Springer Proceedings in Mathematics & Statistics

David H. Bailey · Naomi Simone Borwein ·
Richard P. Brent · Regina S. Burachik ·
Judy-anne Heather Osborn · Brailey Sims ·
Qiji J. Zhu *Editors*

From Analysis to Visualization

A Celebration of the Life and Legacy of Jonathan M. Borwein, Callaghan, Australia, September 2017

Springer

Springer Proceedings in Mathematics & Statistics

Volume 313

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

David H. Bailey · Naomi Simone Borwein ·
Richard P. Brent · Regina S. Burachik ·
Judy-anne Heather Osborn ·
Brailey Sims · Qiji J. Zhu
Editors

From Analysis to Visualization

A Celebration of the Life and Legacy of Jonathan M. Borwein, Callaghan, Australia, September 2017

Springer

Editors

David H. Bailey
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

Richard P. Brent
Mathematical Sciences Institute
Australian National University
Canberra, ACT, Australia

Judy-anne Heather Osborn
School of Mathematical and Physical
Sciences
University of Newcastle
Callaghan, NSW, Australia

Qiji J. Zhu
Department of Mathematics
Western Michigan University
Kalamazoo, MI, USA

Naomi Simone Borwein
Western University
London, ON, Canada

Regina S. Burachik
School of IT and Mathematical Sciences
University of South Australia
Mawson Lakes, SA, Australia

Brailey Sims
School of Mathematical and Physical
Sciences
University of Newcastle
Callaghan, NSW, Australia

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-36567-7 ISBN 978-3-030-36568-4 (eBook)
https://doi.org/10.1007/978-3-030-36568-4

Mathematics Subject Classification (2010): 00B20, 11-06, 26-06, 33-06, 47-06, 49-06, 52-06, 62P05, 91G99, 97-06

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

Jonathan M. Borwein as a high school student, age 14–15

Jonathan M. Borwein at his investiture into the Australian Academy of Science in 2010

Preface

Jonathan Borwein: Mathematician Extraordinaire

Many of us were shocked when our dear colleague Jonathan Michael Borwein of the University of Newcastle, Australia, died in August 2016. Jonathan ranked among the most wide ranging and influential mathematicians of the last half-century. He made significant contributions to a diversity of areas, in the process of which he exploited and greatly expanded the methods of *experimental mathematics*. In the course of this, he wove a network of colleagues and collaborators that spanned six continents and numbered well into the hundreds. After his passing, one immediate priority was to gather together as many of his works as possible. Accordingly, David H. Bailey and Nelson H. F. Beebe of the University of Utah began collecting as many of Borwein's published papers, books, reports and talks as possible, together with book reviews and articles written by others about Jonathan and his work. The current catalogue [1] lists nearly 2000 items. Even if one focuses only on formal, published, peer-reviewed articles, there are over 500 such items. These works are heavily cited—the Google citation tracker finds over 22,000 citations.

What is most striking about this catalogue is the wide range of topics. One bane of modern academic research, in general, and of the field of mathematics, in particular, is that most researchers today focus on a single specialized niche, seldom attempting to branch out into other specialties and disciplines or to forge potentially fruitful collaborations with researchers in other fields. In contrast, Borwein not only learned about numerous different specialities, but, in fact, did significant research in a wide range of fields, including experimental mathematics, optimization, convex analysis, applied mathematics, computer science, scientific visualization, biomedical imaging and mathematical finance. It is hard to think of a single mathematician of the modern era who has published notable research in so many different arenas.

A Portrait of the Man as a Mathematician

Jonathan was a polymath, by its very definition, and a true renaissance scholar. His knowledge base was as expansive as it was detailed. Let us take a moment to paint a fuller picture of this man and his engagement with mathematics, technology and the world around him. Born in St Andrews, Scotland on 20 May 1951, Jon attended the Madras College in Fife as a child, and went to university at 15, becoming at 23 a rather young Postdoctoral Fellow at Dalhousie University where he remained until 1991.

Always steeped in the mathematical world, thanks to his father, David Borwein, Jon started doing AMS Math Monthly problems with his father from an early age. The only time Jon didn't have a math book under his arm, a pen leaking ink in his pocket desperate to be etched across a sheet of crisp white paper—and later an iPad to work on—was the 3 months he travelled along the route of Xenophon's *Anabasis* through Turkey and Greece in the summer of 1973.

Jon won a 1971 Rhodes Scholarship and settled into life at Oxford. His classes allowed him to rub shoulders with the likes of Michael Atiyah, Professor of Geometry, who actually attended a class on mathematical Linguistics with Jon. Atiyah often sat unheeding and would then ask the professor dumb basic questions about the lecture. At a much later time, at a conference, Atiyah laughed and said that Jon had certainly gotten more from the class than he did. Such interactions reinforced the humanity of the great professors and in many ways became a template for his interactions with students, accepting of differences and ready to teach all.

Jon was a member of 'Professors for Peace for the Middle East', a Canadian organization with very dedicated people who went on to be influential in Canadian society. In 1967, a group went to Israel to discuss the problems with both Palestinians and Israelis, from the Mufti of Gaza to Teddy Kollek, the Mayor of Jerusalem. They arrived in May, days after the election that heralded the start of the tenure of Menachem Begin. The authorities they had arranged to meet were so rattled by the result of the election that they actually reported the truth about many things 'Middle East', while filling boxes to depart their offices. In some cases, the group was also able to access the newcomers. This organization still exists despite a name change and continues to press for a peaceful resolution.

In 1985, while on sabbatical, Jon went to Cambridge, England and then Limoges, France. While in Cambridge Jon received a redirected Christmas card from Yasumasa Kanada, a Professor in the Department of Information Science at the University of Tokyo. He looked at the return address and was astonished to see that Kanada was in Cambridge as well. There was a meeting and this led to fruitful discussions about computer modelling algorithms and π. This friendship lasted until Kanada's retirement in 2015. When asked why a computer scientist would come to such a place as Cambridge, Kanada replied that he was looking to imbibe the theoretical underpinnings and classical view of his work. This was something he could not get at home. Next in the sabbatical was a stay in France. Jon and his wife, Judi, spent an idyllic time in a gîte rural on the grounds of a Chateau near Rilhac

Rancon. This was a perfect place to ruminate on the mysteries of math. The time spent here led to a very prolific association with Michel Thera and the French mathematics community, leading eventually to an honorary Doctorate at the University of Limoges.

At this same time, Jon was reunited with Ephraim J. Borowski, an old friend from Oxford, who joined Jon for a while at the gîte while working on a Collins Dictionary. The two sat on the grass outside, or around the table in the fifteenth-century stone cottage with file cards, filling in definitions. This delightful interlude led to the *Collins Dictionary of Mathematics*. Later, Ephraim, burdened down with file cards, came to work with Jon in Halifax, Nova Scotia. They set up shop in Jon's office, situated in the old Halifax Archives, and had a very strict protocol: nothing was to be changed on the Lisa computer unless it was first listed and dated on the white boards. They never misplaced a letter because of this, but even with five levels of certainty, they almost lost an entire section. Jon had a lifelong interest in utilizing technology and pushing its limits in the pursuit of a plethora of academic interests.

There are numerous Biographies of Jon, some of which are also obituaries (see, for instance, [2] and others at [1]). However, these often glom to certain details about his life, influence and output. Citing his own writing about 'The Best Teacher I Ever Had was …', these include his childhood experience of teaching another boy a two-by-two simultaneous equation at the age of 6, and how on arrival at Western University, in his second year, he very nearly decided to major in history —which would have been a loss to the mathematical community. These biographies often unevenly focus on the breadth of his contributions, containing them within the lens of the journal or researcher-author penning the biography. It is pertinent to state that he was so accomplished and had expertise in so many fields, that the editors at Springer could not find an appropriate replacement for him as Editor of the SUMAT Series. (Indeed, he was a founding Co-Editor in Chief of Springer-Verlag's SUMAT Series of Springer Undergraduate Mathematics and Technology books.) Apparently, it normally would have required four editors to cover the fields he knew so intimately.

Overview of this Proceedings

It is the intention of this volume to commemorate Jonathan's remarkable achievements and legacies that were explored at his Commemorative Conference held on 25–29 September 2017, at Noah's On The Beach, Newcastle, NSW—one of Jonathan's favourite spots. Associated events included the Sunday, 24th September Satellite meeting on Mathematics and Education; the Tuesday, 26th September Public Lecture given by Keith Devlin, entitled 'Finding Fibonacci—The Quest to Rediscover the Forgotten Mathematical Genius Who Changed the World' and the Wednesday, 27th September 'An Evening of Mathematics, Music and Art' held at the Harold Lobb Concert Hall of the Conservatorium of Music. The conference was devoted to five main areas in which Jonathan made outstanding contributions; these became leading session themes:

1. *Applied Analysis, Optimisation and Convex Functions*, chaired by Regina S. Burachik and Guoyin Li;
2. *Education*, chaired by Naomi Simone Borwein and Judy-anne Heather Osborn;
3. *Experimental Mathematics and Visualisation*, chaired by David H. Bailey;
4. *Financial Mathematics*, chaired by Qiji (Jim) Zhu and
5. *Number Theory, Special Functions and Pi*, chaired by Richard P. Brent.

Four of these also constitute the sections into which this proceedings is divided. Upon a cursory glance one might be surprised that there isn't also a section dealing with experimental mathematics and visualization. This was indeed an early intention; however, it soon became apparent that the pertinent articles also fitted under one of the other banners, while almost every article in every section exemplified aspects of the experimental approach and in many cases the power of visualization. It seemed natural, therefore, to see these as themes threaded throughout the volume, reflecting how these methodologies grew to colour all Jonathan's mathematical endeavours.

