digitaal en interactief aangeboden, terwijl het persoonlijk contact met docenten zoveel mogelijk gehandhaafd blijft.

Computerondersteund onderwijs kent een groot scala aan verschijningsvormen, te veel om in een kort artikel als dit te beschrijven. De digitale stappen die Noordhoff [14] maakt voor het voortgezet onderwijs (vo), zijn aangestipt in een recent artikel in dit blad [10], tezamen met de ervaringen uit eigen praktijk van de docent Peter Eshuis. Ook ik beperk me grotendeels tot de eigen praktijk. Ik geef aan welke ervaring ik heb met de aanmaak van digitaal wiskundig lesmateriaal, bespreek de voordelen die ik heb gezien van het gebruik ervan en ga in op enkele mogelijke verbeteringen en interessante ontwikkelingen.

Eigen ervaring

Mijn belangstelling voor het onderwerp komt uit de technische hoek. Vanaf het begin van de jaren negentig heb ik aan veel onderzoeken meegewerkt waarin een goede computeromgeving voor exact wiskundig werk centraal stond. In het begin, nog vóór het internet tot stand kwam, vatten we (onder wie Lambert Meertens en Steven Pemberton) op het CWI het plan op om een echt interactief wiskundeboek over Lie-algebra's te maken. De eerste poging, in het begin van de jaren negentig, is niet veel verder gekomen dan een specificatie van de ideeën [6], al was de algoritmische benadering van de theorie van Lie-algebra's [15] een zeer interessant en nuttig bijproduct.

In tegenstelling tot ons wiskundig onderzoek, werd het werk aan interactieve systemen voor wiskunde sterk bepaald door de opkomst van steeds weer krachtiger software en meer geavanceerde standaarden in de informatica. Een typisch voorbeeld is OpenMath [32], waaraan we in de jaren negentig in Eindhoven werkten. Het behelst een manier om elke wiskundige uitdrukking zodanig op te schrijven dat elk softwarepakket (in principe) begrijpen kan wat ermee bedoeld wordt. Om die uitdrukking semantisch eenduidig vast te leggen, wordt gebruik gemaakt van elementaire bouwstenen, symbolen geheten. De betekenis van die symbolen wordt

in aparte documenten, te vergelijken met woordenboeken, vastgelegd. Zo kan het reële getal π onderscheiden worden van een vlak dat π heet, of een variabele met de naam π. Een belangrijke implementatie van OpenMath is in XML [46] gerealiseerd. Min of meer tegelijkertijd kwam de standaard MathML [27] voor presentatie van wiskunde tot stand. De derde versie van MathML maakte het mogelijk aan deze XML-taal 'inhoud' toe te voegen die uit de XML-implementatie van OpenMath bestaat. Zo is OpenMath in zekere zin een onderdeel geworden van de dominante standaard MathML.

Om de doorsnee gebruiker van het web niet op te zadelen met de grote brij punthaken, attributen en andere verwarrende zaken die XML met zich meebrengt, hebben we (onder wie Hans Cuypers, Jan Willem Knopper en Mark Spanbroek) aan de Technische Universiteit Eindhoven een gestructureerde editor geschreven die 'normale' wiskundige uitdrukkingen omzet in OpenMath [23]. Door deze editor in te zetten kunnen we studenten met een minimum aan kennis over hoe je de wiskunde moet invoeren, antwoorden op vragen laten geven die volledig semantisch te analyseren zijn.

Rond 2000 publiceerden Hans Cuypers, Hans Sterk en ik het interactieve boek *Algebra Interactive* [5]. Hierin werd de interactiviteit verzorgd door het computeralgebra-systeem GAP [12]. De student kon interactieve opgaven uitvoeren en voorbeelden aanroepen (gapplets) door in invoervelden GAP-uitdrukkingen te typen. De uitgever verkocht het materiaal als een boek, met daarin een cd die de eigenlijke interactieve versie droeg. Zo werd ervoor gezorgd dat het boek eenvoudig universitaire bibliotheken ingeloodst kon worden en werd voorkomen dat de cd in discotheken terecht kwam.

Kort daarna werd het internet met de onderliggende software zo krachtig dat *Algebra Interactive* tot een nieuw soort interactief boek getransformeerd zou kunnen worden dat via de cloud werkt. Deze overweging motiveerde ons om de bovengenoemde OpenMath-editor te bouwen. Hiermee zouden veel meer computeralgebra-systemen dan alleen GAP aangesproken kunnen worden. Dat dit inderdaad mogelijk is, blijkt uit het gebruik van de editor door de firma SOWISO (een spin-off van de Technische Universiteit Eindhoven),

Figuur 1 Voorbeeld van een opgave in SOWISO.

die daarmee een interactief platform voor onderwijs ontwikkeld heeft.

De afgelopen twee jaar heb ik als schrijver meegewerkt aan wiskundig lesmateriaal in het SOWISO-platform. Het resultaat beslaat calculus en lineaire algebra, dat door een onderwijsinstelling aangeboden kan worden als online cursus of als online onderdeel van een meer klassieke cursus (die daarmee, volgens de moderne terminologie, 'blended' wordt). De docent stelt een selectie van de beschikbare hoofdstukken samen tot een cursus, bepaalt welke opgaven gemaakt moeten worden en beheert het forum — een plaats waar vragen en antwoorden over de lesstof samenkomen. Studenten en docenten kunnen ook reageren, waardoor een digitaal klaslokaal ontstaat. Er worden logs bijgehouden van de opgaven die de student heeft gemaakt. De docenten krijgen heel veel data over het gebruik van het systeem tot hun beschikking. Er bestaan diagnostische toetsen, op grond waarvan bepaalde onderwerpen voor nadere studie aanbevolen kunnen worden, formatieve toetsen, waarin tussenantwoorden mogelijk zijn en bijsturing plaatsvindt, en summatieve toetsen, die gebruikt kunnen worden voor de vaststelling van een eindcijfer. Ook biedt dit platform een auteursomgeving waarin docenten hun eigen interactieve cursussen kunnen schrijven of bestaande cursussen voor eigen gebruik kunnen redigeren.

Dankzij de OpenMath editor wordt het antwoord van een student op een open vraag in het SOWISO-platform (er zijn ook andere typen opgaven) verwerkt als een uitdrukking waarvan de semantiek vastligt. Met behulp van speciaal voor het web ontworpen scripttalen kan de structuur geanalyseerd worden. Met gebruikmaking van een computeralgebra-systeem kan onderzocht worden in hoeverre het antwoord voldoet. Bekijk, bij wijze van voorbeeld, de opgave "Los de vergelijking $\frac{2x+3}{x+8} = 3$ op." Als het gelijkteken niet het topsymbool in het antwoord is of als er andere bewerkingen in het antwoord voorkomen dan nodig om de vergelijking op te lossen, dan zal het systeem dat melden dankzij de analyse met de scripttaal. Het platform is niet alleen in staat om het goede antwoord, $x = -21$, te herkennen, maar ook om de student die een vergelijking intikt als tussenstap (of zelfs start met het invoeren van de gegeven vergelijking door erop te klikken), stapsgewijs met suggesties te begeleiden naar het goede eindantwoord. De auteur van de opgave bepaalt zelf welke regels van toepassing zijn. De opgaven kunnen als oefening maar ook in toetsen worden gebruikt en als voorbeelden aan de theorie worden toegevoegd. Van elk hoofdstuk kan met een druk op de knop een statische versie gecreëerd worden in de vorm van een pdf- of Word-file. Zo kan het platform dus ook klassieke tekstboeken produceren. Wat een systeem als dit meer te bieden heeft dan de tekstboeken, wordt in de volgende paragraaf behandeld.

Voordelen van digitaal lesmateriaal

De laatste jaren werd het optimisme uit de tijd van de Twentse studiezaal met

computers bewaarheid: de eerste toepassingen op grote schaal van digitaal wiskundeonderwijs zijn een feit. In zijn algemeenheid is dit natuurlijk te danken aan de snelle ontwikkelingen in de informatica en elektrotechniek. Maar specifiek voor de wiskunde telt de opkomst van de computeralgebra-systemen ook zwaar mee. De twee meest verbreide pakketten van het eerste uur, Maple en Mathematica, zijn met bijzondere onderwijsproducten gekomen: Maple met MapleTA [22], een manier om open wiskundeopgaven te maken, waarvan de invoer door het pakket Maple getoetst wordt, en Mathematica in eerste instantie met notebooks [25], die hebben laten zien hoezeer cursusmateriaal verlevendigd kan worden door geparametriseerde voorbeelden, animaties, mogelijkheden voor de student om zelf te programmeren, et cetera. Gebruik van zo'n notebook vereist echter een lokale installatie van het Mathematica-pakket [24] op de eigen computer, en enige vertrouwdheid met het systeem. Dit is niet langer nodig voor gebruik van de in 2009 gelanceerde website Wolfram Alpha [45], waartoe een versie van de notebooks geëvolueerd is.

Mathematica [24] is een commercieel product dat uitstekend geschikt is om online cursussen van wiskundige interactiviteit te voorzien. Een oud en veel beperkter systeem als het opensource-pakket Maxima [28] voldoet vaak ook goed voor grote basiscursussen en kan gratis en vrij ingezet worden. Zowel het SOWISO-platform als de Engelse STACK-omgeving [40] werken hiermee. Maar er is zicht op goede toegang tot nog meer universele computeralgebra-systemen. Zo is het pakket Magma dankzij financiële steun van de Simons Foundation [21] gratis beschikbaar voor elke openbare onderwijsinstelling in de Verenigde Staten (idee voor Nederland?) en wordt het gratis beschikbare opensource-systeem SageMath [38] steeds verder geïntegreerd in een grote omgeving van waaruit allerlei geavanceerde systemen zijn aan te roepen; zie bijvoorbeeld het OpenDreamKit-project [32], waarin sprake is van een ecosysteem voor computationele wiskunde.

De techniek achter het web, inclusief de wiskundig georiënteerde programmatuur (zoals MathJax [26] en de eerder genoemde MathML [27]), zorgt ervoor dat de presentatie van wiskunde op het scherm niet meer hoeft afwijken van wat de student gewend is van LaTeX. Dit is van belang als de inhoud dynamisch gepresenteerd wordt, met parameters die random, of afhankelijk van de context, gekozen kunnen worden. Zo zijn heel wat obstakels weggenomen om wiskunde op een boeiende manier online te brengen.

Naar aanleiding van deze verworvenheden hoor je vaak de vraag wat de digitalisering nou echt bijdraagt aan verbetering van het onderwijs. Daarom ga ik in op vijf in het oog springende aspecten: toegankelijkheid, interactiviteit, adaptiviteit, het grote arsenaal aan oefeningen en het gemak voor de docent.

Toegankelijkheid

Het grootste voordeel dat digitaal lesmateriaal (al of niet wiskundig) biedt ten opzichte van klassiek lesmateriaal is mijns inziens de toegankelijkheid. De loftrompet hierover is goed gestoken in de TED-voordracht [18] van de mede-oprichter van Coursera, Daphne Koller. Veel mensen die vroeger door geografische of andere omstandigheden niet in staat waren onderwijs te volgen, komen via het internet vol aan de bak. Het feit dat allerlei cursusmateriaal en gegevens over wiskunde overal ter wereld bereikbaar zijn geworden, maakt het web tot een wiskundige leerschool die meer informatie toegankelijk maakt, betere uitleg geeft, en meer contact met topdocenten en toponderzoekers mogelijk maakt dan ik me ooit had kunnen voorstellen. Eén concreet voorbeeld is de website van Terence Tao [41], waarop niet alleen advies staat aan studenten hoe wiskunde te studeren en vele interessante doorverwijzingen te vinden zijn, maar ook zijn recente opzienbarende stellingen worden toelicht. Universitaire bibliotheken maken een overgang door van fysieke verzamelplaats van boeken en andere informatiedragers naar doorgeefluik (portal) voor digitale varianten op het web. Docenten geven al vaak de beste plaatsen op het web aan om iets over hun onderwerp te leren, maar ook zonder die hulp vinden veel studenten hun eigen weg.

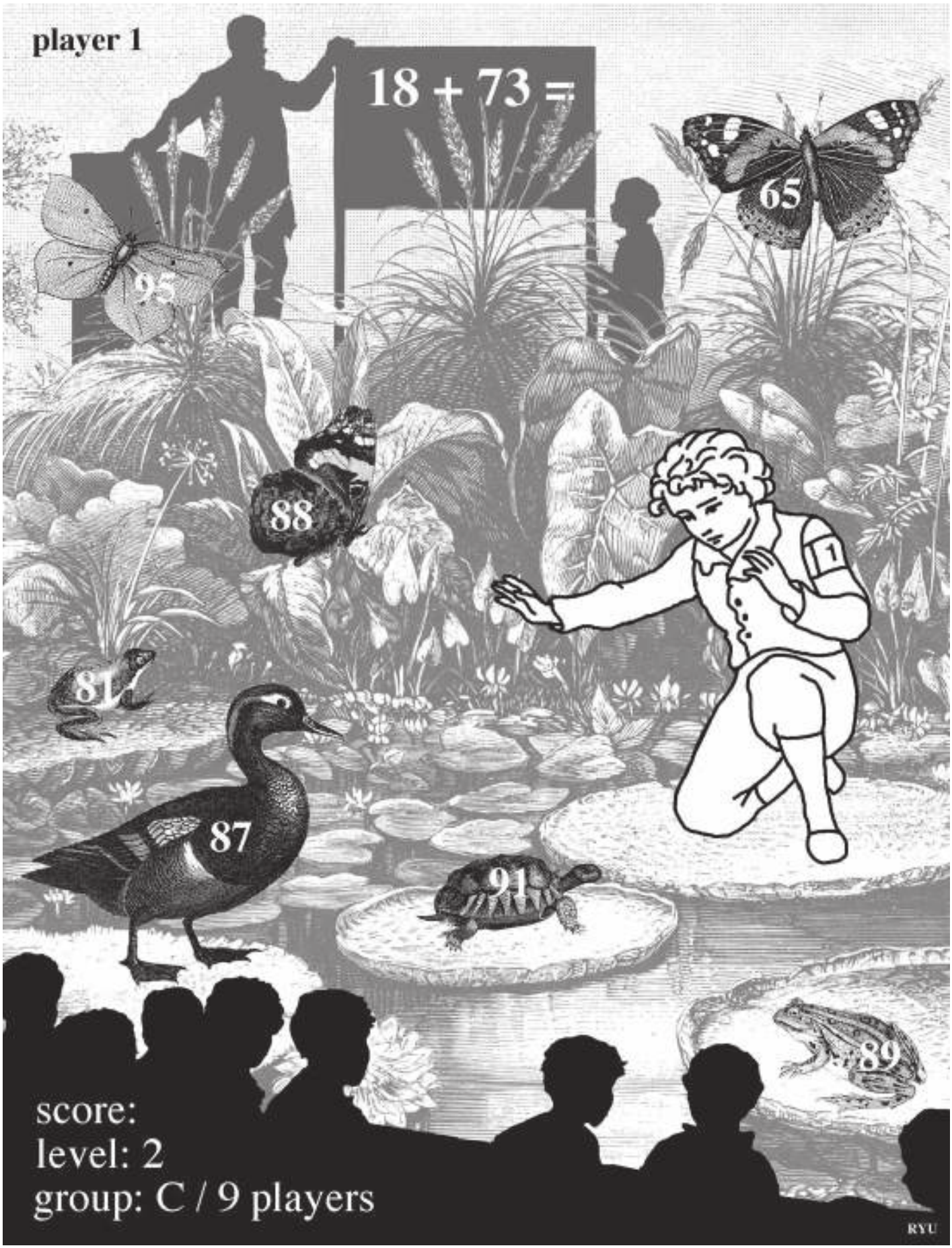
Door de wiskundige formule te googelen waar je meer van zou willen weten, kom je al gauw terecht op websites waar nuttige informatie over de ingetypte uitdrukking staat. Gebruik van WebMathematica [42] maakt het mogelijk bijna alle antwoorden op vragen uit een doorsnee calculustentamen te voorschijn te toveren door simpelweg de vraag of formule op een webpagina in te voeren. Massive Open Online Courses (MOOCs) van bijvoorbeeld Coursera [8] en EdX [9] voorzien in goede videocolleges voor de basisvakken wiskunde, compleet met studiegroepen en begeleiders. Via deze MOOCs, of individuele videopresentaties van docenten elders op het internet (de Khan Academy [17] is hier een bekend voorbeeld van), kunnen studenten uitleg krijgen over punten in het dictaat die ze niet begrepen hebben, of zelfs een hele cursus doorlopen. Ik heb studiegroepjes gekend die door dit soort methoden samen te gebruiken een veel beter dan gemiddeld cijfer op hun tentamen scoorden. Ik laat in het midden wat dit zegt over het college, maar het illustreert wel de goede toegankelijkheid van online lesmateriaal.

Interactiviteit

Ook in interactiviteit gaat digitaal materiaal veel verder dan klassiek lesmateriaal. De term interactief duidt op computersystemen die op een gebruiker reageren om gegevens of opdrachten te verkrijgen en om directe resultaten of bijgewerkte informatie te geven, vaak als substituut voor een menselijk antwoord. De feedback die een student kan krijgen bij het maken van oefenopgaven, valt hier onder. Bij een open opgave als "Bereken het functievoorschrift van de raaklijn in het punt $[1,4]$ aan de grafiek van de functie $f(x)=\sqrt{3x^2+13}$" in het SOWISO-platform, wordt de door de student ingevoerde uitdrukking achtereenvolgens getest op de volgende kenmerken, en voorzien van bijpassende reactie naar de student:

- er komt een andere variabele voor dan x of een ander symbool dan een wortel of een standaard rekenkundige operatie;
- het antwoord is fout volgens een verwacht patroon, bijvoorbeeld $\frac{1}{4}x+\frac{15}{4}$ omdat bij het differentiëren de kettingregel niet naar behoren is toegepast;
- het antwoord is niet van de vorm $a\cdot x+b$;
- de coëfficiënt van x is ongelijk aan $\frac{3}{4}$;
- het punt $[1,4]$ ligt niet op de door de student ingevoerde raaklijn.

Een goede docent heeft ongetwijfeld meer te bieden, maar is lang niet altijd beschikbaar. Het systeem heeft heel veel geduld: ook bij de 123ste poging van de student om de vraag te beantwoorden blijft de



Illustratie: Ryu Tajiri

feedback onvermoeid terugkomen. Ten slotte kan de auteur de verzameling feedback-regels bij elk nieuw geconstateerd obstakel uitbreiden.

Een ander voorbeeld van interactiviteit is te vinden in de effectieve (constructieve of algoritmische) benadering van wiskunde. In het vo-curriculum zie ik weinig aandacht voor deze aanpak. Gezien de rol van de computer in de samenleving en de behoefte aan meer informatica-onderwijs in het vo, lijkt me een algoritmische benadering op zijn plaats. Neem bijvoorbeeld de reële getallen. De vo-leerlingen wordt een reëel getal voorgeschoteld als een vanzelfsprekendheid, een punt op de rechte lijn tussen bepaalde rationale getallen in. Natuurlijk komt de decimale ontwikkeling hier en daar ter sprake, maar nergens wordt het begrip reëel getal duidelijk gemaakt. Dit is tot op zekere hoogte te begrijpen als je ziet hoeveel moeite de docent analyse zich hiervoor in het eerste universitaire wiskundejaar getroost. Maar er is een concreet beeld te geven van een reëel getal als een orakel [20]: het orakel weet precies welk getal bedoeld wordt (denk aan de oneindige decimale ontwikkeling); als mens kun je slechts een eindig beginstuk daarvan verwerken, een benadering door middel van een rationaal getal met een macht van 10 in de noemer. Als je meer geduld hebt, kan het orakel je een betere benadering (een decimaal getal met een hogere macht van 10) geven. Dit beeld sluit perfect aan bij de meetkundige opvatting van een reeel getal dat ingekneld ligt tussen gegeven rationale getallen. Om een voorbeeld te noemen: als vervolg op de behandeling van $\sqrt{2}$ als het positieve getal a met de eigenschap dat $a^2 = 2$, kan het programmaatje besproken worden dat bij invoer door de student van het aantal decimalen n, het grootste natuurlijke getal x zoekt dat voldoet aan $x^2 \leq 2 \cdot 10^{2n}$, om de gevraagde decimale ontwikkeling van $\sqrt{2}$ af te leveren als $x \cdot 10^{-n}$. De interactiviteit kan verwezenlijkt worden door zo'n programmaatje daadwerkelijk te draaien en de werking ervan te visualiseren. De bonus van deze behandeling is het inzicht hoe rekenmachines tot benadering van een getal als $\sqrt{2}$ zouden kunnen komen. Hierbij ligt wel het gevaar op de loer dat het orakel in alle voorbeelden een algoritme is, waardoor de verkeerde indruk kan ontstaan dat dit altijd zo is (cf. [44]).

Als laatste voorbeeld noem ik de interacties die recent gerealiseerd zijn in GeoGebra [13] met driedimensionale technologie. Hierdoor is het mogelijk een op het scherm gevisualiseerd driedimensionaal object, laten we zeggen een hyperboloïde, nader te bekijken. Een student die voor het scherm zit, kan door handbewegingen in de fysieke ruimte een vlak op het scherm oppakken en het daar laten bewegen, terwijl de doorsnede van het vlak met de gegeven hyperboloïde dynamisch gepresenteerd wordt. Zo wordt letterlijk interactief inzicht verschaft.

Adaptiviteit

Hoewel interactiviteit al tot gevolg heeft dat invoer van studenten een passende reactie krijgt, is het vermogen van digitale onderwijssystemen om het lesmateriaal aan te passen aan de individuele gebruiker, de zogeheten adaptiviteit, nog niet sterk ontwikkeld. Technisch zijn er weinig belemmeringen en het nut is groot. Daarom verwacht ik dat adaptiviteit snel meer zal opleveren. Uit de logs van individuele studenten in een interactieve cursus valt af te leiden waar ze sterk en waar ze zwak in zijn, wat ze al beheersen en wat nog niet. De beginnende student kan onderworpen worden aan een diagnostische toets, waardoor hij/zij geleid wordt naar de onderwerpen die het eerst in een bepaalde cursus aan de orde zouden moeten komen. Het niveau kan aangepast worden aan de hand van al bekende gegevens, en de uitleg op een theoriepagina kan zich beperken tot onderwerpen die nog niet bekend verondersteld worden. Ook kan de aanpassing afhangen van individueel aansprekende voorbeelden, cognitieve eigenschappen en leerstijl van de student.

De principes om de studenten aan te spreken op hun individuele kwaliteiten kunnen ook gebruikt worden om degenen die daartoe in staat zijn, een beter beeld te geven van wat wiskunde inhoudt. In het vo voorziet het vak wiskunde D al enigszins in die behoefte; maar is het niet veel effectiever als de verdieping direct geboden wordt waar de student er aan toe blijkt te zijn? Om een voorbeeld te noemen: van de kwadratische vergelijkingen in twee variabelen worden de oplossingsverzamelingen in specifieke gevallen benoemd als parabool, hyperbool of ellips. De classificatie van deze objecten als alle objecten (op ontaardingen na) die je zo kunt krijgen, is geen onderdeel van het vo-curriculum. In een adaptieve omgeving kan de goede en geïnteresseerde student de achterliggende stelling leren kennen en zo de kracht van de wiskunde leren waarderen.

Groter arsenaal aan oefeningen

Niet alleen kan een student er 123 keer over doen voor hij/zij het goede antwoord op een oefenopgave vindt, ook kan de student in een interactieve leeromgeving vaak heel veel verschillende opgaven van hetzelfde type maken. Het is bijvoorbeeld geen probleem om opgaven met random parameters a en b te genereren, zodat $\sqrt{a^2+b^2}$ een geheel getal is, met als oogmerk dat het antwoord op een vraag over de stelling van Pythagoras zonder wortelteken geschreven kan worden. Een tekstboek bevat slechts enkele combinaties van die waarden. Het grote aantal komt niet alleen goed van pas bij studenten die meer oefening behoeven, maar ook bij de productie van meerdere versies van een tentamen bij één zitting.

Gemak voor de docent

Het gemak voor de docent wordt goed gediend door nieuw digitaal lesmateriaal. Het delen van opgaven onder docenten is bijvoorbeeld eenvoudiger geworden als ze over een gemeenschappelijke databank beschikken; dit idee is uitgewerkt in het SURF-project ONBETWIST [30] en wordt ook toegepast door SOWISO. Via de toegangspoorten tot dergelijke databanken word je al gauw lid van een gemeenschap van collega's met overeenkomstige interesses, met alle voordelen van dien.

Aan meerdere universiteiten in Nederland worden inmiddels met succes volledig digitale tentamens wiskunde met open vragen afgenomen. De docent selecteert de vraagstukken uit een databank van geparametriseerde vragen. Het nakijkwerk is geautomatiseerd, zodat de uitslag van het tentamen binnen enkele tellen na afloop van het tentamen bekend is. Aan de Technische Universiteit Eindhoven worden relevante gegevens over de vorderingen van studenten bij online vraagstukken aan de tutoren doorgegeven, zodat de besprekingen met studenten specifieker en intensiever kunnen zijn dan voorheen.

Met behulp van digitaal materiaal kunnen docenten zich dus meer richten op het creatieve deel van het lesgeven.

Mogelijke verbeteringen

Bij vergelijking van de studierendementen doen online cursussen het in het algemeen (nog?) iets minder goed dan de klassieke (face-to-face) cursussen; zie bijvoorbeeld [47]. Gezien de jonge leeftijd van het digitale materiaal is dat geen verrassing. Er zal ongetwijfeld nog veel geëxperimenteerd worden om tot een beter product te komen en het beter in te kunnen zetten.

Hoewel ik me niet op statistisch verantwoord onderzoek kan beroepen, zie ik in het algemeen twee trends bij de aanbieding van online onderwijs. De ene trend is gekoppeld aan succes: de individuele doorzetter lijkt ver te komen. Studenten met een voldoende aantal opgaven achter de kiezen in de interactieve omgeving behalen vrijwel altijd een goed eindresultaat. Dit is op zich al een bemoedigende ontwikkeling.

De andere trend is dat er steeds een aanzienlijke groep studenten blijkt te zijn die vrijwel meteen vastloopt. Dit kan aan de steile leercurve in het lesmateriaal liggen; de digitale versies moeten nog uitrijpen. Maar er zijn ook aanwijzingen (zie bijvoorbeeld [3]) dat hier gebrek aan contact met docenten, motivatie, beschikbare tijd en planning van de deelnemende student een grote rol spelen. Op de eerste twee oorzaken ga ik nader in. Daarna bespreek ik nog twee punten waaraan nog veel te doen is: beveiliging en het leren bewijzen.

Contact met docenten

Het verschil in rendement wordt vaak verklaard door het geringe contact met docenten. Bij een cursus op een specifieke locatie kan gekozen worden voor een vorm van normaal contact tussen student en docent via colleges of practica, ondersteund door digitaal aangeboden oefeningen (al dan niet voorafgaand aan ontmoetingen met een docent). Voor studenten op afstand is het vanzelfsprekend al winst dat zij de cursus überhaupt kunnen volgen. De afwezigheid van het natuurlijke contact met docenten kan ten dele gecompenseerd worden door forum posts, e-mails en chats met medestudenten en/of docenten.

Motivatie

Het hooghouden van de motivatie wordt ook als probleem genoemd bij online cursussen. De vaak eenzame positie van de online student vereist een sterke persoonlijke instelling om door te kunnen zetten. Zo'n motivatie is vaak al aanwezig bij studenten die zich bij een voorbereidende digitale cursus aanmelden voor een opleiding waar ze zich eerder voor hebben ingeschreven. Bij de Masteropleidingen aan de Technische Universiteit Eindhoven en aan een economiefaculteit elders in Nederland kregen in de zomer van 2015 alle van buiten instromende studenten de kans de wiskunde tot op het verwachte niveau te brengen vóór hun fysieke komst in september. Voor de een bleek de cursus een nuttige opfrisser, voor de ander een moeilijke horde, maar in de meeste gevallen is de aansluiting met het bestaande onderwijs verbeterd.

Figuur 2 Voorbeeld van een opgave in Spatchcoq.

MOOC-organisaties als Coursera proberen ook de motivatie hoog te houden door aantrekkelijke mogelijkheden na het met succes volgen van de cursus aan te bieden. Ze zijn druk in de weer om toelating op grond van dat succes als student tot bestaande universiteiten of als medewerker tot grote firma's te vergemakkelijken.

Als remedie voor gebrek aan motivatie worden wel spelelementen ingebouwd: *gamification*. Deze kunnen stimulerend werken, maar kunnen ook indruisen tegen de opvattingen hoe je studenten zou moeten benaderen. Zo is het in Singapore meer gebruikelijk dan in Nederland om te laten zien hoe goed een student het ten opzichte van andere studenten gedaan heeft. Twee mooie milde vormen van gamification zijn te vinden in de Rekentuin [36] (voor elementaire rekenoefeningen), dat de prestaties van de studenten onder andere gebruikt om de opgaven naar moeilijkheidsgraad in te delen en MemRise [29], dat door gebruikers samengestelde geheugenoefeningen aanbiedt, vooral voor vreemde talen maar ook een klein beetje voor wiskunde.

Beveiliging

Bij summatieve toetsen voor een examen, toelating of certificaat zijn veel voorzieningen op het gebied van beveiliging en privacy nodig. Drie voorbeelden van de zorgen die daarbij naar voren komen zijn: het op afstand vaststellen van de identiteit van de maker van de toets, de hulp van anderen bij de toets en het gebruik van software om de toets te maken die niet toegestaan is. Er zijn al heel wat firma's (zie bijvoorbeeld [19, 34, 35, 37, 39]) die examens zo veilig mogelijk afnemen bij studenten met een internet-aansluiting en een webcam.

Leren bewijzen

Voor het digitaal aanleren van wiskundige bewijzen is nog geen goede methode voorhanden. Een van de meest vergaande pogingen in deze richting is een inleiding in de logica voor studenten filosofie aan de Carnegie Mellon University [1]. Daarin wordt een automatische bewijsassistent gebruikt in een venster naast het dictaat om de student zelf met de aangeboden bewijzen te laten werken. Net als bij *Algebra Interactive* [5] wordt de student dus gevraagd de taal van het achterliggende programma, hier een vorm van formele wiskunde, te gebruiken bij de invoer. De

Bruijn [4], een van de grote pioniers van de formele wiskunde, suggereerde een 'vernacular', een taaltje dat weliswaar dicht bij de gebruikelijke taal staat die je in bewijzen tegenkomt, maar dat precies genoeg is om omgezet te kunnen te worden in formele wiskunde. Deze taal is nog niet ontwikkeld en de wiskundige gemeenschap loopt er nog niet warm voor. Toch verwacht ik dat het er binnen afzienbare tijd van gaat komen. Een aantal wiskundigen gaat er serieus mee aan de slag, zoals Ganesalingam en Gowers [11] voor wat betreft het vernacular, en Blok en Hoffman [3], die een interface bouwen om studenten bewijsvoering aan te leren volgens bekende wiskundige gewoonten met gebruik op de achtergrond van de automatische bewijsverificator en ondersteuningssoftware Coq [7].

Tot slot

Van alle mogelijke vormen van digitaal wiskundig lesmateriaal heb ik maar een klein deel aangestipt. Er valt bijvoorbeeld nog veel te zeggen over de mooie software die opgesomd staat op websites als [33] en [16]; daar zit veel goed materiaal bij om online onderwijs te verlevendigen. De learning management systems (LMS), bedoeld voor het beheer van cursussen aan onderwijsinstellingen, zijn helemaal niet genoemd. Maar ik heb duidelijk willen maken dat digitaal ondersteund onderwijs al zeer nuttig is in het geval van basisvakken wiskunde (voor de individueel opererende student, voor de klassen die een blended cursus volgen en voor hun docenten), en dat we uit kunnen zien naar veel meer en beter digitaal lesmateriaal, zelfs als het gaat om bewijsvoering voor wiskundestudenten.

Dankwoord

Hierbij bedank ik Max Cohen, Hans Cuypers, Marc Habbema, Wim Honselaar, Bas Spitters en Hans Sterk voor hun goede suggesties bij het schrijven van dit artikel.

Referenties

1 J. Avigad, Logic and Proof, avigad.github.io/logic_and_proof.
2 R. Blok en C. Hoffman, Spatchcoq, https://spatchcockedblog.wordpress.com.
3 Sean B. Eom, H. Joseph Wen en Nicholas Ashill, The determinants of students' perceived learning outcomes and satisfaction in university online education: an empirical investigation, *Decision Sciences Journal of Innovative Education* 4 (2006), 215–235.
4 N.G. de Bruijn, The mathematical vernacular, a language for mathematics with typed sets, in P. Dybjer e.a., eds., *Proceedings of the Workshop on Programming Languages, Marstrand*, Sweden, 1987.
5 A.M. Cohen, H. Cuypers en H. Sterk, *Algebra Interactive*, Springer, 1999.
6 A.M. Cohen en L. Meertens, The ACELA project: aims and plans, in: N. Kajler, ed., *Computer-Human interaction in Symbolic Computation, Texts and Monographs in Symbolic Computation*, Springer, 1998 pp. 7–23.
7 Coq, bewijsassistent, coq.inria.fr.
8 Coursera, MOOC, www.coursera.org.
9 EdX, MOOC, www.edx.org.
10 P. Eshuis, R. Houtenbos en J. Boertjens, Digitalisering van wiskunde in het voortgezet onderwijs, *Nieuw Archief voor Wiskunde* 5/17 (2016), 10–14.
11 M. Ganesalingam en W.T. Gowers, A fully automatic problem solver with human-style output, arxiv.org/abs/1309.4501.
12 GAP, www.gap-system.org.
13 GeoGebra, Graphing calculator for functions, geometry, algebra, calculus, statistics and 3D mathematics, www.geogebra.org.
14 *Getal & Ruimte* en *Moderne Wiskunde*, Noordhoff, Groningen.
15 Willem de Graaf, *Lie Algebras: Theory and Algorithms,* deel 56 van North-Holland Mathematical Library, 2000.
16 Imaginary, platform for open and interactive mathematics, imaginary.org.
17 Khan Academy, www.khanacademy.org/math.
18 D. Koller, What we're learning from online education, TED-voordracht, 2012, https://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education.
19 Kryterion, www.kryterion.com.
20 L. Lovász, *An Algorithmic Theory of Numbers, Graphs and Convexity*, SIAM, Philadelphia, 1986.
21 Magma Simons Agreement, magma.maths.usyd.edu.au/magma/simons_details.
22 MapleTA, www.maplesoft.com/mapleta.
23 MathDox formula editor, mathdox.org/formulaeditor.
24 Mathematica, www.wolfram.com/mathematica.
25 Mathematica Notebooks, www.wolfram.com/mathematica.
26 Mathjax, www.mathjax.org.
27 MathML, www.mathml.org.
28 Maxima, maxima.sourceforge.net.
29 MemRise, Ed Cooke en Greg Detre, online cursusmateriaal gecreëerd door gebruikers, memrise.com.
30 ONBETWIST: ONderwijs verBETeren met WISkunde Toetsen, SURF-project, www.onbetwist.org.
31 OpenDreamKit project, opendreamkit.org.
32 OpenMath, www.openmath.org.
33 ORMS: Oberwolfach References on Mathematical Software, orms.mfo.de.
34 Proctorexam, proctorexam.com.
35 ProctorU, www.proctoru.com.
36 Rekentuin, online rekenoefenprogramma voor po en vo, www.rekentuin.nl.
37 Respondus, www.respondus.com.
38 SageMath, www.sagemath.org.
39 Software Secure, www.softwaresecure.com.
40 STACK, www.stack.bham.ac.uk.
41 Terence Tao, blog, terrytao.wordpress.com.
42 WebMathematica, http://www.wolfram.com/products/webmathematica'.
43 WebMathematica, http://www.wolfram.com/mathematica/online.
44 Wikipedia entry on Computable number, en.wikipedia.org/wiki/Computable_number.
45 Wolframalpha, www.wolframalpha.com.
46 XML, www.xml.org.
47 Di Xu en Shanna Smith Jaggars, The impact of online learning on students' course outcomes: Evidence from a large community and technical college system, *Economics of Education Review* 37 (2013) 46–57.

Assia Mahboubi
Inria
Palaiseau, France
assia.mahboubi@inria.fr

Machine-checked mathematics

Assia Mahboubi is a permanent researcher at the French research institute Inria. She also works at the Microsoft-Inria Joint Centre. She has contributed to the formalization of the Feit–Thompson theorem in the Coq proof assistant and is currently working on verified computer algebra. In this article she gives an overview about machine-checked mathematics.

Proof assistants are software tools dedicated to the conception of digital libraries of formalized mathematics, that can be *machine-checked* automatically. So far mostly used by researchers in computer science, they have recently started to attract the attention of researchers in pure mathematics, thanks to the successful formalization of recent and major mathematical results. This article discusses the issues and the benefits of this alternate way of doing mathematics with the help of a computer.

Black boards and computers

The popular representation of a researcher is usually a person wearing a white lab coat and playing with an emblematic tool of her discipline: a microscope for a biologist, glasses and test tubes for a chemist, a telescope for the astrophysicist, ... But what does a mathematics laboratory look like? Researchers in mathematics do not need lab coats nor protection glasses. But if you ask them about it, they will probably make a strong requirement about at least their office and seminar rooms being equipped with a black board, and even may be other places like the cafeteria too. There is certainly a black board – white board preference split, just like there is a coffee – tea one, but in any case these vertical surfaces play a fundamental role in the most creative and fun part of the activity of research in mathematics: understanding a proof, a definition, shaking a wannabe theorem by trying to find counter-examples, explain and explore new ideas ...

Since the last half of the twentieth century however, computers have deeply changed the face of research in mathematics. Mathematicians nowadays seldom exchange snail mails, as J.-P. Serre and A. Grothendieck [6, 19] did not so long ago, but emails — see for instance the correspondence of the two French mathematicians C. Villani and C. Mouhot, reprinted in part in C. Villani's novel *Birth of a Theorem* [22]. They write research blog posts, they typeset themselves their articles using scientific document preparation systems, and they access them via the Internet — and less and less often at the library. Computers have also relieved mathematicians from pedestrian calculation work, now handed down to scientific computing software, able to compute more than what would be possible in a life time. This easy access to heavy computations has even motivated new forms of mathematics, like computer algebra or numerical analysis. In various areas of mathematics, there are now theorems whose sole known proofs rely as of today on the execution of a computer program, and on superhuman computational resources. Famous examples include the Four Colour Theorem [1], the existence of the Lorentz attractor [21], the Kepler conjecture [10] or the weak Goldbach conjecture [12], ...

What is a theorem?

The most visible outcome of research in mathematics is the production of new results, called theorems, together with their proofs. What makes a candidate proof become a theorem is a blend of a social process and of a scrutinization: people give talks explaining their new ideas to their peers, they eventually submit papers to journals, and these submissions are carefully reviewed by anonymous colleagues. In the vast majority of cases, everything goes fine and only minor errors subsist in published texts, that can easily be fixed by the target audience of specialists. But sometimes proofs are published that are truly incomplete or essentially incorrect, and nonetheless believed to be true for a while. There are notorious long sagas of trials and errors like the Four Colour Theorem [8] or Hilbert's 16th problem [13] — although the latter example is an emblematic example of how fertile false proofs can turn out to be. It is indeed in some cases very difficult to find reviewers who are at the same time: competent enough in all the areas of expertise required, focused enough to detect all the potential flaws, and insensitive to subjective information like their personal relation to the author or his/her reputation in the com-

munity. On this topic, the interested reader may watch V. Voevodsky explaining his misfortunes in a popular science talk at the Institute of Advanced Study, Princeton (http://video.ias.edu/node/6395). We also recommend the reading of W. Thurston's reflexions [20] on the process of mathematics.

In principle, the whole mathematical literature could be expressed in a non-ambiguous formal language, like set-theory and first-order logic, and proofs could be detailed in an exhaustive manner, making verification (boring to death but) trivial and absolutely objective. But this is not how one communicates mathematics in practice. The language of mathematics is made of a complex apparatus of abstractions, ellipsis and notations which requires some culture, years of training, but which also just makes the communication of ideas possible. Providing all the seemingly missing details to an audience of specialists would not only be utterly pedantic, but in fact it would soon obfuscate the discourse completely, sterilize creativity and miss the point of what makes mathematics beautiful. The famous group of French mathematicians N. Bourbaki make it explicit in *The Architecture of Mathematics* [4]: "What the axiomatic method sets as its essential aim is exactly that which logical formalism by itself cannot provide, namely the profound intelligibility of mathematics." Clarifying the abstract structures, which capture the deep similarities shared by seemingly extremely different objects is what it is about. "To lay down the rules of this language, to set up its vocabulary and to clarify its syntax, all that is indeed extremely useful; indeed this constitutes one aspect of the axiomatic method, the one that can properly be called logical formalism (or 'logistics' as it is sometimes called). But we emphasize that it is but one aspect of this method, indeed the least interesting one." [4]. To summarize, a verbose expansion of all the definitions and arguments of a given proof, down to the initial actions of a logical foundation, would both be too boring a task and fail to help understanding its content and checking its correctness.

Photo: flickr.com. Nelson Minar

W.S. Jevons' Logic Piano

Machines can help

Except may be if one could take benefit from the super computing powers of machines... The idea of building machines in order to mechanize the process of verifying deductions is certainly an old one (see for instance G. Leibniz' Calculus Raciocinator in his 1966 doctoral thesis, or W.S. Jevons' Logic Piano built in 1869), but the advent of computers, modern logic and computer science have turned it into concrete tools that can help for real. Indeed *Proof Assistants* are pieces of software that provide an environment for defining mathematical objects, their properties and the associated candidate proofs in a formal language well suited to this purpose. There are many such proof assistants, just like there are many programming languages, computer algebra systems or typesetting softwares. But all of

them share the same general anatomy. The first ingredient of a proof assistant is the logical foundation used to fix the formal language of mathematics: although most mathematicians are used to set theory and first-order logic, there are many more possible options, some being better-suited to the formalization of mathematical concepts *in practice*. The majority of available proof assistants are actually based on an instance of *type theory*, instead of set theory, and allow quantification on arbitrary objects and functions, instead of being limited to first-order logic. The grammatical constructions and rules of this language are rigid enough so that a program called the *kernel* of the proof assistant can be used to check the correctness of proofs written by the user, by means of a purely mechanical automated process. The kernel is really the cornerstone of a proof assistant: if one trusts this single piece of code, one trusts all the proofs it validates. Formalizing mathematics with a proof assistant means forging *by hand* a formal description of the mathematical definitions, theorem statements and candidate proofs: only the verification of the latter is automated, and not its discovery. Therefore the main difficulty is to bridge the gap between the very low level language that allows for this routine verification and the much higher-level language that make mathematics intelligible to humans. This gap is akin to the distance between the language in which human beings write programs and the patterns of bits corresponding to the physical commands executed by the processor. The bulk of proof assistants is thus to provide tools that help with this matter. The code implementing these tools is carefully separated from the one of the kernel, so that the trusted base of code remains clearly identified.

Formalizing mathematics

Proof assistants have been around for about half a century now, since the pioneering work of N. G. de Bruijn [16] on his Automath system (a project he started in 1967). These systems have been mostly used by computer science inclined users, concerned with obtaining the highest possible confidence in the behavior of *programs*. Describing the properties of the output of an algorithm and the behavior of the program implementing this algorithm in a given language is indeed a notoriously difficult task. Impressive successes have been obtained in this area in the last decade like the verification of the core of an operating system [14] and the implementation of a certified compiler for (a subset of) the C programming language [15]. Another flagship application of formal proofs is the verification of mathematical results that depend on heavy calculations. Some major mathematical results of this kind have also been machine-checked using proof assistants, like the Four Colour Theorem [8] or Hales' proof of the Kepler conjecture [11], the latter completed after over a decade of collaborative work under the direction of Th. Hales. Proof assistants and formal proofs have nonetheless failed so far to catch the interest of the majority of researchers in mathematics. One of the reasons that may explain this situation is a general feeling that the technology is not yet mature enough. In particular, a blatant lack of significant libraries of digitized mathematics has long prevented interested external users to consider starting a formalization project of their own research results, to the notable exception of Th. Hales work on his formal proof of the Kepler conjecture. Formalizing the proof of a theorem with a proof assistant remains immensely more time-consuming than writing it down with a scientific word processing software. And to the best of our knowledge, no existing proof assistant can pretend to offer a digital formal library covering a comprehensive graduate curriculum in pure mathematics.

A formal proof of the Odd Order Theorem

Indeed, finding the most appropriate methodology and techniques to craft formal definitions of mathematical objects in type theory and to construct reusable formal libraries remains to a large extent a research topic. These issues motivated the work of a team of researchers led by G. Gonthier on the formal verification of a landmark result in finite group theory due to W. Feit and J. G. Thompson, called the Odd Order Theorem [7]. This project uses a proof assistant called the Coq system and took six years to be completed [9]. The algebraic structure of group is an abstract object which is at the same time very simple, and very deep: F. Klein, author of the seminal Erlangen research program, defined geometry as the study of properties that remain invariant under the action of a given group of transformations. *Simple groups* are the most elementary, atomic instances of groups and any other finite group can be built from the finite simple ones — like molecules from these atoms. The classification of finite simple groups describes precisely the shape that a finite simple group can take, for each possible cardinal, and the Odd Order Theorem is a corner stone of this edifice:

Theorem [Odd Order Theorem]. *Every finite group of odd order is solvable*.

As a direct consequence, the Odd Order Theorem actually provides a complete description of the structure of simple groups with an odd cardinal: they are necessarily cyclic. The simplicity of the statement of this result strikingly contrasts with the sophistication of its proof, which calls for a combination of arguments of local analysis and of character theory of finite groups. The original proof [7] was published as a 250 page paper, a record at the time, and was later extensively polished and revised [3, 17]. R. Solomon describes its impact by saying that: "This short sentence and its long proof were a moment in the evolution of finite group theory analogous to the emergence of fish onto dry land. Nothing like it had happened before; nothing quite like it has happened since." [18] The validity of the Odd Order Theorem itself has never been really called into question, but the story of the classification of finite simple groups has been

The author's Odd Order Theorem bookshelf

```
Section JacobsonDensity.

Variables (gT : finGroupType) (G : {group gT}) (n : nat).
Variable rG : mx_representation F G n.
Hypothesis irrG : mx_irreducible rG.

Local Notation E_G := (enveloping_algebra_mx rG).
Local Notation Hom_G := 'C(E_G)%MS.

Lemma mx_Jacobson_density : ('C(Hom_G) <= E_G)%MS.
Proof.
apply/row_subP=> iB; rewrite -[row iB _]vec_mxK; move defB: (vec_mx _) => B.
have{defB} cBcE: (B \in 'C(Hom_G))%MS by rewrite -defB vec_mxK row_sub.
have rGnP: mx_repr G (fun x => lin_mx (mulmxr (rG x)) : 'A_n).
  split=> [|x y Gx Gy]; apply/row_matrixP=> i.
    by rewrite !rowE mul_rV_lin repr_mx1 /= !mulmx1 vec_mxK.
  by rewrite !rowE mulmxA !mul_rV_lin repr_mxM //= mxvecK mulmxA.
move def_rGn: (MxRepresentation rGnP) => rGn.
pose E_Gn := enveloping_algebra_mx rGn.
pose e1 : 'rV[F]_(n ^ 2) := mxvec 1%:M; pose U := cyclic_mx rGn e1.
have U_e1: (e1 <= U)%MS by rewrite cyclic_mx_id.
have modU: mxmodule rGn U by rewrite cyclic_mx_module.
pose Bn : 'M_(n ^ 2) := lin_mx (mulmxr B).
suffices U_e1Bn: (e1 *m Bn <= U)%MS.
  rewrite mul_vec_lin /= mul1mx in U_e1Bn; apply: submx_trans U_e1Bn _.
  rewrite genmxE; apply/row_subP=> i; rewrite row_mul rowK mul_vec_lin_row.
  by rewrite -def_rGn mul_vec_lin /= mul1mx (eq_row_sub i) ?rowK.
have{cBcE} cBncEn A: centgmx rGn A -> A *m Bn = Bn *m A.
  rewrite -def_rGn => cAG; apply/row_matrixP; case/mxvec_indexP=> j k /=.
  rewrite !rowE !mulmxA -mxvec_delta -(mul_delta_mx (0 : 'I_1)).
  rewrite mul_rV_lin mul_vec_lin /= -mulmxA; apply: (canLR vec_mxK).
  apply/row_matrixP=> i; set dj0 := delta_mx j 0.
  pose Aij := row i \o vec_mx \o mulmxr A \o mxvec \o mulmx dj0.
  have defAij := mul_rV_lin1 [linear of Aij]; rewrite /= [2]/Aij /= in defAij.
  rewrite -defAij row_mul -defAij -!mulmxA (cent_mxP cBcE) {k}//.
  rewrite memmx_cent_envelop; apply/centgmxP=> x Gx; apply/row_matrixP=> k.
  rewrite !row_mul !rowE !{}defAij /= -row_mul mulmxA mul_delta_mx.
  congr (row i _); rewrite -(mul_vec_lin (mulmxr_linear _ _)) -mulmxA.
  by rewrite -(centgmxP cAG) // mulmxA mx_rV_lin.
suffices redGn: mx_completely_reducible rGn 1%:M.
  have [V modV defUV] := redGn _ modU (submx1 _); move/mxdirect_addsP=> dxUV.
```

A piece of Coq code

much more uneven and controversial, with famous skeptics like J.P. Serre [5]. As of today, the confidence of mathematicians in this complex edifice still rests more upon the reputation of the rare experts who are able to understand this composite proof in extenso than on a genuine assimilation by the community.