As any of us familiar with Jonathan knew well, he was so prolific that at any moment in time he had numerous research projects underway and papers in preparation. Testament to this is the surprising number of papers with Jonathan listed as one of the authors found in this proceedings. All represent lines of research in which he was actively engaged at the time of his passing.

In the following sections of this Preface, we briefly delve into just a few of Jon's contributions that are pertinent to the research areas listed above.

Nonlinear Analysis and Optimization

Some of Jon's most significant contributions were in the area of optimization; indeed, papers in the area of optimization and convex analysis are the single-most numerous category in the catalogue [1].

One notable paper in the optimization arena is [3], which presents what is now known as the Barzilai–Borwein algorithm for large-scale unconstrained optimization. This paper has been cited over 1300 times. There are numerous techniques for this type of problem (unconstrained optimization) in the literature. The standard gradient method, namely, to iterate $x_{k+1} = x_k - \alpha_k g_k(x_k)$, where α_k is typically calculated based on a fixed line search procedure, is fairly simple to use, but it makes no use of second-order information and sometimes zig-zags rather than converging. Newton's method is to iterate $x_{k+1} = x_k - (F_k(x_k))^{-1} g_k(x_k)$, where $F_k = \nabla^2 f(x_k)$ is the Hessian of the system. It utilizes second-order information and typically converges quite rapidly near the solution, but it requires the expensive computation of the matrix $(F_k(x_k))^{-1}$, and for some applications the scheme requires additional custom modifications to ensure convergence.

The Barzilai–Borwein method mimics the gradient method, in that it selects α_k so that $\alpha_k g_k(x_k)$ approximates $(F_k(x_k))^{-1} g_k(x_k)$, but it does not require that one actually compute $(F_k(x_k))^{-1}$. As a result, this scheme often converges nearly as fast as the Newton method, but at significantly lower computational overhead. Due to its simplicity and efficiency, variations of this method have been applied in a variety of applications, including sparse optimization, image analysis and signal processing.

Experimental Mathematics

Jon is perhaps best known for deriving, with his brother Peter, quadratically and higher order convergent algorithms for π, including p-th-order convergent algorithms for any prime p, and similar quadratically convergent algorithms for certain other fundamental constants and functions [4–6]. Here 'quadratically convergent' means that each iteration of the algorithm approximately *doubles* the number of correct digits in the result, with a similar definition for higher order convergence; p-th-order convergent means that the number of correct digits increases approximately by a factor of p with each iteration.

One of their best-known algorithms is the following: Set $a_0 = 6 - 4\sqrt{2}$ and $y_0 = \sqrt{2} - 1$. Then iterate

$$\begin{aligned} y_{k+1} &= \frac{1 - (1 - y_k^4)^{1/4}}{1 + (1 - y_k^4)^{1/4}} \\ a_{k+1} &= a_k(1 + y_{k+1})^4 - 2^{2k+3} y_{k+1}(1 + y_{k+1} + y_{k+1}^2). \end{aligned}$$

Then a_k converges *quadratically* to $1/\pi$: each iteration approximately *quadruples* the number of correct digits. This algorithm, together with a quadratically convergent algorithm due to Brent and Salamin, was employed in several large computations of π by Kanada and others.

But an event of more enduring legacy is his advocacy of *experimental mathematics*, in particular, his championing of the usage of advanced computing technology to discover new principles and formulas of mathematics, not just verify them with mathematical software.

One of many examples of this methodology in action was his analysis (in conjunction with David H. Bailey and the late Richard Crandall) of the following three classes of integrals that arise in mathematical physics: C_n are connected to quantum field theory, D_n arise in Ising theory, while the E_n integrands are derived from D_n:

$$C_n := \frac{4}{n!}\int_0^\infty \cdots \int_0^\infty \frac{1}{\left(\sum_{j=1}^{n}(u_j + 1/u_j)\right)^2}\frac{\mathrm{d}u_1}{u_1}\cdots\frac{\mathrm{d}u_n}{u_n}$$

$$D_n := \frac{4}{n!}\int_0^\infty \cdots \int_0^\infty \frac{\prod_{i<j}\left(\frac{u_i-u_j}{u_i+u_j}\right)^2}{\left(\sum_{j=1}^{n}(u_j + 1/u_j)\right)^2}\frac{\mathrm{d}u_1}{u_1}\cdots\frac{\mathrm{d}u_n}{u_n}$$

$$E_n = 2\int_0^1 \cdots \int_0^1 \left(\prod_{1\le j<k\le n}\frac{u_k-u_j}{u_k+u_j}\right)^2 \mathrm{d}t_2\,\mathrm{d}t_3\ldots\mathrm{d}t_n,$$

where in the last line $u_k = t_2\ldots t_k$ [7].

One early observation was that the C_n integrals can be converted to one-dimensional integrals involving the modified Bessel function $K_0(t)$:

$$C_n = \frac{2^n}{n!}\int_0^\infty tK_0^n(t)\,\mathrm{d}t.$$

It was quickly evident that high-precision numerical values of this sequence, computed using tanh-sinh quadrature, approach a limit. For example:

$$C_{1024} = 0.6304735033743867961220401927108789043545870787\ldots$$

When the first 50 digits of this constant were copied into the online Inverse Symbolic Calculator-2 (ISC-2) at https://isc.carma.newcastle.edu.au (which Jon was instrumental in developing and deploying), the result was

$$\lim_{n\to\infty} C_n = 2e^{-2\gamma},$$

where γ denotes Euler's constant, a result which was then proved.

Subsequently high-precision computations, in conjunction with Ferguson's PSLQ algorithm [8, 9], were applied to find experimental evaluations of numerous other specific instances of these integrals, including

$$
\begin{aligned}
D_3 &= 8 + 4\pi^2/3 - 27\mathrm{L}_{-3}(2)\\
D_4 &= 4\pi^2/9 - 1/6 - 7\zeta(3)/2\\
E_2 &= 6 - 8 log 2\\
E_3 &= 10 - 2\pi^2 - 8 log 2 + 32\, log^2 2\\
E_4 &= 22 - 82\zeta(3) - 24 log 2 + 176\, log^2 2 - 256(log^3 2)/3 + 16\pi^2 log 2 - 22\pi^2/3\\
E_5 &= 42 - 1984\,\mathrm{Li}_4(1/2) + 189\pi^4/10 - 74\zeta(3) - 1272\zeta(3) log 2 + 40\pi^2\, log^2 2\\
&\quad - 62\pi^2/3 + 40(\pi^2 log 2)/3 + 88\, log^4 2 + 464\, log^2 2 - 40 log 2,
\end{aligned}
$$

where $\mathrm{L}_{-3}(2)$ is a Dirichlet L-function constant, $\zeta(x)$ is the Riemann zeta function and $\mathrm{Li}_n(x)$ is the polylogarithm function [7]. The formula for E_5, which was initially found by Borwein (and which he was quite proud of), remained a numerically discovered but open conjecture for several years, but was finally proven in 2014 by Erik Panzer [10]. Resolution of the general case is still open.

Number Theory, Special Functions and Pi

Jon's work on number theory, special functions and π is inextricably linked to his work on experimental mathematics. Typically, using his excellent mathematical intuition supported by experimental mathematics tools such as PSLQ [9], Jon would make a conjecture that was almost certainly true, and in many cases could be proved rigorously.

To give just one example, we mention Jon's work, together with collaborators David H. Bailey, Richard Crandall, Karl Dilcher, Armin Straub, James Wan and others, on the so-called *Mordell–Tornheim–Witten zeta function* [11]. This function is a vast generalization of the Riemann zeta function, and is defined by

$$\omega(s_1, \ldots, s_{K+1}) := \sum_{m_1,\ldots,m_K > 0} \frac{1}{m_1^{s_1} \cdots m_K^{s_K} (m_1 + \cdots + m_K)^{s_{K+1}}}$$

with suitable restrictions on the parameters $s_1, \ldots, s_{K+1}$. For integer values of the parameters, there are many interesting identities satisfied by the ω values and their derivatives.

Further examples are described in the Preface to Part IV 'Number Theory, Special Functions and Pi', and in the various papers contributed to that Part.

Mathematical Finance

A notable example of how Jon ventured into arenas quite far afield from his core research in optimization and computational mathematics is his work in mathematical finance. This began in 2013, when David H. Bailey mentioned to Jon some research he had been doing with Marcos Lopez de Prado, a financial mathematician in New York City. Bailey and Lopez de Prado were concerned about the yawning gap between state-of-the-art mathematical techniques that were being successfully applied in leading quantitative investment funds, on one hand, and the mathematically and statistically naive schemes and practices that were often being promoted to the public and even being presented in presumably peer-reviewed journals. It had become clear, based on the preliminary research, that 'backtest overfitting', namely, the statistical overfitting of historical market data, was rampant in the finance field, and is arguably the principal reason why so many financial strategies and investment fund designs, which look great on paper and in promotional literature, fall flat when actually fielded. David and Marcos were also concerned with the many pseudoscientific techniques and strategies that are mentioned on a daily basis in the financial press.

When they presented some of their findings and thoughts on the topic to Jon, he immediately understood the technical issues, appreciated their gravity and concurred that these issues deserved rigorous treatment. So Bailey, Borwein, Lopez de Prado and Qiji J. Zhu then co-authored a pair of papers with full details. The first paper, entitled 'Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance' (a provocative title that Borwein himself proposed), was published as a feature article in the *Notices of the American Mathematical Society* [12], and has been circulated widely in the financial community. The second paper addressed the probability of backtest overfitting in more technical depth [13].

Jon also urged Bailey, Lopez de Prado and Zhu to start a blog presenting many of these related issues for an even broader audience. The result was the *Mathematical Investor* blog [14], with the provocative subtitle (also proposed by Jon) 'Mathematicians against fraudulent financial and investment advice (MAFFIA)'. Its mission was and is to identify and draw attention to abuses of mathematics and statistics in the financial field, and also to call out the financial mathematics community for their silence on these abuses. These abuses include:

1. Failing to disclose the number of models or variations that were used to develop an investment strategy or fund (which failure makes the strategy or fund highly susceptible to backtest overfitting).
2. Making vague predictions that do not permit rigorous testing and falsification.
3. Misusing probability theory, statistics and stochastic calculus.
4. Suggesting in press reports and promotions that investors can achieve above-market returns via unsophisticated chart-watching techniques (e.g. 'technical analysis', 'Elliott waves', etc.).

5. Using pseudo-mathematical technical jargon: ‘stochastic oscillators’, ‘Fibonacci ratios’, ‘cycles’, ‘waves’, ‘golden ratios’, ‘parabolic SAR’, ‘pivot point’, ‘momentum’, etc., none of which has any rigorous scientific basis.

As Jon and the other authors of the ‘Pseudo-mathematics’ paper explained, ‘Our silence is consent, making us accomplices in these abuses’ [12].

Mathematical Education and Public Communication

Jon’s passion for sharing the joy of mathematical research and communicating this joy to the public was central to his career. He personally mentored scores of graduate students, and taught hundreds of others. Many of these students have, in turn, become notable mathematicians and computer scientists themselves. This alone would be an achievement worthy of acclaim.

Along this line, Jon specifically selected many of his research topics based on their potential for public appeal and inspiring students. This is particularly clear with his interest in π, formulas for π and experimental mathematics in general, which he saw as a powerful vehicle to convey the excitement of modern mathematics to the younger, tech-savvy crowd, and yet basic enough to be comprehensible even to high school and undergraduate students. The depth of Jon’s personal engagement with mathematics education, and experimental mathematics as an educational tool, is explored in great detail in Naomi Simone Borwein and Judy-anne Heather Osborn’s contribution to this Springer volume.

As mentioned above, Jon was an avid blogger, which again was rooted in his passion for communicating with students and the public at large. David H. Bailey is deeply grateful to have been a part of this effort with Jon. Beginning in 2009, when Bailey and Jon Borwein founded the ‘Math Drudge’ blog [15], he and Bailey co-authored over 200 articles on a wide range of topics, covering virtually every facet of modern mathematics, computing and science. A few of the topics they addressed in these blogs include

1. The psychology of mathematics.
2. Pseudoscience and anti-science.
3. The fallacies of creationism and intelligent design.
4. The sad state of math and science education.
5. Global warming and global warming denial.
6. The computation of π.
7. New developments in physics and cosmology.
8. New developments in computer science.
9. Fermi’s paradox.
10. Artificial intelligence.
11. Computer games versus humans.

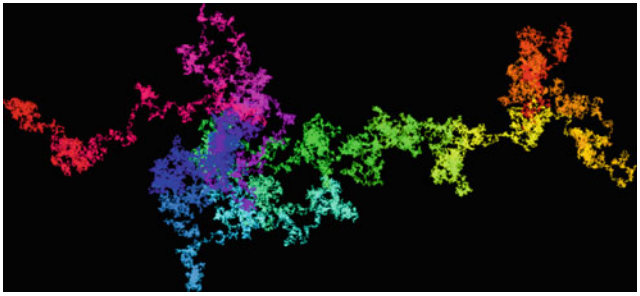
12. Supercomputers.
13. Ancient Indian mathematics.
14. The ancient origins of decimal arithmetic.
15. Moore's law and the future of mathematics.
16. The discovery of the Higgs boson.
17. New ways to visualize the digits of π.
18. DNA and evolution.
19. Pseudoscience from the political left and right.
20. New energy technologies, including LENR and fusion.
21. Fields medalists, Abel Prize recipients and Breakthrough Prize recipients.

It should be emphasized that although Bailey did most of the actual writing, Jon personally proposed, co-wrote and co-edited virtually every one of these blogs. They reflect both his interests and his passions to communicate better with the public. Some of these blogs were subsequently published in venues such as *The Conversation* and *The Huffington Post*. Such expository writing was an extension of Jon's dynamic educational praxis across institutional and popular lines.

Visualization

The title of this volume 'From analysis to visualization' might just as easily have been 'from visualization to analysis' to reflect Jonathan's championship of visualization as an important paradigm for the discovery and dissemination of mathematics.

Jonathan understood visualization as a powerful tool in the arsenal of both the mathematical communicator and the experimental mathematician. One of his oft applied maxims was *it is often easier to see something than to explain it in words*. He saw the use of visual tools as a powerful catalyst in triggering our imagination, revealing patterns and synergies that are otherwise difficult to detect, and often as a highly efficient and informative way to convey large amounts of information, insight and understanding.



A random walk on the first 10 billion base 4 digits of π

A piece of work that employed visualization in a striking and essential way and that Jonathan found especially pleasing was jointly undertaken with Francisco Aragón Artacho, David H. Bailey and his brother Peter [16]. They found much of a number's nature was beautifully exposed by its footprints when, starting from the origin, successive digits of its base 4 expansion were used to generate unit steps in a planar walk. A 0 corresponded to a step to the east, 1 a step north, 2 a step west and 3 a step south. Passage through the spectral hues (red to violet) was employed to indicate progression along the walk.

In Summary

Jonathan Borwein's prodigious output in nonlinear analysis and experimental mathematics is certainly his singular contribution to modern mathematics. But beyond his technical accomplishments, he was a master of mathematical communication (his lectures were always paragons of well-organized and visually appealing mathematics and graphics), mathematical education (part of his interest in π was to bring the joy of mathematical discovery to students), and in promoting science, mathematics and computing to the general public. To this end, he wrote and lectured tirelessly. By one reckoning he presented an average of one talk per week for decades, and wrote hundreds of articles targeted to the general public. His death is a loss to all those who treasure modern mathematics, science and clear thinking.

Berkeley, USA	David H. Bailey
London, Canada	Naomi Simone Borwein
Canberra, Australia	Richard P. Brent
Mawson Lakes, Australia	Regina S. Burachik
Callaghan, Australia	Judy-anne Heather Osborn
Callaghan, Australia	Brailey Sims
Kalamazoo, USA	Qiji J. Zhu

References

1. Jonathan Borwein Memorial Website, https://www.jonborwein.org. The catalogue of Borwein's papers is at https://www.jonborwein.org/jmbpapers/
2. Borwein, J., Borwein, N., Sims, B.: Obituary: Jonathan M. Borwein. Gaz. Aust. Math. Soc. **44** (5), 289–293 (2017)
3. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**, 141–148 (1988)
4. Borwein, J.M., Borwein, P.B.: The arithmetic–geometric mean and fast computation of elementary functions. SIAM Review **26**(3), 351–366 (July 1984)

5. Borwein, J.M., Borwein, P.B.: Cubic and higher order algorithms for pi. Can. Math. Bull. **27**, 436–443 (1984)
6. Borwein, J.M., Borwein, P.B., Bailey, D.H.: Ramanujan, modular equations, and approximations to pi, or how to compute one billion digits of pi. Am. Math. Mon. **96**(3), 201–219 (March 1989)
7. Bailey, D.H., Borwein, J.M., Crandall, R.E.: Integrals of the Ising class. J. Phys. A: Math. Gen. **39**, 12271–12302 (2006)
8. Bailey, D.H., Broadhurst, D.J.: Parallel integer relation detection: Techniques and applications. Math. Comput. **70**(236), 1719–1736 (October 2000)
9. Ferguson, H.R.P., Bailey, D.H., Arno, S.: Analysis of PSLQ, an integer relation finding algorithm. Math. Comput. **68**(225), 351–369 (January 1999)
10. Panzer, E.: Algorithms for the symbolic integration of hyperlogarithms with applications to Feynman integrals. Comput. Phys. Commun. **188**, 148–166 (2015), available at http://arxiv.org/abs/1403.3385
11. Bailey, D.H., Borwein, J.M., Crandall, R.E.: Computation and theory of extended Mordell-Tornheim-Witten sums. Math. Comput. **83**, 1795–1821 (2014)
12. Bailey, D.H., Borwein, J.M., Lopez de Prado, M., Zhu, Q.J.: Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. Not. Am. Math. Soc. **61**(5), 458–471 (May 2014), available at http://www.ams.org/notices/201405/rnoti-p458.pdf
13. Bailey, D.H., Borwein, J.M., Lopez de Prado, M., Zhu, Q.J.: The probability of backtest overfitting. J. Comput. Finan. **20**(4), 39–69 (April 2017)
14. The Mathematical Investor: Mathematicians against fraudulent financial and investment advice (MAFFIA) blog, https://www.financial-math.org. Newer items are at https://www.mathinvestor.org and https://www.maffia.org
15. The Math Drudge blog, https://www.experimentalmath.info/blog. Newer items are at the Math Scholar blog, https://www.mathscholar.org
16. Aragon Artacho, F.J., Bailey, D.H., Borwein, J.M., Borwein, P.B.: Walking on real numbers. Math. Intell. **35**, 42–60 (2013)

Acknowledgements

First and foremost, we must acknowledge the incredible mathematical legacy left by Jonathan Borwein. A legacy that both inspired and informed this commemorative volume. We must also acknowledge our indebtedness to members of Jonathan's family who, in the most painful of times, gave tirelessly of themselves to help reveal just how extraordinary a person Jonathan was.