Teaching proof assistants

The first and foremost motivation of the formalization was not to track petty errors and holes in this proof. The interest of this proof as a case study for the formalization of mathematics, beside the significance of the result, is the broad spectrum of algebraic theories in its prerequisite, like graduate-level linear and multilinear algebra, Galois theory, representation theory and character theory of finite groups, constructions of algebraic closures,... As of today, this realization constitutes the largest corpus of algebraic theories ever formalized. The challenge was hence to represent formally all the mathematical objects that play a role in this proof: starting from the definition of natural numbers, and covering the 250 pages of the proof, plus all the pre-requisite in-between.

One of the main issues in formalization is to model the aforementioned training the mathematician went through to be able to make sense of a mathematical text. Consider for instance the sentence:

> The determinant of a square matrix $A \in M_n(R)$ is $\mathrm{Det}(\mathrm{A}) = \sum_{\sigma \in S_n} \epsilon_\sigma \prod_i a_{\sigma(i),i}$.

The average reader is able to infer without thinking about it that the Greek letters Σ and Π, notations for iterated binary operations on the domain described in subscript, here apply respectively to the addition and to the multiplication of the ring structure which equips R, that the iteration domain of the product is necessarily $1 \leq i \leq n$ and that the signature ϵ_σ should be understood as its embedding in R. Although none of these details is explicitly denoted in the formula, they can all be deduced from the context of the formula. The information seemingly missing in such a string of symbols is however necessary for the computer to make sense of the statement. Yet it would not be reasonable to expect the user of the proof assistant to provide it by hand by decorating of the words constituting the formal sentence. (S)he would soon no more see the forest for the trees. Fortunately the training of the reader can be modeled by implementing and running appropriate algorithms which are able to infer the canonical properties hidden behind standard notations, once these rules and habits have been explicited completely. For this purpose, it is very helpful to work with a formal language based on a flavor of *type* theory. As in usual programming languages, types are labels carrying information like the domain and codomain loci of functions. Well-formed sentences satisfy constraints on their types, which are prescribed by the rules of the formal language. This helps structuring the sentences and ruling out nonsensical ones, hence facilitating the verification. But not all type annotations need to be provided by the user, as some of them can be retrieved by just enforcing the constraints on the types. This mechanism, called *type inference*, also comes from the theory of programming languages and can be used to reconstruct what the trained reader can read behind the notations, by encoding enough information in the types. This has been extensively used in the libraries of the Odd Order Theorem formal proof to preserve the readability of statements, the modularity of the theories and the ability to superimpose several views on a same object. In the end, the code producing the formula

$$\sum_{\sigma \in S_n} \epsilon_\sigma \prod_i a_{\sigma(i),i}$$

in the LaTeX typesetting language is:

```
\sum_{\sigma \in S_n}\epsilon_
{\sigma}\prod_i a_{\sigma (i),i}
```

when an analogue in Coq would be:

```
Definition det (R : ringType) n
(A : 'M[R]_n) : R := \sum_(s : 'S_n)
(-1) ^+ s * \prod_i A i (s i).
```

For a more detailed insight into the use of type inference in the formalization of mathematics, the interested reader can refer to J. Avigad's introduction [2].

Computers and black boards

Writing machine-checked digital libraries of formalized mathematics is not like using a super spell-checker for mathematics, nor does it intend to superseed the work of human reviewers — although it will hopefully assist them in this task. Formalizing mathematics is a creative research activity, not because it produces new, previously unknown theorems, but because it sparks off a reflection on the most appropriate formal representation of mathematical objects, which allows for the coherence of the involved theories and provides an appropriate socle for the upcoming layers of formalization. Revisiting mathematical theories with the help of a proof assistant is not only about chasing errors in proofs: it helps exploring and improving the architecture of a proof and the layout of the theories at stake, with the precious help of a computer. Proof assistants can help teaching mathematics, they could foster collaborative work and may be one day allow the discovery of new objects, of useful abstractions. Although the technology of proof assistants is actually already mature enough to serve the purpose of the verification of any theorem, all the components of their skeleton are still active topics of research: the logical foundations that should provide the most convenient framework for the mechanization of proofs, the automation tools helping best the users, the best user interface,... Hopefully proof assistants will rapidly evolve towards as widely used tools as computer algebra systems and eventually change the reviewing standards in mathematical journals. But there is little doubt that black (or white) board will remain the most vital piece in a mathematician's office.

References

1 Kenneth Appel and Wolfgang Haken, *Every Planar Map is Four Colorable*, Contemporary Mathematics, Vol. 98, American Mathematical Society, 1989. With the collaboration of J. Koch.

2 Jeremy Avigad, Type inference in mathematics, *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* 106 (2012), 78–98.

3 Helmut Bender and Georges Glauberman, *Local Analysis for the Odd Order Theorem*, London Mathematical Society, LNS, Vol. 188, Cambridge University Press, 1994.

4 Nicolas Bourbaki, The architecture of mathematics, *The American Mathematical Monthly* 57(4), 1950.

5 C.T. Chong and Y.K. Leong, An interview with Jean-Pierre Serre, *Math. Intelligencer* 8(4) (1986), 8–13.

6 Pierre Colmez and Jean-Pierre Serre, eds., *Grothendieck–Serre correspondence*, Bilingual Edition, American Mathematical Society – Societe Mathematique de France, 2004.

7 Walter Feit and John G. Thompson, Solvability of groups of odd order, *Pacic Journal of Mathematics* 13(3) (1963), 775–1029.

8 Georges Gonthier, Formal proofIthe four-color theorem, *Notices Amer. Math. Soc.* 55(11) (2008), 1382–1393.

9 Georges Gonthier et al., A machine-checked proof of the odd order theorem, in *Interactive Theorem Proving*, Lecture Notes in Comput. Sci., Vol. 7998, Springer, 2013, pp. 163–179.

10 Thomas C. Hales, A proof of the Kepler conjecture, *Ann. of Math. (2)* 162(3) (2005), 1065–1185.

11 Thomas C. Hales et al., A formal proof of the Kepler conjecture, arxiv.org/abs/1501.02155, January 2015.

12 Harald A. Helfgott, The ternary Goldbach conjecture is true, arxiv.org/abs/1312.7748, 2013.

13 Yu. Ilyashenko, Centennial history of Hilbert's 16th problem, *Bull. Amer. Math. Soc. (N.S.)* 39(3) (2002), 301–354.

14 Gerwin Klein et al., sel4: formal verification of an os kernel, in *SOPS ACM SIGOPS*, 2009, pp. 207–220.

15 Xavier Leroy, Formal certification of a compiler back-end, or: programming a compiler with a proof assistant, in *POPL*, ACM Press, 2006, pp. 42–54.

16 R.P. Nederpelt, J.H. Geuvers and R.C. de Vrijer, eds., *Selected Papers on Automath*, Studies in Logic and the Foundations of Mathematics, Vol. 133, North-Holland, 1994.

17 Thomas Peterfalvi, *Character Theory for the Odd Order Theorem*, London Mathematical Society, LNS, Vol. 272, Cambridge University Press, 2000.

18 Ronald Solomon, A brief history of the classification of the finite simple groups, *Bull. Amer. Math. Soc. (N.S.)* 38(3) (2001), 315–352.

19 John Tate, Correspondance Grothendieck–Serre, *Nieuw Archief voor Wiskunde* 5/5(1) (2004), 42–44.

20 William P. Thurston, On proof and progress in mathematics, *Bull. Amer. Math. Soc. (N.S.)* 30(2) (1994), 161–177.

21 Warwick Tucker, The Lorenz attractor exists, *Comptes Rendus de l'Academie des Sciences, Series I – Mathematics* 328(12) (1999), 1197–1202.

22 Cedric Villani, *Birth of a Theorem: A Mathematical Adventure*, Farrar, Straus and Giroux, April 2015.

Freek Wiedijk
Faculteit FNWI, iCIS
Radboud Universiteit Nijmegen
f.wiedijk@cs.ru.nl

Herman Geuvers
Faculteit FNWI, iCIS
Radboud Universiteit Nijmegen
h.geuvers@cs.ru.nl

Josef Urban
CIIRC
Czech Technical University, Prague
josef.urban@gmail.com

Een wiskundig bewijs correct bewezen: De meest efficiënte manier om bollen op te stapelen

Sommige bewijzen van wiskundige stellingen zijn zo bewerkelijk dat ze alleen met computerondersteuning geverifieerd kunnen worden. Een voorbeeld is het bewijs van de vierkleurenstelling, die zegt dat iedere landkaart met vier kleuren ingekleurd kan worden zonder dat aangrenzende landen dezelfde kleur krijgen. Een recent voorbeeld is het bewijs van het vermoeden van Kepler, dat zegt dat de meest voor de hand liggende stapeling van bollen in de ruimte (zoals de groenteman sinaasappelen stapelt) ook de meest efficiënte is. In dit artikel geven Freek Wiedijk, Herman Geuvers en Josef Urban een overzicht van het wiskundige deel van het bewijs en leggen uit op welke manier de computer bij de verificatie is gebruikt.

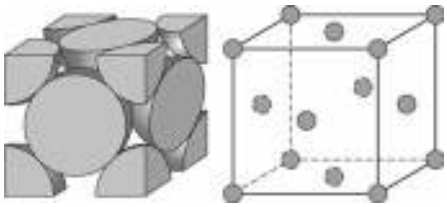
Johannes Kepler beschreef zijn vermoeden in 1611 en in 1998 kondigde de Amerikaan Tom Hales aan het bewezen te hebben. Zijn bewijs reduceert het probleem tot een kleine 20.000 mogelijke tegenvoorbeelden, die vervolgens verworpen worden. Daarvoor worden meer dan 23.000 niet-lineaire vergelijkingen opgelost en wordt aangetoond dat meer dan 43.000 lineaire programma's onoplosbaar zijn. Dat stuk van het bewijs is gedaan met computerprogramma's, en dit deel werd door de wiskundige reviewers niet geaccepteerd. Als reactie daarop is Hales het 'Flyspeck'-project begonnen, met als doel het gehele bewijs met een bewijsassistent te verifieren. Een bewijsassistent is een computerprogramma waarmee een gebruiker interactief een wiskundig bewijs construeert, dat vervolgens door het programma gecheckt wordt. Omdat bewijsassistenten een buitengewoon hoge betrouwbaarheid hebben zijn er daarna geen referenten meer nodig. In 2014 is het Flyspeck-project voltooid.

Het Kepler-vermoeden

Stel we hebben oneindig veel even grote bollen, en we willen die zo dicht mogelijk tegen elkaar in de driedimensionale ruimte plaatsen. Wat is de manier om dat met de grootst mogelijke dichtheid te doen?

Een voor de hand liggende manier is om de rangschikking te kiezen die een groenteboer voor sinaasappelen gebruikt, en die ook gebruikt wordt om kanonskogels naast een kanon op te stapelen. Dit is de zogenaamde *kubische vlakgecentreerde stapeling* (in het Engels: face centered cubic of FCC packing). In deze

Figuur 1 Efficiënt opgestapelde sneeuwballen.



Figuur 2 De FCC-stapeling.

stapeling is

$$\frac{\pi}{\sqrt{18}} = 74{,}0480\ldots\,\%$$

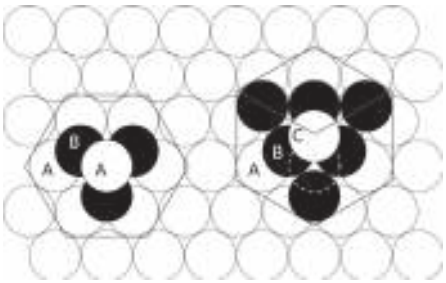
van de ruimte gevuld. Bij deze stapeling liggen de middelpunten van de bollen zowel op de hoekpunten als in het midden van de zijvlakken van de kubussen uit een driedimensionaal kubisch rooster. De roosterpuntafstand is $2r\sqrt{2}$ als r de straal van de bollen is. Zoals Figuur 2 laat zien passen er precies vier bollen in een kubus met ribbe $2r\sqrt{2}$, dus het deel van de kubus dat opgevuld wordt door de bollen is precies $\pi/\sqrt{18}$.

Een stapeling met deze dichtheid is overigens niet uniek. Ook de *hexagonale dichtste stapeling* (hexagonal close-packed of HCP packing) heeft deze dichtheid. Deze wordt bereikt door in het platte vlak een bol te omsluiten door zes andere bollen en zo verder het platte vlak vol te leggen. Zo ontstaat een hexagonaal rooster, als de cellen van een honingraat. Vervolgens wordt hier op dezelfde wijze een tweede laag bovenop gelegd. De derde laag kan nu weer op dezelfde wijze gelegd worden als de eerste, en de vierde weer op dezelfde wijze als de tweede enzovoort. Dit levert HCP, de hexagonale dichtste stapeling. Daarbij blijft het grondvlak altijd zichtbaar: we kunnen door de stapeling heen kijken. De derde laag kan ook gelegd worden over de openingen die er na de eerste twee lagen nog zijn. Als we vervolgens de bollen weer leggen volgens de posities in laag 1, en vervolgens weer in de posities van laag 2, laag 3 enzovoort, dan is er sprake van FCC, de kubische vlakgecentreerde stapeling. Als we in de FCC-stapeling een bol op het derde vlak beschouwen met de zes bollen daar direct onder in het tweede vlak, zien we één uitstekende punt van de kubus van het eerste figuur van de FCC-stapeling. Zie Figuur 3.

Er bestaan zelfs overaftelbaar veel stapelingen met de optimale dichtheid. Ze worden allemaal gevormd door oneindig veel vlakke lagen van bollen op elkaar te plaatsen. Binnen zo'n laag liggen de bollen op een hexagonaal rooster en voor iedere volgende laag zijn er dan twee manieren om die op de laag eronder te plaatsen. Als je hierbij alterneert tussen twee posities (loodrecht gezien) dan krijg je de hexagonale dichtste stapeling. Als je telkens dezelfde afstand opschuift krijg je de kubische vlakgecentreerde stapeling.

Een dichtheid van $\pi/\sqrt{18}$ kan dus op veel manieren bereikt worden. Maar kan het ook beter?

In 1611 schreef Johannes Kepler (1571–1630) in zijn boek *Strena Seu De Nive Sexangula* ('Over de zeshoekige sneeuwvlok') [9] dat dit niet het geval was:

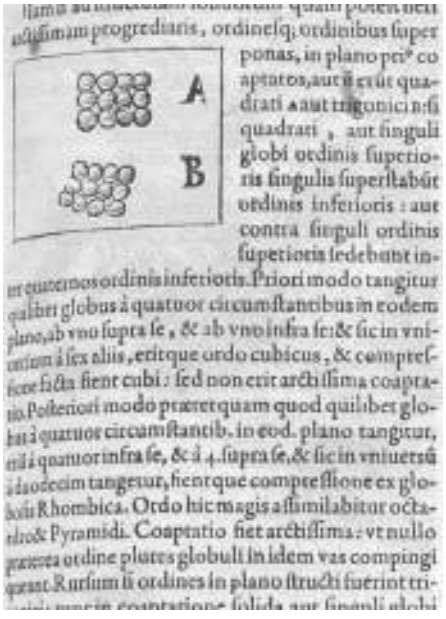


Figuur 3 Laagjes bollen in de HCP en FCC stapelingen.

> "Posteriori modo praeterquam quod quilibet globus a quatuor circumstantibus in eodem plano tangitur, etiam a quatuor infra se et a quatuor supra se, et sic in universum a duodecim tangetur, fientque compressione ex globosis rhombica. Ordo hic magis assimilabitur octaedro et pyramidi. Coaptatio fiet arctissima: ut nullo praeterea ordine plures globuli in idem vas compingi queant."

Dit kan vertaald worden als:

> "In deze rangschikking raakt elke bol niet alleen zijn vier buren in hetzelfde vlak, maar ook vier onder hem, en vier erboven, zodat alle bollen door twaalf andere worden aangeraakt, en de bollen worden samengeperst tot rombische dodecaëders. Deze rangschikking heeft meer de structuur van een octaëder en een piramide. Deze stapeling is de dichtst mogelijke: in geen enkele rangschikking kunnen meer bollen worden geplaatst in dezelfde ruimte."



Figuur 4 Links Johannes Kepler, in het midden een detail van de titelpagina van *Strena Seu De Nive Sexangula*, en rechts de claim van Kepler.

Hoewel Kepler claimt dat deze stapeling het dichtst is, geeft hij geen bewijs. De uitspraak dat de kubische vlakgecentreerde stapeling de grootst mogelijke dichtheid heeft van alle bolstapelingen is daarom bekend geworden als *het Kepler-vermoeden*.

Optimaliteit bewijzen in het geval dat de bollen gerangschikt zijn volgens een rooster was al gedaan door Carl Friedrich Gauss (1777–1855). Evenwel was het lange tijd een open probleem of de kubische vlakgecentreerde stapeling ook optimaal was zonder deze restrictie.

Een bewijs te ingewikkeld om te beoordelen

In 1953 liet László Fejes Tóth (1915–2005) zien dat het bewijzen van het Kepler-vermoeden kon worden gereduceerd tot een groot maar eindig aantal berekeningen [2].

Gebaseerd op deze ideeën gaf Tom (1958–) samen met zijn student Samuel Ferguson een bewijs voor het Kepler-vermoeden. Hij kondigde de voltooiing van dit bewijs in augustus 1998 aan.

Een bijzonderheid van dit bewijs is dat het is gebaseerd op de resultaten van een divers en groot aantal computerberekeningen. Om het bewijs te vertrouwen moet je er dus op kunnen vertrouwen dat de gebruikte software geen bugs heeft en dat de computer tijdens het rekenen nergens een storing heeft gehad. We zullen verderop de structuur van het bewijs schetsen.

Hales probeerde zijn bewijs gepubliceerd te krijgen in de *Annals of Mathematics*. Wat hij ter beoordeling instuurde was een document van 250 pagina's, vergezeld van 3 gigabytes aan computerbestanden. De *Annals* stelde vervolgens een leesgroep van twaalf referenten in met als opdracht de correctheid van het bewijs te controleren. In 2003, na vier jaar werk, gaf deze groep de opdracht terug, zonder de correctheid te hebben vastgesteld. De groep had geen fouten in het bewijs kunnen vinden. Maar aan de andere kant was het bewijs zo complex dat de groep niet in staat was geweest het te beoordelen. De groep meldde dat ze '99 procent zeker' waren van de correctheid van het bewijs, maar dat ze er niet volledig waren uitgekomen.

Er werd vervolgens besloten om het bewijs in tweeën te hakken. Het tekstuele deel werd geaccepteerd in de *Annals* in 2005 [4], terwijl het computergedeelte werd gepubliceerd in twee artikelen in *Discrete and Computational Geometry* in 2006 [7]. Het *Annals*-artikel bevatte een verwijzing naar de computerhelft van het bewijs, maar gaf daarbij de disclaimer van de editors van het tijdschrift dat het hun niet gelukt was de correctheid hiervan te beoordelen.

Tom Hales

Het Flyspeck-project

Computers zijn bezig het karakter van de wiskunde te veranderen. Als je een hulpmiddel hebt gebruik je het ook, en computers zijn voor wiskunde een buitengewoon krachtig hulpmiddel.

Er zijn allerlei manieren waarop je computers kunt gebruiken om wiskunde te doen:

- Ten eerste kun je computers gebruiken om te rekenen. Hierbij kan het gaan om getalsmatige berekeningen, maar ook om grote hoeveelheden gevallen langs te gaan. Dit is het soort computergebruik dat de basis is voor het bewijs van Hales en Ferguson.

  Computers rekenen gewoonlijk met eindige precisie (met *floating point numbers* of drijvendekommagetallen), en maken dus afrondfouten, maar dit hoeft geen reden te zijn dat berekeningen niet wiskundig hard bruikbaar zijn. Zo kun je kiezen naar welke kant je getallen laat afronden, en in plaats van alleen met benaderingen kun je met intervallen rekenen. Bij *interval-aritmetiek* zorg je ervoor dat het reële getal waarover je een uitspraak doet gegarandeerd binnen de intervalbenadering ligt. Interval-aritmetiek is een essentieel onderdeel van het Hales–Ferguson-bewijs.
- Wiskundige berekeningen hoeven niet numeriek te zijn, maar kunnen ook *symbolisch* zijn. Als ik het getal $\sqrt{2}$ beschouw, dan kan ik maar eindig veel cijfers achter de komma opschrijven. Evenwel, door dit getal symbolisch te beschouwen heb ik toch maar eindig veel inkt nodig om het in volledige precisie te kunnen manipuleren.

  Een dergelijk onderscheid bestaat ook in de computer. De *computer algebra* systemen als Mathematica en Maple werken niet met getallen maar met symbolische expressies. Ook deze systemen worden vaak gebruikt als handig hulpmiddel in wiskundig onderzoek.
- Bij de vorige twee manieren om computers te gebruiken gaat het over rekenen, en het *redeneren* over wat de berekeningen voorstellen wordt aan de mens overgelaten. Er zijn ook computersystemen die helpen bij het redeneren [3]. Hierbij kun je onderscheid maken tussen het vinden van redeneringen versus het verifiëren van redeneringen. In het eerste geval gaat het om ATP-systemen (automated theorem provers), en in het tweede geval om ITP-systemen (interactive theorem provers). Een andere term voor een ITP-systeem is een *bewijsassistent*. Hoewel bij ITP het bewijs door de mens wordt aangedragen bieden deze programma's toch een heleboel hulp bij het redeneren.

De eerste twee vormen van computergebruik hebben binnen de wiskunde een hoge vlucht genomen. De laatste is momenteel voornamelijk nog een onderzoeksonderwerp voor informatici, maar was precies wat Hales nodig had.

In januari 2003 besloot Hales zijn bewijs met een ITP-systeem te gaan controleren. Omdat deze systemen een buitengewoon hoge betrouwbaarheid hebben, zouden er daarna geen referenten meer nodig zijn. Hij had met ITP-specialisten gecorrespondeerd (waaronder met ons), en de schatting was dat het ongeveer twintig manjaar zou kosten om het bewijs van het Kepler-vermoeden in een ITP-systeem in te voeren en te verifiëren. Hales besloot dat dit een doenlijk project was. Als naam voor het project [6] koos hij *Flyspeck*, door met het patroon `f*p*k` (voor 'formal proof of Kepler') in een lijst woorden te zoeken. In het Engels

John Harrison, ontwikkelaar HOL Light-bewijsassistent

betekent "to flyspeck" ook zoiets als "in extreem detail op gebreken controleren", dus deze naam was zeer toepasselijk. Als systeem voor het project koos hij het HOL Light-systeem [8] van John Harrison. Harrison is op *interactive theorem proving* (ITP) gepromoveerd in Cambridge, en is als ITP-specialist werkzaam bij Intel in Portland in de Verenigde Staten. Hoewel hij in zijn baan correctheidsbewijzen voor Intel verifieert, die dus over informatica gaan, heeft hij in zijn vrije tijd een indrukwekkende hoeveelheid wiskundige bewijzen in zijn systeem geformaliseerd.

Het HOL Light-systeem is vooral voor wiskunde een van de sterkere systemen, maar wordt in de informatica minder gebruikt. Andere, in de informatica bekendere ITP-systemen zijn het Coq-systeem [1] uit Frankrijk en het Isabelle-systeem [11] uit Engeland/Duitsland. De gebruikers van deze systemen bleken ook zeer in Flyspeck geïnteresseerd ("eindelijk een echte wiskundige die onze systemen nodig heeft!") en hebben ook aan Flyspeck gewerkt. Een deel van het werk met Isabelle is ook in het uiteindelijke Flyspeck-bewijs terechtgekomen (zie verderop), maar het overgrote deel van het Flyspeck-bewijs is geformaliseerd in het HOL Light-systeem.

Hales besloot Flyspeck een extra impuls te geven door het coderen van het bewijzen te 'outsourcen, namelijk in Vietnam. Hij organiseerde daar een groep Vietnamese wiskundigen, leerde ze HOL Light, en zette ze aan het coderen van het bewijs van het Kepler-vermoeden. Tegelijk werkte hij aan een boek dat het bewijs minutieus gedetailleerd weergeeft, het zogeheten 'blueprint'-boek, want de subtitel is 'A blueprint for formal proofs'. Dit boek [5] telde uiteindelijk 334 bladzijden.

Op 10 augustus 2014 werd bekend gemaakt dat het Flyspeck-project was voltooid. Het bewijs was in de tussentijd aangepast en gestroomlijnd (ook om verificatie met de computer te vergemakkelijken), en dit aangepaste bewijs is dus 100 procent zeker correct. Het Kepler-vermoeden kan sinds deze datum als bewezen worden beschouwd.

Zekerheid dat een bewijs correct is

Een wiskundige gaat achter een computer met HOL Light zitten, en claimt: dankzij dit systeem weet ik 100 procent zeker dat deze stelling bewijsbaar is. Hoe kan hij zo zeker van zijn zaak zijn? Het oorspronkelijke Kepler-bewijs van Hales en Ferguson bevatte een groot aantal computerberekeningen, en het was daardoor moeilijk vast te stellen of het allemaal correct was. Het Flyspeck-bewijs heeft net zo goed een groot aantal computerberekeningen (door HOL Light) nodig om te verwerken. Waarom zou dit anders zijn?

Een ITP-systeem als HOL Light implementeert een redeneersysteem uit de mathematische logica dat extreem eenvoudig is. Alle complexiteit van de redenering in het zeer ingewikkelde Kepler-bewijs wordt teruggebracht tot een gigantisch aantal (miljarden) zeer elementaire redeneerstapjes. Stuk voor stuk zijn deze stapjes uiterst simpel. Het gebruikte redeneersysteem heet HOL (voor 'higher order logic', hogere-orde logica), en bestaat uit maar tien zeer eenvoudige redeneerregels. Ieder stapje past één van die regels toe.

Nu komt de controle van de correctheid van het bewijs uitsluitend neer op het controleren van deze redeneerstapjes. Maar omdat het redeneersysteem zo eenvoudig is, is het maar een heel klein deel van de gebruikte software dat dit doet. Deze *logische kern* is ongeveer vierhonderd programmaregels lang (geschreven in de functionele programmeertaal OCaml). Door het gebruik van de techniek van *abstracte datatypen* kan er geen wiskundige incorrectheid optreden als deze kern van het programma foutloos is. En bij een programma van vierhonderd regels is het goed te doen om alle fouten eruit te halen.

In andere woorden: in plaats van het belachelijk ingewikkelde Kepler-bewijs op correctheid te moeten beoordelen, hoeven de referenten nu alleen vierhonderd regels OCaml-code op correctheid te controleren, en de computer doet dan de rest.

Nu zijn er ook wel programmeerfouten in kernen van ITP-systemen gevonden, dus we zouden nog een stap verder willen gaan. In de informatica bestaat er technologie — ook gebaseerd op ITP — om van programma's (in het bijzonder als het geschreven is in een functionele programmeertaal) vast te stellen dat er *nul* fouten in zitten. Deze technologie hebben eerst John Harrison en vervolgens Ramana Kumar (een andere promovendus uit Cambridge) gebruikt om formeel te bewijzen dat de kern van HOL Light foutloos is [10]. Zelfs hier hoeven de referenten zich dus geen zorgen meer over te maken.

Er is hier natuurlijk wel sprake van een kip-en-ei-probleem, en tevens volgt uit de Gödelstelling (die zegt dat een consistent logisch systeem niet van zichzelf kan bewijzen dat het consistent is) dat een systeem niet de correctheid van zijn eigen broncode kan vaststellen. Maar al met al betekent al deze technologie dat de betrouwbaarheid van ITP-systemen, en daardoor van de gecontroleerde bewijzen, extreem hoog is. Onvoorstelbaar veel hoger dan die van bewijzen die alleen door mensen geverifieerd zijn.

Als de implementatie van het ITP-systeem volledig vertrouwd wordt kan er uiteraard in een ander onderdeel van de computer een softwarefout of hardwarefout zitten, waardoor een incorrect bewijs geaccepteerd wordt door de computer. Het is uiterst onwaarschijnlijk dat, bijvoorbeeld, een fout in het besturingssysteem ervoor zou zorgen dat een incorrect bewijs juist wel door de ITP-kern geaccepteerd zou worden. Dit punt wordt ondervangen door de bewijsverificatie op verschillende computers met verschillende software opnieuw te draaien.

Het Hales–Ferguson-bewijs

Het bewijs van het Kepler-vermoeden bestaat uit ruwweg acht stappen, waarbij we hier het bewijs vertellen van het resultaat terugwerkend naar eerdere lemma's. De spil van het bewijs is stap (d), de "lokale annulus-ongelijkheid".

(a) De precieze formulering van het Kepler-vermoeden. Het Kepler-vermoeden gaat over oneindigheid: wat er mogelijk is als

we een oneindige ruimte vullen met oneindig veel bollen met vaste straal r_0. Dit betekent dat we het over een limietproces hebben, want de dichtheid van de bollen is de limiet van de dichtheid van deze bollen binnen een groeiende bol met straal r, waarbij r naar oneindig gaat. Nu hoeft deze limiet niet te bestaan, maar de $\limsup$ bestaat natuurlijk altijd wel. Wat we dus moeten bewijzen is dat dit voor iedere mogelijke configuratie van de bollen nooit groter zal zijn dan $\pi/\sqrt{18}$.

Het is duidelijk dat de $\limsup$ niet afhangt van de keuze van r_0, dus vanaf nu zullen we $r_0 = 1$ nemen. We kijken dus naar een stapeling van eenheidsbollen.

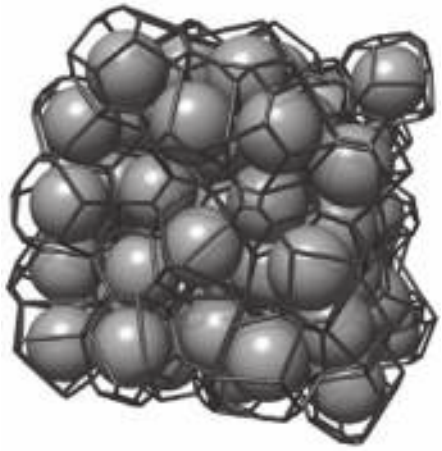
(b) De uitspraak die bewezen wordt in het Flyspeck-project. Flyspeck kijkt niet naar het volume van de doorsnede van de bollen met de grote bol met straal r, maar telt de middelpunten van de kleine bollen die zich binnen de grote bol bevinden. Omdat het volume van de grote bol evenredig is met r^3 en wat we zo verwaarlozen correspondeert met een volume ten hoogste evenredig met r^2, maakt dit voor de $\limsup$ van de dichtheid niet uit.

Voorts heeft Flyspeck het niet over de $\limsup$, maar bewijst het de volgende uitspraak:

Voor iedere stapeling S bestaat een c zodat voor iedere $r \geq 1$ geldt dat:

$$\operatorname{card}(V_S \cap B(0,r)) \leq \frac{\pi}{\sqrt{18}} r^3 + cr^2.$$

Hierin is V_S de verzameling met de middelpunten van de kleine bollen behorend bij stapeling S, en $B(0,r)$ is een groeiende (open) bol met straal r. De eis $r \geq 1$ is duidelijk nodig voor het geval dat de oorsprong in V_S zit. De constante c hangt in deze formulering af van de stapeling, maar Hales claimt dat deze ongelijkheid ook voor een vaste c kan worden bewezen. Deze verfijning zit evenwel niet in Flyspeck.



Afbeelding: G.E. Schröder-Turk e.a., Europhysics Letters 90(3) (2010)

Figuur 5 Voronoi-cellen van een bolstapeling.

In de computercode die de input voor het HOL Light-systeem is, is deze uitspraak gecodeerd als:

```
!V. packing V ==>
  (?c. !r. &1 <= r ==>
    &(CARD(V INTER ball(vec 0,r))) <=
    pi * r pow 3 / sqrt(&18) + c * r pow 2))
```

Hierin is '**!**' de universele kwantor $\forall$, '**?**' de existentiële kwantor $\exists$, en '**&**' de functie die een natuurlijk getal afbeeldt op het corresponderende reële getal.

(c) Verwaarloosbare FCC-compatibele functies. We hebben het ook in deze herformulering nog steeds over een oneindig probleem, en we willen dit graag reduceren naar een eindig probleem, en zo een aanpak mogelijk maken die analyseert wat er rond een enkele bol kan gebeuren.

Hiertoe nemen we een stapeling S en we verdelen de ruimte in de *Voronoi-cellen* van de verzameling van bolmiddelpunten van S, V_S. Voor ieder punt $v \in V_S$ is de Voronoi-cel $\Omega(V_S, v)$ de verzameling punten die dichter bij v ligt dan bij ieder ander punt in V_S.

Een reëelwaardige functie $G(v)$ op de punten van V_S heet *verwaarloosbaar* als er een c_1 bestaat zodat voor alle $r \geq 1$:

$$\sum_{v \in V_S(0,r)} G(v) \leq c_1 r^2.$$

De functie $G(v)$ heet *FCC-compatibel* als voor iedere $v \in V_S$,

$$4\sqrt{2} - \operatorname{vol}(\Omega(V_S, v)) \leq G(v).$$

Het getal

$$4\sqrt{2} = 5{,}55685\ldots$$

is het volume van de Voronoi-cellen van FCC, de kubische vlakgecentreerde stapeling.

Als de stapeling S echt beter is dan FCC, dan zijn de Voronoi-cellen van S kleiner dan die van FCC. De definitie van een verwaarloosbare FCC-compatibele functie $G(v)$ betekent dat de Voronoi-cellen van S niet significant kleiner zijn dan die van FCC: het verschil is maximaal van orde r^2, terwijl het aantal punten in $V_S(0,r)$ van orde r^3 is. Dus in de limiet is de afwijking verwaarloosbaar. Hieruit volgt duidelijk het Kepler-vermoeden.

Het bestaan van zo'n verwaarloosbare FCC-compatibele functie $G(v)$ is stap (f) van deze bewijsschets.

(d) De lokale annulus-ongelijkheid. Dit is het centrale lemma van het bewijs, dat ook voor andere stellingen toepasbaar is gebleken.

Definieer het magische getal h_0 exact als

$$h_0 = 1{,}26$$

Bekijk nu de 'schil' van punten v om een eenheidsbol met middelpunt in de oorsprong met de eigenschap

$$1 \leq h(v) \leq h_0$$

waarbij $h(v)$ de helft van de afstand van punt v tot de oorsprong is. Deze 'schil' heet de *annulus* van de centrale bol. We gaan kijken wat er gebeurt als we bollen om deze centrale bol plaatsen, zodanig dat de middelpunten zich allemaal in deze annulus bevinden.

Voorts tellen we deze bollen gewogen: als ze verder naar buiten zijn tellen ze minder mee. (In dat geval is er meer ruimte tussen de bollen, en we willen een ongelijkheid die het aantal begrenst.) Precies gezegd definiëren we een functie $L(h)$ door

$$L(h) := \begin{cases} \frac{h_0 - h}{h_0 - 1} & \text{if } 1 \leq h \leq h_0, \\ 0 & \text{if } h_0 \leq h. \end{cases}$$

Deze functie vervalt lineair van 1 op het binnenoppervlak van de annulus naar 0 op het buitenoppervlak.

De lokale annulus-ongelijkheid is nu:

$$\sum_{v \in V_S} L(h(v)) \leq 12.$$

Het centrale lemma zegt dat deze ongelijkheid geldt voor iedere stapeling S.

De geldigheid van deze ongelijkheid zal volgen uit het resultaat in stap (g), en is een van de ingrediënten voor het bestaan van een verwaarloosbare FCC-compatibele functie in stap (f).

Dat er maximaal twaalf bollen aan een centrale bol kunnen raken was Kepler al duidelijk. Door de bollen verder naar buiten te bewegen passen er mogelijk meer, maar als je ze minder meetelt, blijft het 'aantal' dus toch hoogstens twaalf.

(e) Marchal-cellen en de cel-cluster ongelijkheid. Nu wordt het bewijs nog technischer, en we worden hier nog schetsmati-

ger in onze beschrijving van het bewijs dan we toch al waren.

De Voronoi-cellen worden verdeeld in *Rogers-simplices* en deze worden weer onderverdeeld in fragmentjes die *Marchal-cellen* worden genoemd. Dit alles wordt gedaan om te komen tot de zogenaamde *cel-cluster-ongelijkheid*. Die heeft te maken met configuraties waarin twee bollen een afstand in een specifiek interval hebben (zo'n configuratie wordt een 'cel-cluster' genoemd). De cel-cluster-ongelijkheid blijkt de cruciale stap om te laten zien dat de functie $G(v)$, die we in de volgende sectie definiëren, verwaarloosbaar is.

(f) Met wat we nu hebben kunnen we, gegeven een stapeling S, een verwaarloosbare FCC-comptibele functie definiëren:

$$G(v) = -\operatorname{vol}(\Omega(V_S, v)) + 8m_1 - \sum 8m_2 L(h(v)).$$

Hierin:

$$\begin{aligned}
&\mathrm{sol}_0 = 3\arccos(1/3) - \pi = 0{,}551285\ldots,\\
&\tau_0 = 4\pi - 20\mathrm{sol}_0 = 1{,}54065\ldots,\\
&m_1 = \mathrm{sol}_0 2\sqrt{2}\,/\tau_0 = 1{,}01208\ldots,\\
&m_2 = (6\mathrm{sol}_0 - \pi)\sqrt{2}\,/\,(6\tau_0) = 0{,}0254145\ldots.
\end{aligned}$$

De som in de definitie van $G(v)$ loopt over alle bolmiddelpunten in V_S uitgezonderd de oorsprong. Als v te ver van de oorsprong ligt geldt $L(h(v)) = 0$, dus dit is een eindige som.

Uit de lokale annulus-ongelijkheid volgt nu dat deze functie FCC-compatibel is:

$$\begin{aligned}
&4\sqrt{2} - \operatorname{vol}(\Omega(V_S, v))\\
&\quad \le -\operatorname{vol}(\Omega(V_S, v)) + 8m_1 - 12 \cdot 8m_2\\
&\quad \le -\operatorname{vol}(\Omega(V_S, v)) + 8m_1 - \sum 8m_2 L(h(v))\\
&\quad = G(v).
\end{aligned}$$

Dat deze functie ook verwaarloosbaar is volgt dus uit de cel-cluster-ongelijkheid.

(g) Het niet-bestaan van een tegenvoorbeeldstapeling. We hebben nog niet beschreven hoe de lokale annulus-ongelijkheid en de cel-cluster-ongelijkheid bewezen worden. We gaan nu eerst kijken hoe het bewijs van de eerste gaat. Over het bewijs van de tweede hebben we het in de volgende paragraaf.

Een verzameling bollen S rond een centrale bol waarvoor de lokale annulus-ongelijkheid niet geldt noemen we een *tegenvoorbeeldstapeling*. In feite eisen we voor een tegenvoorbeeldstapeling nog wat meer, zoals dat het aantal bollen ten hoogste veertien is, en dat de waarde van $\sum_{v \in V_S} L(h(v))$ maximaal is.

We willen laten zien dat er geen tegenvoorbeeldstapeling bestaat, want dan geldt de lokale annulus-ongelijkheid voor alle stapelingen.

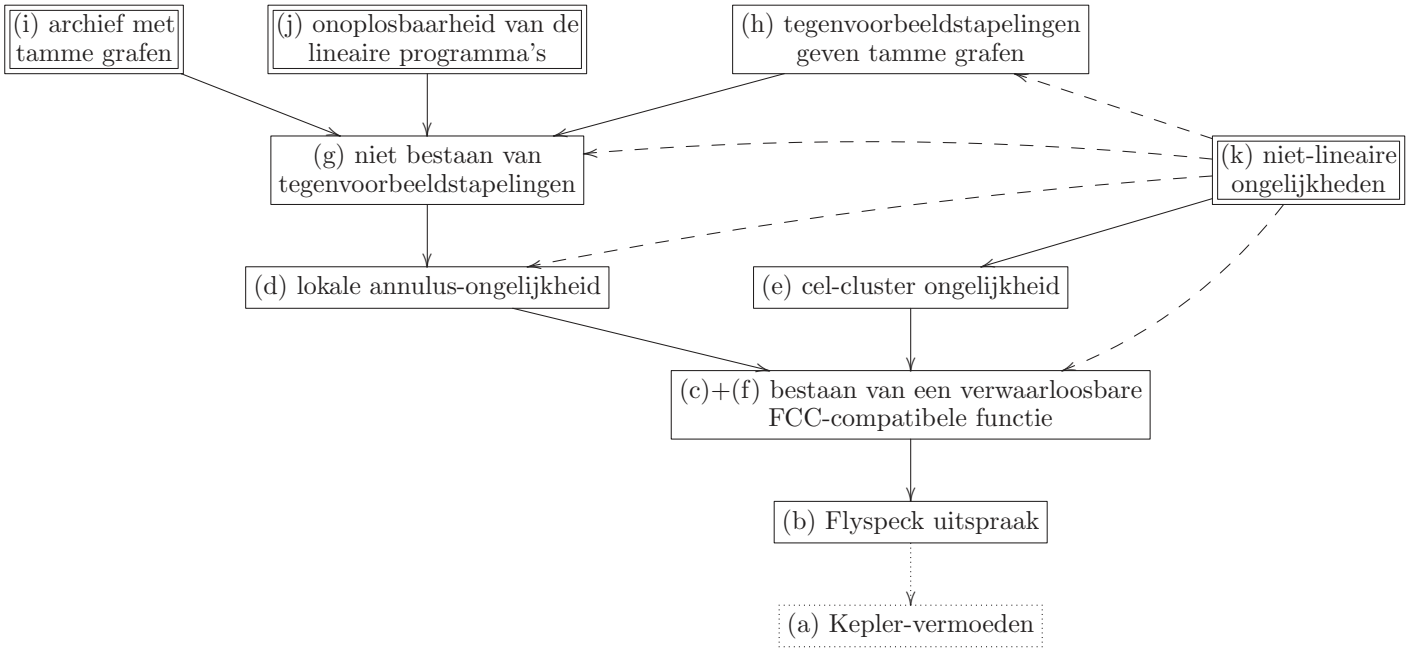
(h) Tamme grafen zijn de grafen van tegenvoorbeeldstapelingen. We definiëren nu bij een tegenvoorbeeldstapeling een graaf door twee middelpunten van bollen te verbinden als de afstand kleiner is dan $2h_0$. Deze grafen voldoen aan een reeks eigenschappen, en alle grafen die aan die eigenschappen voldoen noemen we *tamme* grafen. Deze worden gedefinieerd met een lijst van elf eigenschappen, waarvan de ingewikkeldste is dat er een functie van de vlakken in de graaf naar de reële getallen moet bestaan die aan drie verdere eigenschappen moet voldoen.

Het bewijs wordt nu afgemaakt door eerst te bewijzen dat de graaf van een tegenvoorbeeldstapeling altijd tam is. Dan wordt bewezen dat iedere tamme graaf in een lange lijst van 19.715 grafen voorkomt. En ten slotte wordt voor iedere graaf uit die lijst gekeken of het mogelijk is om de bollen zo te plaatsen dat de lokale annulus-ongelijkheid niet geldt. Dit correspondeert met het zoeken van een oplossing in een hele reeks lineaire programma's. Doordat dat voor geen enkele tamme graaf lukt, volgt hieruit de lokale annulus-ongelijkheid.

De computer helpt

Zoals gezegd rust het Hales–Ferguson-bewijs op een aantal computerberekeningen:

(i) Een programma dat de lijst van 19.715 grafen genereert waar alle tamme grafen



Figuur 6 Het bewijs van het Kepler-vermoeden.

bij zitten. Dit programma was oorspronkelijk in Java geschreven. De versie van dit programma in Flyspeck is geschreven in de taal van de bewijsassistent Isabelle en vervolgens geëxporteerd naar een programma in de programmeertaal ML. De verificatie van dit programma is het enige onderdeel van Flyspeck dat niet in HOL Light is gedaan maar in Isabelle. In de HOL Light-verificatie is het resultaat hiervan, inclusief de lijst grafen, als axioma aangenomen.

(j) Een computerberekening die verifieert dat 43.078 lineaire programma's onoplosbaar zijn. Ieder lineair programma in deze lijst heeft rond de duizend variabelen en een vergelijkbaar aantal vergelijkingen. Deze lineaire programma's waren de hoofdmoot van de 3 gigabyte data uit het oorspronkelijke bewijs. In het HOL Light-bewijs worden ze *on the fly* gegenereerd en de onoplosbaarheid ervan wordt vervolgens bewezen.

(k) Een computerverificatie dat 23.242 niet-lineaire ongelijkheden met hoogstens zes variabelen gelden. Dit is de verificatie waarbij in het oorspronkelijke bewijs interval-aritmetiek wordt gebruikt. Deze ongelijkheden worden op allerlei plaatsen in het bewijs gebruikt. Om een idee te geven: het gaat hierbij om te verifiëren dat ongelijkheden van de volgende vorm gelden binnen een bepaald interval:

$$\frac{\begin{array}{c}-x_1x_3-x_2x_4+x_1x_5+x_3x_6-x_5x_6\\+x_2(-x_2+x_1+x_3-x_4+x_5+x_6)\end{array}}{\sqrt{4x_2\begin{pmatrix}x_2x_4(-x_2+x_1+x_3-x_4+x_5+x_6)+\\x_1x_5(x_2-x_1+x_3+x_4-x_5+x_6)+\\x_3x_6(x_2+x_1-x_3+x_4+x_5-x_6)\\-x_1x_3x_4-x_2x_3x_5-x_2x_1x_6-x_4x_5x_6\end{pmatrix}}}$$
$$< \tan\left(\frac{\pi}{2}-0.74\right).$$

Hoe dit allemaal aan elkaar hangt is in het diagram in Figuur 6 weergegeven. De rechthoeken met dubbele randen zijn de rekenintensieve taken waarvoor speciale programma's zijn geschreven. De pijlen geven de afhankelijkheden tussen de verschillende onderdelen.

Het moeilijkste in de HOL Light-verificatie was het doenlijk krijgen van de verificatie van de computerberekeningen, omdat deze ook allemaal in hele kleine redeneerstapjes moesten worden afgebroken en door de HOL Light-kern geverifieerd. Om dit efficiënt genoeg te doen was een heleboel slimheid nodig. Dit was het onderwerp van het proefschrift van Alexey Solovyev [12], een promovendus van Hales.

Alexey Solovyev

En zelfs hierna kon het project pas worden afgerond toen Microsoft genereus rekentijd in de cloud beschikbaar stelde. Evenwel werd vrij kort daarna de hele verificatie nog eens overgedaan op een van de clusters van de Radboud Universiteit in Nijmegen.

Referenties

1 Coq development Team, The Coq proof assistant, https://coq.inria.fr.

2 L. Fejes Tóth, *Lagerungen in der Ebene auf der Kugel und im Raum*, first edition, Springer, 1953.

3 Georges Gonthier, Thomas C. Hales, John Harrison en Freek Wiedijk, Formal proof, *Notices of the American Mathematical Society* 55(11) (2008), 1370–1414.

4 Thomas C. Hales, A proof of the Kepler conjecture, *Annals of Mathematics* 162 (2005), 1063–1183.

5 Thomas C. Hales, *Dense Sphere Packings: A blueprint for formal proofs*, London Math. Soc. Lecture Note Series, Vol. 400, Cambridge University Press, 2012.

6 Thomas C. Hales, Mark Adams, Gertrud Bauer, Dat Tat Dang, John Harrison, Truong Le Hoang, Cezary Kaliszyk, Victor Magron, Sean McLaughlin, Thang Tat Nguyen, Truong Quang Nguyen, Tobias Nipkow, Steven Obua, Joseph Pleso, Jason Rute, Alexey Solovyev, An Hoai Thi Ta, Trung Nam Tran, Diep Thi Trieu, Josef Urban, Ky Khac Vu en Roland Zumkeller, A formal proof of the Kepler conjecture, *CoRR*, abs/1501.02155, 2015.

7 Thomas C. Hales en Samuel P. Ferguson, Historical overview of the Kepler conjecture; A formulation of the Kepler conjecture; Sphere packing III, Extremal cases; Sphere packing IV, Setailed bounds; Sphere packings VI, Tame graphs and linear programs, *Discrete & Computational Geometry* 36(1) (2006), 5–20; 21–69; 71–110; 111–166; 205–265.

8 J. Harrison, The HOL Light theorem prover, http://www.cl.cam.ac.uk/~jrh13/hol-light.

9 J. Kepler, *The Six-Cornered Snowflake*, Clarendon Press, 1966.

10 Ramana Kumar, Rob Arthan, Magnus O. Myreen en Scott Owens, Self-formalisation of higher-order logic, *Journal of Automated Reasoning* 56(3) (2016), 221–259.