Thanks to all who participated in the Jonathan Borwein Commemorative Conference and especially those who contributed articles for these Proceedings. Thanks also to the Section Chairs who, in the capacity as editors of the four Parts, solicited the articles and steered them through the process of peer review, and to Guoyin Li for his assistance with Part I. Also to Brailey Sims for management of the overall editorial process. A very special acknowledgement is owed to the many (anonymous) referees who made exceptional efforts to ensure that every article met the highest standards of both presentation and content. Also, thanks to Matthew Skerritt whose TeXnical and computing skills ensured everything came together seamlessly. Lastly, very special thanks to Springer, their production team, and in particular to Elizabeth Loew, who championed the publication of this volume from its earliest inception.

Contents

Part I Applied Analysis, Optimisation, and Convexity

Part II Education

Part III Financial Mathematics

Part IV Number Theory, Special Functions, and Pi

Part I
Applied Analysis, Optimisation, and Convexity

Introduction

Regina S. Burachik and Guoyin Li

Our friend and colleague Jonathan Michael Borwein excelled in an incredibly vast range of fields: his work spans from pure analysis, functional analysis and maximal monotone operators, applied optimization, multiobjective optimisation, numerical and computational analysis, mathematics for high performance computing, probability theory, and many more. Jon's mathematical breadth and his naturally inquisitive energy makes a unique, fascinating picture, that extends far beyond traditional areas such as functional analysis and number theory. Hence, it is a challenging (or maybe impossible) task to sketch the totality of his mathematical works.

On the other hand, many of us know of Jon's deep interest in analysis and optimization. Indeed, his groundbreaking work in this field has influenced researchers since the time of his PhD years. Jon's PhD thesis was on "Optimization with Respect to Partial Orderings." Moreover, Jon strongly advocated for convex analysis as a main driving force for many years, and one of his most influential outputs is his now classical book "Convex Analysis and Nonlinear Optimization," jointly written with Adrian Lewis [1].

Jon made numerous fundamental contributions to the area of Applied Analysis, Optimisation, and Convexity. One particular example is the celebrated result in nonsmooth analysis and optimisation which is now widely known as the Borwein–Preiss variational principle [2]: a ubiquitous technique showing that a smooth small

R. S. Burachik (✉)
School of Information Technology and Mathematical Sciences, University of South Australia,
Mawson Lakes, SA, Australia
e-mail: Regina.Burachik@unisa.edu.au

G. Li
Department of Applied Mathematics, University of New South Wales,
Kensington, NSW, Australia
e-mail: g.li@unsw.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_1

perturbation ensures the existence of a minimizer of a possibly nonsmooth function. Another example is the Barzilai–Borwein method [3], an ingenious and highly efficient nonmonotone gradient-based minimization algorithm. Interestingly, Jon explained once to the second author (Guoyin Li) that he invented this method using insights from his research discoveries in the area of number theory. This is only a very small sample of the fundamental contributions made by Jon to the field of our special session, but they give an idea of the fresh and innovative force of his research. Of course, there are too many of these to fit in a limited space, and we invite the reader to discover more examples of Jon's contributions in his personal web pages at https://carma.newcastle.edu.au/jon/, where most of his beautiful papers and talks can be found.

At the Jonathan Borwein Commemorative Conference (JBCC), there were 15 speakers from 6 countries, that presented in the area of "Applied Analysis, Optimisation, and Convex Functions." The speakers reflected on how Jon's research has greatly influenced theirs. We are honored to have gathered the following contributions from the participants of the conference to the topic of our session:

- *Symmetry and the Monotonicity of Certain Riemann Sums* by David Borwein, Jonathan M. Borwein, and Brailey Sims: This was one of the last papers Jon worked on and exemplifies his experimental approach to mathematical discovery. Given a function from [0, 1] to the real line, this contribution considers conditions ensuring the monotonicity of right and left Riemann sums of the function with respect to uniform partitions. Experimentation is used to study the role of symmetry in these monotonicity properties. The authors use the results of this paper to obtain nice applications of Descartes' rule of signs.
- *Risk and Utility in the Duality Framework of Convex Analysis* by R. Tyrrell Rockafellar: This paper gives a comprehensive survey on how risk and utility are in fact closely related, using the tools of convex analysis and in particular, the beautiful and elegant framework of conjugate duality.
- *Characterizations of Robust and Stable Duality for Linearly Perturbed Uncertain Optimization Problems* by N. Dinh, M.A. Goberna, M.A. López and M. Volle: This paper introduces a robust optimization model for an uncertain convex optimization problem, and discusses new robust and stable duality results between the robust model problem and its dual problem.
- *Comparing Averaged Relaxed Cutters and Projection Methods: Theory and Examples* by R. Díaz Millán, Scott B. Lindstrom, and Vera Roshchina: This paper presents a convergence analysis of new projection and reflection type methods for solving convex feasibility problem.

The untimely passing of Jon has deprived us all of a singular and brilliant mind and an inspirational intellectual leader. We have lost an inspired collaborator, a highly motivated teacher and a close personal friend. We hope that this section gives a snapshot of Jon's research work in this particular field.

References

1. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization. Springer, New York (2000)
2. Borwein, J.M., Preiss, D.: A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions. AMS Trans. **303**, 517–527 (1987)
3. Barzilai, J., Borwein, J.M.: Two point step-size methods. IMA J. Numer. Anal. **8**, 141–148 (1988)

Symmetry and the Monotonicity of Certain Riemann Sums

David Borwein, Jonathan M. Borwein and Brailey Sims

This paper is dedicated to the memory and lasting legacy of Jonathan Borwein, son, friend, and colleague.
It was one of the last papers he worked on and exemplifies his experimental approach to mathematical discovery.

1 Introduction

For a bounded function $f : [0, 1] \to \mathbb{R}$ the *left* and *right Riemann sums* of f with respect to the uniform partition $\mathcal{U}_n$ of $[0, 1]$ into n equal intervals are

$$\sigma_n := \sigma_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right), \quad \text{and} \quad \tau_n := \tau_n(f) = \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{k}{n}\right).$$

Jonathan M. Borwein Passed away suddenly and unexpectedly on 2 August 2016.
Was Laureate Professor and Director of the Centre for Computer-Assisted Research Mathematics and its Applications at the University of Newcastle, Australia.

D. Borwein
Department of Mathematics, Western University, London, ON, Canada
e-mail: dborwein@uwo.ca

J. M. Borwein · B. Sims (✉)
School of Mathematical and Physical Sciences, The Centre for Computer Assisted Research Mathematics and its Applications (CARMA), The University of Newcastle, Callaghan, NSW, Australia
e-mail: brailey.sims@newcastle.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_2

Both are linear functionals with $\sigma_n(1) = \tau_n(1) = 1$ and $\sigma_n(f) - \tau_n(f) = \dfrac{1}{n}(f(0) - f(1))$. If f is decreasing (increasing) on [0, 1] then σ_n is the upper (lower), and τ_n the lower (upper), Riemann sum of f with respect to $\mathcal{U}_n$. And, of course, if f is Riemann integrable (as it is in either of the above cases) then both σ_n and τ_n converge to $\int_0^1 f(x)\mathrm{d}x$ (see, for example [1]). Further, for all $n, \sigma_n(f(1-x)) = \tau_n(f(x))$, so if f is symmetric about the midpoint of [0, 1]; that is, $f(x) = f(1-x)$, then $\tau_n = \sigma_n$.

We seek conditions which will ensure the sequence (σ_n), (τ_n), or perhaps that for some other related Riemann sum increases/decreases with n. If for example f is decreasing then $\tau_{2n} \geq \tau_n$, so τ_{2^n} increases monotonically to $\int_0^1 f$, but how does τ_{n+1} compare to τ_n?

In the process of producing [2] one of the current authors gave the following example.

Example 1 (Digital assistance, arctan(1) *and a black box)* Consider for integer $n > 0$ the sum

$$\sigma_n := \sum_{k=0}^{n-1} \frac{n}{n^2 + k^2}.$$

The definition of the Riemann sum means that

$$\begin{aligned}\lim_{n\to\infty} \sigma_n &= \lim_{n\to\infty} \sum_{k=0}^{n-1} \frac{1}{1+(k/n)^2}\frac{1}{n}\\ &= \int_0^1 \frac{1}{1+x^2}\mathrm{d}x\\ &= \arctan(1). \end{aligned} \tag{1}$$

Even without being able to do this *Maple* will quickly tell you that

$$\sigma_{10^{14}} = 0.78539816339746\ldots$$

Now, if you ask for 100 billion terms of most slowly convergent series, a computer will take a long time. So this is only possible because *Maple* “knows” the identity

$$\sigma_N = -\frac{i}{2}\Psi\left(N - iN\right) + \frac{i}{2}\Psi\left(N + iN\right) + \frac{i}{2}\Psi\left(-iN\right) - \frac{i}{2}\Psi\left(iN\right)$$

and has a fast algorithm for computing our new friend the psi (dilogarithm) function of a complex variable. Now `identify`(0.78539816339746) yields $\frac{\pi}{4}$.

We can also note that

$$\tau_n := \sum_{k=1}^{n} \frac{n}{n^2 + k^2}$$

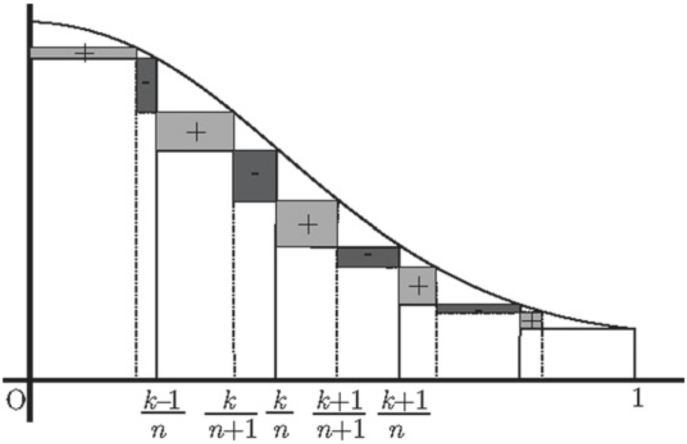


Fig. 1 Difference in the lower Riemann sums for $\frac{1}{1+x^2}$

is another (lower) Riemann sum converging to $\int_0^1 \frac{1}{1+x^2}\mathrm{d}x$. Indeed, $\sigma_n - \tau_n = \frac{1}{2n} > 0$. Moreover, experimentation *suggests* that σ_n decreases, and τ_n increases, to $\pi/4$. As we will see, the validity of this last observation is a consequence of our principal result. $\diamondsuit$

If we enter "monotonicity of Riemann sums" into Google, one of the first entries is http://elib.mi.sanu.ac.rs/files/journals/tm/29/tm1523.pdf which is a 2012 article by Szilárd [4] that purports to show the monotonicity of the two sums for the function

$$f(x) := \frac{1}{1+x^2}.$$

The paper goes on to prove that *if* $f\colon [0,1] \to R$ *is concave, or convex, and decreasing then* $\tau_n := \frac{1}{n}\sum_{k=1}^{n} f(\frac{k}{n})$ *increases and* $\sigma_n := \frac{1}{n}\sum_{k=0}^{n-1} f(\frac{k}{n})$ *decreases to* $\int_0^1 f(x)\,\mathrm{d}x$*, as* $n \to \infty$. Related results for a concave, or convex, and increasing function follow by applying these results to $-f$.

All proofs in [4] are based on looking at the rectangles which comprise the difference between τ_{n+1} and τ_n as in Figure 1 (or the corresponding difference for σ_n). This yields

$$\tau_{n+1}(f) - \tau_n(f) = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{(n+1-k)}{n+1} f\left(\frac{k}{n+1}\right) + \frac{k}{n+1} f\left(\frac{k+1}{n+1}\right) - f\left(\frac{k}{n}\right)\right\}. \quad (2)$$

In the easiest case, each bracketed term

$$\delta_n(k) := \frac{(n+1-k)}{n+1} f\left(\frac{k}{n+1}\right) + \frac{k}{n+1} f\left(\frac{k+1}{n+1}\right) - f\left(\frac{k}{n}\right)$$

has the same sign for all n and $1 \leq k \leq n$ as happens for a function which is concave, or convex, and decreasing.

But in [4] the author mistakenly asserts this applies for $1/(1+x^2)$ which has an inflection point at $1/\sqrt{3}$. Indeed, the proffered proof flounders at the inequality in the last line of [4, p. 115] which fails for instance when $n = 5$ and $k = 1$. This same error invalidates the assertion in [4] that the monotonicity of the corresponding σ_n can be proved by similar reasoning (left to the reader). Below in Corollary 3 and Example 2 we supply a correct proof that $\tau_n = \sum_{k=1}^{n} n/(n^2+k^2)$ increases, but we are unable as of yet to prove that the corresponding σ_n decreases.

It appears, however, on checking in a computer algebra system (CAS), that $\delta_n(k) + \delta_n(n-k) \geq 0$ which if rigorously established would repair the hole in the proof. It also suggests that symmetry may have a role to play.

In our opinion, all of this provides a fine instance of digital assistance in action.

For the convenience of the reader we supply the following proofs of Szilárd's theorems. The proofs are basically his but a bit cleaner. The proofs use telescoping and do not need consideration of the + and − rectangles of Figure 1.

2 Szilárd's Theorems

Theorem 1 *If the function $f : [0,1] \to \mathbb{R}$ is concave and decreasing on the interval* $[0,1]$, *then $\tau_n(f)$ increases and $\sigma_n(f)$ decreases as n increases.*

Theorem 2 *If the function $f : [0,1] \to \mathbb{R}$ is convex and decreasing on the interval* $[0,1]$, *then $\tau_n(f)$ increases and $\sigma_n(f)$ decreases as n increases.*

Before proceeding to the proofs of Theorems 1 and 2 we first give two lemmas.

Lemma 1 *If $f : [0,1] \to \mathbb{R}$ is concave and decreasing on the interval* $[0,1]$, *then for $k = 0, 1, 2, \cdots, n$*

$$f\left(\frac{k+1}{n+1}\right) \geq \frac{n-k}{n} f\left(\frac{k+1}{n}\right) + \frac{k}{n} f\left(\frac{k}{n}\right). \tag{3}$$

Proof Since f is concave on $[0,1]$ we have

$$\begin{aligned}\frac{n-k}{n} f\left(\frac{k+1}{n}\right) + \frac{k}{n} f\left(\frac{k}{n}\right) &\leq f\left(\frac{n-k}{n} \cdot \frac{k+1}{n} + \frac{k}{n} \cdot \frac{k}{n}\right) \\ &= f\left(\frac{nk+n-k}{n^2}\right). \end{aligned} \tag{4}$$

Due to the monotonicity of f on $[0,1]$ and the readily verified inequality

$$\frac{nk+n-k}{n^2} \geq \frac{k+1}{n+1}, \tag{5}$$

we have

$$f\left(\frac{k+1}{n+1}\right) \geq f\left(\frac{nk+n-k}{n^2}\right). \tag{6}$$

Together, inequalities (4) and (6) imply inequality (3). This completes the proof of the lemma.

Lemma 2 *If* $f : [0, 1] \to \mathbb{R}$ *is convex and decreasing on the interval* $[0, 1]$, *then for* $k = 0, 1, 2, \cdots, n$

$$f\left(\frac{k}{n}\right) \leq \frac{n+1-k}{n+1} f\left(\frac{k}{n+1}\right) + \frac{k}{n+1} f\left(\frac{k+1}{n+1}\right). \tag{7}$$

Proof Since f is convex on $[0, 1]$ we have

$$\begin{aligned} \frac{n+1-k}{n+1} f\left(\frac{k}{n+1}\right) + \frac{k}{n+1} f\left(\frac{k+1}{n+1}\right) &\geq f\left(\frac{n+1-k}{n+1} \cdot \frac{k}{n+1} + \frac{k}{n+1} \cdot \frac{k+1}{n+1}\right) \\ &= f\left(\frac{(n+2)k}{(n+1)^2}\right). \end{aligned} \tag{8}$$

Due to the monotonicity of f on $[0, 1]$ and the inequality

$$\frac{(n+2)k}{(n+1)^2} \leq \frac{k}{n}, \tag{9}$$

we have

$$f\left(\frac{(n+2)k}{(n+1)^2}\right) \geq f\left(\frac{k}{n}\right). \tag{10}$$

Together, inequalities (8) and (10) imply inequality (7). This completes the proof of the lemma. □

Proof of Theorem 1 Since for any constant K we have $\tau_n(f + K) = \tau_n(f) + K$ (and the same for σ_n), we may suppose without loss in generality that $f(1) = 0$. Observe that inequality (3) is equivalent to

$$\begin{aligned} \frac{1}{n+1} f\left(\frac{k+1}{n+1}\right) \geq \frac{1}{n} f\left(\frac{k+1}{n}\right) + \frac{1}{n(n+1)} \Bigg(& kf\left(\frac{k}{n}\right) \\ & -(k+1) f\left(\frac{k+1}{n}\right)\Bigg), \end{aligned} \tag{11}$$

for which, when we sum both sides from $k = 0$ to $n - 1$, the right hand side telescopes to yield

$$\frac{1}{n+1} \sum_{k=0}^{n-1} f\left(\frac{k+1}{n+1}\right) \geq \frac{1}{n} \sum_{k=0}^{n-1} f\left(\frac{k+1}{n}\right),$$

or equivalently, noting that $f(1) = 0$,

$$\tau_{n+1}(f) = \frac{1}{n+1}\sum_{k=1}^{n+1} f\left(\frac{k}{n+1}\right) \geq \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{k}{n}\right) = \tau_n(f). \tag{12}$$

This completes the proof of the first part of Theorem 1. The second part can be obtained by applying the first part of Theorem 2 (established below) to $-f(1-x)$.

Proof of Theorem 2 We again suppose without loss in generality that $f(1) = 0$. Observe that inequality (7) is equivalent to

$$\frac{1}{n} f\left(\frac{k}{n}\right) \leq \frac{1}{n+1} f\left(\frac{k}{n+1}\right) + \frac{1}{n(n+1)}\left(kf\left(\frac{k+1}{n+1}\right) - (k-1)f\left(\frac{k}{n+1}\right)\right), \tag{13}$$

from which it follows that

$$\tau_n(f) = \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{k}{n}\right) \leq \frac{1}{n}\sum_{k=1}^{n-1} f\left(\frac{k}{n+1}\right) = \tau_{n+1}(f). \tag{14}$$

This completes the first part of the proof of Theorem 2. The second part can be obtained by applying the first part of Theorem 1 to $-f(1-x)$.