11 L. Paulson, T. Nipkow en M. Wenzel, The Isabelle proof assistant, https://isabelle.in.tum.de.

12 A. Solovyev, *Formal Computations and Methods*, PhD thesis, University of Pittsburgh, 2013.

David de Laat
Centrum Wiskunde & Informatica
Amsterdam
mail@daviddelaat.nl

Frank Vallentin
Mathematisches Institut
Universität zu Köln, Germany
frank.vallentin@uni-koeln.de

A breakthrough in sphere packing: the search for magic functions

This paper by David de Laat and Frank Vallentin is an exposition about the two recent breakthrough results in the theory of sphere packings. It includes an interview with Henry Cohn, Abhinav Kumar, Stephen D. Miller and Maryna Viazovska.

The sphere packing problem asks for a densest packing of congruent solid spheres in n-dimensional space $\mathbb{R}^n$. In a packing the (solid) spheres are allowed to touch on their boundaries, but their interiors should not intersect.

While the case of the real line, $n = 1$, is trivial, the case $n = 2$ of packing circles in the plane was first solved in 1892 by the Norwegian mathematician Thue (1863–1922). He showed that the honeycomb hexagonal lattice gives an optimal packing; see Figure 1.

The first rigorous proof is due to the Hungarian mathematician Fejes Tóth (1915–2005) in 1940. He also proved that this packing is unique (up to rotations, translations, and uniform scaling) among periodic packings. For $n = 3$, the sphere packing problem is known as the Kepler conjecture. It was solved by the American mathematician Hales in 1998 following an approach by Fejes Tóth. Hales' proof is extremely complex, takes more than 300 pages, and makes heavy use of computers. One of the difficulties of the sphere packing problem in three dimensions is the fact that there are uncountably many inequivalent optimal packings. In 2014 a fully computer verified version of Hales' proof was completed; it was a result of the collaborative Flyspeck project, also directed by Hales [7].

Recently, Maryna Viazovska, a postdoctoral researcher from Ukraine working at the Humboldt University of Berlin, solved the eight-dimensional case. On 14 March 2016 she announced her spectacular result in the paper titled 'The sphere packing problem in dimension 8' [10] on the arXiv-preprint server. Only one week later, on 21 March 2016, Henry Cohn, Abhinav Kumar, Stephen D. Miller, Danylo Radchenko and Maryna Viazovska announced a proof for the $n = 24$ case [3], building on Viazovska's work.

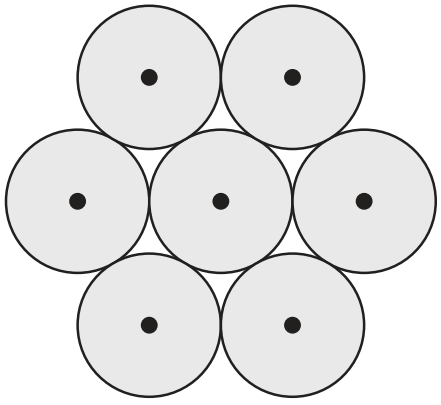


Figure 1 The hexagonal lattice and the corresponding circle packing.

Here we want to illustrate that the optimal sphere packings in dimensions 8 and 24 are very special (in the next section we give constructions of the E_8 lattice and of the Leech lattice Λ_{24}, which provide the optimal sphere packings in their dimensions), and we aim to explain the main ideas of the recent breakthrough results in sphere packing:

Theorem 1. *The lattice* E_8 *is the densest packing in* $\mathbb{R}^8$. *The Leech lattice* Λ_{24} *is the densest packing in* $\mathbb{R}^{24}$. *Moreover, no other periodic packing achieves the same density in the corresponding dimension.*

In the last section we will see that the beautiful proofs of these theorems use ideas from analytic number theory. Viazovska found a 'magic' function for dimension 8, which together with the linear programming bound of Cohn and Elkies,

as explained later on, gives a proof for the optimality of the E_8 lattice. Her method gave a hint how to find a magic function for dimension 24. Although the proof is relatively easy to understand, and basically no computer assistance is needed for its verification, computer assistance was crucial to conjecture the existence of, and to find, these magic functions.

Optimal lattices

In this section we introduce the two exceptional sphere packings in dimension 8 and 24. The book *Sphere Packings, Lattices, and Groups* of Conway and Sloane [5] is the definitive reference on this topic; the Italian-American combinatorialist Rota (1932–1999) reviewed the book saying:

> "This is the best survey of the best work in one of the best fields of combinatorics, written by the best people. It will make the best reading by the best students interested in the best mathematics that is now going on."

Lattice packings

How does one define a packing of unit spheres in n-dimensional space? In general, such a packing is defined by the set of centers L of the spheres in the packing.

We talk about *lattice packings* when L forms a *lattice*. Then there are n linearly independent vectors $b_1,\dots,b_n \in L$, called a *lattice basis* of L, so that L is the set of integral linear combinations of $b_1,\dots,b_n$. For instance, the three lattices

$$\mathbb{Z}, \quad \mathbb{Z}\begin{pmatrix}2\\0\end{pmatrix}+\mathbb{Z}\begin{pmatrix}-1\\\sqrt{3}\end{pmatrix},$$
$$\mathbb{Z}\begin{pmatrix}-\sqrt{2}\\-\sqrt{2}\\0\end{pmatrix}+\mathbb{Z}\begin{pmatrix}\sqrt{2}\\-\sqrt{2}\\0\end{pmatrix}+\mathbb{Z}\begin{pmatrix}0\\\sqrt{2}\\-\sqrt{2}\end{pmatrix}$$

define densest sphere packings in dimensions 1, 2, and 3.

We should also define what we mean when we talk about *density*. Intuitively, the density of a sphere packing is the fraction of space covered by the spheres of the packing. When the sphere packing is a lattice, this intuition is easy to make precise: The density of the sphere packing determined by L is the volume of one sphere divided by the *volume of* L, that is, the volume of a fundamental domain of L. One possible fundamental domain of L is given by the parallelepiped spanned by the lattice basis $b_1,\dots,b_n$, so that we have

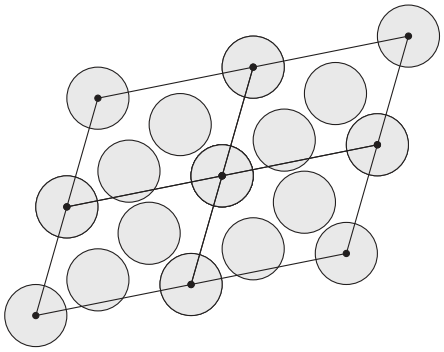


Figure 2 A periodic packing that is not a lattice packing.

$$\operatorname{vol}(\mathbb{R}^n/L) = |\det(b_1,\dots,b_n)|.$$

The density of L is then given by

$$\Delta(L) = \frac{\operatorname{vol}(B_n(r_1/2))}{\operatorname{vol}(\mathbb{R}^n/L)},$$

where r_1 is the shortest nonzero vector length in L. Here, $B_n(r_1/2)$ is the solid sphere of radius $r_1/2$, whose volume is

$$\operatorname{vol}(B_n(r_1/2)) = (r_1/2)^n \frac{\pi^{n/2}}{\Gamma(n/2+1)},$$

where Γ is the gamma function; it satisfies the equation $\Gamma(x+1) = x\Gamma(x)$, and two particularly useful values are $\Gamma(1)=1$ and $\Gamma(1/2)=\sqrt{\pi}$.

The optimal sphere packing density can be approached arbitrarily well by the density of a periodic packing. In a *periodic packing* the set of centers is the union of a finite number m of translates of a lattice L:

$$\{x_1+v : v\in L\} \cup \{x_2+v : v\in L\} \cup \cdots \cup \{x_m+v : v\in L\};$$

see Figure 2 for an example where $m=3$. The density of a periodic packing is $m\cdot\operatorname{vol}(B_n(r))/\operatorname{vol}(\mathbb{R}^n/L)$, where r is the radius of the spheres in the packing. It is possible that in some dimensions the optimal sphere packing is not a lattice packing; for example, the best known sphere packing in $\mathbb{R}^{10}$ is a periodic packing but it is not a lattice packing.

A few last definitions: From a lattice L we can construct its *dual lattice* by

$$L^* = \{y\in\mathbb{R}^n : x\cdot y \in \mathbb{Z} \text{ for all } x\in L\}.$$

It is not difficult to see that the volume of a lattice and its dual are reciprocal values, so that $\operatorname{vol}(\mathbb{R}^n/L)\cdot\operatorname{vol}(\mathbb{R}^n/L^*)=1$ holds. When a lattice equals its dual ($L=L^*$) and when the square of every occurring vector length is an even integer, then L is called an *even and unimodular* lattice.

The E_8 *lattice*

The nicest lattices are those which are even and unimodular. However, they only occur in higher dimensions: one can show that the first appearance of such an even and unimodular lattice is in dimension 8. It is the E_8 lattice, which was first explicitly constructed by the Russian mathematicians Korkine (1837–1908) and Zolotareff (1847–1878) in 1873.

Here we give a construction of the E_8 lattice which is based on lifting binary error correcting codes. For this we define the (extended) *Hamming code* $\mathcal{H}_8$ via a regular three-dimensional tetrahedron: consider the binary linear code $\mathcal{H}_8$, which is the vector space over the finite field $\mathbb{F}_2$ (consisting of the elements 0 and 1) spanned by the rows of the matrix

$$G = (I\,|\,A) = \left(\begin{array}{cccc|cccc}1&0&0&0&0&1&1&1\\0&1&0&0&1&0&1&1\\0&0&1&0&1&1&0&1\\0&0&0&1&1&1&1&0\end{array}\right) \in \mathbb{F}_2^{4\times 8},$$

where I is the identity matrix and where A is the adjacency matrix of the vertex-edge graph of a three-dimensional tetrahedron with vertices v_1, v_2, v_3, v_4. Hence, the Hamming code is a 4-dimensional subspace of the vector space $\mathbb{F}_2^8$. It consists of $2^4=16$ code words:

$$\begin{array}{cccccccc}0000|0000 & 1000|0111 & 1100|1100 & 0111|1000 & 1111|1111\\ & 0100|1011 & 1010|1010 & 1011|0100 & \\ & 0010|1101 & 1001|1001 & 1101|0010 & \\ & 0001|1110 & 0110|0110 & 1110|0001 & \\ & & 0101|0101 & & \\ & & 0011|0011 & & \end{array}$$

It is interesting to look at the occurring *Hamming weights* (the number of non-zero entries) of code words. In $\mathcal{H}_8$, one code word has Hamming weight 0, 14 code words have Hamming weight 4, and one code word has Hamming weight 8. Since all occurring Hamming weights are divisible by four, and four is two times two, the Hamming code $\mathcal{H}_8$ is called *doubly even*.

Let us compute the *dual code*

$$\mathcal{H}_8^{\perp} = \left\{ y\in\mathbb{F}_2^8 : \sum_{i=1}^8 x_i y_i = 0 \pmod 2 \text{ for all } x\in\mathcal{H}_8 \right\}.$$

Squaring the matrix A yields

$$A_{ij}^2 = \sum_{k=1}^4 A_{ik}A_{kj} = |\{k : v_i\sim v_k \text{ and } v_k\sim v_j\}| = \begin{cases}3 & \text{if } i=j,\\ 2 & \text{if } i\neq j.\end{cases}$$

Hence, $A^2 = I \bmod 2$. From this, $GG^T = I + A^2 = 0 \bmod 2$ follows. Hence, we have the inclusion $\mathcal{H}_8 \subseteq \mathcal{H}_8^\perp$ and by considering dimensions we see that $\mathcal{H}_8$ is a *self-dual code*; that is, $\mathcal{H}_8^\perp = \mathcal{H}_8$ holds.

We can define the lattice E_8 by the following lifting construction (which is usually called *Construction A*):

$$\mathsf{E}_8 = \left\{ \frac{1}{\sqrt{2}} x : x \in \mathbb{Z}^8,\ x \bmod 2 \in \mathcal{H}_8 \right\}.$$

Now it is immediate to see that E_8 has 240 shortest (nonzero) vectors:

$16 = 2^4$ vectors:

$$\pm\sqrt{2}\, e_i,\ i = 1, \ldots, 8$$

$224 = 2^4 \cdot 14$ vectors:

$$\frac{1}{\sqrt{2}} \sum_{i=1}^{8} (\pm x_i) e_i,\ x \in \mathcal{H}_8 \text{ and } \operatorname{wt}(x) = 4,$$

where $e_1, \ldots, e_8$ are the standard basis vectors of $\mathbb{R}^8$ and where $\operatorname{wt}(x) = |\{i : x_i \neq 0\}|$ denotes the Hamming weight of x. The shortest nonzero vectors of E_8 have length $\sqrt{2}$. The occurring vector lengths in E_8 are $0, \sqrt{2}, \sqrt{4}, \sqrt{6}, \ldots$, so that E_8 is an even lattice.

From the lifting construction it follows that the density of E_8 is $|\mathcal{H}_4| = 16$ times the density of the lattice $\sqrt{2}\,\mathbb{Z}^8$ which is spanned by $\sqrt{2}\, e_1, \ldots, \sqrt{2}\, e_8$. Thus,

$$\begin{aligned} \operatorname{vol}(\mathbb{R}^8/\mathsf{E}_8) &= \frac{1}{16} \cdot \operatorname{vol}(\mathbb{R}^8/\sqrt{2}\,\mathbb{Z}^8) \\ &= \frac{1}{16} \cdot (\sqrt{2})^8 = 1, \end{aligned}$$

so that E_8 is unimodular. One can show that E_8 is the *only* even unimodular lattice in dimension 8. In general, the lifting construction always yields an even unimodular lattice when we start with a binary code which is doubly even and self-dual.

Next to this exceptional number theoretical property, E_8 also has exceptional geometric properties: In 1979, Odlyzko and Sloane, and independently Levenshtein, proved that one cannot arrange more vectors on a sphere in dimension 8 of radius $\sqrt{2}$ so that the distance between any two distinct vectors is also at least $\sqrt{2}$; the 240 vectors give the unique solution of the kissing number problem in dimension 8 as was shown in by Bannai and Sloane in 1981. Blichfeldt (1873–1945) showed in 1935 that E_8 gives the densest sphere packing among lattice packings. For a long time it has been conjectured that E_8 also gives the unique densest sphere packing in dimension 8, without imposing the (severe) restriction to lattice packings. Now this conjecture has been proved in the breakthrough work of Maryna Viazovska.

The Leech lattice

We turn to 24 dimensions and to the Leech lattice. In 1965 Leech (1926–1992) realized that he constructed a surprisingly dense sphere packing in dimension 24. For his construction, he used the (extended binary) Golay code which is an exceptional error correcting code found by Golay (1902–1989) in 1949. To define the Leech lattice we modify the lifting construction of the E_8 lattice. We replace the Hamming code by the Golay code and apply two extra twists.

For defining the Golay code we replace the regular tetrahedron in the construction of the Hamming code by the regular icosahedron and we apply the first twist. Consider the binary code $\mathcal{G}_{24}$ spanned by the rows of the matrix

$$G = (I \,|\, J - A) = \left(\begin{array}{c|c} 100000000000 & 100000111111 \\ 010000000000 & 010110001111 \\ 001000000000 & 001011100111 \\ 000100000000 & 010101110011 \\ 000010000000 & 011010111001 \\ 000001000000 & 001101011101 \\ 000000100000 & 101110101100 \\ 000000010000 & 100111010110 \\ 000000001000 & 110011101010 \\ 000000000100 & 111001110100 \\ 000000000010 & 111100011010 \\ 000000000001 & 111111000001 \end{array}\right) \in \mathbb{F}_2^{12 \times 24},$$

where we use $J - A$ instead of A, with J the all-ones matrix and A the adjacency matrix of the vertex-edge graph of a three-dimensional icosahedron. This code, the *extended binary Golay code* $\mathcal{G}_{24}$, is a 12-dimensional subspace in $\mathbb{F}_2^{24}$. It contains one vector of Hamming weight 0, 759 vectors of Hamming weight 8, 2576 vectors of Hamming weight 16, 759 vectors of Hamming weight 20, and one vector of Hamming weight 24; $\mathcal{G}_{24}$ is a doubly even and self-dual code.

We define the even unimodular lattice

$$L_{24} = \left\{ \frac{1}{\sqrt{2}} x : x \in \mathbb{Z}^{24}, x \bmod 2 \in \mathcal{G}_{24} \right\}.$$

Since the minimal non-zero Hamming weight occurring in the Golay code is 8, this lattice has 48 shortest vectors $\pm\sqrt{2}\, e_i$, with $i = 1, \ldots, 24$, of length $\sqrt{2}$. To eliminate them we make the second twist, we define

$$\begin{aligned} \Lambda_{24} = &\left\{ x \in L_{24} : \sqrt{2} \sum_{i=1}^{24} x_i = 0 \bmod 4 \right\} \\ \cup &\left\{ (1, \ldots, 1) + x : x \in L_{24}, \sqrt{2} \sum_{i=1}^{24} x_i = 2 \bmod 4 \right\}. \end{aligned}$$

This is the Leech lattice. It is an even unimodular lattice. In Λ_{24} there are 196560 shortest vectors which have length $\sqrt{4}$. The occurring vector lengths in Λ_{24} are $0, \sqrt{4}, \sqrt{6}, \sqrt{8}, \ldots$

In 1969 Conway showed that the Leech lattice again has a remarkable number theoretical property: It is the only even unimodular lattice in dimension 24 which does not have vectors of length $\sqrt{2}$. He used this result to determine the automorphism group (the group of orthogonal transformation which leave Λ_{24} invariant) of the Leech lattice and it turned out the number of automorphisms equals

$$\begin{aligned} |\operatorname{Aut}(\Lambda_{24})| &= 2^{22} \cdot 3^9 \cdot 5^4 \cdot 7^2 \cdot 11 \cdot 13 \cdot 23 \\ &= 8315553613086720000, \end{aligned}$$

and that this group contained three new sporadic simple groups $\mathrm{Co}_1, \mathrm{Co}_2, \mathrm{Co}_3$. The classification theorem of finite simple groups, which was announced in 1980, says that there are only 26 finite simple sporadic groups. They are sporadic in the sense that they are not contained in the infinite families of cyclic groups of prime order, alternating groups and groups of Lie type.

Similar to the eight-dimensional case, by results of Odlyzko, Sloane, Levenshtein, Bannai and Sloane, the 196560 shortest vectors of Λ_{24} give the unique solution of the kissing number problem in dimension 24. In 2004 Cohn and Kumar proved the optimality of the sphere packing of the Leech lattice among lattice packings by a computer assisted proof, see [2] and section 'Producing numerical evidence' further on. However, despite all the similarities of E_8 and Λ_{24}, there is a puzzling difference between E_8 and Λ_{24} when it comes to sphere coverings: Schürmann and Vallentin [9] showed in 2006 that Λ_{24} provides at least a locally thinnest sphere covering in the space of 24-dimensional lattices, whereas Dutour-Sikirić, Schürmann and Vallentin [9] showed in 2012 that one can improve the sphere covering of the E_8 lattice when picking a generic direction in the space of eight-dimensional lattices.

Theta series and modular forms

As already indicated, the class of even unimodular lattices is restrictive, at least in small dimensions. One can show that they only exist in dimensions that are divisible by 8, furthermore for every such dimension n there are only finitely many even uni-

$h_8 = 1$	Mordell, 1938
$h_{16} = 2$	Witt, 1941
$h_{24} = 24$	Niemeier, 1973
$h_{32} \geq 1162109024$	King, 2003

Table 1

modular lattices, this number is denoted by h_n. In Table 1 we summarize the known values of h_n.

A major tool for studying even unimodular lattice are their theta series (first studied by Jacobi (1804–1851)): The *theta series* of a lattice L is

$$\vartheta_L = \sum_{r=0}^{\infty} n_L(r)\, q^r$$
$$\text{with } n_L(r) = \{x \in L : x \cdot x = 2r\},$$

the generating function of the number of lattice vectors of length $\sqrt{2r}$. In order to work with them analytically we set $q = e^{2\pi i z}$ where z lies in the complex upper half plane $\mathbb{H} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$, so that ϑ_L is a function of z. The theta function is periodic mod $\mathbb{Z}$: we have $\vartheta_L(z) = \vartheta_L(z+1)$. The *Poisson summation formula* states

$$\sum_{x \in L} f(x+v) = \frac{1}{\operatorname{vol}(\mathbb{R}^n/L)} \sum_{y \in L^*} e^{2\pi i x \cdot y} \widehat{f}(y),$$

with $v \in \mathbb{R}^n$, where

$$\widehat{f}(y) = \int_{\mathbb{R}^n} f(x)\, e^{-2\pi i y \cdot x}\, dx$$

is the n-dimensional *Fourier transform*. Using the Poisson summation formula one can show that ϑ_L satisfies the transformation law

$$\vartheta_L(-1/z) = (z/i)^{n/2} \frac{1}{\operatorname{vol}(\mathbb{R}^n/L)} \vartheta_{L^*}(z),$$

which in particular shows that ϑ_L is a modular form of weight $n/2$. From this it is not difficult to derive that an even unimodular lattice can only exist when n is a multiple of 8.

What is a modular form? The group

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z},\ ad - bc = 1 \right\},$$

which is generated by the matrices

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

acts on upper half plane $\mathbb{H}$ by fractional linear transformations

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}.$$

The action of the generator S corresponds to the involution $z \mapsto -1/z$ and the action of the generator T corresponds to the translation $z \mapsto z+1$.

A *modular form of weight* k is a holomorphic function $f : \mathbb{H} \to \mathbb{C}$ that satisfies the transformation law

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z)$$
$$\text{for all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),\ z \in \mathbb{H},$$

and which has a power series expansion in $q = e^{2\pi i z}$. Next to theta series, *Eisenstein series*, due to Eisenstein (1823–1852), form an important class of modular forms. For an integer $k \geq 3$ define

$$E_k(z) = \frac{1}{2\zeta(k)} \sum_{(c,d) \in \mathbb{Z}^2 \setminus \{0\}} \frac{1}{(cz+d)^k}, \qquad (1)$$

where ζ is the Riemann zeta function $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$. For even integers $k \geq 3$, the Eisenstein series E_k is a modular form of weight k.

Curiously, a theorem of Siegel (1896–1981) gives a relation between the theta series of even unimodular lattices and Eisenstein series. Let $L_1, \ldots, L_{h_n}$ be the set of even unimodular lattices in dimension n. Define

$$M(n) = \frac{1}{|\operatorname{Aut}(L_1)|} + \cdots + \frac{1}{|\operatorname{Aut}(L_{h_n})|}.$$

Then

$$E_{n/2}(z) = \frac{1}{M(n)} \sum_{j=1}^{h_n} \frac{1}{|\operatorname{Aut}(L_j)|} \vartheta_{L_j}(z).$$

Another striking fact is that one can show that the modular forms form an algebra which is isomorphic to the polynomial algebra $\mathbb{C}[E_4, E_6]$.

When k is even, the Eisenstein series has the Fourier expansion

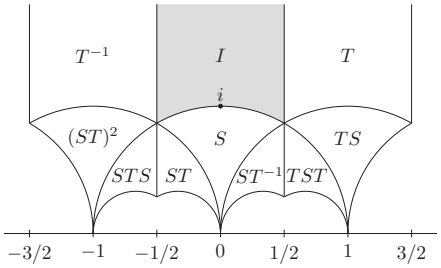


Figure 3 A fundamental domain of the action of $\mathrm{SL}_2(\mathbb{Z})$ on the upper half plane $\mathbb{H}$.

$$E_k(z) = 1 + \frac{2}{\zeta(1-k)} \sum_{n=1}^{\infty} \sigma_{k-1}(n)\, e^{2\pi i n z},$$

where $\sigma_{k-1}(n) = \sum_{d \mid n} d^{k-1}$ is the divisor function. We can express the theta series of the E_8 lattice and the Leech lattice Λ_{24} by E_4 and E_6. For the E_8 lattice we have

$$\begin{aligned} \vartheta_{\mathsf{E}_8}(z) = E_4(z) &= 1 + 240 \sum_{r=1}^{\infty} \sigma_3(r)\, q^r \\ &= 1 + 240 \cdot 1 \cdot q + 240 \cdot 9 \cdot q^2 + 240 \cdot 28 \cdot q^3 + \cdots \\ &= 1 + 240 \cdot q + 2160 \cdot q^2 + 6720 \cdot q^3 + \cdots. \end{aligned}$$

For the Leech lattice we have

$$\begin{aligned} \vartheta_{\Lambda_{24}}(z) &= E_4^3(z) - 720\Delta(z) \\ &= 1 + 196560 \cdot q^2 + 16773120 \cdot q^3 + \cdots \end{aligned}$$

where Δ is defined as

$$\Delta(z) = \frac{E_4(z)^3 - E_6(z)^2}{1728}.$$

When $k = 2$, we can still write down the series as in (1), but then we loose some pleasant properties. For example, the series no longer converges absolutely, so the ordering of summating matters. Then

$$\begin{aligned} E_2(z) &= \frac{1}{2\zeta(2)} \sum_{c \in \mathbb{Z}} \sum_{d \in \mathbb{Z}} \frac{1}{(c+dz)^2} \\ &= 1 - 24 \sum_{n=1}^{\infty} \sigma_1(n)\, e^{2\pi i n z}, \end{aligned}$$

where we of course omit the pair $(c,d) = (0,0)$ in the first sum. This *forbidden Eisenstein series* is not a modular form; instead it satisfies the following transformation law

$$E_2(-1/z) = z^2 E_2(z) - \frac{6iz}{\pi}.$$

It is a *quasi-modular form*.

The LP Bound of Cohn and Elkies

We can use optimization techniques, in particular linear and semidefinite programming, to obtain upper bounds on the optimal sphere packing density.

Let us recall some facts about linear programming. In a *linear program* we want to maximize a linear functional over a polyhedron. For example, we maximize the functional $a \mapsto c \cdot a$ over all (entrywise) nonnegative vectors $a \in \mathbb{R}^d$ satisfying the linear system $Aa = b$, or we maximize the functional over all (nonnegative) vectors a satisfying the inequality $Aa \leq b$. Linear programs can be solved efficiently in practice by a simplex algorithm (which traverses a path along vertices of the polyhedron) or by Karmakar's interior point method, where the latter runs in polynomial time.

The following theorem, the *linear programming bound of Cohn and Elkies* from 2003, can be used to obtain upper bounds on the sphere packing density. In the statement of this theorem we restrict to Schwartz functions because the proof, which we give here as it is simple and insightful, uses the Poisson summation formula. A function $f : \mathbb{R}^n \to \mathbb{R}$ is a *Schwartz function* if all its partial derivatives exist and tend to zero faster than any inverse power of x. There are alternative proofs that do not use Poisson summation and for which the Schwartz condition can be weakened.

Theorem 2. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a Schwartz function and r_1 is a positive number with $\widehat{f}(0) = 1$, $\widehat{f}(u) \geq 0$ for all u, and $f(x) \leq 0$ for $\|x\| \geq r_1$, then the density of a sphere packing in $\mathbb{R}^n$ is at most $f(0) \cdot \mathrm{vol}(B_n(r_1/2))$.*

Proof. Let P be a periodic packing of solid spheres of radius $r_1/2$. This means there is a lattice L and points $x_1, \ldots, x_m$ in $\mathbb{R}^n$ such that

$$P = \bigcup_{v \in L} \bigcup_{i=1}^{m} (v + x_i + B_n(r_1/2)).$$

The density of P is $m \cdot \mathrm{vol}(B_n(r_1/2)) / \mathrm{vol}(\mathbb{R}^n/L)$. By Poisson summation we have

$$\sum_{v \in L} \sum_{i,j=1}^{m} f(v + x_j - x_i) = \frac{1}{\mathrm{vol}(\mathbb{R}^n/L)} \sum_{u \in L^*} \widehat{f}(u) \sum_{i,j=1}^{m} e^{2\pi i u \cdot (x_j - x_i)},$$

and because $\widehat{f}(0) = 1$ and $\widehat{f}(u) \geq 0$ for all u, this is at least $m^2/\mathrm{vol}(\mathbb{R}^n/L)$. On the other hand, by the condition $f(x) \leq 0$ for $\|x\| \geq r_1$, we have

$$\sum_{v \in L} \sum_{i,j=1}^{m} f(v + x_j - x_i) \leq m f(0).$$

Hence, the density of P is at most $f(0) \cdot \mathrm{vol}(B_n(r_1/2))$. The density of any packing can be approximated arbitrarily well by the density of a periodic packing, so this completes the proof. □

We can additionally require either $f(0) = 1$ or $r_1 = 1$, without weakening the theorem. Moreover, we can restrict to *radial function*s, for if a function f satisfies the conditions of the theorem for some r_1, then the function

$$x \mapsto \int_{S^{n-1}} f(\|x\| \xi) d\omega(\xi),$$

where ω is the normalized invariant measure on S^{n-1}, also satisfies these conditions. For $f : \mathbb{R}^n \to \mathbb{R}$ radial, we (ab)use the notation $f(r)$ for the common value of f on the vectors of length r. The (inverse) Fourier transform maps radial functions to radial functions. Moreover, the Gaussian $x \mapsto e^{-\pi \|x\|^2}$ is fixed under the Fourier transform, and, more generally, the sets

$$P_d = \{x \mapsto p(\|x\|^2) e^{-\pi \|x\|^2} : p \text{ is a polynomial of degree at most d}\}.$$

are invariant under the Fourier transform. A *computer-assisted* approach to find good functions for the above theorem is to restrict to functions from P_d for some fixed value of d. Any function from this set that satisfies $f(0) = 1$ can be written as

$$f_a(x) = \left(1 + \sum_{k=1}^{d} a_k k! \pi^{-k} L_k^{n/2-1}(\pi \|x\|^2)\right) \times e^{-\pi \|x\|^2} \quad (2)$$

for some $a \in \mathbb{R}^d$, where $L_k^{n/2-1}$ is the Laguerre polynomial of degree k with parameter $n/2 - 1$ (Laguerre polynomials are a family of orthogonal polynomials). We choose this form for f_a, so that its Fourier transform is

$$\widehat{f_a}(u) = \left(1 + \sum_{k=1}^{d} a_k \|u\|^{2k}\right) e^{-\pi \|u\|^2}, \quad (3)$$

which means that $a \geq 0$ immediately implies $\widehat{f_a}(u) \geq 0$ for all u. Setting $r_1 = 1$ in the above theorem, we see that the optimal sphere packing density is upper bounded by the maximum of the linear functional $a \mapsto f_a(0)$ over all nonnegative vectors $a \in \mathbb{R}^d$ for which the linear inequalities $f_a(r) \leq 0$ for $r > 1$ are satisfied. For every fixed value of d this gives a semi-infinite linear program, which is a linear program with finitely many variables and infinitely many linear constraints.

One approach to solving these semi-infinite programs is to select a finite sample $S \subseteq [1, \infty)$ and only enforce the constraints $f_a(r) \leq 0$ for $r \in S$. For each S this yields a linear program whose optimal solution a can be computed using a linear programming solver. Then we verify that $f_a(r) \leq 0$ for all $r > 1$, or if this is not (almost) true, we run the problem again with a different (typically bigger) sample S. This approach works well in practice for the *spherical code problem*, which is a compact analogue of the sphere packing problem. Here, given a scalar $t \in (-1, 1)$, we seek a largest subset of the sphere $S^{n-1} \subseteq \mathbb{R}^n$ such that the inner product between any two distinct points is at least t. A spherical code corresponds to a spherical cap packing where we center spherical caps of angle $\arccos(t/2)$ about the points in the code. The Cohn–Elkies bound can be seen as a noncompact analogue of a similar bound for the spherical code problem known as the *Delsarte linear programming bound*. However, because of noncompactness this sampling approach does not work well for the sphere packing problem.

Another approach, based on semidefinite programming, does work well for noncompact problems such as the sphere packing problem. In [8] this approach is used to compute upper bounds for packings of spheres and spherical caps of several radii. *Semidefinite programming* is a powerful generalization of linear programming, where we maximize a linear functional over a spectrahedron instead of a polyhedron. That is, we maximize a functional $X \mapsto \langle X, C \rangle$ over all positive semidefinite $n \times n$ matrices X that satisfy the linear constraints $\langle X, A_i \rangle = b_i$ for $i = 1, \ldots, m$. Here $\langle A, B \rangle = \mathrm{trace}(B^\top A)$ denotes the trace inner product. As for linear programs, semidefinite programs can be solved efficiently by using interior point methods.

The usefulness of semidefinite programming in solving the above semi-infinite linear programs stems from the following two observations: Firstly, Pólya and Szegö showed that a polynomial is nonnegative on the interval $[1, \infty)$ if and only if it can be written as $s_1(r) + (r-1)s_2(r)$, where s_1 and s_2 are sum of squares polynomials. Secondly, a sum of squares polynomial of degree $2d$ can be written as $b(r)^\top Q b(r)$, where $b(r) = (1, r, \ldots, r^d)$, and where Q is a positive semidefinite matrix. (To see that a polynomial of this form is a sum of squares polynomial one can use a Cholesky factorization $Q = R^\top R$.) Using these observations we can introduce two positive semidefinite matrix variables Q_1 and Q_2, and replace the infinite set of linear constraints $f_a(r) \leq 0$ for $r > 1$, by a set of $2d + 2$ linear constraints that enforce the identity

$$f_a(r) = (1 - r) b(r)^\top Q_2 b(r) - b(r)^\top Q_1 b(r).$$

In this way we obtain a semidefinite program, which can be solved with a semidefinite programming solver, and whose

optimal value upper bounds the sphere packing density. For each d this finds the optimal function f_a.

Producing numerical evidence

Now we want to understand what has to happen when the Cohn–Elkies bound can be used to *prove the optimality* of the sphere packing given by an even and unimodular lattice L.

Then there exists a *magic function* $f : \mathbb{R}^n \to \mathbb{R}$ and a scalar r_1 that satisfy the conditions of Theorem 2, such that the density of the sphere packing equals $f(0) \cdot \mathrm{vol}(B_n(r_1/2))$. As observed before, we may assume $f(0) = 1$, so that r_1 is the shortest nonzero vector length in L. Under these assumptions on f we can derive extra properties that the function f and its Fourier transform must satisfy. Since $f(x) \leq 0$ and $\widehat{f}(x) \geq 0$ for all $\|x\| \geq r_1$, and $f(0) = (0) = 1$, we have *equality* in the following chain of inequalities

$$1 \leq \sum_{x \in L} \widehat{f}(x) = \sum_{x \in L} f(x) \leq 1.$$

This says that we have to have $f(x) = (x) = 0$ for all $x \in L \setminus \{0\}$. In fact, we can apply this argument to any rotation of L, so that $\widehat{f}(x) = f(x) = 0$ for all x where $\|x\|$ is a nonzero vector length in L. As noted before, we may take f to be radial, and then we have (again abusing notation)

$$f(r) = \widehat{f}(r) = 0$$

for all nonzero vector lengths r in L. This also tells us something about the *orders of the roots*. We have $f(0) = 1$ and $f(r) \leq 0$ for $r \in [r_1, \infty)$, so the roots at the vector lengths that are strictly larger than r_1 must have even order. We have $\widehat{f}(0) = 1$ and $\widehat{f}$ is nonnegative on $[0, \infty)$, so the roots at the nonzero vector lengths must have even order.

If f does not have additional roots, then in [1] it is shown that there is no other periodic packing achieving the same density as L. To apply this it is important that E_8 is the only even unimodular lattice in $\mathbb{R}^8$ and Λ_{24} is the only even unimodular lattice in $\mathbb{R}^{24}$ that does not contain vectors of length $\sqrt{2}$.

The Cohn–Elkies paper

In [1] Cohn and Elkies used this insight about the potential locations of the roots and double roots to derive a numerical scheme to find functions that are *close* to magic functions. They parametrized the function f_a as in (2). Then they required that f_a and $\widehat{f}_a$ have as many roots and double roots as possible, depending on the degree d. Afterwards they applied Newton's method to perturb the roots and double roots in order to optimize the value of the bound. In dimension 8 and 24 they obtained bounds which were too high only by factors of 1.000001 and 1.0007071. This provided the first strong evidence that magic functions exist for these two dimensions. Since the magic functions f have to have infinitely many roots, the degree d has to go to infinity. Could this method, in the limit, actually give the exact sphere packing upper bounds?

The Cohn–Kumar paper

The next step was taken by Cohn and Kumar in [2]. They improved the numerical scheme and by using degree $d = 803$ (with 3000-digit coefficients) they showed that in dimension 24 there is no sphere packing which is $1 + 1.65 \cdot 10^{-30}$ times denser than the Leech lattice. The actual aim of their paper was to show that the Leech lattice is the unique densest lattice in its dimension. For this they used the numerical data together with the known fact that the Leech lattice is a strict local optimum. The proof of Cohn and Kumar is a beautiful example of the symbiotic relationship between human and machine reasoning in mathematics.

The Cohn–Miller paper

For a long time Cohn and Miller were fascinated by the properties of these conjecturally existing magic functions. In their paper [4], submitted 15 March 2016 to the arXiv-preprint server, they gave a 'construction' of the magic functions using determinants of Laguerre polynomials. However, they could not prove that this construction indeed worked. With the use of high precision numerics they experimented with their construction. This resulted in improved bounds, optimality of Λ_{24} within a factor of $1 + 10^{-51}$. Even more importantly, they detected some unexpected rationalities: For instance using their numerical data they conjectured that for $n = 8$, the magic function f and $\widehat{f}$ have quadratic Taylor coefficients $-27/10$ and $-3/2$, respectively. For $n = 24$, the corresponding coefficients should be $-14347/5460$ and $-205/156$.

Viazovska's breakthrough

Viazovska made the spectacular discovery that a magic function indeed exists for dimension $n = 8$. Building on this, Cohn, Kumar, Miller, Radchenko and Viazovska found a magic function for $n = 24$. Viazovska's construction is based on a couple of new ideas, which we want to explain briefly.

Each radial Schwartz function $f : \mathbb{R}^n \to \mathbb{R}$ can be written as a linear combination of radial eigenfunctions of the Fourier transform in $\mathbb{R}^n$ with eigenvalues $+1$ and -1. Viazovska wrote the magic function as a linear combination $f = \alpha f_+ + \beta f_-$, where f_+ is a radial eigenfunction of the Fourier transform with eigenvalue $+1$ and f_- is a radial eigenfunction with eigenvalue -1. The coefficients α and β are determined later on.

She makes the Ansatz that for $r > r_1$, we can write these functions f_+ and f_- as a squared sine function times the Laplace transform of a (quasi)-modular form. That is, she proposes that

$$f_+(r) = -4 \sin(\pi r^2/2)^2 \times \int_0^{i\infty} \psi_+(-1/z) z^{n/2-2} e^{\pi i r^2 z} dz$$

and

$$f_-(r) = -4 \sin(\pi r^2/2)^2 \int_0^{i\infty} \psi_-(z) e^{\pi i r^2 z} dz,$$

where ψ_+ is a quasi-modular form and ψ_- is a modular form.

The $\sin(\pi r^2/2)^2$ factor insures (assuming the above integrals do not have cusps) that the resulting function f (as well as its Fourier transform) have double roots at all but the first occurring vector lengths.

Viazovska noticed that an analytic extension of f_- exists and that it is an eigenfunction of the Fourier transform having eigenvalue -1 when the following modularity relation holds:

$$\psi_-\left(\frac{az+b}{cz+d}\right) = (cz+d)^{2-n/2} \psi_-(z)$$
$$\text{for all} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}), a, d \text{ odd}, b, c \text{ even}.$$

For the explicit definition of f_- we need the theta functions

$$\Theta_{01}(z) = \sum_{n \in \mathbb{Z}} (-1)^n e^{\pi i n^2 z}$$

and

$$\Theta_{10}(z) = \sum_{n \in \mathbb{Z}} e^{\pi i (n+1/2)^2 z}.$$

Similarly, but technically much more involved, the analytic extension of f_+ is an eigenfunction for the eigenvalue $+1$ when ψ_+ is a quasi-modular form. The forbidden Eisenstein series E_2 becomes important here.

Now it needs to be shown that there exists a linear combination $f = \alpha f_+ + \beta f_-$ so that $\widehat{f}(0) = f(0) = 1$ holds, and that the sign conditions $f(r) \leq 0$ for $r \geq r_1$ and $\widehat{f}(u) \geq 0$ for all $u \geq 0$ are fulfilled. For this, and more, we refer to the beautiful original papers. Here we want to end by taking a look at the magic functions: see Table 2.

E_8	Λ_{24}
$\psi_+ = \frac{(E_2E_4 - E_6)^2}{\Delta}$	$\psi_+ = \frac{25E_4^4 - 49E_6^2E_4 + 48E_6E_4^2E_2 + 25E_6^2E_2^2 - 49E_4^3E_2^2}{\Delta^2}$
$\psi_- = \frac{5\Theta_{01}^{12}\Theta_{10}^8 + 5\Theta_{01}^{16}\Theta_{10}^4 + 2\Theta_{01}^{20}}{\Delta}$	$\psi_- = \frac{7\Theta_{01}^{20}\Theta_{10}^8 + 7\Theta_{01}^{24}\Theta_{10}^4 + 2\Theta_{01}^{28}}{\Delta^2}$
$f(x) = \frac{\pi i}{8640} f_+(x) + \frac{i}{240\pi} f_-(x)$	$f(x) = -\frac{\pi i}{113218560} f_+(x) - \frac{i}{262080\pi} f_-(x)$

Table 2 The magic functions.

Interview with Henry Cohn, Abhinav Kumar, Stephen D. Miller and Maryna Viazovska

The interview was conducted by Frank Vallentin using the online communication platform Google Hangouts between 20 May and 14 June 2016.

Computer Assistance

Dear Henry, Abhinav, Steve, and Maryna, First of all let me congratulate you to your breakthrough papers. This issue of the 'Nieuw Archief voor Wiskunde' is a special issue focussing on computer-assisted mathematics. In it we have two articles about sphere packings: One about the formal proof of the Kepler conjecture and one about your recent breakthrough on sphere packings in dimensions 8 and 24. At the moment a proof of the Kepler conjecture without computer-assistance is not in sight, but your proofs in dimension 8 and 24 require almost no computers. How were computers helpful to you when finding the proofs?

Abhinav: "The Cohn–Elkies paper and later the Cohn-Kumar and the Cohn–Miller papers were certainly useful as indicators that the solution was out there waiting for the right functions. In our new Cohn–Kumar–Miller–Radchenko–Viazovska paper at least, the numerical data was quite useful because once we had figured out the right finite dimensional space of modular or quasi-modular forms, the numerics helped us pin down the exact form (up to scaling) of the function. In particular, we matched values and derivatives of f and $\widehat{f}$ at lattice vector lengths to cut down the space by imposing linear conditions. Steve has a Mathematica code to do some of this but we also used PARI/GP and occasionally Maple.

A couple more things — in both the proofs the final inequalities needed for the functions and Fourier transforms are done with computer assistance. There might be more elegant hand proofs, but so far we haven't found them. And we did quite a lot of messing around with q-expansions et cetera, which would have been very painful outside of a computer algebra system."

Steve: "Though I think there will eventually be a slick proof that can be written by hand, computers were completely essential in this story. To me the best example is the appearance of rational numbers that Henry Cohn and I discovered.

I can only speak for myself, but I was completely fascinated by the (then proposed) existence of these 'magic' functions which completely solve sphere packing in special dimensions. To satisfy this curiosity, Henry and I began computing their features to see if we could learn more about them. We found — using some serendipity with the online 'inverse symbolic calculator' website — that their quadratic Taylor coefficients were rational, and furthermore related to Bernoulli numbers. This was a strong hint that modular forms were connected, though we never understood why until we saw Maryna's paper. We found other rationalities (such as the derivatives at certain points) using some theoretical motivation, combined with good numerical approximations. It was a type of 'moonshine', with a fascinating sequence of numbers and an amazing structure we could not otherwise access.

Once Maryna's paper appeared, it took just a few days to combine her insight with our previous numerics in an exact way. It's important to stress that at this point we could derive the magic function for 24 dimensions without using floating point calculations, since we had already extrapolated exact expressions from previous numerics. Maybe we will later understand a way to derive the 24-dimensional functions without such information, but at the time it was highly convenient to leverage them."

Maryna, did you also use computer assistance when you found the function for dimension 8? In your paper you mention in passing that one compute the first hundred terms of Fourier expansions of modular forms in a few second using PARI/P or Mathematica.

Maryna: "Numerical evidence was crucial to believe in the existence and uniqueness of 'magic' functions. I used computer calculations to verify that approximations to the magic function computed from linear equations (similar to the equations considered in the Cohn–Miller paper) converge to the function computed as an integral transform of a modular form. I used Mathematica and PARI/GP for this purpose."

Checking the positivity conditions is the only part of the proof which depends on the use of computers. How do you make sure that your computer proof is indeed mathematically rigorous?

Henry: "In her 8-dimensional proof, Maryna used interval arithmetic. In the 24-dimensional case, we used exact rational arithmetic. Either way, it's not a big obstacle. The inequalities you need have a little slack, which means you can bound everything

Photo: Archives of the Mathematisches Forschungsinstitut Oberwolfach

Henry Cohn

in any number of different ways, without needing to do anything too delicate."
Steve: "To elaborate: Our positivity check involves showing certain power series $f(q)$ in a parameter q are positive, where $q<1$. It is not difficult to bound the coefficients and deduce this positivity for $q<c$ (where c is an effective constant), so the problem reduces to showing positivity for q in the interval $[c,1]$. Numerically one can plot this directly, of course. From such a graph we see $f(q)>b$ for some explicit constant b. Write $f(q)=p(q)+t(q)$, where $p(q)$ is a polynomial consisting of the first several terms in the power series and $t(q)$ the tail, with the number of terms chosen so that $t(q)$ is provably less than $b/2$ for q in $[c,1]$. We are now reduced to showing $p(q)-b/2>0$ for $q\in[c,1]$, and such an inequality can be rigorously established using Sturm's theorem."

Stephen D. Miller

Modular forms

For a long time it has been known that the E_8 lattice and the Leech lattice have strong connections to modular forms, simply by their theta series. However, one thing which puzzles us (we mainly work in optimization) is that you solve an optimization problem, an infinite-dimensional linear program, with tools from analytic number theory, especially using modular forms. How surprising was it to you that using modular forms was one key to the proof?

Henry: "Modular forms are by far the most important class of special functions related to lattices, so in that sense it's not so surprising that they come up. Over the years many people had suggested using them, but it wasn't clear how. For example, many years ago I had tried the Laplace transform of a modular form, but without Maryna's $\sin^2$-factor. Without that, it seemed impossible to get anything like the right roots, and therefore the approach was completely useless. What I find beautiful about her proof is how ingeniously it puts everything together (when I know from personal experience that thinking 'I'd better use modular forms' will not just lead you to this proof).

Before Maryna's proof, I could imagine two possibilities. One was that the right approach would be to solve the LP problem in general, getting the exceptional dimensions just as special cases. This approach might not have involved modular forms at all. The other was that there would be particular special functions in those dimensions. I always hoped for something special in 8 and 24 dimensions (e.g., based on the numerical experiments Steve and I worked on), but I was a little worried that maybe the difficulty of writing these functions down indicated that this might be the wrong approach (while solving the problem in general seemed even harder). It was great to see that everything was as beautiful as we had always hoped."
Steve: "The use of Poisson summation in the Cohn–Elkies paper (and an earlier technique of Siegel) had already brought methods of analytic number theory into the subject. It was clear relatively early on that modular forms must be somehow involved in the final answer. This is both because of the appearance of special Bernoulli numbers (that prominently arise in modular forms) as well as the overall structure of the summation formulas. For example, Henry and I derived a relevant Fourier eigenfunction with simple zeros using Voronoi summation formulas and derivatives of modular forms.

Abhinav Kumar

However, while these ingredients had been on the table for a long time, we didn't know how to combine them until Maryna's paper appeared. At that point many of the various pieces of evidence we had suddenly fit together. Quasi-modular forms (which are derivatives of modular forms) are very crucial to this story."

Collaboration

Maryna's breakthrough paper which solved the sphere packing problem in dimension 8 was submitted on March 14, 2016 to the arXiv-preprint server. Then it took only one week until you submitted the solution of the sphere packing problem in dimension

Photo: Archives of the Mathematisches Forschungsinstitut Oberwolfach

Maryna Viazovska

24 to the arXiv. Working at five different, distant places, how did you collaborate? What were the difficulties when going from 8 to 24 dimensions?

Steve: "That was certainly an exciting and memorable week. Once we had assembled our team, things moved extremely quickly. This is mainly because Maryna's methods are so powerful, but it was also important that certain pairs of us (Henry–myself, Abhinav–Henry, and Danylo–Maryna) were established collaborators that had already worked together well.

In addition to phone calls and email, we used Skype and Dropbox to communicate our ideas. I particularly like drawing mathematics on a tablet PC and sharing the screen on Skype — this allows the others to watch as if I'm writing on a blackboard. As soon as we saw Maryna's paper on Tuesday morning, we tried to make concrete bridges with the numerical observations Henry and I made in our Cohn-Miller paper. After some reformulation of her modular forms as quotients, by Wednesday it was then clear what properties of the q-expansions would be needed to obtain the 24-dimensional functions. We also set up computer programs to match potential candidate functions with the numerical values that we could compute separately.

On Thursday we had the right space of modular forms and a program ready to search in them. By matching the conjectured rationalities, we found the even eigenfunction on Thursday night and the odd eigenfunction on Friday morning. We used Mathematica and PARI/GP for this. We also checked using graphs and q-expansions that the necessary positivity conditions indeed hold, but did not have a completely rigorous proof of this remaining point.