3 Extensions of Szilárd's Theorems

Theorem 3 *If the function* $f : [0, 1] \to \mathbb{R}$ *is convex on the interval* $[0, c]$, *concave on* $[c, 1]$, *and decreasing on* $[0, 1]$, *then* $\tau_n(f)$ *increases and* $\sigma_n(f)$ *decreases as* n *increases.*

Proof Define

$$f_1(x) := \begin{cases} f(x) & \text{for } 0 \leq x \leq c \\ f(c) & \text{for } c < x \leq 1, \end{cases}$$

$$f_2(x) := \begin{cases} f(c) & \text{for } 0 \leq x < c \\ f(x) & \text{for } c \leq x \leq 1. \end{cases}$$

Observe first that $f_1(x)$ is convex and decreasing on $[0, 1]$. It is convex on $[0, 1]$ since if $0 \leq x_1 < c < x_2 \leq 1$, *and* $0 < \alpha < 1$ then $\alpha f_1(x_1) + (1-\alpha) f_1(x_2) = \alpha f(x_1) + (1-\alpha) f(c) \geq f(\alpha x_1 + (1-\alpha)c) = f_1(\alpha x_1 + (1-\alpha)c) \geq f_1(\alpha x_1 + (1-\alpha)x_2)$. Likewise, $f_2(x)$ is concave and decreasing on $[0, 1]$. Observe next that $f(x) +$

$f(c) = f_1(x) + f_2(x)$. It follows from Theorems 2 and 1 that $\tau_n(f_1)$ and $\tau_n(f_2)$ increase while $\sigma_n(f_1)$ and $\sigma_n(f_2)$ decrease. Since $\tau_n(f) + f(c) = \tau_n(f_1) + \tau_n(f_2)$ and $\sigma_n(f) + f(c) = \sigma_n(f_1) + \sigma_n(f_2)$, this yields the desired conclusion. □

Note that we cannot hope to have a version of Theorem 3 with convex and concave interchanged, since for $\chi_{[0,\frac{1}{2}]}$, the characteristic function of the interval $[0, \frac{1}{2}]$, which is concave on $[0, \frac{1}{2}]$ and convex on $[\frac{1}{2}, 1]$, we have $\tau_{2m-1} + \frac{1}{2(m-1)} = \tau_{2m} = \tau_{2m+1} + \frac{1}{2m}$. However, applying Theorem 3 to $-f$ yields.

Theorem 4 *If the function* $f : [0, 1] \to \mathbb{R}$ *is concave on the interval* $[0, c]$, *convex on* $[c, 1]$, *and increasing on* $[0, 1]$, *then* $\tau_n(f)$ *decreases and* $\sigma_n(f)$ *increases as n increases.*

Next, we prove

Theorem 5 *If the function* $f : [0, 1] \to \mathbb{R}$ *is concave on the interval* $[0, 1]$, *with maximum* $f(c)$, $0 < c < 1$, *then*

$$\tau_n(f) - \frac{f(c) - f(0)}{n}$$

increases as n increases.

Proof Define f_1 and f_2 as in the proof of Theorem 3, and note that $f_1(x)$ is concave and increasing on $[0, 1]$ while $f_2(x)$ is concave and decreasing on $[0, 1]$. The concavity of f_1 and f_2 can be verified by the method used in the proof of Theorem 3. It follows from Theorem 2 that $-\sigma_n(f_1)$ decreases and from Theorem 1 that $\tau_n(f_2)$ increases, and hence that

$$\begin{aligned}\sigma_n(f_1) + \tau_n(f_2) &= \tau_n(f_1) - \frac{f(c) - f(0)}{n} + \tau_n(f_2)\\ &= \tau_n(f) - \frac{f(c) - f(0)}{n} + f(c)\end{aligned}$$

increases as n increases. □

Corollary 1 *If the function* $f : [0, 1] \to \mathbb{R}$ *is concave on the interval* $[0, 1]$ *and symmetric about its midpoint, then*

$$\tau_n(f) - \frac{f(1/2) - f(0)}{n}$$

increases as n increases.

3.1 Symmetrisation

The *symmetrization* of $f : [0, 1] \to \mathbb{R}$ about $x = \frac{1}{2}$ is defined to be

$$F(x) := F_f(x) = \frac{1}{2}\left(f(x) + f(1-x)\right). \tag{15}$$

We will make use of F_f throughout the rest of this note and start by observing that such a symmetrization never destroys convexity or concavity and often improves it.

Example 2 (Concavity of the symmetrization of $1/(1+x^2)$*)* Although the function

$$f(x) = \frac{1}{1+x^2} \tag{16}$$

is neither convex or concave on [0, 1] its symmetrization,

$$F_f(x) = \frac{x^2 - x + 3/2}{\left(x^2+1\right)\left(x^2 - 2x + 2\right)} \tag{17}$$

is concave.

To establish this we show that $F_f''(x) \le 0$ on [0, 1]. Since F_f and hence F_f'' are symmetric about $\frac{1}{2}$ we need only show this on $[\frac{1}{2}, 1]$. Moreover, using the change of variable $x := \frac{1}{2}(y+1)$ this is equivalent to showing

$$F_f''\left(\frac{1}{2}(y+1)\right) \le 0, \quad \text{for } 0 \le y \le 1. \tag{18}$$

Now,

$$F_f''\left(\frac{1}{2}(y+1)\right) = \frac{8(y^8 + 44y^6 - 30y^4 - 660y^2 - 125)}{(y^2+2y+5)^3(y^2-2y+5)^3}. \tag{19}$$

The denominator of (19) is always positive while the numerator is a polynomial, say $p(y)$, that is negative both at $y = 0$ and $y = 1$. To show that it is negative throughout [0, 1] we invoke *Descartes' rule of signs*, see http://mathworld.wolfram.com/DescartesSignRule.html, which tells us that

> *for a real polynomial* p*, the number,* $n(p)$*, of zeros on the positive axis does not exceed the number of sign changes,* $s(p)$*, in the nonzero coefficients (in order) and that* $2|(n(p) - s(p))$.

The coefficients of $p(y)$ change signs only once so Descartes' rule of signs tells us that $p(y)$ has at most one positive zero. It follows that $p(y) \le 0$ for all $y \in (0, 1)$, indeed if $p(c) > 0$ for some $0 < c < 1$, then $p(y)$ must have a zero in $(0, c)$ and another zero in $(c, 1)$. This establishes (18) thus proving that $F_f(x)$ is concave on [0, 1]. $\Diamond$

Another example of a class of functions with a concave symmetrization is $f_a : x \mapsto e^{-\frac{1}{2}ax^2}$, on $x \in [0, 1]$. The functions are themselves only concave for

$0 < a \leq 1$, since $f_a''(x) = a\left(ax^2 - 1\right) f_a(x)$, whereas the symmetrization is concave for $0 < a \leq 4$, a fact we invite the reader to verify through an examination of F_{f_a}''.

4 Monotonicity and Symmetrization

Numerical experiments suggest it is very common for f to be such that τ_n and σ_n exhibit monotonicity but it is harder to find applicable conditions that assure this. Thus, we seek verifiable conditions that in particular will apply to $f(x) = 1/(1 + x^2)$. As will soon become apparent, calculations involving symmetric (concave) functions lead us naturally to the introduction of the following *symmetric Riemann sum.*

For $f : [0, 1] \to \mathbb{R}$ we define:

$$\lambda_n := \lambda_n(f) = \frac{1}{n}\sum_{k=0}^{n} f\left(\frac{k}{n}\right) - \frac{1}{n} f\left(\frac{1}{2}\right). \tag{20}$$

For all $n \in \mathbb{N}$, $\lambda_n(f)$ is linear and symmetric in that $\lambda_n(f) = \lambda_n(f(1 - \cdot))$ and so $\lambda_n(f) = \lambda_n(F_f)$, where F_f is the symmetrization of f.

The term involving $f\left(\frac{1}{2}\right)$ ensures that $\lambda_n(1) = 1$ by making a correction to the central term(s) of $\frac{1}{n}\sum_{k=0}^{n} f\left(\frac{k}{n}\right)$; if n is even we simply omit the central term, $\frac{1}{n} f(\frac{1}{2})$, while if n is odd we replace the two central terms by $\frac{1}{n}\left(f(\frac{1}{2} - \frac{1}{2n}) - f(\frac{1}{2}) + f(\frac{1}{2} + \frac{1}{2n})\right)$.

Further,

$$\lambda_n(f) = \frac{\tau_n + \sigma_n}{2} + \frac{1}{2n}\left(f(0) + f(1) - 2f\left(\frac{1}{2}\right)\right) \tag{21}$$

$$= \sigma_n + \frac{1}{n}\left(f(1) - f\left(\frac{1}{2}\right)\right) \tag{22}$$

$$= \tau_n + \frac{1}{n}\left(f(0) - f\left(\frac{1}{2}\right)\right). \tag{23}$$

As an immediate consequence of (23) and Corollary1 we get.

Theorem 6 (**Monotonicity for symmetric concave functions**) *If the function* $f : [0, 1] \to \mathbb{R}$ *is concave on the interval* $[0, 1]$ *and symmetric about its midpoint, then* $\lambda_n(f)$ *increases with n.*

Corollary 2 *If the function* $f : [0, 1] \to \mathbb{R}$ *has a concave symmetrization and* $f(0) > f(1/2)$, *then* τ_n *increases with n.*

Proof Theorem 6 applies to F_f to show that $\lambda_n(f) = \lambda_n(F_f)$ is increasing and the conclusion follows from (23). □

In particular we have.

Corollary 3 (**Monotonicity for decreasing functions with a concave symmetrization**) *If the function* $f : [0, 1] \to \mathbb{R}$ *is decreasing on the interval* $[0, 1]$ *and its symmetrization;* $F_f(x) = \frac{1}{2}(f(x) + f(1-x))$ *is concave, then* τ_n *increases with n, necessarily to* $\int_0^1 f$.