By Friday afternoon I was completely exhausted (and in any event do not work on the Jewish Sabbath). It's important to note that the positivity analysis was a little different in 24 dimensions than in 8 because of an extra pole that occurs."

Going further

Now the sphere packing problem has been solved in 1, 2, 3, 8 and 24 dimensions and, coming close to the end of the interview, it is time to make speculations: Are there candidate dimensions where a solution is in sight? Do you think that your method will be useful for this (or for other problems)?

Henry: "It's hard to say for sure whether there might be further sharp cases in higher dimensions, but it seems unlikely that they would have remained undetected. (At the very least it's not plausible that they could have the same widespread occurrences in mathematics as E_8 or the Leech lattice, since they would presumably have been discovered in the process of classifying the finite simple groups, for example.) Two dimensions, $n = 2$, remains open, although of course a solution is known by elementary geometry; the LP bounds behave rather differently in two dimensions compared with 8 or 24."

Steve: "Yes, dimension two seems very different because of the delicate arithmetic nature of the root lengths."

Henry: "I'm sure Maryna's wonderful approach to constructing these functions is just the tip of an iceberg, and I'm optimistic that humanity will learn more about how LP bounds and related topics work.

One mystery I find particularly intriguing is what's so special about 8 and 24 dimensions. Maryna's methods give a beautiful proof, but I still don't really know a conceptual explanation as to what's different in, say, 16 dimensions (beyond just the fact that the Barnes-Wall lattice isn't as nice as E_8 or the Leech lattice). On a slightly different topic, I hope someone solves the four-dimensional sphere packing problem, but it will require different techniques. Unlike the cases where the LP bounds are sharp, these pair correlation inequalities will not suffice by themselves. But the D_4 lattice is awfully beautiful, and the world deserves a proof of optimality. The only question is how...

Henry, Abhinav, Steve, and Maryna, thank you very much for this interview.

References

1 H. Cohn and N.D. Elkies, New upper bounds on sphere packings I, *Ann. of Math.* 157 (2003), 689–714.

2 H. Cohn and A. Kumar, Optimality and uniqueness of the Leech lattice among lattices, *Ann. of Math.* 170 (2009) 1003–1050.

3 H. Cohn, A. Kumar, S.D. Miller, D. Radchenko and M.S. Viazovska, The sphere packing problem in dimension 24, arXiv:1603.06518 [math.NT], 12 pp.

4 H. Cohn and S.D. Miller, Some properties of optimal functions for sphere packing in dimensions 8 and 24, arXiv:1603.04759 [math.MG], 23 pp.

5 J.H. Conway and N.J.A. Sloane, *Sphere Packings, Lattices and Groups*, Springer, 1988.

6 M. Dutour Sikirić, A. Schürmann and F. Vallentin, Inhomogeneous extreme forms, *Annales de l'institut Fourier* 62 (2012), 2227–2255.

7 T.C. Hales, M. Adams, G. Bauer, D. Tat Dang, J. Harrison, T. Le Hoang, C. Kaliszyk, V. Magron, S. McLaughlin, T. Tat Nguyen, T. Quang Nguyen, T. Nipkow, S. Obua, J. Pleso, J. Rute, A. Solovyev, A. Hoai Thi Ta, T. Nam Tran, D. Thi Trieu, J. Urban, K. Khac Vu and R. Zumkeller, A formal proof of the Kepler conjecture, arXiv:1501.02155 [math.MG], 21 pp.

8 D. de Laat, F.M. de Oliveira Filho, F. Vallentin, Upper bounds for packings of spheres of several radii, *Forum Math. Sigma* 2 (2014), e23, 42 pp.

9 A. Schürmann and F. Vallentin, Local covering optimality of lattices: Leech lattice versus root lattice E_8, *IMRN* 32 (2005), 1937–1955.

10 M.S.Viazovska, The sphere packing problem in dimension 8, arXiv:1603.04246 [math.NT], 22 pages.

Marcel K. de Carli Silva
Instituto de Matemática e Estatística
Universidade de São Paulo, Brazil
mksilva@ime.usp.br

Fernando Mário de Oliveira Filho
Instituto de Matemática e Estatística
Universidade de São Paulo, Brazil
fmario@ime.usp.br

Cristiane Maria Sato
Centro de Matemática, Computação e Cognição,
Universidade Federal do ABC, Brazil
c.sato@ufabc.edu.br

Flag algebras: a first glance

The theory of flag algebras, introduced by Razborov in 2007, has opened the way to a systematic approach to the development of computer-assisted proofs in extremal combinatorics. It makes it possible to derive bounds for parameters in extremal combinatorics with the help of a computer, in a semi-automated manner. In this article Marcel de Carli Silva, Fernando de Oliveira Filho and Cristiane Sato describe the main points of the theory in a complete way, using Mantel's theorem as a guiding example.

Mantel's theorem, perhaps the first result in extremal graph theory, was motivated by a problem proposed by W. Mantel in an issue of the journal *Wiskundige Opgaven*, published by the KWG [10]:

Vraagstuk XXVIII. K 13 a. Er zijn eenige punten gegeven waarvan geen vier in een zelfde vlak liggen. Hoeveel rechten kan men hoogstens tusschen die punten trekken zonder driehoeken te vormen? [W. Mantel]

(**Problem XXVIII.** K 13 a. Given are some points, no four of which lie on the same plane. How many lines at most can one draw between the points without forming triangles?)

In the language of graph theory, Mantel's problem asks for the maximum number of edges that a graph without triangles can have: the restriction that no four points lie on the same plane is there exactly to ensure that only triangles between the given points can be formed when lines are drawn.

A triangle-free graph on n vertices can be constructed as follows: divide the vertex set into two parts of $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ vertices each and add all edges between the parts. The resulting graph is bipartite, and hence in particular triangle-free, and has $\lfloor n^2/4 \rfloor$ edges. Mantel's theorem states that this is an *extremal example*, the best one can do: *every triangle-free graph on n vertices has at most $\lfloor n^2/4 \rfloor$ edges*.

This answer to Mantel's problem appeared in the same issue of *Wiskundige Opgaven*. There it is mentioned that solutions were provided by Mantel and several others; a proof by W.A. Wythoff (1865–1939), a former student of D.J. Korteweg (1848–1941), is included.

The theory of flag algebras allows us to computationally tackle extremal graph theory problems such as Mantel's problem and to obtain results such as Mantel's theorem. To understand how this is done, we first need to define exactly which extremal problems we consider.

The *size* of a graph G is its number of vertices $|V(G)|$ and is denoted by $|G|$. For $U \subseteq V(G)$, we denote by $G[U]$ the subgraph of G *induced* by U, that is, the subgraph of G with vertex set U and all the edges of G between vertices of U. For graphs F and G, let $p(F;G)$ be the probability that a set $U \subseteq V(G)$ with $|U| = |F|$, chosen uniformly at random, is such that $G[U]$ is isomorphic to F. We say that $p(F;G)$ is the *density* of F in G. In other words, if $c(F;G)$ is the number of times F occurs as an induced subgraph of G, then

$$p(F;G) = c(F;G)\binom{|G|}{|F|}^{-1}.$$

Let $\mathcal{H}$ be a collection of graphs. A graph G is *$\mathcal{H}$-free* if no induced subgraph of G is isomorphic to a graph in $\mathcal{H}$. A fundamental problem in extremal graph theory is to determine, for a given graph C, the maximum asymptotic density of C in $\mathcal{H}$-free graphs

$$\operatorname{ex}(C,\mathcal{H}) = \sup_{(G_k)_{k\geq 0}} \limsup_{k\to\infty} p(C;G_k), \qquad (1)$$

where the supremum is taken over all sequences $(G_k)_{k\geq 0}$ of $\mathcal{H}$-free graphs that are *increasing*, i.e., with $(|G_k|)_{k\geq 0}$ strictly increasing.

Mantel's theorem shows that $\operatorname{ex}(\text{—}, \{\triangle\}) \leq \frac{1}{2}$. Together with the extremal example described above, we actually have $\operatorname{ex}(\text{—}, \{\triangle\}) = \frac{1}{2}$.

Let $\mathcal{G}$ be the set of all finite $\mathcal{H}$-free graphs taken up to isomorphism. An increasing sequence $(G_k)_{k\geq 0}$ is *convergent* if $\lim_{k\to\infty} p(F;G_k)$ exists for every $F \in \mathcal{G}$. Every increasing sequence of $\mathcal{H}$-free graphs has a convergent subsequence. Indeed, densities are numbers in $[0,1]$, so for $k \geq 0$ the function $F \mapsto p(F;G_k)$ can be identified with a point in $[0,1]^{\mathcal{G}}$, which is a compact space by Tychonoff's theorem.

In (1) we may therefore restrict ourselves to convergent sequences and this allows us to work with their limits. Call $\phi : \mathcal{G} \to \mathbb{R}$ a *limit functional* if there is a convergent sequence $(G_k)_{k\geq 0}$ of $\mathcal{H}$-free graphs such that

$$\phi(F) = \lim_{k\to\infty} p(F; G_k)$$

for all $F \in \mathcal{G}$ and let Φ denote the set of all limit functionals. *Then computing* $\mathrm{ex}(C, \mathcal{H})$ *is the same as solving an optimization problem over* Φ*:*

$$\mathrm{ex}(C, \mathcal{H}) = \sup\{\phi(C) : \phi \in \Phi\}. \quad (2)$$

This is just a rewording of the original problem, but it emphasizes that the difficulty here lies in understanding Φ. This set may be very complex and computationally intractable, but to get an upper bound for $\mathrm{ex}(C, \mathcal{H})$ we do not need to work with Φ. Instead, we may look for a nice *relaxation* of Φ, that is, a set $\Phi' \supseteq \Phi$ for which we can solve the optimization problem. A first and obvious relaxation would be to take $\Phi' = [0,1]^{\mathcal{G}}$. Solving the optimization problem is then trivial, but we always get the bound $\mathrm{ex}(C, \mathcal{H}) \leq 1$. The difficulty lies in managing the trade-off between the quality of the relaxation and its tractability.

The theory of flag algebras [12], developed by the Russian mathematician Alexander Razborov, winner of the Nevanlinna Prize in 1990 and the Gödel Prize in 2007, gives us computationally-tractable relaxations of Φ that have displayed good quality in practice. We may then use the computer to solve the corresponding optimization problems, thus obtaining upper bounds for $\mathrm{ex}(C, \mathcal{H})$ that are often tight. Perhaps the most attractive feature in the theory is that the whole process is more-or-less automatic: obtaining the relaxation and solving the corresponding problems is basically a computational matter. So the theory of flag algebras allows us to harness computational power and apply it to problems in extremal combinatorics; it can be understood as part of the growing trend for the use of computers in mathematics.

Razborov credits Bondy [3] with a predecessor of the theory of flag algebras. Bondy applies counting techniques to the Caccetta–Häggkvist conjecture and illustrates his idea on Mantel's theorem. (The Cacceta–Häggkvist conjecture states that every simple directed graph on n vertices with outdegree at least r has a cycle with length at most $\lceil n/r \rceil$.) Here is a proof that $\mathrm{ex}($  , {  }$) \leq \frac{1}{2}$ that is a rewording of the proof by Bondy in terms of densities and limit functionals. This proof is a first glance into the theory of flag algebras; in it we will derive by hand some constraints on limit functionals of sequences of triangle-free graphs and then give an explicit simple relaxation of Φ from which Mantel's theorem will follow.

A triangle-free graph may have three different graphs on three vertices as induced subgraphs: the empty graph , the graph with one edge  , and the graph with two edges . (Nonedges are represented by dashed lines.) Let G be a triangle-free graph. Every edge of G belongs to $|G|-2$ induced subgraphs with three vertices, whence

$$p(\,\cdot\,\cdot\,\cdot\,; G) + 2p(\,\cdot\,\cdot\,\cdot\,; G) = 3p(\,\cdot\!-\!\cdot\,; G).$$

This is valid for every triangle-free graph G, hence also for a limit functional ϕ:

$$\phi(\,\cdot\,\cdot\,\cdot\,) + 2\phi(\,\cdot\,\cdot\,\cdot\,) = 3\phi(\,\cdot\!-\!\cdot\,).$$

We have our first constraint satisfied for all $\phi \in \Phi$.

A second constraint comes from the identity

$$p(\,\cdot\,\cdot\,\cdot\,; G) = \binom{|G|}{3}^{-1} \sum_{v\in V(G)} \binom{d(v)}{2},$$

where $d(v)$ is the degree of vertex v. To rewrite the right-hand side above, we need to extend the definition of the density function p to partially-labeled graphs. Say F and G are graphs each having a special vertex labeled 1, and let x_1 be the vertex of G labeled 1. Let $p(F; G)$ be the probability that a set $U \subseteq V(G)\setminus\{x_1\}$ with $|U| = |F| - 1$, chosen uniformly at random, is such that $G[U \cup \{x_1\}]$ is isomorphic to F via a label-preserving isomorphism, that is, an isomorphism that takes the labeled vertex of F to the labeled vertex of G.

For $v \in V(G)$, denote by G^v the labeled graph obtained from G by labeling vertex v with label 1. Let  denote the labeled graph obtained from by labeling the vertex of degree two with label 1; similarly for other graphs the solid vertex will be the labeled vertex. Then for a triangle-free graph G we have

$$\begin{aligned} p(\,\cdot\,\cdot\,\cdot\,; G) &= \binom{|G|}{3}^{-1} \sum_{v\in V(G)} \binom{d(v)}{2} \\ &= \binom{|G|}{3}^{-1} \sum_{v\in V(G)} p(\,\cdot\,\cdot\,\cdot\,; G^v) \binom{|G|-1}{2} \\ &= \frac{3}{|G|} \sum_{v\in V(G)} p(\,\cdot\,\cdot\,\cdot\,; G^v). \end{aligned} \quad (3)$$

Now comes a key observation. As the size of G goes to infinity, $p(\,\cdot\,\cdot\,\cdot\,; G^v)$ goes to $p(\,\cdot\!-\!\cdot\,; G^v)^2$. This is not hard to prove (do it!), but the intuition should be clear: if G is very large, then choosing a subset of $V(G)\setminus\{v\}$ of size 2 uniformly at random is basically the same as choosing two vertices in $V(G)\setminus\{v\}$ independently — the probability of choosing the same vertex twice becomes negligible as $|G|$ grows larger.

So let ϕ be the limit functional of a convergent sequence $(G_k)_{k\geq 0}$ of triangle-free graphs. Then

$$\begin{aligned} \phi(\,\cdot\,\cdot\,\cdot\,) &= \lim_{k\to\infty} p(\,\cdot\,\cdot\,\cdot\,; G_k) \\ &= \lim_{k\to\infty} \frac{3}{|G_k|} \sum_{v\in V(G_k)} p(\,\cdot\,\cdot\,\cdot\,; G_k^v) \\ &= \lim_{k\to\infty} \frac{3}{|G_k|} \sum_{v\in V(G_k)} p(\,\cdot\!-\!\cdot\,; G_k^v)^2. \end{aligned} \quad (4)$$

Now, for any triangle-free graph G the Cauchy-Schwarz inequality gives

$$\sum_{v\in V(G)} p(\,\cdot\!-\!\cdot\,; G^v)^2 \geq \frac{1}{|G|}\left(\sum_{v\in V(G)} p(\,\cdot\!-\!\cdot\,; G^v)\right)^2.$$

Together with (4) and

$$\sum_{v\in V(G)} p(\,\cdot\!-\!\cdot\,; G^v)(|G|-1) = 2p(\,\cdot\!-\!\cdot\,; G)\binom{|G|}{2}$$

we get

$$\phi(\,\cdot\,\cdot\,\cdot\,) \geq \lim_{k\to\infty} 3p(\,\cdot\!-\!\cdot\,; G_k)^2 = 3\phi(\,\cdot\!-\!\cdot\,)^2.$$

So every limit functional ϕ satisfies the constraints

$$\phi(\,\cdot\,\cdot\,\cdot\,) + 2\phi(\,\cdot\,\cdot\,\cdot\,) = 3\phi(\,\cdot\!-\!\cdot\,),$$

$$\phi(\,\cdot\,\cdot\,\cdot\,) \geq 3\phi(\,\cdot\!-\!\cdot\,)^2.$$

What do we get in (2) if we optimize over the set Φ' of all $\phi : \mathcal{G} \to [0,1]$ satisfying the constraints above? Well, suppose $\phi \in \Phi'$. Multiply the second constraint by 2 and subtract it from the first to get

$$\phi(\,\cdot\,\cdot\,\cdot\,) \leq 3\phi(\,\cdot\!-\!\cdot\,) - 6\phi(\,\cdot\!-\!\cdot\,)^2.$$

Since $\phi($  $) \geq 0$, we then have $\phi($$) \leq \frac{1}{2}$. So the optimal value of (2) with Φ' instead of Φ is at most $\frac{1}{2}$, hence $\mathrm{ex}($, {}$) \leq \frac{1}{2}$.

In the following sections the main points of Razborov's theory of flag algebras are developed. Unless otherwise noted, every definition and result presented here can be found in Razborov's original paper [12].

Types and flags

In the introduction, we derived valid inequalities for Φ by combining densities of partially-labeled graphs as in (3). In the

next few sections we will develop Razborov's theory of flag algebras, which automates this process. The discussion will be focused on families of graphs for concreteness, though one of the most attractive features of the theory is that it applies to a whole range of structures, including directed graphs, hypergraphs, and permutations.

For an integer $k \geq 0$, write $[k] = \{1,\ldots,k\}$. Fix a family $\mathcal{H}$ of forbidden subgraphs. A *type* of *size* k is an $\mathcal{H}$-free graph σ with $V(\sigma) = [k]$. We can think of it as a graph with vertices labeled with $1,\ldots,k$, whereas we regard graphs as unlabeled. The empty type is denoted by $\varnothing$.

Let σ be a type of size k and F be a graph on at least k vertices. An *embedding* of σ into F is an injective function $\theta : [k] \to V(F)$ that defines an isomorphism between σ and the subgraph of F induced by $\operatorname{Im}\theta$.

A *σ-flag* is a pair (F,θ) where F is an $\mathcal{H}$-free graph and θ is an embedding of σ into F. So a σ-flag is a partially-labeled graph that avoids $\mathcal{H}$ and whose labeled part is a copy of σ. When the embedding itself is not important, we will drop it, speaking simply of the σ-flag F.

The *labeled vertices* of (F,θ) are the vertices in the image of θ. Note that an $\varnothing$-flag is just *an $\mathcal{H}$-free graph*. Any type σ of size k can also be seen as the σ-flag (σ,θ) where θ is the identity on $[k]$.

Isomorphism between σ-flags is defined just as for graphs, but now the labels should also be preserved by the bijection. More precisely, σ-flags (F,θ) and (G,η) are *isomorphic* if there is a graph isomorphism $\rho : V(F) \to V(G)$ between F and G such that $\rho(\theta(i)) = \eta(i)$ for $i = 1,\ldots,|\sigma|$. Write $(F,\theta) \simeq (G,\eta)$ when (F,θ) and (G,η) are isomorphic, or simply $F \simeq G$ when the embeddings are not important. In the introduction, this notion was used only for σ-flags where σ is the type of size 1. Figure 1 shows some flags of different types.

For $n \geq |\sigma|$, denote by $\mathcal{F}_n^\sigma$ the set of all σ-flags of size n, taken up to isomorphism; denote by $\mathcal{F}^\sigma$ the set of all σ-flags taken up to isomorphism. Note that the set $\mathcal{G}$ of all $\mathcal{H}$-free graphs is simply $\mathcal{F}^\varnothing$. A type σ is *degenerate* if $\mathcal{F}^\sigma$ is finite. If σ is nondegenerate, then $\mathcal{F}_n^\sigma \neq \emptyset$ for all $n \geq |\sigma|$. It is easy to construct a family $\mathcal{H}$ for which there are degenerate types: take for instance $\mathcal{H}$ as the set of all graphs with 1000 vertices containing at least one triangle. Then the triangle itself is $\mathcal{H}$-free, and hence is a type, but there are no $\triangle$-flags of size ≥ 1000.

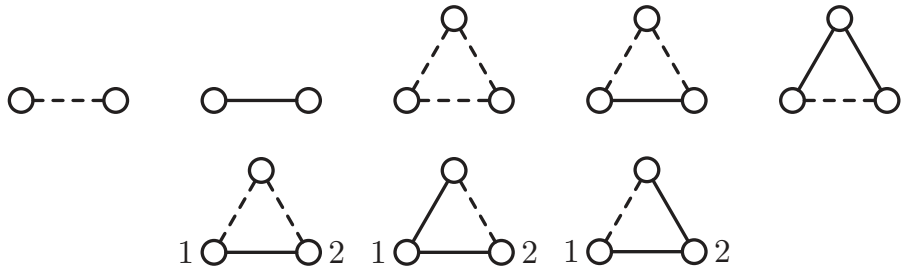


Figure 1 Let $\mathcal{H} = \{\triangle\}$. On the top row we have all $\varnothing$-flags of sizes 2 and 3, up to isomorphism (nonedges are shown as dashed lines); notice that the triangle itself is not a flag. On the bottom row we have all flags of type $\sigma = 1 \bullet\!\!-\!\!\bullet 2$; notice that the last two of these flags are not isomorphic, since the isomorphism has to preserve the labels.

From now on, we assume that all types are nondegenerate. In particular, every time a result about σ-flags is stated, it is implicitly assumed that σ is nondegenerate.

Density

The definition of density given in the introduction can be extended to σ-flags as follows. We say that σ-flags $F_1,\ldots,F_t$ *fit* in a σ-flag G if

$$|G| - |\sigma| \geq (|F_1| - |\sigma|) + \cdots + (|F_t| - |\sigma|).$$

Let $F_1,\ldots,F_t$ and (G,θ) be σ-flags such that $F_1,\ldots,F_t$ fit in G. Consider the following experiment: choose pairwise-disjoint sets $U_1,\ldots,U_t \subseteq V(G) \setminus \operatorname{Im}\theta$ of unlabeled vertices of G with $|U_i| = |F_i| - |\sigma|$ uniformly at random. Let $p(F_1,\ldots,F_t;G)$ be the probability that the σ-flag $(G[U_i \cup \operatorname{Im}\theta], \theta)$ is isomorphic to F_i for $i = 1,\ldots,t$. This is the *density* of $F_1,\ldots,F_t$ in G. For $\varnothing$-flags and $t = 1$, this definition coincides with the usual notion of density for graphs. In the introduction we also extended the definition of density to graphs with one labeled vertex; this corresponds to taking $t = 1$ and the only type of size 1 as σ.

Say $|F| \leq n \leq |G|$. To embed F into G, we may first try to embed F into a σ-flag F' of size n and then embed F' into G. This gives us another way to compute $p(F;G)$:

$$p(F;G) = \sum_{F' \in \mathcal{F}_n^\sigma} p(F;F')\,p(F';G). \quad (5)$$

This identity can be generalized, giving us the *chain rule*:

Theorem 1. *If $F_1,\ldots,F_t$, and G are σ-flags such that $F_1,\ldots,F_t$ fit in G, then for every $1 \leq s \leq t$ and every n such that $F_1,\ldots,F_s$ fit in a σ-flag of size n and a σ-flag of size n together with $F_{s+1},\ldots,F_t$ fit in G, the identity*

$$p(F_1,\ldots,F_t;G) = \sum_{F \in \mathcal{F}_n^\sigma} p(F_1,\ldots,F_s;F)\,p(F,F_{s+1},\ldots,F_t;G)$$

holds.

Recall from the introduction that $p(\triangle;G^v) \to p(\bullet\!\!-\!\!\bullet;G^v)^2$ as $|G| \to \infty$. The argument to see this can be rephrased in two steps as follows. First, since G is triangle-free, then $p(\triangle;G^v) = p(\bullet\!\!-\!\!\bullet,\bullet\!\!-\!\!\bullet;G^v)$. This can be seen directly, but is also a consequence of the chain rule. Indeed, let $\mathcal{H} = \{\triangle\}$ and let $\bullet$ denote the only type of size 1. Then $\bullet$-flags $\bullet\!\!-\!\!\bullet$, $\bullet\!\!-\!\!\bullet$ fit in a $\bullet$-flag of size 3. Since $\mathcal{F}_3^\bullet = \{\triangle, \triangle, \triangle, \triangle, \triangle\}$, the chain rule gives

$$p(\bullet\!\!-\!\!\bullet,\bullet\!\!-\!\!\bullet;G) = \sum_{F' \in \mathcal{F}_3^\bullet} p(\bullet\!\!-\!\!\bullet,\bullet\!\!-\!\!\bullet;F')\,p(F';G) = p(\triangle;G). \quad (6)$$

Second, $p(\bullet\!\!-\!\!\bullet,\bullet\!\!-\!\!\bullet;G^v) \to p(\bullet\!\!-\!\!\bullet;G^v)^2$ as $|G| \to \infty$, that is, density exhibits multiplicative behavior in the limit:

Theorem 2. *If F_1, F_2 are fixed σ-flags, then there exists a function $f(n) = O(1/n)$ such that if F_1, F_2 fit in a σ-flag G, then*

$$|p(F_1,F_2;G) - p(F_1;G)\,p(F_2;G)| \leq f(|G|).$$

Identity (6), that comes from an application of the chain rule, suggests that there is a relation between the pair $(\bullet\!\!-\!\!\bullet, \bullet\!\!-\!\!\bullet)$ and $\triangle$. In the next section, we will use the chain rule to define a product operation on σ-flags, and under this product it will hold that $\bullet\!\!-\!\!\bullet \cdot \bullet\!\!-\!\!\bullet = \triangle$. This product will also commute with the density function in the limit: for σ-flags F_1 and F_2 we will have $p(F_1 \cdot F_2;G) \to p(F_1;G)\,p(F_2;G)$ as $|G| \to \infty$.

Flag algebras

In the introduction, we derived the constraint

$$\phi(\triangle) + 2\phi(\triangle) = 3\phi(\cdot\!-\!\cdot),$$

valid for every $\phi \in \Phi$. If we see $\phi \in [0,1]^{\mathcal{G}}$ as a vector, then this is a linear constraint on the components of ϕ.

To enable the use of tools from optimization, mainly duality, we need to embed our domain into a vector space. We do so by extending ϕ linearly to the space $\mathbb{R}\mathcal{G}$ of formal real linear combinations of graphs in $\mathcal{G}$. We could then rewrite the latter constraint as

$$\phi(\triangle + 2\triangle) = \phi(3\,\cdot\!-\!\cdot),$$

or even

$$\phi(\triangle + 2\triangle - 3\,\cdot\!-\!\cdot) = 0.$$

One of our main goals is to characterize the linear functionals on $\mathbb{R}\mathcal{G}$ that are limit functionals. Instead of describing all the constraints that characterize limit functionals, it is convenient to encode some of them algebraically, that is, by modifying the algebraic structure of $\mathbb{R}\mathcal{G}$. The resulting algebraic object will be the flag algebra, which we construct now for the more general case of σ-flags.

Let $\mathbb{R}\mathcal{F}^\sigma$ be the free vector space over the reals generated by all σ-flags, i.e., $\mathbb{R}\mathcal{F}^\sigma$ is the space of all formal real linear combinations of σ-flags. Let $(A_k)_{k\geq 0}$ be a convergent sequence in $\mathcal{F}^\sigma$ and let

$$\phi(F) = \lim_{k\to\infty} p(F; A_k)$$

be the pointwise limit of the functions $p(\cdot; A_k)$. Extend ϕ linearly to $\mathbb{R}\mathcal{F}^\sigma$, obtaining a linear functional. We say that ϕ is the *limit functional* of the convergent sequence $(A_k)_{k\geq 0}$ or, when the sequence itself is not relevant, that it is a *limit functional*.

For any limit functional ϕ, the chain rule in its form (5) implies that for every σ-flag F and $n \geq |F|$ we have

$$\phi(F) = \phi\left(\sum_{F' \in \mathcal{F}_n^\sigma} p(F; F')F'\right),$$

that is,

$$F - \sum_{F' \in \mathcal{F}_n^\sigma} p(F; F')F' \qquad (7)$$

is in the kernel of ϕ. Instead of enforcing these infinitely many relations, we might as well just quotient them out. So let $\mathcal{K}^\sigma$ be the linear span of vectors of form (7) and define $\mathcal{A}^\sigma = \mathbb{R}\mathcal{F}^\sigma/\mathcal{K}^\sigma$. This is a nontrivial vector space, since for every σ-flag F we have $p(\sigma; F) = 1$, and hence σ is itself not in $\mathcal{K}^\sigma$. Since $\mathcal{K}^\sigma$ is contained in the kernel of every limit functional, every limit functional is also a linear functional of $\mathcal{A}^\sigma$.

The main advantage of working with $\mathcal{A}^\sigma$ instead of $\mathbb{R}\mathcal{F}^\sigma$ is that it is possible to define a product on $\mathcal{A}^\sigma$, turning it into an algebra. This product will conveniently encode the asymptotic multiplicative behavior of densities described in Theorem 2: for every limit functional ϕ and $f, g \in \mathcal{A}^\sigma$we will have $\phi(f \cdot g) = \phi(f)\phi(g)$.

For σ-flags F and G, let n be any integer such that F, G fit in a σ-flag of size n and set

$$F \cdot G = \left(\sum_{H \in \mathcal{F}_n^\sigma} p(F, G; H)H\right) + \mathcal{K}^\sigma. \qquad (8)$$

This defines a function from $\mathcal{F}^\sigma \times \mathcal{F}^\sigma$ to $\mathcal{A}^\sigma$ and one may show that the definition is independent of the choice of n for each pair (F, G) of σ-flags. Now, extend this function bilinearly to $\mathbb{R}\mathcal{F}^\sigma \times \mathbb{R}\mathcal{F}^\sigma$. It is possible to prove that if $f \in \mathcal{K}^\sigma$ and $g \in \mathbb{R}\mathcal{F}^\sigma$, then $f \cdot g = \mathcal{K}^\sigma$, whence the bilinear extension is constant on cosets, and therefore defines a symmetric bilinear form on $\mathcal{A}^\sigma$, that is, a commutative product.

This turns $\mathcal{A}^\sigma$ into an algebra, the *flag algebra of type* σ. The product on $\mathcal{A}^\sigma$ is now defined, and we will use henceforth the natural correspondence $f \mapsto f + \mathcal{K}^\sigma$ between $\mathbb{R}\mathcal{F}^\sigma$ and $\mathcal{A}^\sigma$ without further notice, i.e., we will omit $\mathcal{K}^\sigma$ and write f instead of $f + \mathcal{K}^\sigma$ for an element of $\mathcal{A}^\sigma$. Sometimes, namely in the last section, it is important to work with explicit representatives of each coset; in such cases we will clearly distinguish between cosets and their representatives.

Under the product just defined for $\mathcal{A}^\sigma$, the type σ, taken as a σ-flag, is the identity element. The identity σ can be decomposed in many different ways using relations (7). Indeed, for any $n \geq |\sigma|$, we have

$$\sigma = \sum_{F \in \mathcal{F}_n^\sigma} p(\sigma; F)F = \sum_{F \in \mathcal{F}_n^\sigma} F.$$

It now follows from Theorem 2 that limit functionals are multiplicative, i.e.,

$$\phi(f \cdot g) = \phi(f) \cdot \phi(g)$$

for $f, g \in \mathcal{A}^\sigma$. Since by construction $\phi(\sigma) = 1$, every limit functional ϕ is an algebra homomorphism between $\mathcal{A}^\sigma$ and $\mathbb{R}$. We denote the set of all algebra homomorphisms between $\mathcal{A}^\sigma$ and $\mathbb{R}$ by $\mathrm{Hom}(\mathcal{A}^\sigma, \mathbb{R})$.

As an example, recall the discussion at the end of the previous section. When $\mathcal{H} = \{\triangle\}$, if we expand the product $\cdot\!-\!\cdot \cdot \cdot\!-\!\cdot$ as a linear combination of $\bullet$-flags of size 3, then $\cdot\!-\!\cdot \cdot \cdot\!-\!\cdot = \triangle$. Hence every limit functional ϕ satisfies $\phi(\triangle) = \phi(\cdot\!-\!\cdot \cdot \cdot\!-\!\cdot) = \phi(\cdot\!-\!\cdot)^2$.

Every limit functional ϕ lies in $\mathrm{Hom}(\mathcal{A}^\sigma, \mathbb{R})$. Another obvious constraint that every limit functional ϕ must satisfy is $\phi(F) \geq 0$ for every σ-flag F, which is not necessarily true of all homomorphisms. Call $\phi \in \mathrm{Hom}(\mathcal{A}^\sigma, \mathbb{R})$ *positive* if $\phi(F) \geq 0$ for every σ-flag F, and let $\mathrm{Hom}^+(\mathcal{A}^\sigma, \mathbb{R})$ denote the set of all positive homomorphisms.

It turns out that these are all the essential properties of a limit functional. It is clear that every limit functional is a positive homomorphism. The following theorem of Razborov [12] establishes the converse, and so positive homomorphisms are precisely the limit objects of convergent sequences of flags. In particular, the linear extension of the set Φ is precisely $\mathrm{Hom}^+(\mathcal{A}^\varnothing, \mathbb{R})$.

Theorem 3. *Every limit functional is a positive homomorphism and every positive homomorphism is a limit functional.*

Finally, notice that types and flags are defined in terms of the family $\mathcal{H}$ of forbidden subgraphs, so this family is encoded in the construction of the flag algebra $\mathcal{A}^\sigma$.

Downward operator

We are really interested in working with $\varnothing$-flags, that is, unlabeled graphs, so why consider other types altogether? Most times, in order to obtain results for $\varnothing$-flags, it is necessary to use other types. In the introduction, to obtain Mantel's theorem, it was not enough to work with unlabeled graphs: at some point, we had to introduce labeled graphs, namely to get (3).

The downward operator maps σ-flags into $\varnothing$-flags, in such a way that we can derive valid inequalities for densities of $\varnothing$-flags from valid inequalities for densities of σ-flags. If types can be seen as a form of lifting, then the downward operator is a projection back to our space of interest.

If F is a σ-flag, then $\downarrow F$ is the $\varnothing$-flag obtained from F simply by forgetting the em-

bedding, that is, by forgetting the vertex labels. For a σ-flag F, let $q_\sigma(F)$ be the probability that an injective map $\theta:[k]\to V(F)$ taken uniformly at random is such that $(\downarrow F,\theta)$ is a σ-flag isomorphic to F and set

$$[\![F_\sigma]\!]=q_\sigma(F)\downarrow F,$$

then extend $[\![\cdot]\!]_\sigma$ linearly to $\mathbb{R}\mathcal{F}^\sigma$ to obtain a linear map from $\mathbb{R}\mathcal{F}^\sigma$ to $\mathbb{R}\mathcal{F}^\varnothing$. One key property of this map is that $[\![\mathcal{K}^\sigma]\!]_\sigma\subseteq\mathcal{K}^\varnothing$, and hence $[\![\cdot]\!]_\sigma$ gives a linear map from $\mathcal{A}^\sigma$ to $\mathcal{A}^\varnothing$, which we call *downward operator*. The main tool used in the proof of this result is the following lemma, which relates densities in the labeled and in the unlabeled cases by taking an average.

Lemma 4. *Let F be a σ-flag and G be an $\varnothing$-flag with $|G|\geq|F|$ and $p(\downarrow\sigma;G)>0$. If θ is an embedding of σ into G chosen uniformly at random, then $p(F;(G,\theta))$* is a random variable and

$$\mathbb{E}\left[p(F;(G,\theta))\right]=\frac{q_\sigma(F)\,p(\downarrow F;G)}{q_\sigma(\sigma)\,p(\downarrow\sigma;G)}.$$

Note that equation (3) in the introduction follows trivially from this lemma. Indeed, take $\sigma=\bullet$ as the type of size 1 and let $F=$  . Then $\downarrow F=$, $q_\sigma(F)=\frac{1}{3}$, $q_\sigma(\sigma)=1$, and $p(\downarrow\sigma;G)=1$ for any graph G. Thus, by Lemma 4,

$$\begin{aligned}\frac{1}{|G|}\sum_{v\in V(G)}p(\text{};G^v)&=\mathbb{E}\left[p(F;(G,\theta))\right]\\&=\frac{q_\sigma(F)\,p(\downarrow F;G)}{q_\sigma(\sigma)\,p(\downarrow\sigma;G)}\\&=\frac{1}{3}p(\text{};G).\end{aligned}$$

Conic programming

For $f\in\mathcal{A}^\sigma$ and a linear functional ϕ in the dual space $(\mathcal{A}^\sigma)^*$ of $\mathcal{A}^\sigma$, write $(\phi,f)=\phi(f)$. The *semantic cone* of type σ is the set

$$\mathcal{S}^\sigma=\left\{f\in\mathcal{A}^\sigma:(\phi,f)\geq 0\text{ for all }\phi\in\mathrm{Hom}^+(\mathcal{A}^\sigma,\mathbb{R})\right\}.$$

This is a convex cone and its dual cone

$$(\mathcal{S}^\sigma)^*=\left\{\phi\in(\mathcal{A}^\sigma)^*:(\phi,f)\geq 0\text{ for all }f\in\mathcal{S}^\sigma\right\}$$

contains every nonnegative multiple of functionals in $\mathrm{Hom}^+(\mathcal{A}^\sigma,\mathbb{R})$. So, given a graph C,

$$\begin{aligned}&\max\left\{(\phi,C):\phi\in\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})\right\}\\&\leq\max\left\{(\phi,C):\phi\in(\mathcal{S}^\varnothing)^*\text{ and }(\phi,\varnothing)=1\right\}.\end{aligned}\quad(9)$$

(Here we may write '$\max$' instead of '$\sup$' because $\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})$ is compact. Actually, equality holds by the bipolar theorem.)

The optimization problem on the right-hand side above is a *conic programming problem*. It asks us to maximize a linear function $\phi\mapsto(\phi,C)$ over the intersection of a cone, namely $(\mathcal{S}^\varnothing)^*$, and an affine subspace, in our case determined by the linear equation $(\phi,\varnothing)=1$.

This conic programming problem has a *dual problem*, namely

$$\min\left\{\lambda:\lambda\varnothing-C\in\mathcal{S}^\varnothing\text{ and }\lambda\in\mathbb{R}\right\},\quad(10)$$

where the optimization variable is λ. (We may write '$\min$' instead of '$\inf$' because the feasible region is a closed half-line in $\mathbb{R}$.)

Weak duality holds: any feasible solution of the dual has larger or equal objective value than any feasible solution of the primal. Indeed, if $\phi\in(\mathcal{S}^\varnothing)^*$ is such that $(\phi,\varnothing)=1$ and $\lambda\in\mathbb{R}$ is such that $\lambda\varnothing-C\in\mathcal{S}^\varnothing$, then

$$0\leq(\phi,\lambda\varnothing-C)=\lambda-(\phi,C).$$

Actually, it is easy to show that there is no duality gap, that is, that primal and dual have the same optimal value. Even more: the problem on the left-hand side of (9) has the same optimal value of the dual problem (10), and so all three optimization problems in (9) and (10) have the same optimal value. Indeed, notice that the maximum on the left-hand side of (9) is equal to

$$\min\left\{\lambda:(\phi,C)\leq\lambda\text{ for all }\phi\in\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})\right\}.$$

Now, $\lambda\geq(\phi,C)$ for all $\phi\in\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})$ if and only if $(\phi,\lambda\varnothing-C)\geq 0$ for all $\phi\in\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})$ if and only if $\lambda\varnothing-C\in\mathcal{S}^\varnothing$, as we wanted.

To find an upper bound for $\mathrm{ex}(C,\mathcal{H})$ we work with the dual problem (10). One advantage is that we do not need to solve this problem to optimality to find an upper bound, since any feasible solution provides an upper bound. Solving (10) to optimality is the same as solving the primal problem to optimality, which is the same as computing $\mathrm{ex}(C,\mathcal{H})$.

One way to simplify the dual problem (10) is to replace $\mathcal{S}^\varnothing$ with a cone $\mathcal{C}\subseteq\mathcal{S}^\varnothing$ for which it is easier to solve the resulting problem. Obviously, we still get a valid upper bound. We seem to have taken a tortuous path since the introduction, where we stated our goal of finding a relaxation of Φ, of which $\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})$ is the linear extension, but that is exactly what we achieved, albeit via the dual:

$$\begin{aligned}\mathrm{Hom}^+(\mathcal{A}^\varnothing,\mathbb{R})\subseteq\big\{&\phi\in(\mathcal{A}^\varnothing)^*:(\phi,f)\geq 0\\&\text{for all }f\in\mathcal{C}\text{ and }(\phi,\varnothing)=1\big\}.\end{aligned}$$

What are some $f\in\mathcal{A}^\sigma$ that belong to the semantic cone $\mathcal{S}^\sigma$? Since a positive homomorphism ϕ is by definition nonnegative on every σ-flag F, then any conic combination of σ-flags is in the semantic cone. Another class of vectors in the semantic cone is the class of vectors that are sums of squares. We say that $f\in\mathcal{A}^\sigma$ is a *sum of squares* if there are $g_1,\ldots,g_t\in\mathcal{A}^\sigma$ such that $f=g_1^2+\cdots+g_t^2$. Then for any positive homomorphism ϕ (actually, for any homomorphism) we have $\phi(f)=\phi(g_1)^2+\cdots+\phi(g_t)^2\geq 0$. The class of sum-of-squares vectors is particularly interesting because it is computationally tractable, as we will soon see. Finally, the downward operator maps the semantic cone $\mathcal{S}^\sigma$of type σ into the semantic cone $\mathcal{S}^\varnothing$ of type $\varnothing$:

Theorem 5. *The image of $\mathcal{S}^\sigma$ under $[\![\cdot]\!]_\sigma$ is a subset of $\mathcal{S}^\varnothing$.*

This gives yet another way to obtain vectors in $\mathcal{S}^\varnothing$, by first considering a type σ, then obtaining a vector in $\mathcal{A}^\sigma$ (a sum-of-squares vector, for instance), and then using the downward operator.

The semidefinite programming method

Semidefinite programming is conic programming over the cone of positive semidefinite matrices. Using sum-of-squares vectors in $\mathcal{A}^\sigma$ and the downward operator, we may define a family of tractable cones contained in $\mathcal{S}^\varnothing$. Then using semidefinite programming it is possible to write down optimization problems that provide upper bounds to (10). This approach is known as the *semidefinite programming method*. Its main advantages are that writing down the semidefinite programming problems is mostly a mechanical affair, that can even be automated (and has been; see for instance flagmatic [5]), and solving the resulting problems can be done with a computer.

There is a well-known relation between sums-of-squares polynomials and positive semidefinite matrices (see e.g. the exposition by Laurent [9]). We now establish the analogous relation between sums-of-squares vectors in $\mathcal{A}^\sigma$ and positive semidefinite matrices. The *degree* of a vec-

tor $f \in \mathbb{R}\mathcal{F}^\sigma$ is the largest size of a flag appearing with a nonzero coefficient in the expansion of f; by convention, the degree of 0 is -1. The notion of degree can be extended to $\mathcal{A}^\sigma$, by setting the degree of $f + \mathcal{K}^\sigma \in \mathcal{A}^\sigma$ to be the smallest degree of any $g \in f + \mathcal{K}^\sigma$. For a type σ and $n \geq |\sigma|$, let $v_{\sigma,n} : \mathcal{F}_n^\sigma \to \mathcal{A}^\sigma$ be the canonical embedding, i.e., $v_{\sigma,n}(F) = F$ for all $F \in \mathcal{F}_n^\sigma$.

Theorem 6. *If $f \in \mathcal{A}^\sigma$ and $n \geq |\sigma|$, then there are vectors $g_1, \ldots, g_t \in \mathcal{A}^\sigma$ for some $t \geq 1$, each of degree at most n, such that $f = g_1^2 + \cdots + g_t^2$ if and only if there is a positive semidefinite matrix $Q : \mathcal{F}_n^\sigma \times \mathcal{F}_n^\sigma \to \mathbb{R}$ such that $f = v_{\sigma,n}^\top Q v_{\sigma,n}$.*

Proof. Suppose that there are vectors $g_1, \ldots, g_t$ as described. Modulo $\mathcal{K}^\sigma$, every σ-flag of size m can be written as a linear combination of σ-flags of any fixed size greater than m. So by hypothesis we can take from each coset $g_i + \mathcal{K}^\sigma$ a representative $\hat{g}_i \in \mathbb{R}\mathcal{F}^\sigma$ which is a linear combination of σ-flags of size n.

Let c_i be the vector of coefficients of $\hat{g}_i$, in such a way that $\hat{g}_i = c_i^\top v_{\sigma,n}$. Then

$$\begin{aligned}\hat{g}_1^2 + \cdots + \hat{g}_t^2 &= \sum_{i=1}^t \left(c_i^\top v_{\sigma,n}\right)^2 \\ &= \sum_{i=1}^t v_{\sigma,n}^\top c_i c_i^\top v_{\sigma,n},\end{aligned}$$

and we may take $Q = c_1 c_1^\top + \cdots + c_t c_t^\top$.

For the converse, say there is a positive semidefinite matrix Q as described. Then for some t there are vectors $c_1, \ldots, c_t$ such that $Q = c_1 c_1^\top + \cdots + c_t c_t^\top$. But then $g_i = c_i^\top v_{\sigma,n}$ has degree at most n in $\mathcal{A}^\sigma$. Moreover, $f = g_1^2 + \cdots + g_t^2$, as we wanted. □

Let us describe the semidefinite programming method by applying it to Mantel's theorem. Fix $\mathcal{H} = \{$  $\}$. We have the following $\varnothing$-flags of sizes 2 and 3:  ,  ,  , and  . There is also only one type of size 1, namely the graph on one vertex, which we denote by $\bullet$. These are the $\bullet$-flags of sizes 2 and 3:  ,  ,  , ,  , , and .

Write $v = v_{\bullet,2}$, so that in vector notation we have $v = ($  ,  $)$. From Theorem 6, if $Q : \mathcal{F}_2^\bullet \times \mathcal{F}_2^\bullet \to \mathbb{R}$ is a positive semidefinite matrix, then $v^\top Q v$ belongs to the semantic cone $\mathcal{S}^\bullet$ of type $\bullet$, and hence from Theorem 5 we have that $[\![v^\top Q v]\!]_\bullet$ belongs to the semantic cone $\mathcal{S}^\varnothing$ of type $\varnothing$. Since any conic combination r of $\varnothing$-flags belongs to the semantic cone $\mathcal{S}^\varnothing$, we have that

$$r + [\![v^\top Q v]\!]_\bullet \in \mathcal{S}^\varnothing$$

for every conic combination r of $\varnothing$-flags and every positive semidefinite matrix Q.

So, recalling (10), any feasible solution of the following optimization problem gives an upper bound to $\mathrm{ex}($$, \{$$\})$:

$$\begin{aligned}&\min \lambda \\ &\lambda \varnothing - \text{} = r + [\![v^\top Q v]\!]_\bullet, \\ &r \text{ is a conic combination of } \varnothing\text{-flags}, \\ &Q : \mathcal{F}_2^\bullet \times \mathcal{F}_2^\bullet \to \mathbb{R} \text{ is positive semidefinite.}\end{aligned} \tag{11}$$

This problem is not quite a semidefinite programming problem: the first identity above is an identity between vectors in $\mathcal{A}^\varnothing$, not a linear constraint on λ and the entries of Q. This identity can be translated, however, into several linear constraints, as follows.

If A and B are $n \times n$ matrices, write $\langle A, B \rangle = \operatorname{tr} A^\top B = \sum_{i,j=1}^n A_{ij} B_{ij}$. Then

$$[\![v^\top Q v]\!]_\bullet = [\![\langle v v^\top, Q \rangle]\!]_\bullet = \langle [\![v v^\top]\!]_\bullet, Q \rangle .$$

Here, notice that $vv^\top$ is a matrix. The downward operator, when applied to the matrix $vv^\top$, is applied entrywise and yields a matrix of the same dimensions as the result.

So the first constraint in (11) can be rewritten as

$$\lambda \varnothing - \text{} = r + \langle [\![v v^\top]\!]_\bullet, Q \rangle, \tag{12}$$

which is still an identity between elements of $\mathcal{A}^\varnothing$. To test the above identity, we may choose a large enough N and use the chain rule to expand both left and right-hand sides as linear combinations of $\varnothing$-flags of size N. If the coefficients coincide, then equality holds. This is only a sufficient condition however: for a fixed N, equality may hold in $\mathcal{A}^\varnothing$ even though the coefficients differ, but it is not hard to show that there is always some N for which equality holds if and only if the coefficients coincide.

To make things precise, we have to choose for , r, and every element of $\mathcal{A}^\varnothing$ in $vv^\top$ a representative in $\mathbb{R}\mathcal{F}^\varnothing$. As a representative of $\in \mathcal{A}^\varnothing$ we may choose $\in \mathbb{R}\mathcal{F}^\varnothing$. For $vv^\top$ proceed as follows: use the definition of product in $\mathcal{A}^\bullet$ to get

$$vv^\top = \begin{pmatrix} \text{} + \text{} & \frac{1}{2}(\text{} + \text{}) \\ \frac{1}{2}(\text{} + \text{}) & \text{} \end{pmatrix}$$

and then apply the downward operator to get

$$[\![vv^\top]\!]_\bullet = \begin{pmatrix} \text{} + \frac{1}{3}\text{} & \frac{1}{3}(\text{} + \text{}) \\ \frac{1}{3}(\text{} + \text{}) & \frac{1}{3}\text{} \end{pmatrix}.$$

We will deal with r below in a different way (actually, we will get rid of it). Notice we could have chosen different representatives. For instance, we could have expanded the products in $vv^\top$ using $\bullet$-flags of size 6, say. All that matters, however, is to choose representatives, and it is usually a good idea to choose representatives of smallest possible degree.