Example 3 *(Monotonicity of* τ_n *for* $1/(1+x^2)$*)* Consider the function $f(x) := 1/(1+x^2)$ for which

$$\tau_n := \sum_{k=1}^{n} \frac{n}{n^2 + k^2}.$$

Clearly f is decreasing on $[0, 1]$ and we already observed in Example 2 that its symmetrization $F_f(x) := \frac{1}{2}(f(x) + f(1-x))$ is concave, so Corollary 3 applies to show that τ_n is increasing. ♢

Similarly, for $f_a(x) := e^{-\frac{1}{2}ax^2}$ we see by calculating f_a' and F_{f_a}'' that $\tau_n(f_a)$ increases with n for all a in the range $0 \le a \le 4$.

Remark 1 *(Variations on the theme)*

Let $f : [0, 1] \to \mathbb{R}$. Noting from their linearity that $\tau_n(-f) = -\tau_n(f)$ and similarly for σ_n, and also observing that $\sigma_n(f(x)) = \tau_n(f(1-x))$, we can deduce the following variants of the results above.

(i) If f is symmetric and convex, then λ_n is decreasing. [Apply Theorem 6 to $-f$.]

(ii) If $f(0) < f(1/2)$ (in particular, if f is increasing) and has a convex symmetrization, then τ_n is decreasing. [Apply Corollary 2 to $-f$.]

(iii) If $f(1/2) < f(1)$ (in particular, if f is increasing) and has a concave symmetrization, then σ_n is increasing. [Apply Corollary 2 to $f(1-x)$.]

(iv) If $f(1/2) > f(1)$ (in particular, if f is decreasing) and has a convex symmetrization, then σ_n is decreasing. [Apply Corollary 2 to $-f(1-x)$.]

Since the symmetrization of f is concave (convex) if f is concave (convex) we observe that Corollary 2 and part (iv) extend the final two theorems in [4]; our theorems 1 and 2. ♢

5 Analysis of the Function $\frac{1}{1-bx+x^2}$

As a way of highlighting the subtleties in a seemingly innocent question, we finish by analyzing a one-parameter class of functions to which our results sometimes apply.

We consider the the family of functions

$$f_b : [0, 1] \to \mathbb{R}, \quad \text{where } f_b(x) := \frac{1}{x^2 - bx + 1} \tag{24}$$

in the parameter range $|b| < 2$ so that each f_b assumes only positive values.

The symmetrization of f_b about $1/2$ is

$$F_b(x) := \frac{x^2 - x + (3-b)/2}{\left(x^2 - bx + 1\right)\left(x^2 - (2-b)x + (2-b)\right)}. \tag{25}$$

Then $f_0(x) = 1/(1+x^2)$ while $f_1(x) = F_1(x) = 1/(x^2 - x + 1)$. Now F_0, F_1 and $F_{3/2}$ are concave on $[0, 1]$, while F_{-1} is convex and

$$F_2(x) = \frac{(1-x)x + 1/2}{(1-x)^2\, x^2}$$

is convex as an extended value function from $[0, 1]$ into $(-\infty, \infty]$. By contrast $F_{5/4}$ and $F_{7/4}$ are neither convex nor concave on the unit interval (for more details see Remark 2 below).

In passing we compute for $|b| < 2$ that

$$\int_0^1 \frac{\mathrm{d}x}{x^2 - bx + 1} = \frac{2}{\sqrt{4-b^2}}\left(\arctan\left(\frac{b}{\sqrt{4-b^2}}\right) + \arctan\left(\frac{2-b}{\sqrt{4-b^2}}\right)\right).$$

When $b \to -2$ we arrive at $\int_0^1 \frac{\mathrm{d}x}{x^2+2x+1} = \frac{1}{2}$.

With a view to applying Corollary 2 or Corollary 3, we begin by noting that $f_b(x)$ is decreasing on $[0, 1]$ for $b \le 0$ and increasing only for $b \ge 2$, however, $f_b(0) > f_b(1/2)$ whenever $b < 1/2$.

We next prove that F_b is concave for $0 \le b \le 1$; again we employ Descartes' rule of signs.

Theorem 7 (**Concavity of** F_b) *The function F_b given by* (25) *is concave on* $[0, 1]$ *for* $0 \le b \le 1$.

Proof To establish concavity of F_b we show that F_b'' is negative on $[0, 1]$, see Figure 2 and to do this we need only show its numerator polynomial, n_b, is negative, as the denominator is always positive.[1]

Further, since F_b and hence F_b'' are symmetric about $\frac{1}{2}$ we need only show this on $[1/2, 1]$. Moreover, using the change of variable $x := (y+1)/2$ allows us to use Descartes' rule of signs to detect roots of $n_b(x)$ for $x \ge 1/2$ (that is, for $y \ge 0$).

Now, the numerator of $F_b''((y+1)/2)$ is

$$\begin{aligned} n_b(y) :=& \; 24\, y^8 + 32\, \left(b^2 - 6\,b + 11\right) y^6 + 48\, (2\,b - 5)\left(6\,b^2 - 10\,b + 1\right) y^4 \\ & - 96\, (2\,b-5)\left(4\,b^2 - 2\,b - 11\right)(b-1)^2\, y^2 - 8\, \left(4\,b^2 - 6\,b - 1\right)(2\,b-5)^3 . \end{aligned} \tag{26}$$

[1] Note in Figure 2 how much clearer the situation is made by also plotting the horizontal plane.

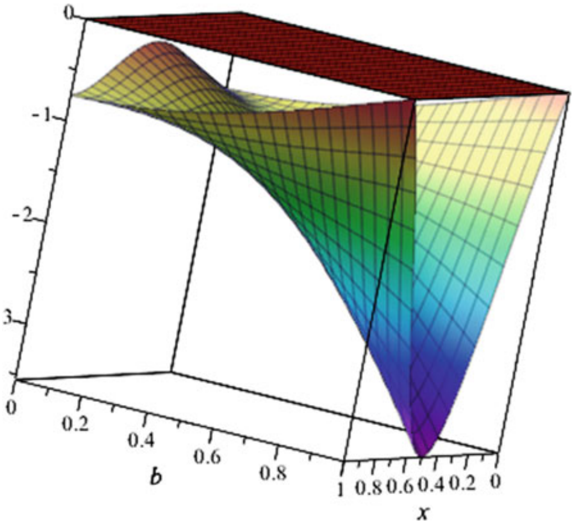


Fig. 2 The second derivative of F_b for $0 \leq b, x \leq 1$.

For $0 < b < 1$ the first two terms in (26) are always positive and the final two are negative, so that irrespective of the sign of the coefficient of y^4 (it in fact has three zeroes, at $5/2$ and $(5 \pm \sqrt{19})/6$) Descartes' rule of signs applies to show the numerator has one positive real zero (including multiplicity). This zero must lie to the right of the point 1 except for $b = 1$ when it equals 1, as illustrated in Figure 3. (Note how close to one the inflection point is for $b = 5/4$.)

For $0 \leq b < 1$ we have

$$n_b(0) = 8\left(4\,b^2 - 6\,b - 1\right)(5 - 2\,b)^3 < 0$$

and

$$n_b(1) = -1024\,(b-2)\,(b-1)\left(b^3 - 3\,b^2 + 3\right) < 0.$$

Thus, when $0 \leq b \leq 1$ the numerator is nonpositive for $y \in [-1, 1)$ and so $F_b(x)$ is concave on $[0, 1]$. □

This proof of concavity for F_b was discovered by examining animations of the behavior of n_b and then getting a computer algebra system to provide the requisite expressions after shifting the symmetry to zero so that Descartes' rule was applicable. Some snapshots of the animation are illustrated in Figure 3. The animation makes it clear that the solution of $n_b(y) = 1$ decreases monotonically with b.

Remark 2 (Convexity properties throughout the range $|b| < 2$*)* In this range the function provides further interesting applications of Descartes' rule.

A careful analysis of the coefficients a_k of y^{2k} for $k = 0, 1, 2, 3$ in (26) and of the signs of $n_b(0)$ and $n_b(1)$ [see Figure 3 where we plot $n_b(0)$ and $n_b(1)$ with $n_0(b)$ a

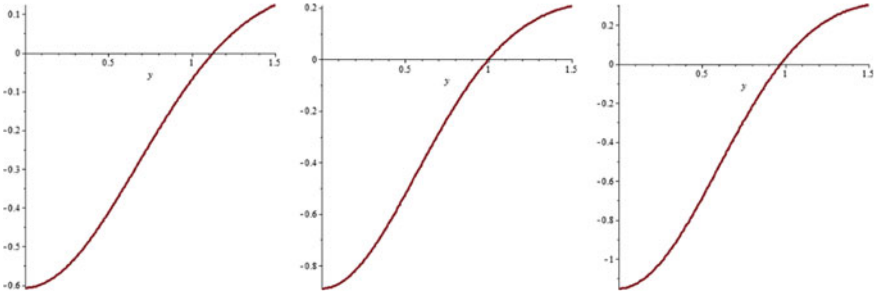


Fig. 3 Graph of $n_b(y)$ on [0, 3/2] for $b = 3/4$ (L), $b = 1$ (M), and $b = 5/4$ (R)

Table 1 Table of signs

b	$[-2, \delta_-]$	$[\delta_-, \alpha_-]$	$[\alpha_-, \beta_-]$	$[\beta_-, \gamma_-]$	$[\gamma_-, 1]$	$[1, \alpha]$	$[\alpha, \gamma_+]$	$[\gamma_+, \beta_+]$	$[\beta_+, \delta_+]$	$[\delta_+, 2)$
a_4	+	+	+	+	+	+	+	+	+	+
a_3	+	+	+	+	+	+	+	+	+	+
a_2	−	−	−	−	+	+++	+	−	−	−
a_1	+	−	−	−	−	−	−	−	−	+
a_0	+	+	+	−	−	−	−	−	+	+
$n_b(0)$	+	+	+	−	−	−	−	−	+	+
$n_b(1)$	+	+	−	−	−	+	−	−	−	−
#	2	2	2	1	1	1	1	1	2	2
$F_b(x)$	conv	conv	infl	conc	conc	infl	conc	conc	infl	infl

dashed line], coupled with reasoning similar to that in the proof of Theorem 7 allows us to extend the results of that theorem to the whole parameter range $|b| < 2$.