Now we are working exclusively with representatives in $\mathbb{R}\mathcal{F}^\varnothing$. For a given $N > 0$ and fixed $G \in \mathcal{F}_N^\varnothing$, extend $F \mapsto p(F; G)$ linearly to $\mathcal{F}_N^\varnothing$. If for every $G \in \mathcal{F}_N^\varnothing$ we have

$$\begin{aligned}&p(\lambda \varnothing - \text{}; G) \\ &\quad = p(r; G) + p(\langle [\![vv^\top]\!]_\bullet, Q \rangle; G),\end{aligned} \tag{13}$$

then (12) holds. Conversely, if (12) holds, then for some $N > 0$ (13) holds for every $G \in \mathcal{F}_N^\varnothing$ (this requires a short argument though).

Now, $p(r; G)$ is the coefficient of G in r; then, since r is a conic combination, $p(r; G) \geq 0$ for every $G \in \mathcal{F}_N^\varnothing$. Together with linearity this implies that we may rewrite (13) equivalently as

$$\lambda - p(\text{}; G) \geq \langle p([\![vv^\top]\!]_\bullet; G), Q \rangle, \tag{14}$$

where $p(\cdot; G)$ is applied entrywise to $vv^\top$. Notice that $p($$; G)$ is a number and $p([\![vv^\top]\!]_\bullet; G)$ is a matrix of numbers, so for each $G \in \mathcal{F}_N^\varnothing$ the above inequality is a linear constraint on λ and the entries of Q.

In our case, we may take $N = 3$. Then (14) gives rise to one linear constraint for each of the $\varnothing$-flags of size 3:

$\varnothing$-flag	constraint
	$\lambda \geq \left\langle \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, Q \right\rangle$,
	$\lambda - \frac{1}{3} \geq \left\langle \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 \end{pmatrix}, Q \right\rangle$,
	$\lambda - \frac{2}{3} \geq \left\langle \begin{pmatrix} 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix}, Q \right\rangle$.

In this way we may rewrite problem (11), obtaining a semidefinite programming problem that gives an upper bound to the optimal value of (11), and hence also to $\mathrm{ex}($$, \{$$\})$. This problem is not necessarily equivalent to (11), since for a given N equal-

ity in the algebra may hold even though the linear constraints are not satisfied.

Now, it is easy to check that $\lambda = \frac{1}{2}$ and $Q = \frac{1}{2}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ form a feasible solution of this semidefinite programming problem (and hence also of (11)), and so we have Mantel's theorem.

All the steps of the semidefinite programming method are contained in the example we worked out above. In general, however, one may choose a finite set $\mathcal{T}$ of types instead of only one type and consider the vectors in $\mathcal{S}^{\varnothing}$ given by

$$r + \sum_{\sigma \in \mathcal{T}} \left[\!\left[v_{\sigma, n_\sigma}^{\top} Q_\sigma v_{\sigma, n_\sigma} \right]\!\right]_\sigma,$$

where r is a conic combination of $\varnothing$-flags, $n_\sigma \geq 2|\sigma|$, and each Q_σ is a positive semidefinite matrix. Choosing more types makes the problem larger, but also potentially stronger.

Summary

The theory of flag algebras provides a powerful, unifying approach for extremal problems involving a host of combinatorial structures. Its novelty is that it allows the formulation of relaxations for such problems using conic programming, which can be further relaxed to semidefinite programming problems, thus enabling the use of a computer to obtain bounds. Most importantly, the computed bounds are often tight. Hence, the theory yields relaxations that achieve the desired trade-off of computational tractability and high-quality bounds.

We have only scratched the surface of the theory of flag algebras. Many optimization aspects of the semidefinite method, such as the use of complementary slackness to obtain further constraints on the optimal solutions for (9), were left out. Complementary slackness can be useful to show properties of all increasing sequences $(G_k)_{k \geq 0}$ that attain $\mathrm{ex}(C, \mathcal{H})$, an important issue in extremal combinatorics. Razborov [12] further developed other methods involving flag algebras, such as the differential method and the inductive method.

Techniques involving flag algebras have been used to obtain many significant new results such as: computing the minimal number of triangles in graphs with given edge density [11, 13], computing the maximum number of pentagons in triangle-free graphs [6, 8], and obtaining new advances towards the Cacceta-Häggkvist conjecture [14]. Besides being applied in the context of graphs and digraphs, flag algebras have also been successfully used in the setting of colored graphs [1, 4] and of permutations [2]. For many more references, see the thesis of Grzesik [7].

References

1 R. Baber and J. Talbot, A solution to the 2/3 conjecture, *SIAM Journal on Discrete Mathematics* 28 (2014), 756–766.

2 J. Balogh, P. Hu, B. Lidický, O. Pikhurko, B. Udvari and J. Volec, Minimum Number of Monotone Subsequences of Length 4 in Permutations, *Combinatorics, Probability and Computing* 24(4) (2015), 658–679.

3 J.A. Bondy, Counting subgraphs: a new approach to the Caccetta-Häggkvist conjecture, *Discrete Mathematics* 165/166 (1997), 71–80.

4 J. Cummings, D. Král', F. Pfender, K. Sperfeld, A. Treglown and M. Young, Monochromatic triangles in three-coloured graphs, *Journal of Combinatorial Theory, Series B* 103 (2013), 489–503.

5 V. Falgas-Ravry and E.R. Vaughan, Applications of the semi-definite method to the Turán density problem for 3-graphs, *Combinatorics, Probability and Computing* 22 (2013), 21–54.

6 A. Grzesik, On the maximum number of five-cycles in a triangle-free graph, *Journal of Combinatorial Theory, Series B* 102 (2012), 1061–1066.

7 A. Grzesik, *Flag Algebras in Extremal Graph Theory*, PhD Thesis, Jagiellonian University, 2014.

8 H. Hatami, J. Hladký, D. Král', S. Norine and A. Razborov, On the number of pentagons in triangle-free graphs, *Journal of Combinatorial Theory, Series A* 120 (2013), 722–732.

9 M. Laurent, Semidefinite programming in combinatorial and polynomial optimization, *Nieuw Archief voor Wiskunde* 5/9 (2008), 256–262.

10 W. Mantel, Vraagstuk XXVIII, *Wiskundige Opgaven* 10 (1910), 60–61.

11 O. Pikhurko and A. Razborov, Asymptotic structure of graphs with the minimum number of triangles, *Combinatorics, Probability & Computing*, First View (2016), 1–23.

12 A. Razborov, Flag algebras, *Journal of Symbolic Logic* 72 (2007), 1239–1282.

13 A. Razborov, On the minimal density of triangles in graphs, *Combinatorics, Probability & Computing* 17(4) (2008), 603–618.

14 A. Razborov, On the Caccetta-Häggkvist Conjecture with Forbidden Subgraphs, *Journal of Graph Theory* 74 (2013), 236–248.

William Kalies
Department of Mathematics
Florida Atlantic University, Boca Raton, USA
wkalies@fau.edu

Robert Vandervorst
Department of Mathematics
VU University, Amsterdam
r.c.a.m.vander.vorst@vu.nl

Computational Conley theory

In this article William Kalies and Robert Vandervorst describe Conley theory. Conley theory studies the dichotomy between gradient-like and non-gradient-like (recurrent) behavior of dynamical systems. It is a topological theory and one is compelled to ask the question whether this theory is computable. Numerical simulation has proved to be a common technique for analyzing dynamics, but does not give a complete picture of global behavior. By discretizing a dynamical system in time and space one can build a combinatorial dynamical system which carries dynamical information of the original system within a given resolution. The novelty of computational Conley theory is the combination of algebra, topology and combinatorics which produces computable tools that can be used to prove rigorous statements about global dynamic behavior.

Many evolutionary processes in physics, engineering, biology, et cetera are modeled by dynamical systems. Examples of dynamical systems range from very simple models such as the motion of a pendulum to extremely complex systems such as for instance meteorological models. Dynamical systems are mostly defined via systems of differential equations, but also as iterations of continuous maps. In practice these systems are too complex to study without any computational tools. A first insight into the workings of a dynamical system is often provided by numerical simulation in order to understand for example long term behavior of a system, which has been extremely successful in the application of dynamical systems as a prediction tool. Such analysis does not give much insight into global behavior in general, nor does it provide rigorous results about the dynamical structure of a system. Computational Conley theory makes various aspects of global dynamics computable using algebra, combinatorics and algebraic topology.

Going beyond local analysis such as linearized behavior of equilibrium points, or periodic orbits, is often very hard for even low dimensional systems let alone complex high dimensional systems. The development of the concepts of Conley theory into computable tools is of great benefit to the study of global dynamics. One of the key theorems in Conley theory is *Conley's decomposition theorem* which states that global dynamics can be categorized in (chain-)recurrent and gradient-like dynamics. For example consider the motion of a natural pendulum. As the pendulum swings, it cannot return to its original position with the same velocity because it loses energy due to friction — gradient-like dynamics, cf. Figure 1. In the absence of friction the system returns to its initial state, and its motion is periodic. In this case the system is said to exhibit recurrence. Understanding this dichotomy between (chain-)recurrent and nonrecurrent or gradient-like behavior is central to the study of dynamical systems and an important aspect of Conley theory. The theory was initially developed as a generalization of Morse theory for flows on smooth manifolds, see [4]. Conley theory also addresses the dependence of a system on parameters which is a very natural question to ask from the point of view of applications. Over the past thirty years this theory has been extended to cover most types of (deterministic) dynamical systems such as flows, discrete time systems and infinite dimensional dynamical systems, cf. [20–24].

The chain-recurrent and gradient-like dynamics of a given system is computable within a given resolution. We will explain

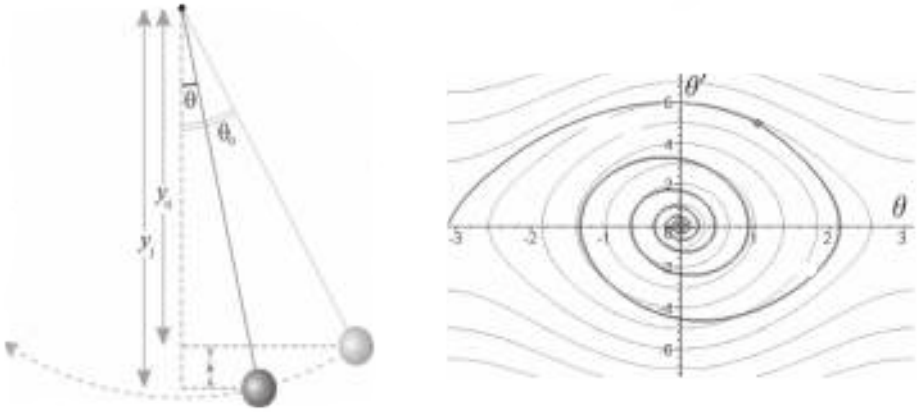


Figure 1 The phase portrait of a natural pendulum.

that these concepts can be alternatively defined so that they are more tangible for the purpose of rigorously computing them. This becomes a matter of understanding the key algebraic and combinatorial structures that play a role and their connection to the topology of the system. The new direction and novelty of this approach lies in the fact that the study of dynamical systems is reduced to the study of (finite) combinatorial dynamical systems in combination with (computable) algebraic structures and invariants. Algebraic topology takes continuum objects and reduces them to algebraic objects which are typically finitely generated and hence can be manipulated by the computer. The algebra carries well understood information about the continuum structure. The methods described here build on ideas that have emerged over last decade due to many authors in which combinatorialization and algebraic topology are the key ingredients. The first steps in this direction have proved very promising. We mention the early work of Eidenschink and Boczko, Kalies and Mischaikow [3, 10, 18], see also [7–9, 11, 12, 19, 24] and [14–17].

On a more philosophical note modern science is all about collecting and processing data. Most of the data comes from dynamical processes that are traditionally modeled in terms of continua, i.e. differential equations, manifolds, differentiable maps, et cetera. The process of collecting data is inherently a discretization. The ideas presented here may have major impact as science moves more and more towards data based representations of science and away from analytical model based representations of science. These developments make a computational Conley theory a valuable asset as a tool in applied dynamical systems. A successful computational theory will allow us to go beyond studying dynamical systems in low dimensions and will open the door to better understanding complex systems via rigorous statements.

Dynamical systems and algebraic structures

In mathematical terms a dynamical system consists of the following ingredients; a state space, or phase space X (metric space) and a family of maps φ_t, parametrized by a time parameter t, mapping X (on)to itself. For instance for the physical pendulum the state space consists of the rotation angles and is an example of a one-dimensional phase space. In more complicated systems such as biological or meteorological models the state space may be high dimensional. The time parameter t is either discrete, $t \in \mathbb{Z}^+$, or continuous, $t \in \mathbb{R}^+$. Dynamical systems describe deterministic processes; for an initial state x the state $\varphi_t(x)$ at time t is determined entirely by the initial state x. This is captured by the (semi-)group property for φ_t:

$$\begin{cases} \varphi_0(x) = x, \\ \varphi_{s+t}(x) = \varphi_s(\varphi_t(x)) \quad \text{for all } s,t \geq 0, \end{cases}$$

and for all $x \in X$. For many aspects of Conley theory it is not important whether the time variable is discrete or continuous. In order to keep exposition transparent we restrict to discrete time dynamical systems. If $t \in \mathbb{Z}^+$, then

$$\varphi_t = f^t = \underbrace{f \circ \cdots \circ f}_{t \text{ times}}, \quad f = \varphi_1,$$

where f is called the *generator*, or *generating map* and the discrete time dynamical system is denoted by (X, f). We emphasize that the generator $f = \varphi_1$ is not necessarily invertible! (For continuous time this allows for semi-flows.)

Robust structures

Consider the two-dimensional state space $X = [0,1] \times [0,1]$ with $x = (\xi, \eta) \in X$ and the mapping $f : X \to X$ given by

$$f(\xi, \eta) = \left(\frac{\xi}{2 - \xi}, \frac{\eta}{2 - \eta} \right), \tag{1}$$

Figure 2 depicts the dynamics of f. We will refer to as Example 1.

We first introduce terminology to describe asymptotic behavior of orbits. A *complete orbit* through x is denoted by γ_x and the forward and backward portions are denoted by γ_x^+ and γ_x^- respectively. The *omega-limit set* of a set $N \subset X$ is defined as

$$\omega(N) = \bigcap_{k \geq 0} \mathrm{cl}\left(\bigcup_{n \geq k} f^n(N) \right),$$

and the *alpha-limit set* is defined as

$$\alpha(N) = \bigcap_{k \leq 0} \mathrm{cl}\left(\bigcup_{n \leq k} f^n(N) \right).$$

The *orbital alpha-limit* and *orbital omega-limit* set are defined as $\alpha_o(\gamma_x^-) = \bigcap_{t \leq 0} \mathrm{cl}\, \gamma_x((-\infty, t])$ and $\omega(x) = \omega(\{x\})$ respectively.

The square $X = [0,1] \times [0,1]$ in Example 1 is invariant, i.e. f maps X onto X. To be more precise a subset $S \subset X$ is *invariant* for the dynamical system (X, f) if $f(S) = S$. In the above example many invariant sets can be found. For instance, the corner points, the boundary, et cetera. An invariant set $A \subset X$ is called an *attractor* if there exist a compact neighborhood $N \subset X$ containing A such that $f(N) \subset N$, $f^k(N) \subset \mathrm{int}(N)$ for some $k > 0$ and $A = \omega(N) = \mathrm{Inv}(N, f)$, where $\mathrm{Inv}(N, f)$ indicates the maximal invariant set inside N. The set $N \supset A$ is called a *trapping region* for A. The choices $N = \varnothing$ and $N = X$ are trivial trapping regions. Examples of attractors for are $A_1 = (0,0)$, $A_2 = \{(\xi, 0) \mid \xi \in [0,1]\}$, and $A_3 = \{(0, \eta) \mid \eta \in [0,1]\}$. Also the empty set and the state space X are attractors in Example 1. Dual to an attractor A is its *dual repeller* defined by

$$A^* = \{x \in X \mid \omega(x) \cap A = \varnothing\}.$$

For example $A_2^* = \{(\xi, 1) \mid \xi \in [0,1]\}$, $X^* = \varnothing$ and $\varnothing^* = X$. A dual repeller may also be defined as $A^* = \alpha(N^c)$ *where* N^c *is called a* repelling region. *A pair* (A, A^*) is called an *attractor-repeller pair* for (X, f) and is a robust structure. Attractor-repeller pairs give a very rough insight into the dynamics and they are not unique in general. Example 1 shows six different attractor-repeller pairs. Given an attractor-repeller pair (A, A^*) then for any $x \in X \backslash \{A \cup A^*\}$ it holds that $\alpha_o(\gamma_x^-) \in A^*$ and $\omega(x) \in A$ for any complete orbit γ_x through x. The idea of decomposing the dynamics into invariant sets and points which converge to theses sets can be further generalized and leads to the notion of Morse representation.

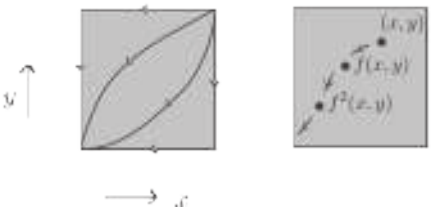


Figure 2 Complete orbits of f (left) and the iterates of f (right).

Figure 3 Three Morse representations. The finest representation consisting of equilibrium points (left), an attractor-repeller pair (middle), and a 3-set Morse representation (right).

Definition 1. A finite poset $(\mathsf{M}, \leq)$ consisting of non-empty, pairwise disjoint, compact invariant subsets of X is called a *Morse representation* for (X, f) if for every $x \in X \setminus (\bigcup_{M \in \mathsf{M}} M)$, there exist $M' < M''$, with $M', M'' \in \mathsf{M}$, such that

$$\omega(x) \subset M' \text{ and } \alpha_0(\gamma_x^-) \subset M'',$$

for any complete orbit γ_x through x.

Figure 4 The four non-trivial attractors in Example 1 (left), and its lattice structure (right).

(We use the terminology *poset* to indicate a partially ordered set.)

The sets $M \in \mathsf{M}$ are called *Morse sets*. In general Morse sets occur as intersections of attractors and repellers, i.e.

$$M = A \cap A'^*,$$

for some attractor A and repeller A'^* dual to some attractor A'. In the next subsection we will explain more of the systematics behind Morse representations.

Via a Morse representation $(\mathsf{M}, \leq)$ *we obtain a combinatorial description — a poset — of the global structure of a dynamical system within a given resolution.*

Morse representations can be further refined by finding attractor-repeller pair decompositions (or Morse representations) of Morse sets, see Figure 3 middle and right.

For some systems this process terminates after finitely many steps and a finest Morse representation exists as in the above example. If not, the intersection $\bigcap (A \cup A^*)$, ranging over all attractors in the system, leads the so-called chain-recurrent set $\mathsf{R} \subset X$. This set is a countable union of chain-connected components R_i which are defined as follows. Two points $x, x' \in \mathsf{R}$ lie in the same connected component R_i if for any attractor-repeller pair (A, A^*) either both $x, x' \in A$ or both $x, x' \in A^*$. Conley's decomposition theorem states that the dynamics outside R is gradient-like. The term chain-recurrence is inherited from its original definition via ε-chain recurrence. The latter states that the components R_i have the property that each point $x \in \mathsf{R}_i$ comes back to itself under the dynamics arbitrary close allowing small errors — ε-chains. The chain-recurrent set R is obtained as a limit. From the point of view of computation this definition is not very practical, cf. [14].

Lattices and Morse representations

The set of attractors (and repellers) in a dynamical system has the algebraic structure of a distributive lattice, cf. [15, 21]. A *distributive lattice* is a set L equipped with two binary operation $\vee$ and $\wedge$. The operations $\vee$ and $\wedge$ play the role of union and intersection and they satisfy the same axioms as union and intersection for subsets. A distributive lattice has a poset structure and $a \leq b$ if and only if $b = a \vee b$ (or equivalently $a = a \wedge b$). Let $\mathsf{Att}(X, f)$ be the set of all attractors of a dynamical system (X, f) and define the binary operations

$$A \vee A' = A \cup A', \quad A \wedge A' = \mathrm{Inv}(A \cap A', f).$$

With respect to these binary operations $\mathsf{Att}(X, f)$ is a bounded distributive lattice, see [15]. In general $\varnothing$ and $\omega(X)$ play the role of 0 and 1, respectively. (The term *bounded* means that $\inf(\mathsf{Att})$ and $\sup(\mathsf{Att})$ exist. As a matter of fact $\inf(\mathsf{Att}) = \varnothing$ and $\sup(\mathsf{Att}) = \omega(X)$.) The associate repellers form an (anti-)isomorphic lattice called $\mathsf{Rep}(X, f)$ with binary relations $\vee = \cup$ and $\wedge = \cap$. The trapping regions and repelling regions are also bounded distributive lattices with binary relations intersection and union and are denoted by $\mathsf{TrapR}(X, f)$ and $\mathsf{RepR}(x, f)$, respectively. Duality between the lattices is expressed in the following fundamental commuting diagram:

$$\begin{array}{ccc} \mathsf{TrapR}(X,f) & \overset{c}{\longleftrightarrow} & \mathsf{RepR}(X,f) \\ \downarrow \omega & & \downarrow \alpha \\ \mathsf{Att}(X,f) & \overset{*}{\longleftrightarrow} & \mathsf{Rep}(X,f) \end{array} \qquad (2)$$

where both ^{c} and * are involutions, and are the respective duality mappings. Figure 4 shows the attractors in Example 1 together with the lattice structure of $\mathsf{Att}(X, f)$. The lattice structure of $\mathsf{Att}(X, f)$ is intimately related to the poset structure of Morse representations. To explain this relation we introduce some elementary notions of poset theory.

Given a finite poset P then a *down-set* I consists of those elements $p \in \mathsf{P}$ such that $q \leq p$ implies that $q \in I$. Denote the set of all down-sets of P by $\mathsf{O}(\mathsf{P})$ which is a finite distributive lattice with respect to $\vee = \cup$ and $\wedge = \cap$. Beside the down-sets one can also define *up-sets* J which consists of those elements $p \in \mathsf{P}$ such that $q \geq p$ implies that $q \in J$. The up-sets form the distributive lattice $\mathsf{U}(\mathsf{P})$ dual to $\mathsf{O}(\mathsf{P})$ via set-complement. The *join-irreducible elements* in a finite distributive lattice L are the elements that are preceded by a unique predecessor, cf. Figure 5. A classical result by G. Birkhoff [2] provides a representation theorem for finite distributive lattices which is formulated as follows. Let L be a finite distributive lattice and let P be a finite poset. By $\mathsf{J}(\mathsf{L})$ we denote the poset of join-irreducible elements where the order is defined by set-inclusion. Then,

$$\mathsf{L} \cong \mathsf{O}(\mathsf{J}(\mathsf{L})), \quad \mathsf{P} \cong \mathsf{J}(\mathsf{O}(\mathsf{P})). \qquad (3)$$

Figure 5 gives a schematic account of Birkhoff's theorem in the case of Example 1. The operations O and J are contravariant functors and lattice monomorphisms $f : \mathsf{L} \rightarrowtail \mathsf{K}$ are equivalent to order-surjections $\mathsf{J}(f) : \mathsf{J}(\mathsf{K}) \twoheadrightarrow \mathsf{J}(\mathsf{L})$, and similarly lattice epimorphisms are equivalent to order-embeddings. The same correspondence can also be formulated via the O functor using Birkhoff's representation theorem, cf. [6].

Morse representations and finite sublattices of attractors in $\mathsf{Att}(X, f)$ are equivalent as a consequence of Birkhoff's representation theorem. This relation can be seen as follows, cf. Figure 5. Let $\mathsf{A} \subset \mathsf{Att}(X, f)$ be a finite sublattice. For every join-irreducible attractor $A \in \mathsf{J}(\mathsf{A})$ let $\overleftarrow{A} \prec A$ be the unique predecessor and define $M = A \cap \overleftarrow{A}^*$. It can be shown that the set $\mathsf{M}(\mathsf{A}) = \{A \cap \overleftarrow{A}^*\}$, with $A \in \mathsf{J}(\mathsf{A})$ and the order-relation induced by $\mathsf{J}(\mathsf{A})$, is a Morse representation and the sets $M = A \cap \overleftarrow{A}^*$ are Morse sets. In general $M = A_I \cap A_J^*$, where the representation only depends on the set-difference $I \setminus J = \{M\}$, with $I, J \in \mathsf{O}(\mathsf{M})$. For the lattice of attractors $\mathsf{A} = \mathsf{Att}(X, f)$ in Example 1 we have $\mathsf{J}(\mathsf{A}) = \{a, b, c, X\}$ and the associated poset $\mathsf{M}(\mathsf{A}) = \{M_1, \ldots, M_4\}$ is the finest Morse representation in Example 1 consisting of the rest points. The above procedure

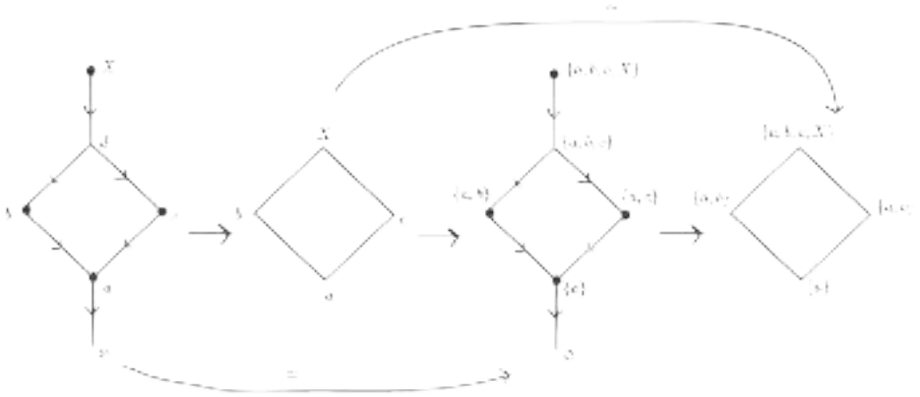


Figure 5 The join-irreducible elements form a poset $\mathsf{J}(\mathsf{A})$ and the down-sets of J form a lattice $\mathsf{O}(\mathsf{J}(\mathsf{A}))$ which is isomorphic to A.The solid dots in the diagrams denote the join-irreducible elements.

yields an isomorphism between $(\mathsf{M}(\mathsf{A}), \leq)$ and $(\mathsf{J}(\mathsf{A}), \subseteq)$. Conversely, by Birkhoff's theorem the lattice of down-sets in $(\mathsf{M}(\mathsf{A}), \leq)$ is isomorphic to A. The isomorphism is quantified as follows. If $I \in \mathsf{O}(\mathsf{M})$, then the sets $A_I = \bigcup_{M \in I} W^u(M)$ are attractors which build the lattice A. (The set $W^u(M)$ denotes the *unstable set* of M and is defined as: $W^u(M) = \{x \in X \mid \exists \gamma_x^- \ni \alpha_0(\gamma_x^-) \subset M\}$.)

Morse decompositions

Let $\mathsf{A} \subset \mathsf{Att}(X, f)$ be a finite sublattice (including $\varnothing$ and $\omega(X)$) and let $\mathsf{A}^* \subset \mathsf{Rep}(X, f)$ be the associated dual and let P be a finite poset such that $\mathsf{O}(\mathsf{P})$ maps onto A (as a lattice homomorphism). Consider the commuting diagram

$$\begin{array}{ccc} \mathsf{O}(\mathsf{P}) & \overset{c}{\longleftarrow\!\!\shortmid} & \mathsf{U}(\mathsf{P}) \\ \downarrow & & \downarrow \\ \mathsf{A} & \overset{*}{\longleftrightarrow} & \mathsf{A}^* \end{array} \tag{4}$$

Since Birkhoff's representation theorem is functorial the vertical lattice-surjections are dual to the order-injection $\mathsf{J}(\mathsf{A}) \rightarrowtail \mathsf{P}$. Using the Morse representation for $\mathsf{J}(\mathsf{A})$ as described in the previous section we obtain the order-injection $\mathsf{M}(\mathsf{A}) \rightarrowtail \mathsf{P}$, which is referred to as a *Morse decomposition*.

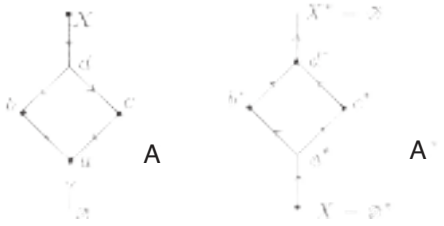


Figure 6 A lattice of attractors and the anti-isomorphic lattice of dual repellers.

The poset $(\mathsf{P}, \leq)$ may be regarded as a combinatorial model for the dynamics of (X, f) if we interpret P as directed graph, or *digraph*. The order-injection $\mathsf{M}(\mathsf{A}) \rightarrowtail \mathsf{P}$ signifies the dynamically relevant sets (Morse sets).

Consider an embedding $\mathsf{O}(\mathsf{P}) \rightarrowtail \mathsf{TrapR}(X, f)$ and denote the associated sublattices (images) by N. If we define $\omega(\mathsf{N}) = \mathsf{A} \subset \mathsf{Att}(X,f)$, then Diagram (2) yields

$$\begin{array}{ccc} \mathsf{N} & \overset{c}{\longleftarrow\!\!\shortmid} & \mathsf{N}^c \\ {\scriptstyle\omega}\downarrow & & \downarrow{\scriptstyle\alpha} \\ \mathsf{A} & \overset{*}{\longleftrightarrow} & \mathsf{A}^* \end{array} \qquad \mathsf{M}(\mathsf{A}) \rightarrowtail \mathsf{T}(\mathsf{N}), \tag{5}$$

where $\mathsf{T} = \mathsf{T}(\mathsf{N})$ is a representation of $\mathsf{J}(\mathsf{N}) \cong \mathsf{P}$ via the identification: $T_p = N_I \cap N_J^c = N_I \setminus N_J$ which depends only on $I \setminus J = \{p\}$ — compare the construction of a Morse representation. The poset $\mathsf{T}(\mathsf{N})$ is called a *Morse tiling* and the order-embedding $\mathsf{M}(\mathsf{A}) \rightarrowtail \mathsf{T}(\mathsf{N})$ is called a *tesselated Morse decomposition*.

By construction $\mathsf{P} \cong \mathsf{T}(\mathsf{N})$ and the elements of $\mathsf{T}(\mathsf{N})$, called *Morse tiles*, tesselate the space X. If the poset set structure of $\mathsf{T}(\mathsf{N})$ is regarded as a digraph, then a tesselated Morse decomposition is a first example of a combinatorialization of a dynamical system, i.e. the order relation indicates where points in X may map under the dynamics of f. The embedding $\mathsf{M}(\mathsf{A}) \rightarrowtail \mathsf{T}(\mathsf{N})$ marks the dynamically relevant tiles, i.e. the tiles that contain a Morse set.

The theory of Morse representations and decompositions as a combinatorial description of dynamics is a first motivation for the algebraic treatment of attractors via the theory of distributive lattices.

Combinatorial dynamics

The existence of a finite sublattice $\mathsf{N} \subset \mathsf{TrapR}(X, f)$ yields a tesselated Morse decomposition, cf. the previous subsection. If we start with a finite sublattice $\mathsf{A} \subset \mathsf{Att}(X, f)$ it is much harder to find a sublattice $\mathsf{N} \subset \mathsf{TrapR}(X, f)$ such that $\omega(\mathsf{N}) = \mathsf{A}$. The latter is called the lifting problem. In [15, Theorem 1.2] a lifting theorem of finite sublattices in $\mathsf{Att}(X, f)$ is proved: For every finite sublattice $\mathsf{A} \subset \mathsf{Att}(X, f)$ there exists a finite sublattice $\mathsf{N} = k(\mathsf{A}) \subset \mathsf{TrapR}(X, f)$, called a *lift*, such that $\omega(\mathsf{N}) = \mathsf{A}$. The lift $\mathsf{N} := k(\mathsf{A}) \subset \mathsf{TrapR}(X, f)$ provides the Morse tiling $\mathsf{T}(\mathsf{N}) \cong \mathsf{J}(\mathsf{A})$ as described in the previous section. Since finding Morse representations is one of the objectives we cannot a priori use them in order to combinatorialize dynamics. In the next subsection we describe an explicit procedure to combinatorialize a dynamical system.

Discretization of space

Let $\mathcal{X}$ be a finite 'labeling set' with labels $\xi \in \mathcal{X}$ and $|\xi| \subset X$ are subsets of X satisfying the following axioms:

1. $X = \bigcup_{\xi \in \mathcal{X}} |\xi|$;
2. $|\xi| = \mathrm{cl}(\mathrm{int}(|\xi|))$ for all $\xi \in \mathcal{X}$;
3. $|\xi| \cap \mathrm{int}(|\eta|) = \varnothing$ for all $\xi \neq \eta \in \mathcal{X}$.

The set $\{|\xi|\}_{\xi \in \mathcal{X}}$ is called a *grid* on X and the elements $|\xi|$ are called *grid-elements*. The labeling set is also referred to as a grid in X. If we consider the Boolean algebra of *regular closed sets* on X, denoted $\mathsf{RG}(X)$, then choosing a grid on X is equivalent to selecting a finite subalgebra in $\mathsf{RG}(X)$. The latter is similar to the relation between attractor lattices and Morse representations. The subalgebra is the equivalent of the sublattice and the join-irreducible elements in the subalgebra are called the *atoms*. As with Morse tiles, the grid elements are constructed in the same way. The only difference is that $\mathcal{X}$ is a trivial poset by construction, i.e. no order relation, cf. [16].

In order to mimic the dynamics of (X, f) we define a multivalued mapping $\mathcal{F} : \mathcal{X} \rightrightarrows \mathcal{X}$ on $\mathcal{X}$. The multivalued mapping $\mathcal{F} : \mathcal{X} \rightrightarrows \mathcal{X}$ is linked to the dynamics of f via the condition

$$f(|\xi|) \subset \mathrm{int}\,|\mathcal{F}(\xi)|, \tag{6}$$

for all $\xi \in \mathcal{X}$, cf. [24]. Such multivalued mappings are called *outer approximations* for (X, f). The grid $\mathcal{X}$ together with a mul-

tivalued mapping $\mathcal{F}$ is called a *combinatorial dynamical system*; notation $(\mathcal{X},\mathcal{F})$. If we regard the elements in $\mathcal{X}$ as vertices and the action of $\mathcal{F}$ by edges, then $(\mathcal{X},\mathcal{F})$ may be regarded as a digraph; $\xi_i \in \mathcal{X}$ are vertices and e_{ij} is an edge from ξ_i to ξ_j if $\xi_j \in \mathcal{F}(\xi_i)$.

Combinatorial dynamical systems

Multivalued mappings regarded as combinatorial dynamical systems share many similarities with dynamical systems (X, f). Except for multivaluedness most of the dynamical concepts carry over to combinatorial systems. A subset of vertices $\mathcal{S} \subset \mathcal{X}$ is called *invariant* if $\mathcal{S} \subset \mathcal{F}(\mathcal{S})$ and $\mathcal{S} \subset \mathcal{F}^{-1}(\mathcal{S})$. A subset $\mathcal{A} \subset \mathcal{X}$ is called an *attractor* if $\mathcal{F}(\mathcal{A}) = \mathcal{A}$, and similarly a subset $\mathcal{R} \subset \mathcal{X}$ is called a *repeller* if $\mathcal{F}^{-1}(\mathcal{R}) = \mathcal{R}$. A *dual repeller* to an attractor $\mathcal{A}$ is defined as $\mathcal{A}^* = \boldsymbol{\alpha}(\mathcal{X}\backslash\mathcal{A})$. (Recall the definition of alpha and omega limit set form $(\mathcal{X},\mathcal{F})$: $\boldsymbol{\omega}(\mathcal{U}) = \bigcap_{k\geq 0}\bigcup_{n\geq k}\mathcal{F}^n(\mathcal{U})$ and $\boldsymbol{\alpha}(\mathcal{U}) = \bigcap_{k\leq 0}\bigcup_{n\leq k}\mathcal{F}^n(\mathcal{U})$, cf. [18].) As before attractors and repeller form (finite) distributive lattices denoted as $\mathsf{Att}(\mathcal{X},\mathcal{F})$ and $\mathsf{Rep}(\mathcal{X},\mathcal{F})$, respectively. A subset $\mathcal{M} \subset \mathcal{X}$ is called a Morse set if $\mathcal{M} = \mathcal{A} \cap \mathcal{R}$ for some attractor $\mathcal{A}$ and some repeller $\mathcal{R}$. Even though attractors and repellers are not necessarily invariant, Morse sets are. A poset $(\mathsf{M}, \leq)$ consisting of non-empty, pairwise disjoint, invariant subsets $\mathcal{M} \subset \mathcal{X}$ is called a *Morse representation* for $(\mathcal{X},\mathcal{F})$ if for every complete orbit $\{\xi_n\}$ such that $\xi_{n'} \in \mathcal{M}'$ and $\xi_{n''} \in \mathcal{M}''$ for some $n' < n''$, then $\mathcal{M}' < \mathcal{M}''$. The theory of Morse representations and distributive lattices of attractors remains unchanged for combinatorial systems. The combinatorial theory can best be summarized as a generalized version of Birkhoff's theorem for finite digraphs, cf. [13, 14].

As before attractors and repellers may be defined via larger sets in the digraph. Forward invariant sets $\mathcal{U}$ satisfy the property $\mathcal{F}(\mathcal{U}) \subset \mathcal{U}$ and form the lattice $\mathsf{Invset}^+(\mathcal{X},\mathcal{F})$. Similarly, backward invariant sets form the lattice $\mathsf{Invset}^-(\mathcal{X},\mathcal{F})$. This yields the analogue of Diagram (2):

$$\begin{array}{ccc} \mathsf{Invset}^+(\mathcal{X},\mathcal{F}) & \overset{c}{\longleftrightarrow} & \mathsf{Invset}^-(\mathcal{X},\mathcal{F}) \\ \Big\downarrow{\scriptstyle \omega} & & \Big\downarrow{\scriptstyle \alpha} \\ \mathsf{Att}(\mathcal{X},\mathcal{F}) & \overset{*}{\longleftrightarrow} & \mathsf{Rep}(\mathcal{X},\mathcal{F}) \end{array} \qquad (7)$$

Applying Birkhoff's representation theorem yields representations for $\mathsf{J}(\mathsf{Invset}^+(\mathcal{X},\mathcal{F}))$ and $\mathsf{J}(\mathsf{Att}(\mathcal{X},\mathcal{F}))$ given by the posets $\mathsf{SC}(\mathcal{X},\mathcal{F})$ and $\mathsf{RC}(\mathcal{X},\mathcal{F})$ respectively. The elements in $\mathsf{SC}(\mathcal{X},\mathcal{F})$ are obtained by intersecting forward and backward invariant sets as described before, and the elements in $\mathsf{RC}(\mathcal{X},\mathcal{F})$ are Morse sets in $\mathcal{F}$ which are obtained by intersecting attractors and repellers. In the language of digraph theory the poset $\mathsf{SC}(\mathcal{X},\mathcal{F})$ is the (acyclic) digraph of strongly connected components and the poset $\mathsf{RC}(\mathcal{X},\mathcal{F})$ is the subgraph of cyclic strongly connected components. (A digraph is *strongly connected* if there exists a path between each ordered pair of vertices. A *strongly connected component* of a digraph is maximal strongly connected subgraph. The *cyclic strongly connected components*, or *recurrent components* are the strongly connected components that either consist of a single vertex with a self-connection, or consist of multiple points.) From the functoriality we have the canonical order inclusion $\mathsf{RC}(\mathcal{X},\mathcal{F}) \rightarrowtail \mathsf{SC}(\mathcal{X},\mathcal{F})$. The acyclic strongly connected components in $\mathsf{SC}(\mathcal{X},\mathcal{F}) \setminus \mathsf{RC}(\mathcal{X},\mathcal{F})$ consists of one grid element ξ and for such elements $\mathrm{Inv}(|\,\xi\,|, f) = \varnothing$. For the components in $\mathsf{RC}(\mathcal{X},\mathcal{F})$ this is not necessarily true and the maximal invariant inside such a component may not be empty. Tarjan's algorithm provides an efficient linear time algorithms in the number of vertices that computes the strongly connected components of a digraph and thus a finest Morse representation, see cf. [5]. The idea of discretizing space and combinatorializing dynamics is best explained by an example.

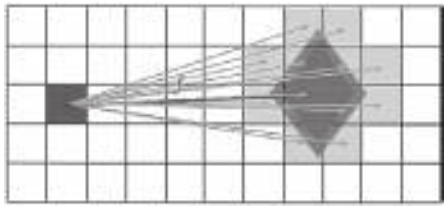


Figure 7 An example of a grid and a multivalued map $\mathcal{F}$; for any $\xi \in \mathcal{X}$, $\mathcal{F}(|\,\xi\,|) = \{\eta \in \mathcal{X} \mid B_\varepsilon(f(|\,\xi\,|)) \cap |\,\eta\,| \neq \varnothing\}$, cf. [1].

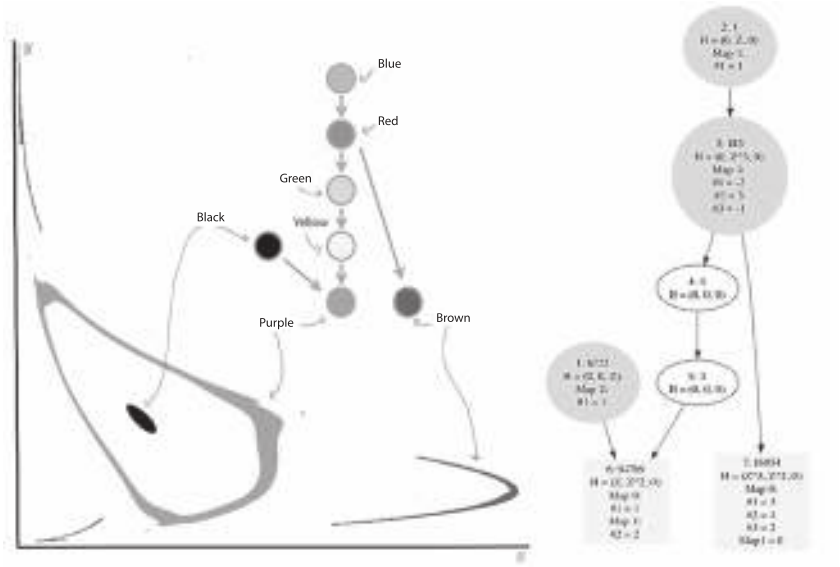


Figure 8 The strongly connected components $\mathsf{SC}(\mathcal{X},\mathcal{F})$ for $\mathcal{F}$ realized in the plane (left), the reduced graph $\mathsf{RC}(\mathcal{X},\mathcal{F})$ with cyclic strongly connected components for $\mathcal{F}$ (middle) and the Conley index computations (right), cf. [1].

The following model is used in population dynamics to describe the interaction of two species x and y. The dynamics is given by the mapping

$$f(x,y;\mu_1,\ldots,\mu_4) = \begin{pmatrix} (\mu_1+\mu_2)\,e^{-\mu_3(x+y)} \\ \mu_4 x \end{pmatrix}, \qquad (8)$$

where $\mu_1,\ldots,\mu_4$ are positive parameters in the model, see [25, 26]. Choose the parameters to be $\mu_1 = 22.5$, $\mu_2 = 25$, $\mu_3 = 0.1$ and $\mu_4 = 0.7$. Observe that the square $X = \{(x,y) \mid 0 \leq x \leq a, 0 \leq y \leq b\}$, with $a \geq \mu_1+\mu_2$ and $b \geq \mu_4(\mu_1+\mu_2)$, is forward invariant with respect to f and therefore consider $f : X \to X$.

We choose a rectangular grid on X with elements of size $\mathrm{diam}(|\,\xi\,|) \leq d$ and construct a multivalued mapping $\mathcal{F}$ that maps grid elements to sets of grid elements, cf Figure 7. In Example 9 we choose

$$\mathcal{F}(\xi) = \{\eta \in \mathcal{X} \mid B_\varepsilon(f(|\,\xi\,|)) \cap \eta \neq \varnothing\}, \qquad (9)$$

factoring in a numerical error ε, cf. Figures 7 and 8, cf. [1].

Computability and interpretation

The dynamics captured by an outer approximation $\mathcal{F}$ has a direct interpretation to the actual dynamics of (X, f). As we indicated before the multivalued mapping $\mathcal{F}$ may be interpreted as a digraph. Figure 8 gives a realization of the strongly connected components of the digraph, color coded (left). By identifying the cyclic strongly connected components with vertices, a schematic picture of the dynamics emerges via $\mathsf{RC}(\mathcal{X}, \mathcal{F})$ in terms of the digraph Figure 8 (middle).

Realization

Define the *evaluation mapping* $|\cdot| : \mathsf{Set}(\mathcal{X}) \to \mathsf{RG}(X)$ via $\mathcal{U} \mapsto \bigcup_{\xi \in \mathcal{U}} |\xi|$. A trapping region N is called an *attracting block* if $f(N) \subset \operatorname{int}(N)$. Attracting blocks that are also regular and closed form the lattice denoted by $\mathsf{ABlock}_{\mathsf{RG}}(X, f) \subset \mathsf{RG}(X)$. If $\mathcal{U} \in \mathsf{Invset}^+(\mathcal{X}, \mathcal{F})$ is a lattice homomorphism. As a matter of fact the following commuting diagram exists, cf. [16]:

$$\begin{array}{ccc} \mathsf{Invset}^+(\mathcal{X}, \mathcal{F}) & \xrightarrow{|\cdot|} & \mathsf{ABlock}_{\mathsf{RG}}(X, f) \\ \downarrow{\scriptstyle \omega} & & \downarrow{\scriptstyle \omega} \\ \mathsf{Att}(\mathcal{X}, \mathcal{F}) & \xrightarrow{\omega(|\cdot|)} & \mathsf{Att}(X, f) \end{array} \tag{10}$$

The dual diagram for backward invariant sets, repelling blocks and repellers is obtained via the duality mapping

$$^{\#} : \mathsf{ABlock}_{\mathsf{RG}}(X, f) \longleftrightarrow \mathsf{RBlock}_{\mathsf{RG}}(X, f),$$

where $N^{\#} = \operatorname{cl}(N^c)$, the 'complement' in $\mathsf{RG}(X)$. Let $\mathsf{N}_{\mathcal{F}}$ and $\mathsf{A}_{\mathcal{F}}$ be the images of $\mathsf{Invset}^+(\mathcal{X}, \mathcal{F})$ and $\mathsf{Att}(\mathcal{X}, \mathcal{F})$ under the mappings $|\cdot|$ and $\omega(|\cdot|)$, respectively. This yields the commuting diagram

$$\begin{array}{ccc} \mathsf{Invset}^+(\mathcal{X}, \mathcal{F}) & \xleftrightarrow{|\cdot|} & \mathsf{N}_{\mathcal{F}} \\ \downarrow{\scriptstyle \omega} & & \downarrow{\scriptstyle \omega} \\ \mathsf{Att}(\mathcal{X}, \mathcal{F}) & \xrightarrow{\omega(|\cdot|)} & \mathsf{A}_{\mathcal{F}} \end{array} \tag{11}$$

By Birkhoff duality and functoriality we obtain a commuting diagram linking the strongly connected components in $\mathcal{F}$ to tesselated Morse decompositions:

$$\begin{array}{ccc} \mathsf{SC}(\mathcal{X}, \mathcal{F}) & \xleftrightarrow{|\cdot|} & \mathsf{T}(\mathsf{N}_{\mathcal{F}}) \\ \uparrow & & \uparrow \\ \mathsf{RC}(\mathcal{X}, \mathcal{F}) & \xleftarrow{\pi} & \mathsf{M}(\mathsf{A}_{\mathcal{F}}) \end{array} \tag{12}$$

We conclude that $\mathsf{RC}(\mathcal{X}, \mathcal{F})$ yields a tesselated Morse decomposition

$$\begin{aligned} \pi : \mathsf{M}(\mathsf{A}_{\mathcal{F}}) &\mapsto \mathsf{RC}(\mathcal{X}, \mathcal{F}) \\ &\mapsto \mathsf{SC}(\mathcal{X}, \mathcal{F}) \longleftrightarrow \mathsf{T}(\mathsf{N}_{\mathcal{F}}). \end{aligned}$$

The mapping $U \mapsto \operatorname{Inv}(U, f)$ acts as a left inverse. In particular, $\mathsf{M}(\mathsf{A}_{\mathcal{F}}) = \{\operatorname{Inv}(|\mathcal{M}|, f) \neq \varnothing \mid \mathcal{M} \in \mathsf{RC}(\mathcal{X}, \mathcal{F})\}$ and $\mathsf{T}(\mathsf{N}_{\mathcal{F}}) = \{|\mathcal{U}| \mid \mathcal{U} \in \mathsf{SC}(\mathcal{X}, \mathcal{F})\}$. From the theory of Morse decompositions as described before we have that the composition $\pi : \mathsf{M}(\mathsf{A}_{\mathcal{F}}) \to \mathsf{T}(\mathsf{N}_{\mathcal{F}})$ is a Morse decomposition.

What does this approach tell us about the example given by (8)? The middle graph $\mathsf{RC}(\mathcal{X}, \mathcal{F})$ in Figure 8 gives a candidate for a Morse representation. However, it is not straightforward to determine whether the evaluated Morse sets $|\mathcal{M}|$ contain actual Morse sets for f. A sophisticated topological tool called the Conley index provides sufficient conditions which imply that $\operatorname{Inv}(|\mathcal{M}|, f) \neq \varnothing$, cf. Figure 8 (right). By construction a Morse set $\mathcal{M}$ is given by $\mathcal{M} = \mathcal{U} \setminus \overleftarrow{\mathcal{U}}$, where $\mathcal{U} \in \mathsf{J}(\mathsf{Att}(\mathcal{X}, \mathcal{F}))$ and $\overleftarrow{\mathcal{U}}$ is the unique predecessor. The version of the Conley index used in Figure 8 is the set of all non-zero eigenvalues of

$$\widehat{f}_* : H_*(N/\overleftarrow{N}, [\overleftarrow{N}]) \to H_*(N/\overleftarrow{N}, [\overleftarrow{N}]),$$

restricted to the torsion free part of $H_*(N/\overleftarrow{N}, [\overleftarrow{N}])$, where $N = |\mathcal{U}|$, $\overleftarrow{N} = |\overleftarrow{\mathcal{U}}|$, and

$$\widehat{f}([x]) = \begin{cases} [f(x)] & \text{if } x, f(x) \in N \setminus \overleftarrow{N}, \\ [\overleftarrow{N}] & \text{otherwise,} \end{cases}$$

is continuous by construction. The above arguments provide a second reason why the algebra of distributive lattices plays a crucial role in determining non-trivial Morse sets.

If we employ the Conley index we conclude for the model given in that there exist at least two attractors, two non-trivial (saddle-like) Morse sets and one repeller. The green and yellow vertices in Figure 8 (middle) may contain no invariant dynamics. The poset $\mathsf{RC}(\mathcal{X}, \mathcal{F})$ combinatorializes the dynamics between the Morse sets. The draw back is that the elements $|\mathcal{M}|$, $\mathcal{M} \in \mathsf{RC}(\mathcal{X}, \mathcal{F})$, do not tile the space X in general. The elements in $\mathsf{SC}(\mathcal{X}, \mathcal{F})$ do, but the number of elements $\{\xi\} \in \mathsf{SC}(\mathcal{X}, \mathcal{F}) \setminus \mathsf{RC}(\mathcal{X}, \mathcal{F})$ is many orders of magnitude larger than $\mathsf{RC}(\mathcal{X}, \mathcal{F})$. If we can construct a sublattice $\mathsf{U} \subset \mathsf{Invset}^+(\mathcal{X}, \mathcal{F})$ which is isomorphic to $\mathsf{Att}(\mathcal{X}, \mathcal{F})$ then the sets $|\mathcal{U} \setminus \overleftarrow{\mathcal{U}}|$, with $\mathcal{U} \in \mathsf{J}(\mathsf{U})$, tile the space X. This task is much harder and is related to the question of realizability and convergence: can every structure be realized provided the grid is sufficiently fine? This question is addressed in the next subsection.

Convergence

By refining a grid — subdividing, or choosing a smaller grid-size — more detailed information about the dynamics of (X, f) may be obtained. Keeping in mind the example in let $(\mathcal{X}_n, \mathcal{F}_n)$ be a sequence of refinements by subdividing the grid elements and by choosing the sequence of multivalued mappings accordingly, cf. (9). In [16] it was proved that for a refining sequence of outer approximations $\mathcal{F}_n : \mathcal{X}_n \rightrightarrows \mathcal{X}_n$ and any finite sublattice $\mathsf{A} \subset \mathsf{Att}(X, f)$, there exists an $n_{\mathsf{A}} > 0$ such that A lifts into $\mathsf{Invset}^+(\mathcal{X}_n, \mathcal{F}_n)$ for all $n \geq n_{\mathsf{A}}$, i.e. there exist embeddings $\mathsf{A} \mapsto \mathsf{Invset}^+(\mathcal{X}_n, \mathcal{F}_n)$ such that $\omega(|\mathsf{U}|) = \mathsf{A}$, where $\mathsf{U} \subset \mathsf{Invset}^+(\mathcal{X}_n, \mathcal{F}_n)$ is the isomorphic image of A. This convergence result has consequences for the existence of Morse tilings for given a Morse representation. Let $\mathcal{F} = \mathcal{F}_0$ and consider $\mathsf{A}_{\mathcal{F}}$ and the associated Morse representation $\mathsf{M}(\mathsf{A}_{\mathcal{F}})$. Let U_n be the image of $\mathsf{A}_{\mathcal{F}} \mapsto \mathsf{Invset}(\mathcal{X}_n, \mathcal{F}_n)$, $n \geq n_{\mathsf{A}_{\mathcal{F}}}$. Then, the isomorphism

$$\omega : \mathsf{N}_n = |\mathsf{U}_n| \longleftrightarrow \mathsf{A}_{\mathcal{F}},$$

yields a tesselated Morse decomposition $\mathsf{M}(\mathsf{A}_{\mathcal{F}}) \longleftrightarrow \mathsf{T}(\mathsf{N}_n)$. From the point of view of computation this result does not provide practical ways to find Morse tilings and tesselated Morse decompositions. A more practical method may be derived as follows.

The Morse decomposition $\pi : \mathsf{M}(\mathsf{A}_{\mathcal{F}}) \mapsto \mapsto \mathsf{RC}(\mathcal{X}, \mathcal{F})$ is the starting point, which yields the tesselated Morse decomposition $\mathsf{M}(\mathsf{A}_{\mathcal{F}}) \mapsto \mathsf{T}(\mathsf{N}_{\mathcal{F}}) \cong \mathsf{SC}(\mathcal{X}, \mathcal{F})$. The objective is to coarsen the poset $\mathsf{SC}(\mathcal{X}, \mathcal{F})$ such we obtain a Morse tiling T which is isomorphic to $\mathsf{RC}(\mathcal{X}, \mathcal{F})$. From Birkhoff duality it follows that the existence of a lift k in Diagram (13),

$$\begin{array}{ccc} \mathsf{O}(\mathsf{SC}(\mathcal{X}, \mathcal{F})) & \longleftrightarrow & \mathsf{Invset}^+(\mathcal{X}, \mathcal{F}) \\ \downarrow & \overset{k}{\nearrow} & \downarrow{\scriptstyle \omega} \\ \mathsf{O}(\mathsf{RC}(\mathcal{X}, \mathcal{F})) & \longleftrightarrow & \mathsf{Att}(\mathcal{X}, \mathcal{F}) \end{array} \tag{13}$$

is equivalent to the existence of an order-

surjection

$$j : \mathsf{SC}(\mathcal{X}, \mathcal{F}) \twoheadrightarrow \mathsf{RC}(\mathcal{X}, \mathcal{F}), \qquad (14)$$

with $j \circ i = \mathrm{id}$ on $\mathsf{RC}(\mathcal{X}, \mathcal{F})$. If the grid size is not small enough then such an order-surjection j need not exist! However, convergence helps out in this case. Let $\mathcal{X}_n$ be a sequence of refining grids with $\mathrm{diam}(\mathcal{X}_n) \to 0$ and let $\mathcal{F}_n : \mathcal{X}_n \rightrightarrows \mathcal{X}_n$ be a sequence of outer approximations as given in with $\mathcal{F} = \mathcal{F}_0$. Then, there exists an $n_{\mathsf{A}_\mathcal{F}} > 0$ such that $\mathsf{RC}(\mathcal{X}, \mathcal{F})$ allows order-surjections $j_n : \mathsf{SC}(\mathcal{X}_n, \mathcal{F}_n) \twoheadrightarrow \mathsf{RC}(\mathcal{X}, \mathcal{F})$ for all $n \geq n_{\mathsf{A}_\mathcal{F}}$, cf. [13]. The mappings j_n provide the rules for coarsening $\mathsf{SC}(\mathcal{X}_n, \mathcal{F}_n)$ and we obtain tesselated Morse decompositions

$$\pi : \mathsf{M}(\mathsf{A}_\mathcal{F}) \twoheadrightarrow \mathsf{RC}(\mathcal{X}, \mathcal{F}) \hookrightarrow \mathsf{T}(\mathsf{N}_n), \qquad (15)$$

where $\mathsf{N}_n = |\, k_n(\mathsf{RC}(\mathcal{X}, \mathcal{F}))\,|$ and

$$k_n : \mathsf{O}(\mathsf{RC}(\mathcal{X}, \mathcal{F})) \twoheadrightarrow \mathsf{Invset}(\mathcal{X}_n, \mathcal{F}_n).$$

We do not expatiate on methods of computing j_n. Given j_n, for n sufficiently large, provides a complete combinatorial picture of the dynamics of (X, f) within a given resolution: every point $x \in X$ is contained in a tile $T \in \mathsf{T}(\mathsf{N}_n)$ and the order relation on $\mathsf{T}(\mathsf{N}_n)$ provides the possible Morse sets to which the dynamics converges as $t \to \pm\infty$. The graph $\mathsf{RC}(\mathcal{X}, \mathcal{F})$ in Figure 8 (left) now provides a complete description of the dynamics within the chosen resolution since in the tesselated Morse decomposition $\mathsf{M}(\mathsf{A}_\mathcal{F}) \twoheadrightarrow \mathsf{T}(\mathsf{N}_n)$, the poset $\mathsf{T}(\mathsf{N}_n)$ is a Morse tiling. The above arguments provide a third reason why the algebraic approach provides the appropriate framework for computational combinatorializations of the dynamics.

The relation between parameter dependence in systems and combinatorial systems is still unclear. Since the way of encoding dynamical information into a combinatorial system is insensitive under small perturbations the question arises how the algebraic structures behave under large parameter variations — continuations. A good understanding of the algebra involved may give algebraic criteria for detecting certain types of global bifurcations. The ideas in [1] about studying dynamical systems with a number of parameters could be comprised in a database with information on for instance Morse decompositions for a finite number of parameter values. With the appropriate algebra at hand various properties of the system can be found in combination with variation in parameters. We believe that the answer should come from the algebra of finite distributive lattices and their morphisms. This requires a formulation of local continuation of Morse representations in terms sheafs of Morse representations and attractor lattices.

References

1 Zin Arai, William Kalies, Hiroshi Kokubu, Konstantin Mischaikow, Hiroe Oka and Paweł Pilarczyk, A database schema for the analysis of global dynamics of multiparameter systems, *SIAM Journal on Applied Dynamical Systems* 8(3) (2009), 757–789.

2 Garrett Birkhoff, *Lattice Theory*, American Mathematical Society Colloquium Publications, Vol. 25, American Mathematical Society, third edition, 1979.

3 Erik Boczko, William D. Kalies and Konstantin Mischaikow, Polygonal approximation of flows, *Topology Appl.* 154(13) (2007), 2501–2520.

4 Charles Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series in Mathematics, Vol. 38, American Mathematical Society, 1978.

5 Thomas H. Cormen, Charles E. Leiserson and Ronald L. Rivest, *Introduction to Algorithms*, The MIT Electrical Engineering and Computer Science Series, MIT Press, 1990.

6 B.A. Davey and H.A. Priestley, *Introduction to Lattices and Order*, Cambridge University Press, second edition, 2002.

7 S. Day, *A Rigorous Numerical Method in Infinite Dimensions*, Ph.D. dissertation, 2003.

12 Michael Dellnitz, Gary Froyland and Oliver Junge, The algorithms behind GAIO-set oriented numerical methods for dynamical systems, in *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, Springer, 2001, pp. 145–174, 805–807.

11 M. Dellnitz and O. Junge, Sofware package GAIO, http://www-math.uni-paderborn.de/206agdellnitz/gaio, 1998.

10 M. Eidenschink, *Exploring Global Dynamics: A Numerical Algorithm Based on Conley Index Theory*, Ph.D. dissertation, 1995.

11 Tomasz Kaczynski, Konstantin Mischaikow and Marian Mrozek, *Computational Homology*, Applied Mathematical Sciences, Vol. 157, Springer, 2004.

12 William D. Kalies and H. Ban, A computational approach to conley's decomposition theorem, *Journal of Computational and Nonlinear Dynamics* 1(1) (2006), 312–319.

13 W.D. Kalies, D. Kasti and R.C.A.M. Vandervorst, Binary relations and Birkhoff's representation theorem, Preprint, 2014.

14 W.D. Kalies, K. Mischaikow and R.C.A.M. Vandervorst, An algorithmic approach to chain recurrence, *Found. Comput. Math.* 5(4) (2005), 409–449.

15 W.D. Kalies, K. Mischaikow and R.C.A.M. Vandervorst, Lattice structures for attractors I, *J. Comput. Dyn.* 1(2) (2014), 307–338.

16 W.D. Kalies, K Mischaikow and R.C.A.M. Vandervorst, Lattice structures for attractors II, *Found. Comput. Math.* 1(2) (2015), 1–41.

17 W.D Kalies, K. Mischaikow and R.C.A.M. Vandervorst, Lattice structures for attractors III, Preprint, 2016.

18 Konstantin Mischaikow, Topological techniques for efficient rigorous computation in dynamics, *Acta Numer.* 11 (2002), 435–477.

19 Marian Mrozek, An algorithm approach to the Conley index theory, *J. Dynam. Differential Equations* 11(4) (1999), 711–734.

20 Joel W. Robbin and Dietmar Salamon, Dynamical systems, shape theory and the Conley index, *Ergodic Theory Dynam. Systems* 8 (Charles Conley Memorial Issue) (1988), 375–393.

21 Joel W. Robbin and Dietmar A. Salamon, Lyapunov maps, simplicial complexes and the Stone functor, *Ergodic Theory Dynam. Systems* 12(1) (1992), 153–183.

22 Krzysztof P. Rybakowski, *The Homotopy Index and Partial Differential Equations*. Universitext, Springer, 1987.

23 Dietmar Salamon, Connected simple systems and the Conley index of isolated invariant sets, *Trans. Amer. Math. Soc.* 291(1) (1985), 1–41.

24 Andrzej Szymczak, A combinatorial procedure for finding isolating neighbourhoods and index pairs, *Proc. Roy. Soc. Edinburgh Sect. A* 127(5) (1997), 1075–1088.

25 Ilie Ugarcovici and Howard Weiss, Chaotic attractors and physical measures for some density dependent Leslie population models, *Nonlinearity* 20(12) (2007), 2897–2906.

26 Ilie Ugarcovici and Howard Weiss, Chaotic dynamics of a nonlinear density dependent population model, *Nonlinearity* 17(5) (2004), 1689–1711.

Jan Bouwe van den Berg
Afdeling Wiskunde
Vrije Universiteit Amsterdam
janbouwe.vanden.berg@vu.nl

Chris Groothedde
Afdeling Wiskunde
Vrije Universiteit Amsterdam
c.m.groothedde@vu.nl

Ray Sheombarsing
Afdeling Wiskunde
Vrije Universiteit Amsterdam
r.s.s.sheombarsing@vu.nl

Simulatie + Contractie = Bewijs

In hoeverre is het verantwoord om de resultaten van een simulatie van een probleem als benadering van een oplossing van het probleem te interpreteren? Een van de manieren om deze vraag te benaderen is door middel van zogenaamde computer-assisted (door de computer ondersteunde) bewijsmethoden. In dit artikel presenteren Jan Bouwe van den Berg, Chris Groothedde en Ray Sheombarsing een methode die, met als uitgangspunt een zorgvuldig geproduceerde numerieke simulatie, de mogelijkheid biedt om het probleem op een wiskundig exacte manier op te lossen met behulp van een computer.

Sinds het Manhattan project in de Tweede Wereldoorlog, maar vooral vanaf de opkomst van de op transistors gebaseerde computers in de jaren zestig zijn computersimulaties niet meer weg te denken uit de wiskunde. In de toegepaste wiskunde (en ver daarbuiten) vormen simulaties vaak het eindstadium van een oplossingsstrategie voor een gemodelleerd probleem. Daarbij bestaat het model bijvoorbeeld uit een groot stelsel gekoppelde (differentiaal) vergelijkingen of een niet-lineaire partiële differentiaalvergelijking, zoals de Navier–Stokes-vergelijking voor vloeistofstromingen. In de zuivere wiskunde dienen simulaties vooral om de wiskundige te helpen haar probleem te visualiseren en om inspiratie op te doen.

Een cruciale vraag die bij het gebruik van computersimulaties vrijwel altijd om de hoek komt kijken is hoe betrouwbaar deze resultaten zijn. Computer-assisted bewijzen zijn bij uitstek geschikt om die betrouwbaarheid onomstotelijk vast te stellen.

De problemen die wij met onze methode zullen behandelen zijn vrijwel allemaal afkomstig uit het vakgebied van de niet-lineaire dynamica, zoals het vinden van oplossingen van differentiaalvergelijkingen of banen van iteratieve afbeeldingen. In veel situaties kan het al lonen om slechts één of enkele speciale oplossingen te vinden, zoals periodieke oplossingen of verbindende banen. Door middel van zogenaamde *forcing theorems* kan het bestaan van dit soort speciale oplossingen gebruikt worden om stellingen te bewijzen die uitspraken doen over globale eigenschappen van het systeem, zoals de aanwezigheid van chaos of het bestaan van speciale families van oplossingen. Een beroemd voorbeeld is de stelling "periode drie impliceert chaos" voor iteraties in één dimensie. We komen daar later op terug.

De problemen in de dynamische systemen zijn eigenlijk zonder uitzondering niet-lineair. Kwalitatieve, grove beschrijvingen van robuste eigenschappen van dit soort systemen kunnen vaak gegeven worden met behulp van topologische en variationele methoden. Bovendien kan in sommige gevallen een 'kleine' parameter uitgebuit worden om oplossingen in limietgevallen via asymptotische analyses te beschrijven. Maar door het niet-lineaire karakter van de problemen kunnen bewijzen van stellingen die kwantitatieve uitspraken doen over dynamische systemen voor 'gewone' parameterwaardes in vrijwel geen enkel geval via analytische weg gevonden worden. Daarom wordt bij bestudering van dit soort problemen, die zowel in de wiskunde zelf als in toepassing erg vaak voorkomen, gewoonlijk teruggevallen op numerieke simulaties. Tegenwoordig kunnen zulke simulaties, dankzij hoogstaande algoritmes, software en hardware vaak eenvoudig en snel ontwikkeld en 'gedraaid' worden. Echter, deze simulaties leveren in de meeste gevallen geen daadwerkelijk bewijs voor het bestaan van een oplossing of een garantie dat een numerieke methode een oplossing goed benadert.

Aan de basis van de methode die we in dit artikel beschrijven ligt het idee dat dit soort numerieke simulaties, ondanks deze beperkingen, toch gebruikt kunnen worden om een wiskundig exact bewijs te leveren. Om precies te zijn zullen wij laten zien dat in de buurt van een numeriek gesimuleerde 'oplossing' (dat wil zeggen, een uitkomst van een numerieke berekening waarvan we hopen/verwachten dat het een goede benadering van een oplossing is) een unieke wiskundig exacte oplossing bestaat. Sterker nog, onze methode vertelt ons expliciet hoe dicht de ware oplossing bij de gesimuleerde benadering ligt en hoe groot de omgeving is waarin deze uniek is.

De allereerste stap van de methode is om het probleem, of het nu afkomstig is uit bijvoorbeeld een stelsel gewone differentiaalvergelijkingen of bestaat uit een partiële differentiaalvergelijking met randvoorwaarden, te herformuleren als

$$F(x) = 0, \tag{1}$$

waarbij F een afbeelding is tussen, laten we zeggen, twee vectorruimten. Die vectorruimten zullen in het algemeen *oneindig* dimensionaal zijn, en dan kiezen we doorgaans Banachruimten, maar in eenvoudige gevallen kunnen ze ook eindig-dimensio-

naal zijn. Het is bijzonder lastig om met een computer gelijkheden zoals in (1) te controleren. In plaats daarvan zullen we daarom gebruik maken van een dekpuntstelling. Dit doen we door het probleem $F(x)=0$ om te schrijven naar een dekpuntprobleem door middel van een variant van de methode van Newton. In de methode die we hier beschouwen, vervult de Banach contractiestelling de rol van dekpuntstelling, maar er bestaan ook aanpakken die werken met de dekpuntstellingen van Brouwer en Schauder, zie bijvoorbeeld [5].

We zullen vervolgens laten zien dat op een goed gekozen omgeving van de numeriek gesimuleerde oplossing (benadering) aan de voorwaarden van de dekpuntstelling van Banach is voldaan. Dan garandeert deze stelling namelijk het bestaan en de uniciteit van een (echte) oplossing. Via pen-en-papier-wiskunde zijn de voorwaarden van de Banach-contractiestelling te formuleren als een *eindig aantal ongelijkheden*, waarin de numerieke simulatie en analytische eigenschappen van het probleem verenigd worden. Om te laten zien dat er aan deze ongelijkheden voldaan wordt, zal er gewerkt moeten worden met de data die eerder door de computer geproduceerd zijn. Omdat gesimuleerde resultaten honderden of zelfs duizenden datapunten bevatten, is het praktisch om deze ongelijkheden met een computer te verifiëren. Het is cruciaal dat *ongelijkheden* (die contractie garanderen) met een computer veel makkelijker te controleren zijn dan gelijkheden zoals (1). Niettemin betekent het dat de computer op een *rigoureuze* manier (zonder afrondfouten, of althans met rigoureuze controle van de afrondfouten) met de data moet omgaan.

Al sinds de jaren tachtig worden dit soort *computer-assisted proofs* toegepast binnen het gebied van dynamische systemen, zowel voor discrete problemen (zoals iteraties van afbeeldingen) als voor continue problemen (zoals gewone en partiële differentiaalvergelijkingen). Een bekend voorbeeld is het eerste bewijs van het Feigenbaum-vermoeden over de universaliteit van de Feigenbaum-constante [8]. Ook binnen de continue dynamische systemen werd in de jaren negentig al gebruik gemaakt van *computer-assisted proofs*, hetgeen uitmondde in het bewijs van het bestaan van de chaotische attractor in het beroemde Lorenz-systeem [12, 13].

Net als bij de methode die we in dit artikel beschrijven berusten de bewijzen van bovengenoemde historische stellingen op het verifiëren van ongelijkheden. Om dit met een computer te kunnen doen wordt gebruik gemaakt van interval-aritmetiek. Interval-aritmetiek is een methode waarmee computersoftware op een geautomatiseerde manier afrondfouten kan bijhouden, zodat ongelijkheden op een exacte manier gecontroleerd kunnen worden. Het belang van afrondfouten moet wel enigszins gerelativeerd worden. In de praktijk vallen ze in het niet in vergelijking met de analytische schattingen die vooral bij oneindig-dimensionale problemen de crux (en het struikelblok) voor het oplossen van het probleem zijn. Niettemin is het essentieel dat naast het deel van het bewijs dat door een brein van vlees en bloed is geconstrueerd, ook het deel van het bewijs dat aan een brein van silicium is uitbesteed foutloos is. Daarom besteden we in de volgende paragraaf kort aandacht aan de vraag hoe we door toepassing van interval-aritmetiek de computer kunnen inschakelen om ongelijkheden wiskundig exact te controleren.

De laatste jaren groeit de interesse in *computer-assisted proofs* in de dynamische systemen snel. Dit komt vooral doordat de ontwikkelingen in algoritmen, software en hardware ervoor gezorgd hebben dat we niet langer een supercomputer en veel geduld nodig hebben om de berekeningen en bewijzen uit te voeren, maar met een laptop al een heel eind komen. Hierdoor kunnen nieuwe ideeën veel makkelijker uitgeprobeerd worden, en gaan de ontwikkelingen in het vakgebied overeenkomstig sneller. Aan het eind van dit artikel geven we enkele voorbeelden van problemen waar tegenwoordig de aandacht naar uitgaat (zie de laatste paragraaf).

Interval-aritmetiek

Een van de aspecten van een *computer-assisted proof* is de uitdaging dat computers niet zomaar kunnen werken met de reële getallen, maar zich moeten beperken tot een eindig aantal (binaire representeerbare) getallen. Deze eindige verzameling getallen, die we noteren als $\mathbb{F}\subset\mathbb{R}$, worden de *floating-point* getallen of *floats* genoemd.

Het grootste nadeel van deze getallen is dat er afrondfouten optreden wanneer er bewerkingen op worden uitgevoerd (zoals optelling, vermenigvuldiging, et cetera). Dit betekent dat, zelfs in eindig veel dimensies, simulaties en numerieke berekeningen op basis van *floating-point* getallen afwijkingen met zich mee brengen. Om toch op een rigoureuze manier om te gaan met *floats* wordt gebruikt gemaakt van zogenaamde *interval-aritmetiek*. Interval-aritmetiek is een manier om, binnen een eindige verzameling van getallen $\mathbb{F}$ (zoals de *floats*), door stelselmatig naar boven én naar beneden af te ronden toch op mathematisch gegarandeerde resultaten uit te komen.

Als voorbeeld bekijken we de operatie vermenigvuldiging $*:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$. Doorgaans zal het niet zo zijn dat uit $a,b\in\mathbb{F}$ volgt dat ook $a*b\in\mathbb{F}$ geldt. In plaats van de gewone vermenigvuldiging gebruiken processoren dus een andere, numerieke, vermenigvuldiging $\circledast:\mathbb{F}\times\mathbb{F}\to\mathbb{F}$. Deze numerieke vermenigvuldiging wordt door computerchips zo uitgevoerd dat deze de ware vermenigvuldiging benadert: $|a*b-a\circledast b|<\epsilon$, waarbij ϵ de 'machine precision' wordt genoemd; gewoonlijk geldt dat $\epsilon\approx 10^{-16}$. Met andere woorden, het getal $a\circledast b$ is een afronding van het getal $a*b$. Het is echter a priori niet bekend of dit getal naar boven of naar beneden wordt afgerond, waardoor er nog meer informatie verloren gaat.

Door systematisch naar boven én naar beneden af te ronden, kunnen we dit soort verlies van informatie tegengaan. In plaats van te rekenen met losse *floats*, zullen we rekenen met intervallen met eindpunten in $\mathbb{F}$:

$$\mathcal{I}_{\bar{\mathbb{F}}} \stackrel{\text{def}}{=} \{[a,b]\subset\bar{\mathbb{R}} : a,b\in\bar{\mathbb{F}}\}.$$

Hierin is $\bar{\mathbb{R}}\stackrel{\text{def}}{=}\mathbb{R}\cup\{-\infty,+\infty\}$ de uitgebreide reële lijn en zijn $\bar{\mathbb{F}}\stackrel{\text{def}}{=}\mathbb{F}\cup\{-\infty,+\infty\}$ de uitgebreide *floating-point* getallen. Voor elke $x\in\mathbb{R}$ kunnen we dus een interval $[a,b]\in\mathcal{I}_{\bar{\mathbb{F}}}$ vinden zodat $x\in[a,b]$.

Het idee is nu om een vermenigvuldiging $\circledast:\mathcal{I}_{\bar{\mathbb{F}}}\times\mathcal{I}_{\bar{\mathbb{F}}}\to\mathcal{I}_{\bar{\mathbb{F}}}$ tussen intervallen te definiëren, zodanig dat voor elke I_1 en I_2 in $\mathcal{I}_{\bar{\mathbb{F}}}$ geldt dat

$$\{x_1*x_2\in\bar{\mathbb{R}} : x_1\in I_1 \text{ en } x_2\in I_2\} \subset I_1\circledast I_2\in\mathcal{I}_{\bar{\mathbb{F}}}.$$

In de praktijk kan de ondergrens van $I_1\circledast I_2$ gevonden worden door al de uiteinden van I_1 en I_2 met elkaar te vermenigvuldigen, deze naar beneden af te ronden (zodat we in $\bar{\mathbb{F}}$ belanden) en vervolgens het kleinste resultaat te kiezen. Analoog

kan de bovengrens gevonden worden door naar boven af te ronden en het grootste resultaat te kiezen.

Op soortgelijke wijze kunnen we de andere elementaire operaties $\oplus, \ominus, \circledcirc$: $\mathcal{I}_{\overline{\mathbb{F}}} \times \mathcal{I}_{\overline{\mathbb{F}}} \to \mathcal{I}_{\overline{\mathbb{F}}}$ definiëren. Het op deze wijze rekenen met intervallen noemen we interval aritmetiek. Minder elementaire functies zoals $\exp$, $\log$, $\sin$, et cetera, kunnen met behulp van bijvoorbeeld de stelling van Taylor zo geïmplementeerd worden dat ze aan de cruciale inclusie-eigenschap

$$\{f(x) : x \in I\} \subset f(I) \in \mathcal{I}_{\overline{\mathbb{F}}} \quad \text{voor alle } I \in \mathcal{I}_{\overline{\mathbb{F}}} \tag{2}$$

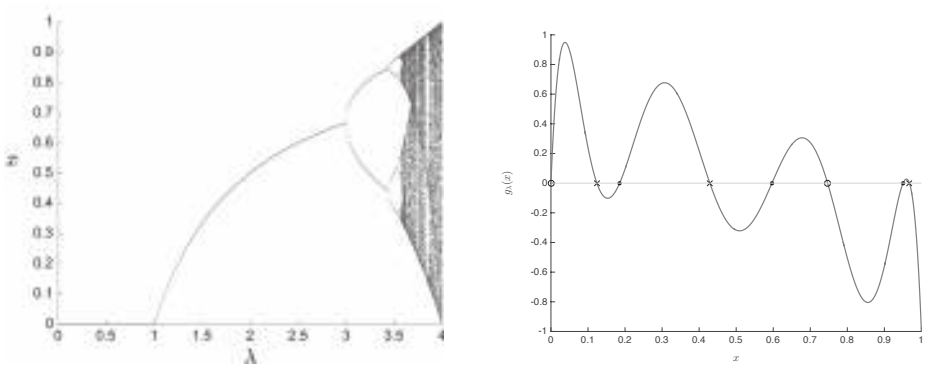
voldoen. Het is goed om te beseffen dat operaties in interval aritmetiek niet uniek gedefinieerd zijn, maar dat dit geen probleem is. Voor verschillende implementaties van interval aritmetiek kan $f(I)$ een verschillend resultaat opleveren, maar de inclusie (2) *moet altijd gegarandeerd zijn*. Op die manier kunnen we interval aritmetiek dus gebruiken om ongelijkheden te bewijzen.

Voor het gebruik van interval aritmetiek zijn verscheidene software-pakketten in omloop. Veel gebruikt zijn bijvoorbeeld de Intlab library voor Matlab en de Profil/BIAS library voor C++. Daarnaast bieden ook symbolische software pakketten als Maple en Mathematica ondersteuning voor interval aritmetiek.

Voorbeeld: de logistische afbeelding

Als voorbeeld bekijken we de in Figuur 1(a) gerepresenteerde logistische afbeelding $f_\lambda : [0,1] \to [0,1]$, gegeven door $f_\lambda(x) \stackrel{\text{def}}{=} \lambda x(1-x)$ met $\lambda \in [0,4]$ een parameter. De dynamica $x_{k+1} = f_\lambda(x_k)$ gegenereerd door f_λ is relatief eenvoudig te beschrijven voor $\lambda \in [0, \lambda_c]$, waarbij $\lambda_c \approx 3{,}57$. Namelijk, alle banen convergeren naar een evenwichtstoestand of een periodieke baan. Voor $\lambda > \lambda_c$ is de dynamica niet meer eenvoudig te beschrijven: het systeem vertoont chaotisch gedrag. Maar hoe bewijs je dit?

Zoals besproken in de introductie bestaan er stellingen, *forcing theorems*, die het bestaan van speciale banen gebruiken om uitspraken te doen over het globaal gedrag van een dynamisch systeem. Een beroemd voorbeeld van een *forcing theorem* voor discrete dynamische systemen op de reële lijn is dat het bestaan van een periodieke baan van periode drie chaos impliceert [11]. Met andere woorden, we kunnen voor een



(a) (b)

Figuur 1 (a) Het beroemde bifurcatiediagram van Feigenbaum. Het asymptotisch gedrag van de logistische afbeelding is in kaart gebracht (te zien op de verticale as) voor $\lambda \in [0,4]$ met behulp van de computer. De numerieke simulatie doet vermoeden dat de dynamica chaotisch is voor $\lambda > \lambda_c \approx 3{,}57$. (b) De grafiek en nulpunten van g_λ voor $\lambda = 3{,}95$. De De open cirkels zijn de twee evenwichtspunten van f_λ. De kruisjes en zwarte punten corresponderen met twee verschillende (niet triviale) periodieke banen van periode drie.

specifieke $\lambda > \lambda_c$ aantonen dat de dynamica van de logistische afbeelding chaotisch is door te bewijzen dat er zo'n periodieke baan van periode drie bestaat. Het is in het algemeen lastig (of zelfs onmogelijk) om een periodieke baan van een niet-lineair probleem met een pen-en-papier-analyse te vinden, en dit is typisch een situatie waarin numerieke simulaties handig zijn en *computer-assisted* bewijzen goed tot hun recht komen.

Een voor de hand liggende strategie om een periodieke baan van periode drie te vinden is het oplossen van de vergelijking $g_\lambda(x) = 0$, waarbij

$$g_\lambda(x) \stackrel{\text{def}}{=} f_\lambda(f_\lambda(f_\lambda(x))) - x,$$

zie Figuur 1(b). Om een eenvoudig voorbeeld van een *computer-assisted* bewijs te geven merken we op dat het probleem $g_\lambda(x) = 0$ kan worden opgelost met behulp van interval-aritmetiek én de tussenwaardestelling: door aan te tonen dat $g_\lambda(x_1) \in I_1$ en $g_\lambda(x_2) \in I_2$ voor zekere $x_1 < x_2$, waarbij $I_1 \subset [0,\infty]$ en $I_2 \subset [-\infty, 0]$ (of andersom), garandeert de tussenwaardestelling het bestaan van een nulpunt van g_λ in $[x_1, x_2]$. In dit speciale geval zijn die berekeningen zelfs met enig geduld nog wel met de hand te doen. Een nadeel van deze specifieke aanpak is dat de methode gebruik maakt van puur eendimensionale argumenten die lastig te generaliseren zijn naar hoger dimensionale problemen.

Een alternatief voor de bovenstaande methode is om een nulpunt te vinden van de afbeelding $F_\lambda : [0,1]^3 \to [0,1]^3$ gegeven door

$$F_\lambda(x_1, x_2, x_3) \stackrel{\text{def}}{=} \begin{pmatrix} f_\lambda(x_1) - x_2 \\ f_\lambda(x_2) - x_3 \\ f_\lambda(x_3) - x_1 \end{pmatrix}. \tag{3}$$

In dit geval is het niet meer mogelijk om de tussenwaardestelling te gebruiken. In plaats daarvan combineren we het gebruik van interval-aritmetiek met de methode van Newton–Kantorovich (zie Stelling 1 verderop). Het grote voordeel van deze aanpak is dat deze makkelijk te generaliseren is naar willekeurige, en in het bijzonder oneindige, dimensies.

De methode van Newton–Kantorovich

We concentreren ons eerst op het eindigdimensionale geval: we willen het bestaan van een nulpunt van een afbeelding $F : \mathbb{R}^n \to \mathbb{R}^n$ aantonen met behulp van de computer. De eerste stap bestaat uit het vinden van een goede numerieke benadering voor een nulpunt van F (bijvoorbeeld met de klassieke methode van Newton). De tweede stap bestaat uit een combinatie van analyse op papier en interval-aritmetiek. Het idee is om te bewijzen dat er een exact nulpunt van F bestaat in de buurt van de numerieke benadering door de condities van de stelling van Newton–Kantorovich te controleren.

In het Newton-algoritme wordt een nulpunt benaderd met het iteratieve proces $x_k = N(x_{k-1})$, $k \in \mathbb{N}$. Hierbij is $N : \mathbb{R}^n \to \mathbb{R}^n$ de klassieke Newton-afbeelding gegeven door

$$N(x) \stackrel{\text{def}}{=} x - [DF(x)]^{-1} F(x), \tag{4}$$

en $x_0 \in \mathbb{R}^n$ is een benadering (een gok) voor het nulpunt van F.

Een belangrijke aanname in de methode van Newton is dat de afgeleide DF van F in het exacte nulpunt bestaat, en continu en inverteerbaar is. Dit betekent in het bijzonder dat $DF(x)$ inverteerbaar is in een kleine omgeving van het exacte nulpunt. In deze omgeving geldt dat $F(x)=0$ dan en slechts dan als $N(x)=x$. Met andere woorden, nulpunten van F zijn vaste punten van N en andersom.

De methode van Newton is in praktijk bijzonder effectief. Dit komt doordat de Newton-afbeelding een contractie is in een omgeving van het exacte nulpunt. Bovendien is de contractiesterkte van N 'oneindig' goed in de buurt van het exacte nulpunt. Om preciezer te zijn, een afbeelding $T: X \to X$ op een genormeerde vectorruimte X is een contractie op $U \subset X$ als er een constante $L \in (0,1)$ bestaat zodat

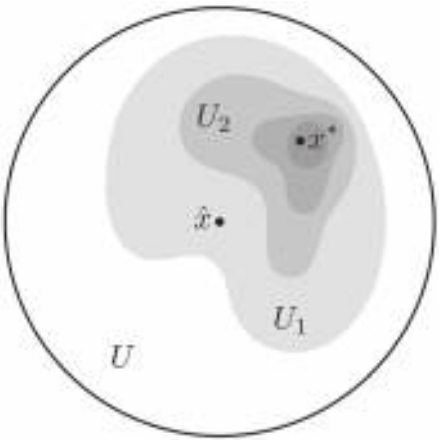
$$|T(x)-T(y)| \leq L|x-y| \quad \text{voor alle } x,y \in U,$$

waarbij $|\cdot|$ de norm op X is. De constante L wordt doorgaans een Lipschitz-constante voor T op U genoemd. In het geval dat T differentieerbaar is (en U convex) kan de middelwaardestelling worden gebruikt om een Lipschitz-constante te bepalen: kies L zodat $\|DT(x)\| \leq L$ voor alle $x \in U$, waarbij $\|DT(x)\|$ de operator-norm van de lineaire operator $DT(x)$ is. Hoe kleiner L, hoe sterker de contractie.

De dekpuntstelling van Banach stelt dat een contractie op een volledige ruimte een uniek dekpunt x^* heeft. Met andere woorden, als U volledig is en T een contractie zodat $T(U) \subset U$, dan bestaat er een uniek element $x^* \in U$ zodat $T(x^*)=x^*$. In het bijzonder geldt $\lim_{k\to\infty} T^k(x)=x^*$ voor elke startwaarde $x \in U$, zie Figuur 2.

Om nu terug te komen op de methode van Newton, die werkt zo goed omdat $DN(x^*)=0$. Dit heeft als gevolg dat de Lipschitz-constante voor N willekeurig klein wordt als we maar dicht genoeg in de buurt van x^* zijn.

Stel nu dat we een goede numerieke benadering $\hat{x} \in \mathbb{R}^n$ voor een nulpunt van F hebben gevonden. Het idee is om deze numerieke benadering te gebruiken om te bewijzen dat er een uniek nulpunt van F bestaat in een omgeving van $\hat{x}$ middels de Banach-contractiestelling. In het ideale geval zouden we een expliciete Lipschitz-constante voor N uitrekenen om aan te tonen dat de afbeelding een contractie is in een omgeving van $\hat{x}$. Echter, een dergelijke analyse van N is voor niet-lineaire afbeeldingen F doorgaans moeilijk, omdat de analyse van de inverse $[DF(x)]^{-1}$ in (4) lastig is vanwege de afhankelijkheid in x.



Figuur 2 Een schematische weergave van de dekpunt-stelling van Banach. We starten met een omgeving $U=\overline{B_r(\hat{x})}$ van een numerieke benadering $\hat{x}$. Omdat T een contractie is die U in zichzelf afbeeld is $U_1 \stackrel{\text{def}}{=} T(U) \subset U$ een verzameling met een kleinere diameter. De verzameling U_1 wordt vervolgens door T afgebeeld op een nog 'kleinere' verzameling $U_2 \stackrel{\text{def}}{=} T(U_1) \subset U_1$, enzovoort. Uit het uniform kleiner worden van het rijtje verzamelingen U_k (en de volledigheid van U) volgt het bestaan van een uniek punt $x^* \in U$ (het dekpunt van T) zodat $\lim_{k\to\infty} T^k(x)=x^*$ voor elke $x \in U$.

Omdat de contractiesterkte van de Newton-afbeelding blijft toenemen naarmate we dichter in de buurt van het nulpunt komen, hebben we de vrijheid om kleine wijzigingen in N aan te brengen zonder de contractie-eigenschap te verliezen. In het bijzonder kunnen we de inverse $[DF(x)]^{-1}$ vervangen door een numerieke benadering $A \approx [DF(\hat{x})]^{-1}$ en daarmee de afhankelijkheid in x vermijden (hier evalueren we DF dus in een vast gekozen punt $\hat{x}$). Dit levert een nieuwe Newton-achtige afbeelding $T: \mathbb{R}^n \to \mathbb{R}^n$ op, gegeven door

$$T(x) \stackrel{\text{def}}{=} x - AF(x).$$

Onder de aanname dat de matrix A injectief is zullen de vaste punten van T precies de nulpunten van F zijn. De contractiesterkte van T is weliswaar zwakker dan die van N, maar het gebruik van deze afbeelding maakt het mogelijk om de analyse daadwerkelijk uit te voeren.

Zoals al eerder aangegeven is het doel nu om een omgeving $U=\overline{B_r(\hat{x})}$ te vinden (een gesloten bol om $\hat{x}$ met straal $r>0$) waarop T een contractie is. Hier hebben we twee ingrediënten voor nodig: de kwaliteit van de numerieke benadering $\hat{x}$ en een Lipschitz-constante voor T. De kwaliteit van de numerieke benadering kan worden bepaald middels een afschatting van het residu $|T(\hat{x})-\hat{x}|$. De contractie sterkte kan worden bepaald door $\|DT(x)\|$ te begrenzen op $\overline{B_r(\hat{x})}$. Omdat we a priori niet weten voor welke straal r de afbeelding T een contractie is, houden we deze variabel in de analyse op papier. We maken vervolgens gebruik van de computer om r te bepalen.

De straal r moet namelijk enerzijds groot genoeg zijn om ervoor te zorgen dat het exacte nulpunt bevat is in $\overline{B_r(\hat{x})}$. Anderzijds moet de straal klein genoeg zijn om de contractiviteit van T te garanderen (zie Figuur 2). Door op zoek te gaan naar een zo klein mogelijke straal waarvoor T een contractie is kunnen we op een wiskundig exacte manier uitspraken doen over de kwaliteit van de numerieke simulatie. Aan de andere kant, door op zoek te gaan naar een zo groot mogelijke straal kunnen we uitspraken doen over de grootte van het gebied waarin de exacte oplossing uniek is.

De precieze formulering van de hierboven beschreven methode is samengevat in de volgende stelling en staat bekend als de geparametriseerde Newton–Kantorovich-methode (het bewijs is niet ingewikkeld, zie bijvoorbeeld [7]).

Stelling 1. *Laat T een differentieerbare afbeelding van een Banachruimte X naar zichzelf zijn, en laat $\hat{x} \in X$. Neem aan dat er een getal $Y>0$ en een functie $Z: \mathbb{R}^+ \to \mathbb{R}^+$ bestaan zodat*

$$|T(\hat{x})-\hat{x}| \leq Y,$$
$$\sup_{v \in \overline{B_1(0)}} \|DT(\hat{x}+rv)\| \leq Z(r).$$

Als er een straal $\hat{r}>0$ bestaat zodat

$$Y + Z(\hat{r})\hat{r} < \hat{r}, \tag{5}$$

dan is $T: \overline{B_{\hat{r}}(\hat{x})} \to \overline{B_{\hat{r}}(\hat{x})}$ een contractie.

Aangezien dekpunten van T corresponderen met nulpunten van F kunnen we dus aantonen dat F een uniek nulpunt in de buurt van de numerieke benadering $\hat{x}$ heeft door de ongelijkheid te controleren. Het werk is nu 'verplaatst' naar het vinden van goede schattingen Y en $Z(r)$; in het bijzonder moet minstens gelden dat Z strikt kleiner dan 1 is.

We illustreren dit aan de hand van het voorbeeld uit de vorige paragraaf. We kiezen $\lambda = 3{,}95$ en proberen een nulpunt te vinden van F_λ, gegeven in (3). We gebruiken een benadering

$$\hat{x} = \begin{pmatrix} 0{,}19 \\ 0{,}60 \\ 0{,}95 \end{pmatrix}$$

van een periodieke baan van periode drie en een bijna-inverse

$$A \stackrel{\text{def}}{=} \begin{pmatrix} 0{,}47 & -0{,}62 & 0{,}17 \\ 0{,}17 & -1{,}54 & 0{,}43 \\ -0{,}13 & 0{,}17 & -0{,}33 \end{pmatrix}.$$

De volgende stap is om de begrenzingen Y en $Z(r)$ uit te rekenen.

Om het residu te begrenzen is het voldoende om in te zien dat

$$T(\hat{x}) - \hat{x} = -AF_\lambda(\hat{x}).$$

De bovenstaande term kan worden begrensd met behulp van interval aritmetiek. Dit levert $Y \stackrel{\text{def}}{=} 5{,}7 \cdot 10^{-3}$ als maat voor het residu (hoe kleiner hoe beter) van de numerieke benadering $\hat{x}$. Om de afschatting $Z(r)$ te bepalen, laat $v \in \mathbb{R}^3$ met $|v| \leq 1$, waarbij $|\cdot|$ de Euclidische norm is. Dan geeft de driehoeksongelijkheid

$$\begin{aligned} \| DT(\hat{x}+rv) \| &= \| I - ADF_\lambda(\hat{x}+rv) \| \\ &\leq \| (I - ADF_\lambda(\hat{x})) \| \\ &\quad + \| A[DF_\lambda(\hat{x}+rv) - DF_\lambda(\hat{x})] \|. \end{aligned} \tag{6}$$

De eerste term in (6) is opnieuw eenvoudig te begrenzen met behulp van interval-aritmetiek, en we vinden $0{,}06$ als maat voor de kwaliteit (weer: hoe kleiner hoe beter) van de bijna-inverse A. De tweede term in (6) kunnen we niet direct afschatten met behulp van de computer, want die heeft een functionele afhankelijkheid van r. Er moet dus eerst enige analyse op papier worden uitgevoerd:

$$DF_\lambda(\hat{x}+rv) - DF_\lambda(\hat{x}) = -2\lambda r \begin{pmatrix} v_1 & 0 & 0 \\ 0 & v_2 & 0 \\ 0 & 0 & v_3 \end{pmatrix}. \tag{7}$$

Omdat we hebben aangenomen dat $|v| \leq 1$, geldt

$$\left\| \begin{pmatrix} v_1 & 0 & 0 \\ 0 & v_2 & 0 \\ 0 & 0 & v_3 \end{pmatrix} \right\| \leq 1.$$

We kunnen dus de tweede term in (6) afschatten met

$$\| A(DF_\lambda(\hat{x}+rv) - DF_\lambda(\hat{x})) \| \leq 2\lambda r \| A \|,$$

en met behulp van interval aritmetiek een begrenzing voor $\| A \|$ (en dus de bovenstaande term) uitrekenen. We vinden dat we $2\lambda \| A \|$ kunnen afschatten met $14{,}05$. Door de bovenstaande berekeningen te combineren concluderen we dat $Z(r) \stackrel{\text{def}}{=} 14{,}05\, r + 0{,}06$, en de ongelijkheid wordt

$$14{,}05 r^2 - 0{,}94 r + 5{,}7 \cdot 10^{-3} < 0. \tag{8}$$

Met interval-aritmetiek is dan eenvoudig aan te tonen dat voor elke r in het interval $[0{,}007, 0{,}06]$ aan (8) wordt voldaan, en daarmee is het bestaan van een unieke periodieke baan van periode drie in $\overline{B_r(\hat{x})}$ bewezen. Dus de logistische afbeelding is chaotisch voor $\lambda = 3{,}95$ op grond van de *forcing theorem*.

Oneindige dimensies

Zoals we gezien hebben in bovenstaand voorbeeld is het mogelijk om met behulp van Stelling 1 en interval-aritmetiek het bestaan van nulpunten van $F : \mathbb{R}^n \to \mathbb{R}^n$ aan te tonen. Om deze methode ook op differentiaalvergelijkingen toe te kunnen passen moeten we deze uitbreiden naar een *oneindig*-dimensionale context.

Als voorbeeld bekijken we het beginwaardeprobleem

$$\begin{cases} x'(t) = f(x(t)) & \text{met } t \in [0,1], \\ x(0) = x_0. \end{cases}$$

Allereerst zullen we dit herschrijven als een probleem van de vorm $F(x) = 0$ op een oneindig-dimensionale ruimte. Een natuurlijke keuze is om de differentiaalvergelijking te integreren en dan de afbeelding $F : C([0,1]) \to C([0,1])$ te definiëren door

$$F(x)(t) \stackrel{\text{def}}{=} x(t) - x_0 - \int_0^t f(x(s))\, ds. \tag{9}$$

Immers, $F(x)(0) = x(0) - x_0$ en $F(x)'(t) = x'(t) - f(x(t))$, en dus geldt dat $F(x) = 0$ dan en slechts dan als $x'(t) = f(t)$ én $x(0) = x_0$.

Een alternatieve strategie is om x als een machtreeks $x(t) = \sum_{k=0}^{\infty} a_k t^k$ te schrijven. Dit is een tweede natuurlijke keuze in het geval f reëel analytisch is. In dat geval kunnen we $f(x(t))$ representeren met Taylor-coëfficiënten $b_k(a_0, \ldots, a_k)$, zodat

$$f(x(t)) = f\left(\sum_{k=0}^{\infty} a_k t^k \right) = \sum_{k=0}^{\infty} b_k(a_0, \ldots, a_k)\, t^k.$$

Omdat $x'(t) = \sum_{k=0}^{\infty} (k+1)\, a_{k+1} t^k$, is het beginwaardeprobleem equivalent met het oneindige stelsel van algebraïsche vergelijkingen

$$(k+1)\, a_{k+1} = b_k(a_0, \ldots, a_k), \quad k = 0, 1, 2, \ldots,$$

aangevuld met $a_0 = x_0$. In dit geval kunnen we F dus definiëren door

$$F : \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_k \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} a_0 - x_0 \\ a_1 - b_0(a_0) \\ 2a_2 - b_1(a_0, a_1) \\ 3a_3 - b_2(a_0, a_1, a_2) \\ \vdots \\ k a_k - b_{k-1}(a_0, \ldots, a_{k-1}) \\ \vdots \end{pmatrix}. \tag{10}$$

Er zijn in de praktijk (door het kiezen van andere bases dan de machten t^k) nog vele andere keuzes voor dit soort rijtjesruimtes mogelijk. Zo wordt er ook vaak gebruikt gemaakt van de Fourier-, Legendre- of Chebyshev-basis. De precieze keuzes die worden gemaakt voor de afbeelding $F : X \to X'$ en de ruimten X en X' verschillen sterk per probleem en hangen vaak nauw samen met hoe de bijna-inverse A geconstrueerd wordt.

Al deze methoden hebben één ding gemeen: F is nu een afbeelding tussen *oneindig*-dimensionale ruimten. Op het eerste gezicht zou dat weinig moeten veranderen, want Stelling 1 is ook waar voor oneindig-dimensionale ruimten. Echter, omdat we nu te maken hebben met oneindig-dimensionale afbeeldingen is het doorgaans niet meer mogelijk om de functie F expliciet uit te rekenen. In het geval dat F een afbeelding is tussen rijtjesruimten is dit meteen duidelijk: F heeft immers oneindig veel termen. Ook het kiezen van de benadering A voor de inverse $DF(\hat{x})^{-1}$ is minder makkelijk.

Niettemin, in veel gevallen kunnen we F schrijven als $F = F^n + F^\infty$, waarbij F^n alleen afhangt van een n-dimensionale deelruimte van X en afbeeldt op een n-dimensionale deelruimte van X'. Hierbij willen we F^n zodanig kiezen dat het al het 'interessante' gedrag van F bevat. Als verder de staart F^∞ voldoende 'klein' is, dan kunnen we dus een bijna-nulpunt $\hat{x}$ van F vinden simpelweg door een bijna-nulpunt van de eindig-dimensionale afbeelding F^n te berekenen. Voor de verdere analyse van het probleem willen we de oneindig-dimensionale component F^∞ zo kiezen dat deze analytisch 'makkelijk' te bestuderen is. Het maken van een juiste keuze is doorgaans moeilijk wanneer F niet-lineaire termen bevat. Er moet in alle niet-lineaire gevallen (dus eigenlijk in alle interessante gevallen) dus goed nagedacht worden over de gevolgen van de keuzes die gemaakt worden voor de functie F, de ruimtes X en X', en de projecties op de eindige dimensionale deelruimtes (waarin de numerieke simulatie wordt uitgevoerd).

De lineaire operator $A: X' \to X$, die dienst doet als bijna-inverse van $DF(\hat{x})$, heeft in het oneindig-dimensionale geval ook een aparte behandeling nodig. Ook die moet worden gesplitst in een eindig-dimensionaal deel A^n (een matrix) en een oneindig-dimensionaal 'staartstuk' A^∞. Aangezien F^n geacht wordt al het interessante gedrag van F te bevatten, kunnen we voor de matrix A^n een numerieke benadering voor de inverse van de Jacobiaan $DF^n(\hat{x})$ kiezen. Het oneindig-dimensionale deel A^∞ is (hopelijk) klein, maar mag niet een soort nulafbeelding zijn, want A moet wel injectief gekozen worden om correspondentie tussen dekpunten van T en nulpunten van F te garanderen.

In het voorbeeld van de afbeelding gedefinieerd in (10) is de term ka_k dominant voor grote k. Daarom kunnen we proberen om $DF(\hat{x})$ te benaderen door

$$DF(\hat{x}) \approx \left(\begin{array}{c|cc} DF^n(\hat{x}) & 0 & \\ \hline & n & \\ 0 & & n+1 \\ & & & \ddots \end{array}\right). \quad (11)$$

In dit geval kunnen we A als volgt construeren. Eerst berekenen we (numeriek) een bijna-inverse voor $DF^n(\hat{x})$, namelijk $A^n \approx DF^n(\hat{x})^{-1}$. Vervolgens kiezen we

$$A \stackrel{\text{def}}{=} \left(\begin{array}{c|cc} A^n & 0 & \\ \hline & \frac{1}{n} & \\ 0 & & \frac{1}{n+1} \\ & & & \ddots \end{array}\right).$$

Er geldt dan dat $ADF(\hat{x})$ een goede benadering van de identiteit is (hoe goed, dat moet expliciet worden afgeschat).

Grote delen van de eindig-dimensionale analyse uit de vorige paragraaf kunnen nu worden gegeneraliseerd naar het oneindig-dimensionale geval. Elk van de af te schatten termen omvat nu wel twee aparte analyses, namelijk die van het n-dimensionale gedeelte (dat net als in het eindig-dimensionale geval interval aritmetisch berekend kan worden) en die van het staartstuk. Dat laatste gedeelte is oneindig-dimensionaal (dus ongeschikt voor een directe computerberekening) en moet met behulp van een combinatie van functionaal-analytische ingrediënten en intervalaritmetiek afgeschat worden.

De kern van het probleem zit hem echter toch vooral in de benadering van de (oneindig-dimensionale) afgeleide $DF(\hat{x})$ door een eenvoudigere lineaire operator (waarvan A dan weer een bijna inverse is). Het vinden van een scherpe afschatting van het verschil tussen $DF(\hat{x})$ en zijn benadering is het lastigste gedeelte van de hele analyse. Bovendien is het helaas niet voor elk probleem zo eenvoudig om het probleem op een diagonaal dominante manier te formuleren (zoals in ons voorbeeld). En zelfs als dat wel kan, is het nog maar de vraag of er een eindig-dimensionale projectie van het probleem gevonden kan worden die groot genoeg is om het essentiële gedrag van het oneindig-dimensionale niet-lineaire probleem te 'vangen' (zodat we schattingen kunnen vinden die aan de voorwaarden van Stelling 1 voldoen), maar die anderzijds klein genoeg is om mee te kunnen rekenen op een computer. Tot slot van dit artikel beschrijven we nog enkele toepassingen waarin het gelukt is om deze uitdagingen het hoofd te bieden.

Toepassingen

In deze laatste sectie geven we twee (omwille van de leesbaarheid relatief eenvoudige) voorbeelden van toepassingen. Er is een heel scala aan problemen uit de dynamische systemen waar momenteel aan gewerkt wordt: evolutionaire partiële differentiaalvergelijkingen [1], *delay* vergelijkingen [9], sterk indefiniete problemen [6], verbindende banen [2,10], domeindecompositie [4], driedimensionale periodieke patronen, en natuurlijk het ontwikkelen van algemeen bruikbare software.

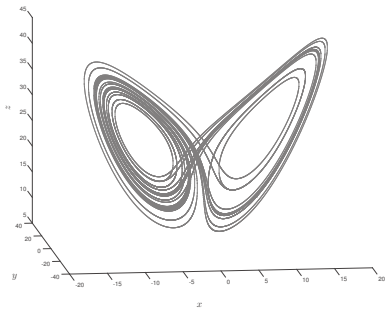
Periodieke banen in het Lorenz-systeem

In dit voorbeeld zoeken we periodieke banen in het klassieke Lorenz systeem:

$$\begin{cases} \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} 10(y-x) \\ x(28-z)-y \\ xy - \frac{8}{3}z \end{pmatrix}, \quad t \in [0, L], \\ x(0) = x(L),\, y(0) = y(L),\, z(0) = z(L). \end{cases} \quad (12)$$

Zowel de periodieke baan $t \mapsto (x(t), y(t), z(t))$ als de bijbehorende periode $L > 0$ zijn onbekend (dus het bepalen van de periode is onderdeel van het probleem).

Als $t \mapsto (x(t), y(t), z(t))$ een oplossing is van (12), dan beschrijft de 'verschoven' afbeelding $t \mapsto (x(t+\tau), y(t+\tau), z(t+\tau))$ dezelfde periodieke baan (voor elke $\tau \in \mathbb{R}$), en is dus ook een oplossing. Door deze translatie-symmetrie is de oplossing *niet* lokaal uniek, dus een contractiestelling is in deze formulering niet toepasbaar. We



Figuur 3 Een gevalideerde periodieke baan van periode $L \approx 25{,}03$ in het Lorenz systeem.

voegen daarom een extra conditie aan (12) toe, een zogenaamde fase-conditie, om de tijd-parametrisatie van de periodieke baan uniek vast te leggen. We eisen daartoe dat de baan op $t = 0$ in een vooraf gespecificeerd vlak in de fase-ruimte $\mathbb{R}^3$ ligt.

Aangezien het gaat om een periodieke oplossing ligt het voor de hand om een startformulering $F(x) = 0$, vergelijk (1), te kiezen in termen van Fourier-coëfficiënten:

$$x(t) = \sum_{k \in \mathbb{Z}} a_k e^{ik\omega t},$$
$$y(t) = \sum_{k \in \mathbb{Z}} b_k e^{ik\omega t},$$
$$z(t) = \sum_{k \in \mathbb{Z}} c_k e^{ik\omega t},$$

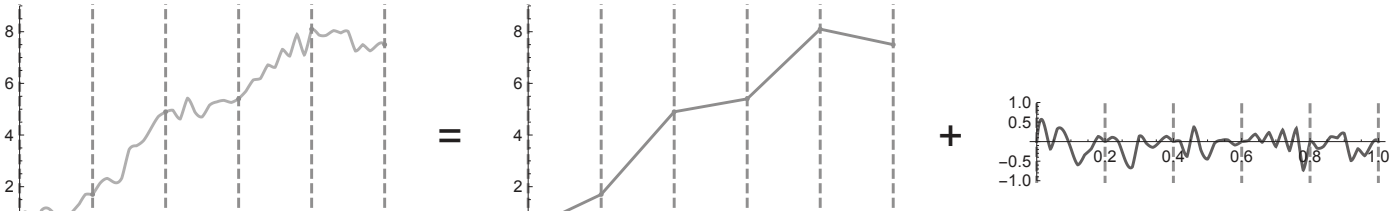
waarbij $\omega = 2\pi/L$. Het invullen van de bovenstaande expansies in (12) en de fase-conditie levert een oneindig stelsel van algebraïsche vergelijkingen op voor de onbekende rijtjes coëfficiënten a, b, c en de frequentie ω. We schrijven dat stelsel hier niet expliciet op, maar het is in hoge mate analoog aan (10). We passen nu de in de vorige paragraaf ontwikkelde methode toe op dit probleem en kunnen daarmee bijvoorbeeld bewijzen dat de kromme (het resultaat van een simulatie) in Figuur 3 gegarandeerd een periodieke baan van (12) representeert.

Een radiaal symmetrische PDV

We kunnen de methode ook toepassen om oplossingen van partiële differentiaalvergelijkingen (PDVs) te vinden. Als voorbeeld kijken we naar gelokaliseerde oplossingen van de stationaire niet-lineaire Schrödinger vergelijking

$$\Delta\psi(x) - \psi(x) + \psi(x)^3 = 0, \quad (13)$$

waarbij $\psi: \mathbb{R}^3 \to \mathbb{R}$. Wanneer we voor deze PDV op zoek gaan naar een radiaal symmetrische oplossing schrijven we $\psi(x) = f(|x|)$ voor een $f: [0, \infty) \to \mathbb{R}$ met



Figuur 4 Een afbeelding op $[0,1]$ kan worden opgedeeld in een spline deel (met 6 roosterpunten in dit voorbeeld) en een restant.

de eigenschap dat $\lim_{r\to\infty} f(r) = 0$. Dit reduceert de PDV (13) tot

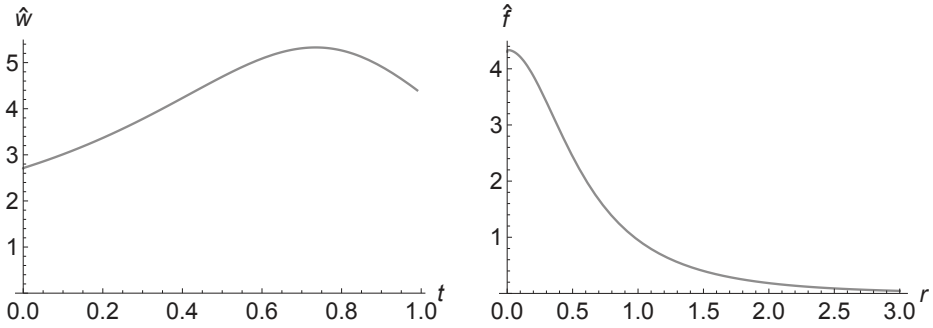
$$f''(r) + \frac{2}{r} f'(r) - f(r) + f(r)^3 = 0, \quad (14)$$

een singuliere niet-autonome gewone differentiaalvergelijking (GDV) op $(0,\infty)$. Numeriek rekenen op een oneindig domein is lastig, maar door deze GDV te transformeren naar het interval $(0,1)$ en de methode van Green toe te passen, kan aangetoond worden dat (14) een begrensde oplossing heeft wanneer er een functie $w \in C[0,1])$ bestaat die voor alle $t \in [0,1]$ aan de volgende integraalvergelijking voldoet:

$$\begin{aligned} F(w) \stackrel{\text{def}}{=} w(t) &- \frac{1}{2}\frac{1+t}{t^2}\int_0^t \frac{s^3(1-s)^3}{\log(s)^2} w(s)^3 ds \\ &- \frac{1}{2}\frac{1}{1-t}\int_t^1 \frac{s(1+s)(1-s)^4}{\log(s)^2} w(s)^3 ds \\ &= 0. \end{aligned} \quad (15)$$

De relatie tussen de oplossing w van de integraalvergelijking (15) en de oplossing f van de GDV wordt gegeven door $f(r) = \frac{1}{r} e^{-r}(1-e^{-r})\, w(e^{-r})$, zie Figuur 5. We merken op dat (15) op hoofdlijnen dezelfde vorm heeft als de formulering (9).

Voor dit probleem kiezen we als eindig-dimensionale benadering $\hat{w}$ van de oplossing een stuksgewijs lineaire *spline*. Elke functie op $C[0,1]$ kan worden opgedeeld in zo'n *spline*-deel met m roosterpunten, en een restant, waarbij het restant kleiner wordt naarmate het aantal roosterpunten m toeneemt, zie Figuur 4. De numeriek gevonden simulatie $\hat{w}$, waarbij voor een uniform rooster is gekozen met $m = 1000$ punten, is afgebeeld in Figuur 5, evenals de bijbehorende functie $\hat{f}$. Het bewijs voor het bestaan van een (echte) oplossing van $F(w) = 0$ vlakbij $\hat{w}$, en daarmee van een radiaal-symmetrische gelokaliseerde oplossing van de PDV (13), volgt nu, na enige zorgvuldige analyse en met behulp van een computationeel gedeelte [3], uit de methode zoals beschreven in de vorige paragraaf.



Figuur 5 Links de functie $\hat{w}(t)$ op $[0,1]$. Rechts de corresponderende functie $\hat{f}(r)$ op $(0,\infty)$.

Referenties

1 G. Arioli en H. Koch, Integration of dissipative partial differential equations: a case study, *SIAM J. Appl. Dyn. Syst.* 9(3) (2010), 1119–1133.

2 J.B. van den Berg, A. Deschenes, J.P. Lessard en J.D. Mireles-James, Stationary Coexistence of Hexagons and Rolls via Rigorous Computations, *SIAM J. Appl. Dyn. Syst.* 14(2) (2015), 942–979.

3 J.B. van den Berg, C.M. Groothedde en J.F. Williams, Rigorous computation of a radially symmetric localized solution in a Ginzburg–Landau problem, *SIAM J. Appl. Dyn. Syst.* 14 (2015), no. 1, 423–447.

4 J.B. van den Berg en R.S.S. Sheombarsing, Rigorous numerics for ODEs using Chebyshev series and domain decomposition, 2015. Preprint.

5 CAPD: Computer assisted proofs in dynamics, a package for rigorous numerics, http://capd.ii.uj.edu.pl.

6 R. Castelli, M. Gameiro en J.-P. Lessard, Rigorous numerics for ill-posed PDEs: periodic orbits in the Boussinesq equation, 2015. Preprint.

7 S. Day, J.-P. Lessard en K. Mischaikow, Validated continuation for equilibria of PDEs, *SIAM J. Numer. Anal.* 45(4) (2007), 1398–1424.

8 O.E. Lanford, A computer-assisted proof of the Feigenbaum conjectures, *Bull. Amer. Math. Soc.* 6 (1982), 427–434.

9 J.P. Lessard, Recent advances about the uniqueness of the slowly oscillating periodic solutions of Wright's equation, *J. Differential Equations* 248(5) (2010), 992–1016.

10 J.P. Lessard, J.D. Mireles-James en C. Reinhardt, Computer assisted proof of transverse saddle-to-saddle connecting orbits for first order vector fields, *J. Dynamics and Differential Equations* 26(2) (2014), 267–313.

11 T. Li en J.A. Yorke, Period three implies chaos, The American Mathematical Monthly 82 (1975), no. 10, 985–992.

12 K. Mischaikow en M. Mrozek, Chaos in the Lorenz equations: a computer assisted proof, *Bull. Amer. Math. Soc.* 32(1) (1995), 66–72.

13 W. Tucker, The Lorenz attractor exists, *C. R. Acad. Sci. Paris* 328(12) (1999), 1197–1202.

Pieter Collins
Department of Data Science and Knowledge Engineering
Maastricht University
pieter.collins@maastrichtuniversity.nl

Model checking dynamic systems

Pieter Collins started work on model checking hybrid systems as a postdoctoral research fellow at the Centrum Wiskunde & Informatica in the group of Prof. Jan H. van Schuppen, working on an EU-funded project 'Control and Computation'. He contributed to the software project Ariadne for reachability analysis and verification of hybrid systems, developed by the company Parades (now Ales) in Rome, and the group of Tiziano Villa in Udine (now Verona). This project led to a Vidi grant on 'Computational Topology for Systems and Control'. Since 2011 he has been a lecturer in the Department of Data Science and Knowledge Engineering and Maastricht University.

Engineers design and build things: mechanical engineers build machines, genetic engineers design lifeforms, and software engineers make computer programs. And of course, these things should serve some purpose and properly perform the tasks they were designed to do. Trains on a railway track should not crash into each other, radiotherapy machines should not deliver a lethal dose, and your computer's operating systems should never display a 'blue screen of death'. To this end, designs are analysed, prototypes tested, and the final product trialled. This process is known as *verification*, and is one of the most important and challenging parts of engineering.

Model checking [5, 11] is a computational approach to verification, in which a (mathematical) model of the system is analysed to determine whether it satisfies its specification as expressed by some temporal logical formula. The analysis must be performed entirely rigorously to avoid the possibility of dangerous or costly errors. Working with models is often cheaper than working with physical prototypes, and may be more powerful, since exhaustive testing may not be feasible, or scenarios may occur which had not been considered during testing.

I first started work in model checking as a postdoctoral researcher in the group of Jan H. van Schuppen at the Centrum Wiskunde & Informatica in Amsterdam, on an EU-funded project 'Control and Computation'. A part of this project was the development of a model checking tool Ariadne for *hybrid systems*, led by the group of Tiziano Villa in Udine, and a small Rome-based company Parades with links to the automotive industry. Hybrid systems [45] first appeared in [18, 41], and are characterised by continuous evolution interspersed by discrete *events* causing an instantaneous jump in the system state, or a switch in the *mode* of behaviour. They typically arise in the control of a physical system via discrete sensors and actuators, but are also encountered in systems with *impacts*, such as robot locomotion, or as electrical systems with switches and/or diodes.

The problem of model checking hybrid systems is a very rich domain, encompassing many of the important parts of continuous mathematics: real numbers, continuous functions, open and compact sets, algebraic and differential equations, and constraint feasibility problems. There are important theoretic questions of which properties of the evolution are even *computable* [48], and practical problems in the development of tools for efficient and *rigorous* numerical computation [25]. As I had already developed an interest in these issues in the context of chaos theory, I saw collaboration on Ariadne as a perfect opportunity to actually develop software, both as an industrially-viable tool for the model checking problem itself, and a flexible, general-purpose package for rigorous numerical computation.

My work on Ariadne led to a rich vein of both theoretical research and practical software development, and to a successful Vidi grant proposal on 'Computational Topology for Systems and Control'. An initial version of Ariadne focused solely on reachability analysis of hybrid systems using affine sets as a fundamental internal representation, and rigorous linear algebra and linear programming to perform computations. While this provided fast computations with reasonable accuracy, the under-

lying technology was only appropriate for fairly coarse safety properties. Subsequent versions of Ariadne featured a full rigorous numerical calculus for numbers, vectors, functions and sets.

Of course, there were already tools for rigorous numerical computation, including the interval arithmetic package INTLAB [43], differential equation solvers AWA [31], VNODE [39] and COSY Infinity [33], and dynamical systems analysers CAPD-Lib [38] and GAIO [14]. Hybrid system solvers included d/dt [2] and HyTech [20] for linear continuous dynamics, and HyperTech [21] and Checkmate [10] for nonlinear dynamics. With Ariadne we aimed to extend the scope and flexibility of these tools, in particular to provide an extensible framework where sets and functions could be manipulated as first-class objects, and to study highly nonlinear and/or nondeterministic systems. Later hybrid systems tools include HSolver [42], Phaver [16], SpaceEx [17], Flow* [9] and AERN [15], with the latter three featuring a similar rigorous numerical calculus to Ariadne.

Model checking hybrid systems

Model checking was originally developed for verifying the flow-of-control of software systems. System models are given by *discrete automata* with a finite state set X specifying the set of possible *executions* $(x_0, x_1, x_2, \ldots)$. Specifications are given by a *temporal logic* formula describing the allowable executions. A *linear temporal logic* formula is built up from *atomic propositions* a, which are sets of states, basic logical operations, and *temporal operators*: *next* (denoted $\bigcirc$), *always* ($\Box$), *eventually* ($\Diamond$), *until* and *release*. For example, the property $\Diamond a \vee \bigcirc\Box b$ (eventually a or next always b) means that either property a holds at some time in the future, or that after one time step, property b always holds. Of particularly importance are *safety* properties $\Box a$ (always a). These can be checked by computing the *reachable set* of states which it is possible to attain from a starting state.

There are many tools for model checking of linear temporal logic for finite automata, such as [23]. They either prove that the specification is satisfied for all executions of the system, or output an execution not satisfying the specification. This *counterexample* may indicate a flaw in the system design, or that the specification has not been properly thought-out.

The main practical difficulty in model checking is state-space explosion caused by the so-called *curse of dimensionality* — for a moderately-sized system with 100 on–off switches, there are $2^{100} \sim 10^{30}$ states, which would take approximately the age of the universe (13.8 billion years) to analyse even when looking at a trillion states per second! Discrete model checkers therefore use a variety of data representations and algorithmic tricks to speed up the calculation. Since the result must be rigorously correct, randomised algorithms cannot be used.

As soon as the number of states becomes infinite, such as by introducing integer variables, then it turns out that model checking linear temporal logic is *undecidable*. A a yes/no question is *decidable*, or a function computing a string or numerical value *computable*, if it can be implemented on a *Turing machine*, which can be thought of as a program running on a digital computer with unlimited memory. It is easy to see that there must be undecidable problems and uncomputable functions, since there are uncountably many functions $\mathbb{N} \to \{0,1\}$, but only countably many programs. The *halting problem*, of determining whether a computer program will terminate or run forever, is the most famous undecidable problem, and can actually be formulated as a model checking problem. Note that undecidability of a problem *class* does not mean that particular *instances* of a problem are unsolvable; for example, termination of a program can always be proved by running it long enough (so halting is *verifiable*), and the infinite loop `while(true) {...}` clearly never terminates.

Hybrid automata were seen by computer scientists as a natural extension of discrete automata to include continuous dynamics. They consist of finitely many *discrete states* $q_1, \ldots, q_n$, in each of which the *continuous state* $x_i \in X_i$ satisfies a differential equation $\dot{x}_i(t) = f_i(x_i(t))$. When the continuous state satisfies a *guard condition* $g_{i,k}(x_i(t)) \geq 0$, a *discrete event* e_k occurs, with the discrete state changing to q_j, and the continuous state being *reset* to $x_j(t) = r_{i,k}(x_i(t))$. Hybrid automata are often described by a graph with state dynamics in the vertices, and guards/re-

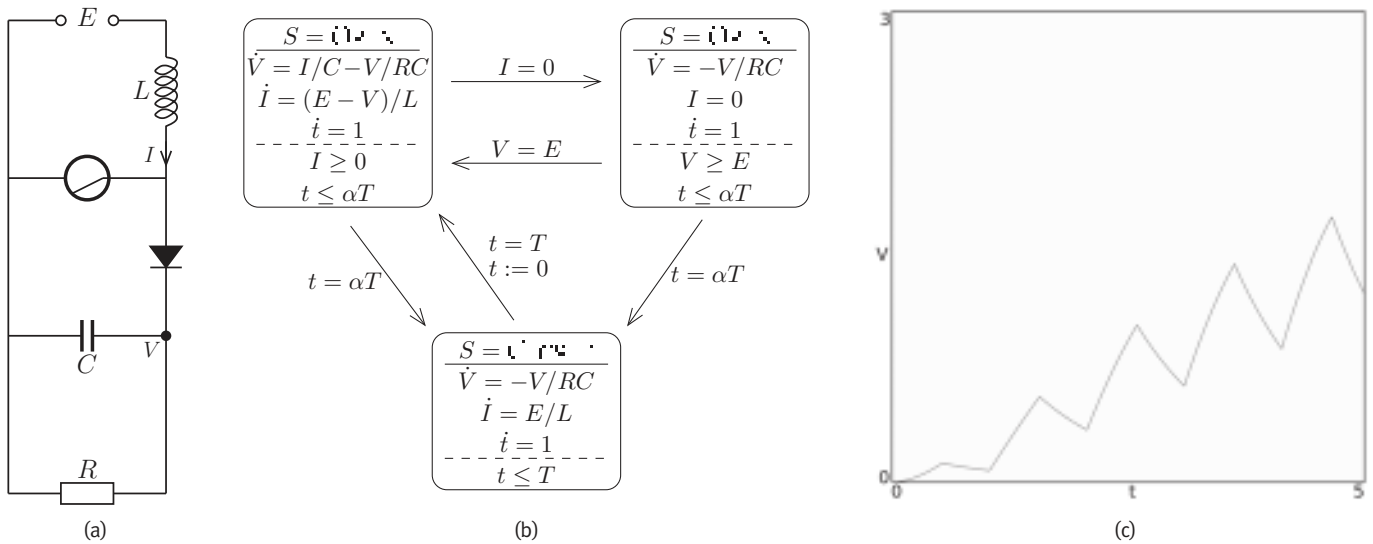


Figure 1 (a) Circuit diagram for a boost power converter. (b) A hybrid automaton representing the boost power converter. (c) Time-evolution of the capacitor voltage.

sets on the edges. An example of a hybrid automaton and its dynamics is shown in Figure 1.

For a simple class of hybrid systems, the *timed automata*, in which the only continuous dynamics are *clocks* satisfying $\dot{t}_i = 1$, model checking was shown to be decidable [1]. This result led to the development of tools such as UPPAAL [29] for the analysis of timed automata. For slightly more complicated classes of hybrid system, including saturated linear systems [7], model checking becomes undecidable [3, 22]. Clearly model checking more general nonlinear hybrid systems is also undecidable; the real question is whether we can verify any such systems at all. Since properties of continuous systems may be highly sensitive to small changes; for example, a small change in a parameter may cause a *bifurcation* or *catastrophe*, we should not expect to be able to say very much unless the property we are considering is *robust* with respect to parameter uncertainty and noise.

Computable analysis

To answer the question of what is possible to verify for nonlinear hybrid systems, we need a theory of (rigorous) computation for continuous mathematics, a subject known as *computable analysis* [26, 48]. The roots of such a theory lie in constructive mathematics, and computable mathematics can be derived from a suitable constructive mathematics using realization theory [6]. However, while in Brouwer's intuitionistic mathematics [12] and other constructive logics, axioms are admitted depending on whether one believes them to be self-evident, in computational mathematics, we have a much clearer criterion: *can we implement it?* Hence the principle of the excluded middle is not accepted (as there are statements that can be neither proved nor disproved), but the axiom of dependent choice (that we can make an unending sequence of choices, each depending on the one before) is accepted. Computable mathematics is uniform in the way that constructive/intuitionistic mathematics is not.

The main difficulty in continuous mathematics is that objects lie in *uncountable* sets, which means we can only describe them exactly using an infinite *stream* of data. The key idea of computable analysis is that it should be possible to extract useful information about a value from a *finite* piece of its describing stream. Computation is still performed using Turing machines, but a program runs forever, reading data from its input streams and writing to its output stream.

It turns out that there is a deep link between topology and computation; a representation of a set of objects by streams induces a natural topology on the set [46], and only continuous functions can be computable. Indeed, the set of all functions with uncountable domain has higher than continuum cardinality, so cannot be described by streams. Two representations of the same set are equivalent if each can be computably converted into the other. A *type* is defined to be a set with an equivalence class of representations.

We start developing computable analysis by constructing the type of real numbers. We are used to describing real numbers either by algebraic formulae, such as $1/7$, π or $\sin(\pi/7)$, or as a decimal expansion, such as $0.\dot{1}42857\dot{}$, or $3.14159\cdots$. Since the real numbers are uncountable, only infinite representations such as decimal expansions can represent them all.

Unfortunately, there is also a problem with decimal expansions: They do not support arithmetic! What is the value of $3\times 0.33333\cdots$? Clearly, if the second number is exactly $\frac{1}{3}$, then the result is $1.00000\cdots$, or equivalently, $0.99999\cdots$. But we can never output $1.0\cdots$, because maybe we have $3\times 0.333332\cdots$, which is definitely less than 1. And we can never output $0.9\cdots$, because maybe we have $3\times 0.333334\cdots$. Hence we can never give the first digit of the answer!

Thankfully, there are ways around this conundrum. The most straightforward way is to use a *signed-digit representation*, which allows negative digits such as $\bar{2}$ (representing '-2') as well as positive digits. The signed-digit representation is equivalent to allowing an error of ± 1 in the last given digit, since $3.1416\cdots$ could represent

$$3.1416\dot{\bar{9}} = 3.1415,$$

$$3.1416\dot{9} = 3.1417,$$

or any number in the interval

$$[3.1415 : 3.1417] = 3.141\,[5:7] = 3.141\,[6\pm 1].$$

Use of signed digits resolves our arithmetical dilemma, since

$$\begin{aligned} &3\times 0.33333\cdots \\ &\quad\in 3\times[0.3333\bar{3}\bar{9}\bar{9}\cdots : 0.3333399\cdots] \\ &\quad= 3\times[0.33332 : 0.33334] \\ &\quad= [0.99996 : 1.00002] \\ &\quad= [1.0000\bar{4} : 1.00002] \\ &\quad\subset 1.0000\cdots. \end{aligned}$$

There are other representations of the real numbers which are equivalent to the signed-digit representation. The *Cauchy* representation describes a real number x by a sequence of rationals q_n with $|x-q_n|\leq 2^{-n}$, and the *Dedekind* representation by a sequence of rational intervals $[a_n, b_n]$ with $a_n \leq a_{n+1} \leq b_{n+1} \leq b_n$ and $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = x$. Note that the Cauchy representation focusses on the *metric* structure of the real numbers, whereas the Dedekind representation focusses on the *order* structure.

A real number x is computable, as per Turing's definition [47], if there is a computer program which outputs its signed-digit (or an equivalent) representation. All algebraic numbers are computable, as are important constants such as e and π, but only countably many real numbers are computable.

The usual arithmetical operators $+$, $-$, $\times$, $\div$ are computable, as are elementary functions such as $\exp$, $\log$, $\sin$ and $\arctan$. Further, it is possible to take the limit of a sequence x_n with a known a rate of convergence ϵ_n i.e. $|x_{n_1} - x_{n_2}| \leq \epsilon_{\min\{n_1,n_2\}}$.

Comparison of real numbers is undecidable, since we can never tell that a number x actually equals 0 from finite piece of its signed-digit expansion $0.0000\cdots$. For this reason, we need *Kleenean* logic for comparison $x \lesssim y$, with three values $\mathbb{K} = \{\top, \bot, \uparrow\}$ representing *true*, *false* and *indeterminate* ("don't know"), the latter for the case x and y are equal. Note that while a symbolic approach may allow one to decide whether $x = 0$ in some cases e.g. $\sin(3\pi) = 0$, it is an open problem as to whether a number defined in terms of rationals and elementary functions can always be tested for being zero.

To define more complicated types, we can fall back on some very general constructions. For types $\mathbb{X}_1, \ldots, \mathbb{X}_n$, there is a *product type* $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$ of tuples $(x_1, \ldots, x_n)$ with $x_i \in \mathbb{X}_i$. For types $\mathbb{X}, \mathbb{Y}$, there is an *exponent* type $\mathbb{Y}^{\mathbb{X}}$ which consists of *sequentially continuous* functions $\mathbb{X} \to \mathbb{Y}$. The operations of *evaluation* $(f, x) \mapsto f(x)$

and *composition* $(g,f)\mapsto g\circ f$ are both computable. A foundation of functional programming languages is that the types $\mathbb{X}\times\mathbb{Y}\to\mathbb{Z}$ and $\mathbb{X}\to(\mathbb{Y}\to\mathbb{Z})$ are equivalent. Hence if for every x, the function f_x is computable, then the function f defined by $f(x,y):=f_x(y)$ is also computable.

For real-valued functions $\mathbb{X}\to\mathbb{R}$, it follows from computability of composition that pointwise arithmetic operations $f_1,f_2\mapsto f_1\star f_2$ are computable. For more complicated operators, we may need to return to first principles to prove computability. In particular, solutions of the algebraic equation $f(x)=0$ for $f:\mathbb{R}^n\to\mathbb{R}^n$ are computable if, and only if, they are isolated and have non-zero *index*. The solution of a differential equation $\dot{x}=f(t,x)$ starting at x_0 is computable if it is unique [44].

We define the type of *open* subsets $U\in\mathcal{O}(\mathbb{X})$ in terms of the *membership* function $\chi_U:\mathbb{X}\to\mathbb{S}$, where $\mathbb{S}=\{\top,\uparrow\}$ is the *Sierpinskian* type. Computable open sets therefore have a logical characterisation as sets for which membership is verifiable. The type of *compact* sets is defined as the subtype of $\mathcal{O}(\mathbb{X})\to\mathbb{S}$ satisfying $C\subset(V_1\cap V_2)\Leftrightarrow(C\subset V_1)\wedge(C\subset V_2)$. Computable compact sets as those for which being a subset of an open set is verifiable. It is easy to show computablility of unions and intersections of open and compact sets, of the preimage $f^{-1}(V)$ of an open set V under a continuous function f, and the image $F(C)$ of a compact set C under a compact-valued function F.

Rigorous numerics

Computability theory addresses the problem of what it is possible to compute with a digital computer. To show something is possible, we give an algorithm, but this algorithm is usually given to be theoretically clean and need not be particularly efficient. In order to implement a practical model checker, we also need fast *rigorous numerical* algorithms.

In almost any computational environment, if you type in `x=0.6+0.3+0.1` you will get the answer `x=0.9999999999999999` (and if you get something different, it will probably be `1.00000`, but then `x-1` yields `-1.1102e-16`). The reason for this is that most numerical computational tools use (double-precision) *floating-point* arithmetic, in which numbers are represented by a binary expansion with 53 significant (binary) digits, and results of any arithmetical operation are rounded to the nearest representable value. For many applications, the roundoff-errors are so small as to have negligable effect on the answer. But we don't know this for sure.

In *interval arithmetic* [25, 35, 36], finite-accuracy approximation to a real number is represented by an *interval*, either with lower and upper bounds $x\in[\underline{x},\overline{x}]$or a midpoint and error, $x=\check{x}\pm e$. Ariadne provides intervals with endpoints which can be either builtin double-precision floating-point numbers, or multiple-precision floating-point numbers (as provided by the MPFR library [37]). Operations on intervals are implemented by rounding the results of finite-precision calculations *outward*. For example, working to three decimal places

$$\begin{aligned}\pi\times e&\in[3.141{:}3.142]\,\hat{\times}\,[2.718{:}2.719]\\&=[8.537{:}8.544]\\&\supset[8.537238{:}8.543098].\end{aligned}$$

Interval arithmetic illustrates the core idea of rigorous numerical computation, namely to work with sets of values which are guaranteed to contain the correct result. The sets used are called *basic* sets, and ideally are kept small. This means, for example, that if $\hat{x}$ is a basic set containing x and $\hat{y}\ni y$, then a validated version $\hat{\star}$ of a binary operator $\star$ satisfies $\hat{x}\,\hat{\star}\,\hat{y}\ni x\star y$. In a metric space, basic sets of the form $\{x\mid d(x,\check{x})\leq e\}$ are most useful, where $\check{x}$ lies in a countable dense subset $\check{X}$ of $\mathbb{X}$.

Continuous functions $\mathbb{R}^n\to\mathbb{R}$ can be approximated uniformly accurately by polynomials on bounded domains. In Ariadne, we provide basic sets of continuous functions on domains which are coordinate-aligned *boxes*

$$\prod_{i=1}^{n}[a_i{:}b_i]=[a_1{:}b_1]\times[a_2{:}b_2]\times\cdots\times[a_n{:}b_n].$$

A simple class of representation are *Taylor models* [34], in which basic sets of functions over a box domain $D\subset\mathbb{R}^n$ are given as

$$\max_{x\in D}|\,f(x)-p(s^{-1}(x))\,|\leq e,$$

where $p=\sum_\alpha c_\alpha x_1^{\alpha_1}\cdots x_n^{\alpha_n}$ is a polynomial on $[-1{:}+1]^n$ with floating-point coefficients c_α, and $s{:}[-1{:}+1]^n\to D$ is a coordinate scaling function, as shown in Figure 2. The scaling to the unit box domain improves the accuracy of operations on the polynomial p, and allows for easy simplification of the representation by *sweeping* small terms into the error bound e. Other polynomial model representations are possible by using a different basis. The Taylor basis of monomials is easiest to perform arithmetic on, but the Chebyshev basis or Bernstein bases may yield more accurate evaluation. For problems in partial differen-

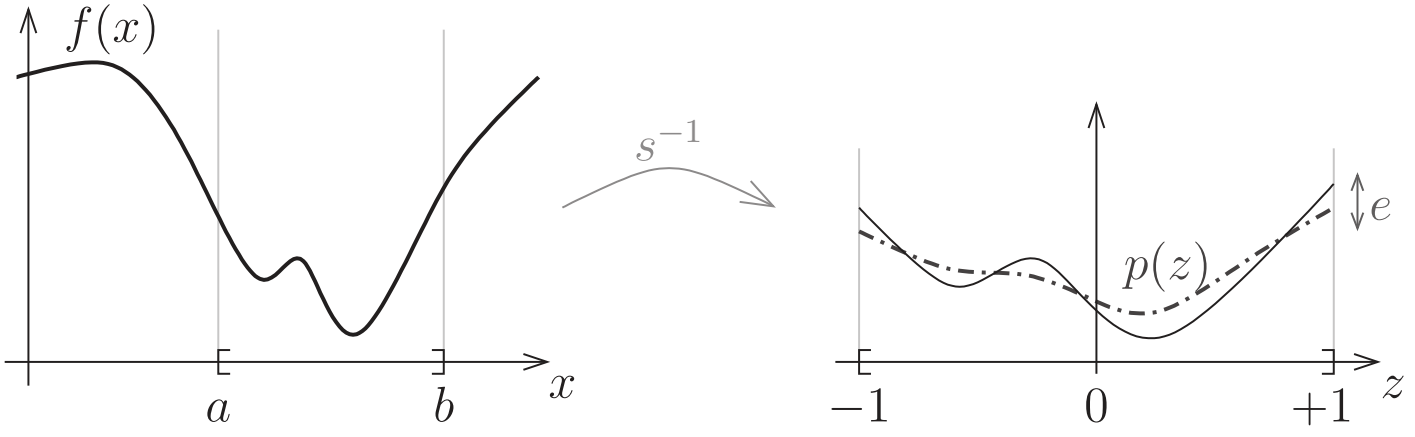
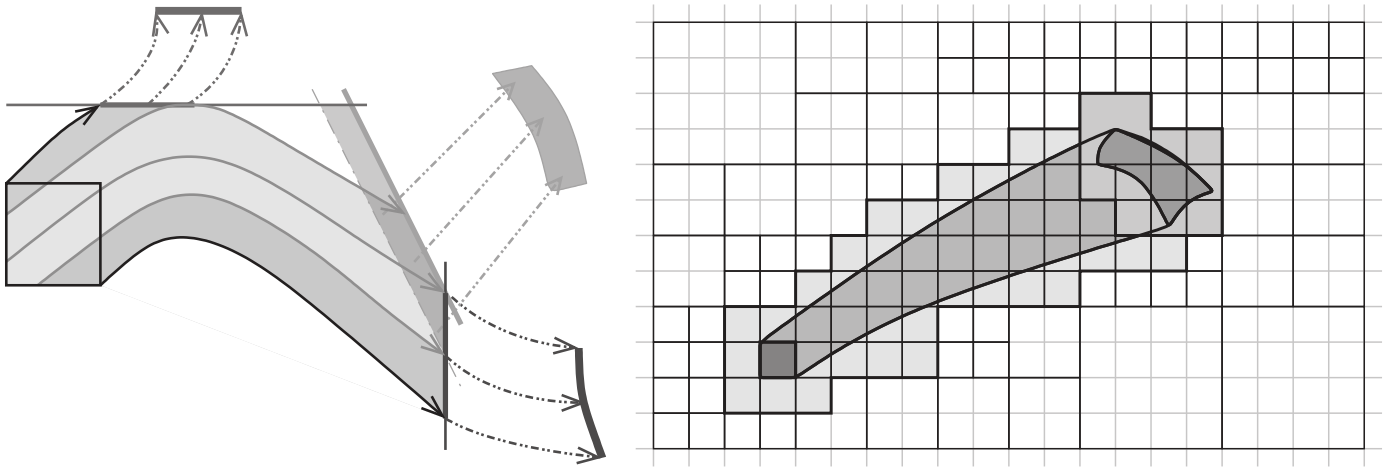


Figure 2 A Taylor model.



(a) (b)

Figure 3 (a) Behaviour of a hybrid system. (b) Discretising a flow tube.

tial equations, a Fourier basis of $\cos(n\pi x)$, $\sin(n\pi x)$ may be used [13].

Parameterised algebraic equations of the form $f(x,h(x))=0$ with $x\in D$ and $h(x)\in E$, are solved in Ariadne using a version of the efficient *interval Newton method* [36]. Differential equations of the form $\dot{\phi}(x,t)=f(\phi(x,t))$ can be solved in Ariadne either using an interval version of the *Picard operator* for the corresponding integral equation or by computing the Taylor series of the flow with error bounds [49].

In Ariadne, we provide two different kinds of sets. Accurate representations are given in terms of functions, with *constraint sets*

$$S=g^{-1}(C)=\{x\in\mathbb{R}^n \mid g(x)\in C\}$$

being regular/open, and *constrained image sets*

$$\begin{aligned} S &= h(D\cap g^{-1}(C)) \\ &= \{h(w) \mid w\in D \mid g(w)\in C\} \end{aligned}$$

being located/compact. The intersection of a constraint set with a constrained image set, or the image of a constrained image set, can both be realised using function composition. Coarser representations are given by *pavings*, which are unions of boxes with disjoint interiors, and support efficient intersection, union, and subset testing. An over-approximation of a constrained image set by a paving is computed be testing emptiness of sets $h(D\cap g^{-1}(C))\cap B$, which reduces to the constraint satisfaction problem $\exists\, z\in D{:}g(z)\in C\wedge h(z)\in B$.

We now have all the operators we need to verify hybrid systems in Ariadne. We start by considering computation of the evolution $\psi(x,t)$ from at a point x. The continuous dynamics can be computed by solving the differential equation in each mode $\dot{\psi}(x,t)=f_i(\psi(x,t))$, the times $\tau(x)$ of the discrete events can be computed by solving the equation $g_j(\psi(x,\tau(x)))=0$, and the discrete transitions can be computed by composition $\psi'(x,\tau(x))=r_k(\psi(x,\tau(x)))$. The only difficulty which may occur is in computing the crossing time, for if a trajectory $\psi(x,t)$ *grazes* a guard set $g_j(y)=0$, then the further evolution is discontinuous in the initial condition. In this case, we need to consider both the case that a transition occurs, and that a transition does not occur, leading to multivalued dynamics.

By computing the evolution $\psi(x,t)$ as a function of both the initial point and time, we can compute the *reach sets* $\psi(B,[0,t])$ starting from a set of initial states B over a time interval $[0,t]$, and the *evolved sets* $\psi(B,t)$ attained at time t, both as constrained image sets, as shown in Figure 3(a). To verify a safety property, we *partition* the state space X into boxes, and construct an over-approximation of the infinite-time reachable set by discretising the reach- and evolve-sets on this partition, as shown in Figure 3(b), continuing until no more reachable boxes are found. If safety has not been verified, we check for an unsafe evolution, or refine the partition. It can be shown that safety is verifiable if, and only if, the *chain-reachable set*, which is an over-approximation to the set of points which the state of the system may attain, is a subset of the safe set.

Implementation

At the beginning of the project, a decision was made to use C++ as the implementation language, as this is a powerful and efficient language widely used in industry. Later, I added an interface in the scripting language Python to facilitate interactive use of the tool. Considerable attention has been devoted to designing a clean and understandable interface; indeed, most of my time working on Ariadne has been spent on issues of interface design and software engineering rather than algorithm development.

To organise the data types, each class of objects has an underlying mathematical type, such as `Number`, `VectorFunction` or `CompactSet`, and a tag specifying to kind of *information* provided about the quantity. The information may be `Exact`, `Effective`, meaning arbitrarily-accurate approximations can be computed, `Validated`, which means finitely-accurate bounds can be computed, or `Approximate`, which means no information about accuracy is given. These abstractions are implemented by concrete classes, such as `FloatMPBounds` (an interval with multiple-precision floating-point bounds) implementing a `ValidatedNumber`. Although this separation is conceptually clean, how to implement it causes problems, especially regarding automatic conversions which are prevalent in C++. For although any `Effective` description contains more information than a 'Validated' description, so should be convertible to validated, any conversion requires a *precision* or *accuracy* parameter, which must either be given ex-

plicitly, or implicitly from the environment, neither solution being entirely satisfactory.

A major challenge in software engineering is due to the C++ language itself: there are many ways of supporting the data abstractions used in Ariadne, and different approaches require changes to the underlying code. Indeed, a 'good' language would be one in which the underlying numerical, functional and geometric kernels could be implemented in a clean way! Together with Michal Konečný, the developer of [27], another package for rigorous numerics, I have been experimenting with the use of the functional language Haskell in order to find a good semantics.

Since a model checking tool is supposed to prove correct functioning of a dynamic system, we also need to know that the tool itself works correctly, without mistakes in the mathematical algorithms or their implementation. Proven-correct code for real arithmetic exists in Coq [40] and Haskell [30], and for differential equations in Coq [32] and Isbell/HOL [24]. A C implementation of the solution of the wave equation was proved correct in [8], using several different tools including the Frama-C code analyser platform and the Coq theorem-prover. However, this is already a huge effort for a fairly simple C programme, and verified correct implementations of a tool such a Ariadne are a long way off! More promising from the point of view of provably-correct code, is to demonstrate correctness of the algorithms using a theorem-prover such as Coq or Isabelle, and automatically generate code in an efficient and user-friendly language such as the functional language ML.

The future of model checking

Practical use of model checking for large-scale hybrid systems arising in industry will require many advances in algorithm design. A particularly important technical challenge, and currently one of the major bottlenecks in Ariadne, is to improve the core algorithms for working with functions and sets, notably in simplifying the representation of sets while preserving accuracy of an over-approximation.

It will also be important to develop algorithms to handle uncertain systems efficiently. Ultimately, any mathematical analysis of a model of reality is only as good as the model itself, and whenever we make models, there are approximations built-in. Even fundamental physical constants, such as the fine structure constant $\alpha^{-1} = 137.0359991\,[39 \pm 31]$, is only known to finite precision. Systems arising in engineering have larger uncertainties (such as the exact resistence of an electrical component), and our knowledge of biological systems is even less precise. This means that in order to validate properties of real-life systems, we also need to take this uncertainty into account.

Probability is the most widely-used mathematical theory of uncertainty, but in many cases, we cannot assign exact probabilities to the events, and independence requirements needed for stochastic systems models are not met. We therefore need uncertainty models based on what is *possible*, leading to *differential inclusions* [4]. Ellipsoidal methods for their solution are described in [28] and higher-order methods in [19]. Although non-stochastic modelling is sometimes seen as being 'overly pessimistic', in the absence of more information about the system, it is the only way of guaranteeing correctness. Other frameworks, such as *imprecise probability* give more modelling flexibility, but will be even harder to reason about.

Perhaps the main challenge lies in detecting and exploiting system structure. In particular, since model checking has exponential time complexity in the state-space dimension, any methods to reduce the effective dimensionality, or split the problem into sub-problems of lower dimensionality, will be invaluable. Differences in time- and length-scales provide further opportunities for reduction by splitting into fast and slow behaviour, or averaging out spacial fluctuations. Modularity provides the greatest opportunities for dimension reduction, by splitting a system into parts which communicate via external inputs and outputs. Next-generation model checking tools will need to automatically exploit possible reductions in order to make fundamental breakthroughs in engineering practise.

Concluding remarks

The main goal of the Ariadne project was to develop a tool for model checking nonlinear hybrid automata, and the current version is capable of verifying well-conditioned, low-dimensional nonlinear systems. However, I feel that rather than immediately develop software for hybrid systems, it would have been better to first focus on the considerably simpler case of continuous-time systems. This would have allowed testing and showcasing the tool on industrially-relevant problems with considerably less technical demands, leading to more robust software and more opportunity to attract broader scientific interest. The idea that moving straight from discrete automata to hybrid systems may have been a challenge too far, and an incremental approach would have perhaps been better, is also seen in recent workshops explicitly focussing on continuous as well as hybrid systems.

My other main vision with Ariadne was to have general-purpose rigorous numerics package, and the current version does provide a broad interoperative functionality for numbers, functions and sets. Further, the combination of a solid theoretical foundation based on computable analysis, clean generic abstractions and extensible concrete data types is unique amongst rigorous numerical tools. My immediate goals for future development and to stabilise the interface, and sollicit user feedback on naming conventions, data abstractions and documentation.

I still find it odd that in the digital computer, we have a device of astonishing speed, accuracy and reliability, yet in mathematics, a subject justly proud of its application of rigour, most computations are approximative with no guarantees on correctness of the answer! Of course, there are reasons why this situation arose: for many problems, the approximations are usually good enough, speed is a bigger issue than accuracy, and fixed-size floating-point representations can be handled most efficiently with current hardware, and rigorous numerical algorithms are *much* harder to develop. Even so, existing work in rigorous numerics, still a rather niche field of research, shows that we can have the benefits of mathematical rigour with a reasonable efficiency. I believe that in the long-term, rigorous computations will eventually take over from approximate methods, just as formal rigorous mathematics eventually dominated. To facilitate this process, we need to provide tools that are powerful enough for industrial use, and simple enough to train undergraduate students in mathematics, science and engineering. Ariadne is one important step in this direction, and I plan to continue development work in the future to further realise these aims.

References

1 Rajeev Alur and David L. Dill, A theory of timed automata, *Theoretical Computer Science* 126 (1994), 183–235.

2 E. Asarin, T. Dang and O. Maler, The d/dt tool for verification of hybrid systems, in *Proc. 14th International Conference on Computer Aided Verification*, LNCS, Vol. 2404, Springer, 2002, pp. 365–370.

3 Eugene Asarin, Oded Maler and Amir Pnueli, Reachability analysis of dynamical systems having piecewise-constant derivatives, *Theoretical Computer Science* 138(1) (1995), 35–65, Hybrid Systems.

4 Jean-Pierre Aubin and Arrigo Cellina, Differential inclusions: Set-valued maps and viability theory, *Grundlehren der Mathematischen Wissenschaften*, No. 364, Springer, 1984.

5 Christel Baier and Joost-Pieter Katoen, *Principles of Model Checking*, MIT Press, 2008.

6 Andrej Bauer, *The Realizability Approach to Computable Analysis and Topology*, PhD thesis, Carnegie Mellon University, 2000.

7 Vincent D. Blondel, Olivier Bournez, Pascal Koiran, Christos H. Papadimitriou and John N. Tsitsiklis, Deciding stability and mortality of piecewise affine dynamical systems, *Theoretical Computer Science* 255(1) (2001), 687–696.

8 Sylvie Boldo, François Clément, Christophe Jean-Filliâtre, Micaela Mayero, Guillaume Melquiond and Pierre Weis, Wave equation numerical resolution: a comprehensive mechanized proof of a C program, *Journal of Automated Reasoning* 50(4) (2013), 423–456.

9 Xin Chen, Erika Ábrahám and Sriram Sankaranarayanan, Flow*: An analyzer for non-linear hybrid systems, in *Computer Aided Verification*, Springer, 2013, pp. 258–263.

10 E. Clarke, A. Fehnker, Z. Han, B. Krogh, J. Ouaknine, O. Stursberg and M. Theobald, Abstraction and counterexample-guided refinement in model checking of hybrid systems, *Internat. J. Found. Comput. Sci.* 14(4) (2003), 583–604.

11 Edmund M. Clarke, Orna Grumberg and Doron Peled, *Model Checking*, MIT Press, 1999.

12 Dirk van Dalen, Poincaré and Brouwer on intuition and logic, *Nieuw Archief voor Wiskunde* 5/13(3) (2012), 191–195.

13 S. Day, O. Junge and K. Mischaikow, A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems, *SIAM Journal on Applied Dynamical Systems* 3(2) (2004), 117–160.

14 M. Dellnitz, G. Froyland and O. Junge, The algorithms behind GAIO-set oriented numerical methods for dynamical systems, in *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pages 145–174. Springer, 2001.

15 Jan Duracz, Amin Farjudian, Michal Konečný, and Walid Taha, Function interval arithmetic, in *Mathematical Software–ICMS 2014*, Springer, 2014, pp. 677–684.

16 G. Frehse, PHAVer: algorithmic verification of hybrid systems past HyTech, *Int. J. Softw. Tools Technol. Trans.* 10 (2008), 263–279.

17 G. Frehse, C. Le Guernic, A. Donzé, S. Cotton, R. Ray, O. Lebeltel, R. Ripado, A. Girard, T. Dang and O. Maler, SpaceEx: Scalable verification of hybrid systems, in *Proc. 23rd International Conference on Computer Aided Verification*, LNCS, Vol. 6806, 2011, pp. 379–395.

18 Aleks Göllü and Pravin Varaiya, Hybrid dynamical systems, in *Proceedings of the 28th IEEE Conference on Decision and Control*, IEEE, 1989, pp. 2708–2712.

19 Sanja Gonzalez Živanovič and Pieter Collins, Higher order methods for differential inclusions, Technical Report MAC-1007, Centrum Wiskunde & Informatica, 2010.

20 Thomas A. Henzinger, Pei-Hsin Ho and Howard Wong-Toi, HyTech: A model checker for hybrid systems, *Software Tools for Technology Transfer* 1 (1997), 110–122.

21 Thomas A. Henzinger, Benjamin Horowitz, Rupak Majumdar and Howard Wong-Toi, Beyond HyTech: Hybrid systems analysis using interval numerical methods, in N. Lynch and B. Krogh, eds., *Hybrid Systems: Computation and Control*, LNCS, Vol. 1790, Springer, 2000, pp. 130–144.

22 Thomas A Henzinger, Peter W Kopke, Anuj Puri and Pravin Varaiya, What's decidable about hybrid automata? *Journal of Computer and System Sciences* 57 (1998), 94–124.

23 Gerard J. Holzmann, *The Spin Model Checker: Primer and Reference Manual*, Addison-Wesley, 2003.

24 Fabian Immler, A verified enclosure for the Lorenz attractor (rough diamond), in Christian Urban and Xingyuan Zhang, eds., *Proceedings of the 6th International Conference on Interactive Theorem Proving*, Springer, 2015, pp. 221–226.

25 Luc Jaulin, Michel Kieffer, Olivier Didrit and Éric Walter, *Applied Interval Analysis*, Springer, 2001.

26 Ker-I Ko, Complexity theory of real functions, *Progress in Theoretical Computer Science*, Birkhäuser, 1991.

27 Michal Konečný, Aern-rntorm: Arbitrary-precision arithmetic of multivariate piecewise polynomial enclosures, 2008.

28 Alexander B. Kurzhanski and Istvåan Vályi, Ellipsoidal calculus for estimation and control, *Systems & Control: Foundations & Applications*, Birkhäuser, 1997.

29 Kim G. Larsen, Paul Pettersson and Wang Yi, UPPAAL in a nutshell, *Intern. J. Software Tools for Technology Transfer* 1(1–2) (1997), 134–152.

30 David R. Lester, The world's shortest correct exact real arithmetic program? *Information and Computation* 216, (2012), 39–46. Special Issue: 8th Conference on Real Numbers and Computers.

31 R.J. Lohner, AWA – software for the computation of guaranteed bounds for solutions of ordinary initial value problems, Technical report, Institut für Angewandte Mathematik, 2006.

32 Evgeny Makarov and Bas Spitters, The =Picard algorithm for ordinary differential equations in Coq, in *Interactive Theorem Proving*, Springer, 2013, pp. 463–468.

33 K. Makino and M. Berz, COSY INFINITY Version 9, *Nuclear Instruments and Methods* A558 (2006), 346–350.

34 Kyoko Makino and Martin Berz, Taylor models and other validated functional inclusion methods, *Int. J. Pure Appl. Math* 4(4) (2003), 379–456.

35 R. Moore, *Interval Analysis*, Prentice-Hall, 1966.

36 Ramon E. Moore, R. Baker Kearfott and Michael J. Cloud, *Introduction to Interval Analysis*, SIAM, 2009.

37 MPFR Library, 2000, www.mpfr.org.

38 M. Mrozek et al., CAPD Library, 2007.

39 Ned Nedialkov, VNODE-LP, Technical report, McMaster University, 2006. Technical Report CAS-06-06-NN.

40 Russell O'Connor, Certified exact transcendental real number computation in Coq, in Otmane Ait Mohamed, César Muñoz, and Sofiéne Tahar, eds., *Proceedings of the 21st International Conference on Theorem Proving in Higher Order Logics*, Springer, 2008, pp. 246–261.

41 Philippos Peleties and Raymond DeCarlo, A modeling strategy with event structures for hybrid systems, in *Proceedings of the 28th IEEE Conference on Decision and Control*, IEEE, 1989, pp. 1308–1313.

42 S. Ratschan and Z. She, Safety verification of hybrid systems by constraint propagation based abstraction refinement, *ACM Trans. Embedded Comput. Sys.* 6(1), 2007.

43 S.M. Rump, INTLAB – INTerval LABoratory, in Tibor Csendes, ed., *Developments in Reliable Computing*, Kluwer Academic Publishers, 1999. www.ti3.tuhh.de/rump, pp. 77–104.

44 Keijo Ruohonen, An effective Cauchy-Peano existence theorem for unique solutions, *Int. J. Found. Comput. Sci.* 7(2) (1996), 151–160.

45 Arjan van der Schaft and Hans Schumacher, An introduction to hybrid dynamical systems, *LNCIS*, Vol. 251, Springer, 2000.

46 Matthias Schröder, *Admissible Representations for Continuous Computations*, PhD thesis, Fachbereich Informatik, FernUniversität Hagen, 2002.

47 A.M. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proc. London Math. Soc. (2)* 43(1), 1937, 230–265.

48 Klaus Weihrauch, Computable analysis – An introduction, *Texts in Theoretical Computer Science*, Springer, 2000.

49 Daniel Wilczak and Piotr Zgliczyński, Cr-Lohner algorithm, *Schedae Informaticae* 20 (2011), 9–42.

Jan Beuving
cabaretier, Zeist
janbeuving@gmail.com

Het keerpunt van Jaap van den Herik

Schaken is een heel serieus beroep, geen spel

Jaap van den Herik studeerde wiskunde in Amsterdam, promoveerde in Delft op onderzoek naar computerschaak en werd later hoogleraar informatica in Maastricht en Tilburg en hoogleraar recht en informatica aan de Universiteit van Leiden. Hij begeleidde in die jaren meer dan zeventig aio's. "Als je echt een goeie wiskundige wilt zijn, moet je er dag en nacht mee bezig zijn."

Schaakt u nog weleens?
"Zelden. Ik ben daar eigenlijk in 1987 mee opgehouden toen ik hoogleraar in Maastricht werd. Ik heb nog wel jarenlang een computerschaakrubriek gehad in *Schakend Nederland*, en ik was tot afgelopen december hoofdredacteur van het tijdschrift van de International Computer Games Association.

De reden dat ik niet meer schaak is mijn ambitie: ik wil graag winnen. Ik schaak niet om te verliezen, ik wil niet afbouwen op een bijveld. Als ik meedoe, wil ik op het hoofdveld spelen, maar daar heb ik geen tijd voor."

Waar bent u wel mee bezig?
"In 2012 hebben Jos Vermaseren, Aske Plaat en ik van het European Research Council een zogeheten Advanced Grant gekregen. Jos Vermaseren is de hoofdonderzoeker, en heeft mij gevraagd om op de bagagedrager te zitten, maar het ligt in mijn aard om mee te trappen. Vermaseren is aan het NIKHEF verbonden, en een van de meest onderschatte wetenschappers die ik ken. Hij heeft onder andere het programma FORM geschreven, waarmee een computer grote, ingewikkelde formules snel kan verwerken. FORM wordt nog steeds veel gebruikt, maar na twintig jaar was die taal verouderd, en was nieuw onderzoek nodig. Daar hebben we een voorstel voor geschreven, en in dat kader begeleid ik nu twee aio's."

Doet u nog veel wiskunde?
"Ja en nee. Het onderzoek dat ik nu doe is veelal toegepast, en daar komt natuurlijk veel wiskunde bij kijken. Als je bijvoorbeeld dit artikel over 'Nonparametric Bayesian Line Detection' bekijkt [de interviewer krijgt iets erg ingewikkelds voorgelegd, JB], zie je dat de formules je om de oren vliegen. Maar als je echt een goeie wiskundige wilt zijn, moet je er dag en nacht mee bezig zijn. Dat ben ik niet. Ik ben dag en nacht bezig met informatica en intelligente systemen. Dit soort artikelen zijn daar uitlopers van, maar niet meer mijn hoofdonderzoek."

Jaap van den Herik

Toch bent u ooit wel voor de wiskunde gevallen. Waarom ging u het destijds studeren?
"Ik woonde als kind in Rotterdam, waar ik naar het Marnix Gymnasium ging. Ik had in mijn eigen stad economie kunnen gaan studeren, maar ik wist zeker dat mijn vader dan zou zeggen: dan kun je wel thuis blijven wonen. Dat was precies wat ik niet wilde.

Ik was behalve in wiskunde ook goed in scheikunde, en ging daarom kijken in Delft en in Amsterdam. Ik zag al snel dat je bij scheikunde slaaf was van je agenda, omdat je altijd afhankelijk was van de laboratoriumuren om je onderzoek te doen. Wiskunde kan dag en nacht, en dat bood ruimte voor wat ik echt wilde: schaken. Amsterdam was in die dagen het schaakcentrum van Nederland, daar wilde ik naartoe. Toen is het dus wiskunde geworden, een studie die ik makkelijker om mijn schaakpartijen heen kon plooien."

Wat stond er voor u op één? Schaken of studeren?
"Studeren stond op één, schaken op twee. Maar, ik sloot niet uit dat ik, als

ik goed genoeg was, die volgordes om wilde draaien. Toen ik in Amsterdam ging studeren kreeg ik Rien Kaashoek als mentor toegewezen. Hij was een strenge maar begripvolle man. Toen ik mijn eerste jaar met succes binnen de gestelde normen had doorlopen, zei ik tegen hem dat ik graag een jaar fulltime wilde schaken. 'Als je dat wilt, dan moet je dat doen', zei hij toen. Hij zorgde dat ik niet in militaire dienst hoefde en dat mijn beurs behouden bleef, als ik mijn kandidaats dan maar wel binnen de gestelde vier jaar zou halen. En, hij vond dat ik wel iets moest bereiken: 'Je moet wel werken, geen bier drinken!' 'Professor, dat is afgesproken', zei ik. Een jaar later, in 1968, won ik het Nederlands Studentenkampioenschap. Vervolgens werd ik uitgezonden naar de Studenten Wereldkampioenschappen in Ybbs, Oostenrijk, maar daar haalde ik nul punten uit de eerste drie partijen, en vervolgens werd het de C-groep. Eindresultaat was 4,5 uit 10. Ik was wel goed, maar niet top. Ik concludeerde dat schaken een heel serieus beroep is, geen spel. Je moet, net als in de wiskunde, meters maken. En talent hebben natuurlijk.

Na een jaar vroeg Kaashoek aan me hoe het ging, waarop ik antwoordde: 'Maakt u zich geen zorgen professor, per september studeer ik weer!' "

Waar ging uw afstudeerwerk over?
Spectra van differentiaaloperatoren, een onderwerp uit de numerieke analyse. Dat was een vakgebied dat destijds vijftig jaar jong was. Ik hield me bezig met het opsporen van defectruimtes, en de dimensies daarvan. Ik deed het onderzoek bij Gerke Nieuwland, die mij na het voltooien van de scriptie nog een aanstelling van bijna een jaar gaf om door te werken aan mijn afstudeeronderzoek. 'Maar,' zo zei hij, 'je kunt hier niet promoveren. Je moet niet afstuderen, promoveren en hoogleraar worden op één universiteit. Dat is de pest voor het wetenschappelijk onderzoek! Je moet je op meer plekken ontplooien, en overal wat van meenemen.' Ik ben hem tot op de dag van vandaag dankbaar voor dat advies, en ik heb het ook altijd aan al mijn eigen studenten gegeven."

En dus ging u naar Delft.
"Ja, maar als wetenschappelijk medewerker; in die tijd was promoveren geen usance. Mijn aanstelling was tweeledig; ik gaf onderwijs in analyse en lineaire algebra, en de andere helft van de tijd mocht ik het splinternieuwe informaticaonderzoek op poten zetten. Dat vond ik geweldig.

Op zekere dag was er een voormalig student, ir. Barend Swets, die mee wilde doen aan de Europese Computerschaakkampioenschappen. Hij wilde zijn programma ergens draaien, en aan mij werd gevraagd of ik dat wilde begeleiden. Door hem ben ik echt geïnteresseerd geraakt in computerschaak. Ik besloot me te concentreren op eindspelen, omdat de computers daar destijds dramatisch slecht in waren. Een computer kon niet eens mat zetten met koning en dame tegen koning. Later heb ik technieken ontwikkeld om mat zetten met loper en paard tegen koning te formaliseren, en mijn promotieonderzoek aan het schaken gewijd. Dat was het moment dat ik van mijn hobby mijn beroep heb gemaakt, en van mijn beroep mijn hobby."

Totdat u naar Maastricht ging.
"Ik werd gevraagd om daar de informaticatak op poten te zetten. Dat heb ik met heel veel liefde gedaan. In Maastricht werden we uiteindelijk twee jaar op rij tot beste kunstmatige intelligentieopleiding van Nederland verkozen — onderwijs én onderzoek — en dat vond ik met mijn competitieve karakter natuurlijk leuk. Na twee jaar werd die onderscheiding afgeschaft. Ondertussen was ik ook in Leiden bijzonder hoogleraar informatica en recht geworden. Ook dat was een heel nieuw onderzoeksveld in die tijd."

Kan een computer rechtspreken?
"Dat doen ze al. Bij bijvoorbeeld alimentatiezaken en nationalisatieverzoeken volgt de rechter beslissingen van de computer. Behalve bij Máxima natuurlijk, dat moest buiten het programmeren om."

U heeft dankzij de opkomst van het computertijdperk veel pionierswerk kunnen doen.
"Dat was fantastisch om mee te maken. Je moet wel tegen kritiek kunnen. Ik ben mijn hele leven tegengesproken. Het was idioot wat ik deed, het zou niet kunnen. Toen ik voorspelde dat computers sterker zouden gaan spelen dan de beste grootmeesters werd ik voor gek verklaard. Ik heb toch gelijk gehad."

Kunnen computers slimmer worden dan mensen? Of alleen maar beter?
"Nu zijn ze alleen maar beter. Maar ik ben ervan overtuigd dat ze ook slimmer worden dan wij. Ik riep altijd dat dat in 2080 het geval zou zijn. Dat is nog 65 jaar. Het zou ook nog 165 jaar kunnen duren, pin me daar niet op vast, maar ik weet zeker dat het gaat gebeuren. Het is een kwestie van tijd voor een computer ook ethische beslissingen kan nemen."

Maar zal er niet altijd een mens nodig zijn? Een computer kan zich bijvoorbeeld niet voortplanten.
"Waarom zou een computer dat niet kunnen? Hij kan een nieuwe versie van zichzelf bouwen!"

Is dat geen angstaanjagende gedachte?
"Nee. Maar ik ben een optimist. Kijk, mieren hebben ook alle reden om bang te zijn voor mensen. Mensen zijn veel groter en sterker. Toch hebben ze betrekkelijk weinig last van ons. Een kind zal eens een mierenhoop slopen, maar verder kunnen we naast elkaar bestaan. Als die computer echt zo verstandig wordt, laat hij de mensen lekker met rust."

Wordt de computer niet eerder slim dan ethisch volwassen?
"Ja, ik denk dat hij eerder slim zal zijn. Daarom moeten we het ook goed in de gaten houden. De UvA heeft nu al een *ethical board* om over dit soort problemen na te denken."

Welke ontwikkeling hoopt u nog mee te maken?
"Ik hoop allereerst op een doorbraak in het proefschriftonderzoek van Ben Ruijl, een van mijn aio's. Ik heb me altijd verheugd in de successen van mijn promovendi. Daar doe ik het voor."

Goede suggesties voor een Nederlandse wiskundige met een keerpunt in zijn of haar carrière zijn welkom via keerpunt@nieuwarchief.nl.

Wim Caspers
Lyceum Ypenburg, Den Haag, en
Faculteit EWI en Lerarenopleiding, TU Delft
w.t.m.caspers@tudelft.nl

Onderwijs Bespreking examen vwo wiskunde B 2016

Moeilijk of ongemakkelijk?

Het eindexamen wiskunde B op het vwo is alweer een poos geleden afgenomen. Inmiddels zijn de geslaagde kandidaten getransformeerd tot studenten en bevolken ze de collegezalen. Waar zijn ze destijds aan blootgesteld? De N-term, een getal dat vastlegt hoe uit het aantal scorepunten het cijfer bepaald wordt en bedoeld is voor het compenseren van verschillen in moeilijkheidsgraad van examens door de jaren heen, werd vastgesteld op $2{,}0$. Om de gedachte te bepalen: met N-term $1{,}0$ levert een score van 50 procent een $5{,}5$ op als cijfer en met N-term $2{,}0$ levert het een $6{,}5$ op. Wim Caspers bespreekt een aantal moeilijkheden die wellicht hebben bijgedragen aan deze hoge N-term.

Blijkbaar was het examen 2016 moeilijk. Dat wordt ook bevestigd door het overzicht in de WiskundE-brief waarin scores en cijfers van de examens van de afgelopen zeven jaar op een rij gezet zijn [1]. De score is met 55 procent de laagste van de afgelopen jaren. De hoge N-term levert een gemiddeld cijfer van $6{,}9$ op, wat in lijn is met de afgelopen jaren. De WiskundE-brief meldt ook dat een grote meerderheid van de docenten het examen te lang en moeilijk of te moeilijk vond. Maar eigenlijk is ongemakkelijk meer van toepassing dan moeilijk. Hieronder staat een aantal opgaven om dat te illustreren.

Kettinglijn

De openingsopgave had als onderwerp de kettinglijn, beschreven door

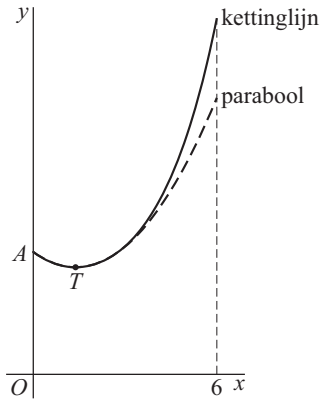
$$f(x) = \tfrac{1}{2}e^{\frac{1}{2}x} + 2e^{-\frac{1}{2}x} + 1\tfrac{1}{2}$$

op het domein $[0, 6]$. Gevraagd werd exact de waarde van de x-coördinaat van het laagste punt van de grafiek te berekenen. Het bepalen van de nulpunten van de afgeleide ging de meeste leerlingen goed af. Bij het tweede onderdeel werd gevraagd om in de situatie dat de kettinglijn de vorm beschrijft van een kabel tussen twee palen te bepalen of een aan een kant losschietende kabel de grond raakt. Duidelijk is dat de lengte van de kabel berekend moet worden. De bijbehorende integraal mag benaderd worden met de grafische rekenmachine, maar dat is een heel gedoe om foutloos in te voeren. In het laatste onderdeel van de opgave komt een benadering met een parabool aan de orde. De vraag is in Figuur 1 weergegeven.

De vraag is niet moeilijk, maar voelt ongemakkelijk aan. Om te beginnen is de vorm van een vergelijking van de parabool gegeven: $y = a(x-b)^2 + c$. Dat is beslist handig als je beseft dat het snijpunt met de y-as en de top van de parabool bekend zijn. Maar de vorm $y = ax^2 + bx + c$ is populairder, misschien wel vanwege de abc-formule waarin ook de x-coördinaat van de top verstopt zit. In de vraagstelling komt die vergelijking echter niet meer voor, dus welke vorm je gebruikt is vrij gelaten. Dat soort vrijheid zijn leerlingen — moet helaas gezegd worden — niet gewend en velen zullen verwacht hebben als vraag: "Bereken a, b en c." In plaats daarvan wordt gevraagd naar de plek waar het hoogteverschil gelijk is aan 1 (waarom 1 gekozen is wordt overigens niet duidelijk). Daarbij dient weer een beroep gedaan te worden op de grafische rekenmachine, compleet met tijdrovend invoergedoe.

In figuur 3 zijn de grafiek van de functie f en de parabool door A met top T getekend.
In deze figuur is te zien dat de parabool de kettinglijn aanvankelijk goed benadert, maar dat voor grotere waarden van x de benadering minder goed wordt.
Van de parabool door A met top T kan een vergelijking van de vorm $y = a(x-b)^2 + c$ worden opgesteld.

figuur 3



6p **3** Bereken de waarde van x waarvoor het (verticale) hoogteverschil tussen de kettinglijn en deze parabool gelijk is aan 1. Rond je antwoord af op één decimaal.

Figuur 1 Het laatste onderdeel van de eerste opgave van het examen.

Afleiding bij het differentiëren

En dat was dan de eerste opgave. Als je kijkt naar de oplossing, dan is die niet moeilijk te noemen, maar voor een openingsopgave kent hij wel veel ongemakkelijke elementen. Menig leerling zal op de klok hebben gekeken en geconstateerd hebben dat er al veel tijd verstreken was. De tweede opgave zal niet veel minder tijd gekost hebben. Die had als onderwerp de automotor en in het bijzonder de beweging van de zuiger. Een van de bijbehorende illustraties is in Figuur 2 weergegeven. De afstand van D tot B wordt s genoemd. Het eerste onderdeel blijkt achteraf het best gemaakte onderdeel van het examen; laten zien dat in de linker situatie geldt:

$$s = 5 - \cos(\alpha) - \sqrt{16 - \sin^2(\alpha)}$$

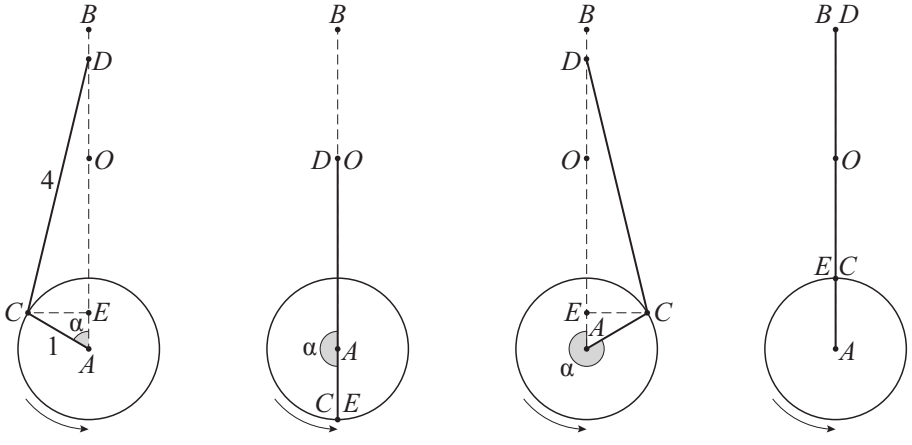
met $0 \le \alpha \le 2\pi$. Vervolgens wordt s benaderd met de formule

$$z = 1 - \cos(\alpha) + \tfrac{1}{8}\sin^2(\alpha)$$

en er wordt gevraagd naar het maximale verschil. Weer een opgave met gepriegel in de grafische rekenmachine om het antwoord $0{,}002$ te vinden.

In het correctievoorschrift worden doorgaans punten toegekend voor het aangeven hoe het antwoord gevonden is met de grafische rekenmachine. In dit geval is dat nog wel zinvol, omdat een kandidaat dan geacht wordt te praten over $|s - z|$. Bij het benaderen van de integraal is het toekennen van een punt voor het vertellen dat er een formule ingevoerd is en op de knop integreren of iets dergelijks is gedrukt misschien wat minder op zijn plaats.

In het laatste onderdeel van deze opgave werd het volgende gevraagd: "Stel een formule voor de afgeleide van z op en bereken hiermee de maximale zuigersnelheid. Rond je antwoord af op twee decimalen." Het is voor minder oplettende leerlingen verleidelijk om nulpunten van de gevonden afgeleide te zoeken in plaats van extremen van die afgeleide. Maar afgezien daarvan dringt zich een ongemakkelijk gevoel op gezien de context, die afleidend kan werken. Ik ben geen expert in automotoren en raadpleeg dus graag vakmensen, in dit geval mijn broer die immers gepromoveerd is aan de TU Eindhoven, faculteit Werktuigbouwkunde. Zijn reactie is weergegeven in het kader op de volgende bladzijde. Elegant laat hij de conclusie over wat er wel en niet in een wiskunde-examen hoort over aan wiskundigen. Duidelijk is wel de bedoeling van de opgave: bepaal het maximum van de afgeleide naar α van z. Als z geïntroduceerd zou zijn als een functie $z(\alpha)$ was dit al iets duidelijker geweest. Door te spreken over de maximale zuigersnelheid wordt de context onverbiddelijk de opgave ingetrokken, lijkt de variabele tijd een rol te gaan spelen ($\alpha(t)$, α als functie van de tijd) en zouden er vereenvoudigende veronderstellingen gedaan moeten worden. Geen prettige opgave en toch gemiddeld gemaakt, blijkt achteraf.



Figuur 2 Illustratie bij de tweede opgave van het examen.

Bijzondere formuleringen

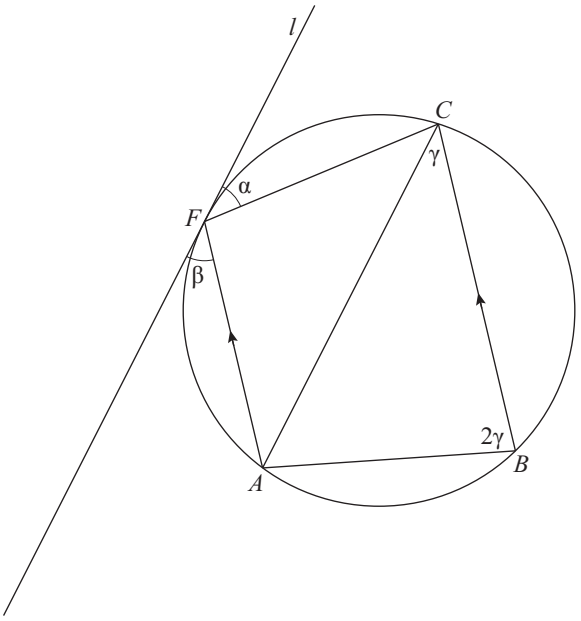
Na deze eerste twee opgaven waren de grootste problemen wel achter de rug, maar het zal een doorsnee leerling veel tijd hebben gekost. De meetkundeopgaven waren overzichtelijk, tenminste wanneer je de zin "Z is het snijpunt van de lijn door C en het snijpunt E van MD en AB met de cirkel" kunt verteren. Het lukte wonderwel. Het onderdeel waarvan de illustratie in Figuur 3 is weergegeven, leverde veel meer problemen op. Het is het op twee na slechtste onderdeel van het examen, misschien omdat veel leerlingen er niet aan toegekomen zijn? Het is ook een opgave waar meer informatie in zit dan er gevraagd wordt. Lijn l is evenwijdig aan AC, maar α, β en nog een paar hoeken zijn gelijk aan γ. Wie weet heeft dat verwarrend gewerkt.

Zuigersnelheid

Het bepalen van een afgeleide van z is per definitie onduidelijk, omdat niet verteld wordt wat voor afgeleide bedoeld wordt: naar de tijd, naar hoek α, of naar... De te gebruiken term zou mijns inziens tijdsafgeleide moeten zijn, of afgeleide naar α, wat dan ook bedoeld wordt. Omdat expliciet over zuigersnelheid wordt gesproken, zal wel de tijdsafgeleide bedoeld zijn. Voor de bepaling van de tijdsafgeleide moet er een verband bekend zijn of verondersteld kunnen worden tussen α en de tijd die in de formule voor z als functie van α kan worden ingevuld. Omdat de opgave een verbrandingsmotor betreft, kan niet zonder meer gesteld worden dat de rotatiesnelheid ($\frac{d\alpha}{dt}$) constant is. De rotatieas (krukas in termen van een verbrandingsmotor) is doorgaans voorzien van een vliegwiel, dat mede zorgdraagt voor een rustige loop van de motor. Ondanks dit vliegwiel zal de rotatiesnelheid afnemen tijdens de zogenaamde compressieslag (als het brandstofmengsel boven de zuiger in de cilinder wordt gecomprimeerd), om, na ontsteking van het brandstofmengsel boven de zuiger weer toe te nemen. Dit geldt voor brandstofmotoren van zowel het tweetakt- als het viertaktprincipe. Vrijwel altijd beschikt een brandstofmotor van een auto over méér dan één cilinder, in het algemeen tussen de 2 en de 8. De betreffende zuigers zijn verbonden aan dezelfde krukas, waarmee de snelheid van de rotatie-as weliswaar gelijkmatiger, maar nog altijd niet constant is. Ook zijn er invloeden van de verdere aandrijfketen te verwachten op het snelheidsverloop als gevolg van bijvoorbeeld elastische en dempingseigenschappen, maar deze worden evenmin duidelijk in de opgave.

Al deze en overige invloeden zijn niet in de opgave beschreven. Verdisconteering hiervan zou ver uitgaan boven de beoogde moeilijkheidsgraad van de opgave. Maar ook is geen veronderstelling gedaan in de opgave over de afwezigheid van variatie in de rotatiesnelheid. Dit wordt aan de leerling overgelaten.

Wiskundigen mogen beoordelen of een dergelijke veronderstelling van een leerling verwacht mag worden tijdens een wiskunde-examen.



5p **14** Bewijs dat l evenwijdig is aan AC.

Figuur 3 Het op twee na slechtst scorende onderdeel van het examen.

Iets soortgelijks geldt misschien voor de slechtst scorende opgave, die over raaklijnen aan de twee parabolen $y = x^2 + 3$ en $y = -x^2 - 1$. Er is veel minder informatie nodig om de opgave tot een goed einde te brengen dan op het eerste gezicht lijkt. Het helpt wel als je weet dat het gemiddelde van -1 en 3 gelijk is aan 1 en niet 2, maar dan is het een kwestie van het bepalen van een raaklijn aan een parabool door het punt $(0, 1)$. Het is een standaardopgave waarbij veel leerlingen een recept kunnen toepassen, ware het niet dat er nu opeens twee grafieken en twee raaklijnen in het spel zijn. Het op een na slechtst scorende onderdeel heeft u nog tegoed, dat was het laatste onderdeel. Tegen de tijd dat men daar aanbeland was bleek vermenigvuldigen ten opzichte van de x-as niet meer onderscheiden te kunnen worden van vermenigvuldigen ten opzichte van de y-as, want dat werd gevraagd. Overigens met de wonderlijke toevoeging; "Schrijf je antwoord als één breuk."

Mag je een breuk opschrijven met in de noemer een breuk, mag je $\frac{1}{\sqrt{2}}$ in je antwoord laten staan en wat te doen als iemand $\ln(1)$ niet ontmaskert als 0? Discussies daaromtrent zijn er nog wel in examentijd maar gelukkig is die laatste toevoeging, schrijf als één breuk, niet typerend voor dit examen of het bijbehorende correctievoorschrift. Je mag immers verwachten dat docenten die bijna dagelijks correctiewerk verrichten goed in staat zijn om zelf beslissingen te nemen als het gaat om dergelijke details.

Wiskundige denkactiviteiten

Veel belangrijker is het dat het examen leerlingen de mogelijkheid biedt om te laten zien wat ze met wiskunde voor elkaar kunnen krijgen. Dat kan doordat ze opgaven herkennen (de opgaven die hier niet besproken zijn), maar ook door ongemakkelijkere opgaven die meer een beroep doen op inzicht en begrip van de stof aan te pakken. Het examen 2016 bood die mogelijkheid dus zonder meer, maar met beide categorieën in verrassende volgorde. Het gaat waarachtig richting wiskundige denkactiviteiten. In het nieuwe examenprogramma staan die voor het eerst expliciet benoemd, wiskundige vaardigheden waaronder modelleren, ordenen en structureren, analytisch denken en probleemoplossen. Het zou een verschuiving in de soort examenopdrachten teweeg moeten brengen, dus misschien heeft het examen 2018 nog veel meer verrassingen in petto. Eerst 2017 nog.

Referentie

1 WiskundE-brief 746, 3 juli 2016. Zie http://www.wiskundebrief.nl/746.htm.

Problemen

| Problem Section

This Problem Section is open to everyone; everybody is encouraged to send in solutions and propose problems. Group contributions are welcome.
For each problem, the most elegant correct solution will be rewarded with a book token worth € 20. At times there will be a Star Problem, to which the proposer does not know any solution. For the first correct solution sent in within one year there is a prize of € 100. When proposing a problem, please either include a complete solution or indicate that it is intended as a Star Problem.
Please send your submission by e-mail (LaTeX is preferred), including your name and address to problems@nieuwarchief.nl.
The deadline for solutions to the problems in this edition is 1 December 2016.

Problem A (folklore)

Show that a group G is torsion-free if and only if for all integers $n \geq 2$ and finite subsets $S, T \subseteq G$ with $\#S = \#T = n$ we have $\#\{st : s \in S, t \in T\} > n$.

Problem B (proposed by Hendrik Lenstra)

Show that for all groups G the commutator subgroup $[G,G] = \langle\, xyx^{-1}y^{-1} : x, y \in G \,\rangle$ of G has order at most 2 if and only if every conjugacy class in G has at most 2 elements.

Problem C (proposed by Carlo Pagano and Mima Stanojkovski)

A subgroup H of a group G is said to be *solitary* if no other subgroup of G is isomorphic to H. A group G is said to be *totally solitary* if all of its subgroups are solitary. Show that a group G is totally solitary if and only if it is isomorphic to a subgroup of $\mathbb{Q}/\mathbb{Z}$.

Edition 2016-1 We received solutions from Raymond van Bommel, Rik Bos, Alexandros Efthymiadis, Pieter de Groen, Alex Heinis, Thijmen Krebs, Hendrik Reuvers, Toshihiro Shimizu, Djurre Tijsma and Martijn Weterings.

Problem 2016-1/A (proposed by folklore)

Let $a, b > 1$ be integers such that $2a \leq b$. Does there exist a map

$$f : \{a, a+1, \ldots, b\} \to \{a, a+1, \ldots, b\}$$

without fixed points, such that for all $n \in \{a, a+1, \ldots, b\}$ we have $f^{f(n)}(n) = n$? Here, for a positive integer k, f^k denotes the k-fold composition

$$\underbrace{f \circ f \circ \cdots \circ f}_{k \text{ times}}$$

of f.

Solution We received solutions from Raymond van Bommel, Rik Bos, Alexandros Efthymiadis, Pieter de Groen, Alex Heinis, Thijmen Krebs, Toshihiro Shimizu and Martijn Weterings. The book token is awarded to Rik Bos. Most of the solutions received are similar, and the following one is based on that.
The answer is no. Suppose for a contradiction that such an f does exist.
First note that f is surjective and therefore a permutation of $\{a, a+1, \ldots, b\}$; write f as a product of disjoint cycles. Note that the cycle containing n is also the cycle containing $f^{-1}(n)$, and by $f^n(f^{-1}(n)) = f^{f^{-1}(f(n))}(f^{-1}(n)) = f^{-1}(n)$, we see that this cycle has length dividing n. It follows that the length of a cycle divides every element contained therein.

Redactie:

Gabriele Dalla Torre
Christophe Debry
Jinbi Jin
Marco Streng
Wouter Zomervrucht

Problemenrubriek NAW
Mathematisch Instituut
Universiteit Leiden
Postbus 9512
2300 RA Leiden

problems@nieuwarchief.nl
www.nieuwarchief.nl/problems

Oplossingen | Solutions

Moreover, as f has no fixed points, every cycle has length greater than 1.
By Bertrand's postulate, there exists a prime p such that $\frac{1}{2}b < p < b$. Then note that the cycle containing p has length p, and that therefore every element in this cycle is divisible by p. Therefore this cycle contains an element that is at least $2p$ and therefore greater than b; this is the required contradiction.

Problem 2016-1/B (proposed by folklore)

Let $n \geq 3$ be an integer. Two players play the following game. Starting with a sheet of paper with the numbers 1 and 2 on them, the players take turns writing down a new number from 1 to n that is the sum of two numbers already on the sheet. The player who writes down the number n wins.
For which n does the first player have a winning strategy?

Solution We received solutions from Raymond van Bommel, Pieter de Groen, Alex Heinis and Thijmen Krebs. The book token is awarded to Thijmen Krebs. The following solution is based on that of Thijmen Krebs and that of Raymond van Bommel.
Consider the equivalent game where the first player starts by writing 1 on an empty sheet. We say a number $1 \leq m < n$ is *safe* if m and $n-m$ are distinct and not on the sheet. Note that the first player who writes down an unsafe number loses, as the next player can write down the number n in his next turn. We show that the first player wins if and only if $n \equiv 3, 4 \bmod 4$.
The strategy for the winning player is simply to write down the smallest safe number, if it exists. This is always possible, as we show below.
Suppose that at any point, the number m is the smallest safe number. If $m = 1, 2$, then it is clear that m can be written down. So suppose that $m \geq 3$. Then at least $m-1$ turns have passed, so there are at least $m-1$ numbers on the sheet, and as the winning player has written down the smallest safe number in each of his previous turns, he has written down at least $\frac{1}{2}(m-1)$ numbers that are at most $m-1$. Moreover, as $m \geq 3$, the losing player has written down at least one number that is at most $m-1$, namely 1 or 2. Therefore there are at least $\frac{1}{2}(m+1)$ numbers on the sheet that are at most $m-1$. By the pigeonhole principle, among these there are at least 2 (possibly the same) that add up to m, so m can be written down.
Now note that with this strategy, the number of safe numbers decreases by 2 every turn (or the other player loses immediately). As the number of safe numbers in the beginning is $2\left\lfloor \frac{1}{2}(n-1) \right\rfloor$, which is 2 modulo 4 if and only if $n \equiv 3, 4 \bmod 4$, we find that the first player can win if and only if $n \equiv 3, 4 \bmod 4$.

Problem 2016-3/C (proposed by Hendrik Lenstra)

Determine all two-sided infinite sequences of positive integers in which each number is the Euler-phi of the next.

Solution We received solutions from Raymond van Bommel, Alex Heinis, Thijmen Krebs, Hendrik Reuvers, Toshihiro Shimizu, Djurre Tijsma and Martijn Weterings. The book token is awarded to Raymond van Bommel, whose solution the following one is based on.
We will make use of the following. Let ϕ denote the Euler-phi function, and let $v_2(a)$ denote the number of prime factors 2 in the prime factorisation of a positive integer a.

Lemma. *Let a be a positive integer. If $\phi(a)$ contains a prime factor of at least 5, then so does a. Moreover, in this case we have $v_2(\phi(a)) \geq v_2(a)$, with equality if and only if a is of the form $2^e p^f$ with $p \geq 5$ prime congruent to 3 modulo 4.*

Proof. If a does not contain a prime factor of at least 5, then neither does $\phi(a)$. If $a = 2^e p_1^{f_1} \cdots p_s^{f_s}$ with $p_1, \ldots, p_s$ distinct odd primes and $f_1, \ldots, f_s > 0$, then $v_2(\phi(a)) = e + \sum_{i=1}^{s} v_2(p_i - 1) - 1$. As for all i we have $v_2(p_i - 1) > 0$, it follows that $v_2(\phi(a)) \geq v_2(a)$, with equality if and only if $s = 1$ and $p_1 \equiv 3 \bmod 4$. □

Oplossingen

| Solutions

Now let $(a_n)_{n\in\mathbb{Z}}$ be a sequence in which each number is the Euler-phi of the next, so that $a_n = \phi(a_{n+1})$ for all $n \in \mathbb{Z}$. We first show that no a_n can have a prime factor of at least 5.
Suppose the contrary: let $N \in \mathbb{Z}$ be such that a_N has a prime factor of at least 5. By the lemma, we see that all for all $n \geq N$ the number a_n has a prime factor of at least 5, and that therefore the sequence $(v_2(a_n))_{n\geq N}$ is decreasing. Hence this sequence stabilises, and we assume (by replacing N by a larger number if necessary) that $(v_2(a_n))_{n\geq N}$ is constant.
We see that for all $n > N$, we have $a_n = 2^{e_n}p_n^{f_n}$ for some prime $p_n \geq 5$ congruent to 3 modulo 4, and $f_n > 0$. As

$$2^{e_n}p_n^{f_n} = a_n = \phi(a_{n+1}) = 2^{e_{n+1}-1}p_{n+1}^{f_{n+1}-1}(p_{n+1}-1),$$

with $v_2(p_{n+1}-1) = 1$ and $p_{n+1}-1 \geq 4$, we see that $p_{n+1}-1$ contains an odd prime factor, which therefore must be p_n by uniqueness of prime factorisation. As therefore $p_{n+1} > p_n$, it follows that $f_{n+1} = 1$ and that $p_{n+1} - 1 = 2p_n$.
Now note that the map $F : \mathbb{Z}/p_{N+1}\mathbb{Z} \to \mathbb{Z}/p_{N+1}\mathbb{Z}$, $x \mapsto 2x+1$ is bijective as p_{N+1} is odd, and therefore has a finite order as permutation on $\mathbb{Z}/p_{N+1}\mathbb{Z}$. Therefore there exists an integer $M > 0$ such that $p_{N+1+M} \equiv 0 \bmod p_N$, which contradicts the fact that p_{N+1+M} is a prime greater than p_{N+1}. Hence no a_n can have a prime factor of at least 5.
Write therefore, for all $n \in \mathbb{Z}$, $a_n = 2^{e_n}3^{f_n}$. Then $e_n > 0$ unless $a_n = 1$, since otherwise a_n is not the Euler-phi of any positive integer. We then have

$$a_n = \phi(a_{n+1}) = \begin{cases} 2^{e_{n+1}}3^{f_{n+1}-1} & \text{if } e_{n+1}, f_{n+1} > 0; \\ 2^{e_{n+1}-1} & \text{if } e_{n+1} > 0 \text{ and } f_{n+1} = 0; \\ 1 & \text{if } e_{n+1} = f_{n+1} = 0. \end{cases}$$

Hence we see that (a_n) must have one of the following forms.

– If there exists an a_n containing a prime factor 3, then the sequence is of the form

$$\ldots, 1, 1, 1, 2, \ldots, 2^{e-1}, 2^e, 2^e 3, 2^e 3^2, \ldots$$

with $e > 0$.
– If no a_n contains a prime factor 3, but there exists one containing a prime factor 2, then the sequence is of the form

$$\ldots, 1, 1, 1, 2, 2^2, \ldots$$

– Otherwise, the sequence is

$$\ldots, 1, 1, 1, \ldots$$