The analysis and conclusions are summarized in Table 1, wherein we denote

$$\begin{aligned}
\alpha_- &= \text{the negative root of } b^4 - 3b^2 + 3 \approx -0.8794\\
\alpha &= \text{the smallest positive root of } b^4 - 3b^2 + 3 = 1 + \sqrt{3}\sin(2\pi/9) \approx 1.3473\\
\alpha_+ &= \text{the largest root of } b^4 - 3b^2 + 3 \approx 2.5231\\
\beta_-,\ \beta_+ &= \text{the roots of } 4b^2 - 6b - 1 = (3 \pm \sqrt{13})/3 \approx -0.1539,\ 1.6514\\
\gamma_-,\ \gamma_+ &= \text{the roots of } 6b^2 - 10b + 1 = (5 \pm \sqrt{19})/6 \approx -0.1069,\ 1.5598\\
\delta_-,\ \delta_+ &= \text{the roots of } 4b^2 - 2b - 11 = (1 \pm \sqrt{45})/4 \approx -1.4271,\ 1.9271\\
\text{and}&\\
\# &= \text{the number of positive roots of } n_b((y+1)/2)
\end{aligned}$$

The conclusion that F_b is convex for $-2 < b \le \alpha_-$ requires the observation that in this range $n_b(x)$ is negative for values of $x > 1$, so neither positive root can lie within the interval [0,1].

Putting all this together we are able to conclude that the sequence $\tau_n(f_b)$ is increasing for $b \in [\beta_-, 1/2]$ and $\sigma_n(f_b)$ is decreasing for $b \in [-2, \alpha_-]$.

A similar analysis in the cases $|b| > 2$ is left to the interested reader. ♢

6 Concluding Remarks

The story we have told highlights the many accessible ways that the computer and the Internet can enrich mathematical research and instruction. The story would be even more complete if we could also deduce that $\sigma_n(1/(1+x^2))$ was decreasing.

References

1. Apostol, T.: Mathematical Analysis. Addison-Wesley, Boston (1974)
2. Borwein, J.M.: The life of modern homo habilis mathematicus: experimental computation and visual theorems. In: Monaghan, J., Troche, L., Borwein, J. (eds.) Tools and Mathematics: Instruments for Learning. Springer, Berlin (2015)
3. Borwein, J.M., Zhu, Q.J.: Variational methods in the presence of symmetry. Adv. Nonlinear Anal. **2**(3), 271–307 (2013). https://doi.org/10.1515/anona-2013-1001
4. Szilárd, András: Monotonicity of certain Riemann type sums. Teach. Math. **15**(2), 113–120 (2012)

Risk and Utility in the Duality Framework of Convex Analysis

R. Tyrrell Rockafellar

1 Preferences Under Uncertainty

Utility functions were developed in economics for the purpose of describing the preferences of individuals among various bundles of goods. Their basis in axioms involved choosing between "lotteries" in which the quantities in a bundle would be obtained with different probabilities. Utility theory, from that angle, was always connected with preferences under uncertainty, but in more recent times it has been important also in guiding financial decisions about portfolios of stocks and bonds. A portfolio put together on a given date at a given cost will have, at a targeted future date, a monetary value that can be viewed as a random variable. Preferences among various portfolio choices come down then to preferences among random variables of returns, and an important approach to that has been to make comparisons in terms of the expected utility of those returns with respect to a utility function for money.

Utility has not been the only way of looking at preferences in finance, however. An increasingly popular alternative has been the quantification of the risk inherent in a random variable, most conveniently oriented in this case not on the attractiveness of rewards but rather on the distaste for losses. There is by now a well-developed theory of such quantification, which like utility theory can capture how the attitudes of one decision-maker may differ from those of another.

Are these two notions, risk and utility, in competition, or do they fit together harmoniously in some larger pattern? Convex analysis provides a mathematical framework in which they not only fit together but also interact productively. Ben-Tal and Teboulle [5] were the first to recognize this on a fundamental level, but the subject has since been pushed into broader territory by Rockafellar and Uryasev in [20].

R. T. Rockafellar (✉)
Department of Mathematics, University of Washington, Box 354350,
Seattle, WA 98195-4350, USA
e-mail: rtr@uw.edu
URL: http://www.math.washington.edu

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_3

Here we offer a consolidated presentation which emphasizes conjugate duality and provides some new results. Duality famously ties risk to the contemplation of sets of probability measures instead of just a single specified probability measure, as first was demonstrated by Artzner et al. in [2]. Our framework is able from that perspective to coordinate such stochastic ambiguity, as applied to expected utility, with particular forms of risk, and this is one of the new contributions made here. It furthermore offers a different perspective on utility which is better suited to comparisons to a benchmark than the customary approach in finance, as seen for example in the treatment in references like the book of Föllmer and Schied [7].

To begin, we have to set the stage with a probability space $(\Omega, \mathcal{F}, P_0)$, where Ω is the set of "scenarios" or "future states" ω, $\mathcal{F}$ is a field of subsets of Ω, and P_0 is a probability measure on $\mathcal{F}$. The reason for the subscript 0 is that we will come to comparing other probability measures P to P_0, which itself may in general just be a nominal choice, perhaps arising empirically or from subjective guesswork. Any measurable function $X : \Omega \to \boldsymbol{R}$ can be interpreted as a *random variable* for which the cumulative distribution function F_X from $(-\infty, \infty) \to [0, 1]$ is given by

$$F_X(\tau) = \text{prob}[X(\omega) \leq \tau] = P_0\big[\big\{\, \omega \in \Omega \,\big|\, X(\omega) \leq \tau \big\}\big].$$

The traditional spaces of random variables are

$$\mathcal{L}^p = \mathcal{L}^p(\Omega, \mathcal{F}, P_0) = \begin{cases} \big\{ X \,\big|\, E[|X|^p] < \infty \big\} & \text{for } \ 1 \leq p < \infty, \\ \big\{ X \,\big|\, \sup |X| < \infty \big\} & \text{for } \ p = \infty, \end{cases}$$

where the expectation E of a random variable is its integral with respect to the probability measure P_0, and sup refers to the essential supremum of a function. But which of these might be "best" to work with?

The answer is not simple here because of the usual duality of $\mathcal{L}^p$ versus $\mathcal{L}^q$. The problem is that through consideration of alternative probability measures P we are led to restricting P to be absolutely continuous with respect to P_0 and expressing the expected value of X with respect to P in terms of the Radon–Nikodym derivative dP/dP_0 as

$$E_P[X] = E\Big[X \frac{dP}{dP_0}\Big] = \int_\Omega X(\omega) \frac{dP}{dP_0}(\omega) dP_0(\omega). \tag{1}$$

For $X \in \mathcal{L}^1$ this dictates restricting P to the class of probability measures such that $dP/dP_0 \in \mathcal{L}^\infty$, which is severe. Embracing the broadest generality in probability measures by having the derivatives dP/dP_0 be in $\mathcal{L}^1$, on the other hand, would seem to limit the random variables X to be essentially bounded, also perhaps not a very good idea.

In the face of this dilemma, we will follow [20] in taking $\mathcal{L}^2$ as the space to work with, the inner product there being

$$\langle X, Q \rangle = E[XQ] = \int_\Omega X(\omega) Q(\omega) dP_0(\omega).$$

For random variables $X \in \mathcal{L}^2$, of course, both the mean and variance,

$$\mu(X) = E[X], \qquad \sigma^2(X) = E[(X - \mu(X))^2] = E[X^2] - E[X]^2,$$

are well defined and finite; and in fact this characterizes our choice. The probability measures P that will come into considerations as alternatives to P_0 will be those that are absolutely continuous with respect to P_0 and have densities dP/dP_0 belonging to $\mathcal{L}^2$. The set of all such density functions forms the "unit simplex" $\mathcal{P}_0$ of $\mathcal{L}^2$,

$$\begin{aligned} \mathcal{P}_0 &= \left\{ Q \in \mathcal{L}^2 \,\middle|\, Q \geq 0,\ E[Q] = 1 \right\} \qquad (2) \\ &= \left\{ dP/dP_0 \in \mathcal{L}^2 \,\middle|\, P \ \text{ absolutely continuous w.r.t. } \ P_0 \right\} \end{aligned}$$

A sort of misalignment in our topic between minimization and maximization comes from the orientation of risk toward losses and utility toward gains. This will be reconciled later by inserting "regret" as an anti-utility, or disutility, suitable for minimization instead of maximization, but for now we will proceed just with risk.

2 How Should "Risk" Be Understood?

In thinking of a random variable X as standing for loss, or cost, or various other things that a decision-maker would want to be lower[1] rather than higher, with negative values interpreted as "good," we come to the idea of quantifying risk as opposed to uncertainty in a random variable X. Risk definitely is involved with uncertainty, but there is an important distinction. An example is afforded by a lottery that offers the loss of \$1,000,000 with probability .99 and a loss of \$-1 (a reward of \$1) with probability .01. The uncertainty is very small, but the risk of participating in the lottery, as we are thinking about it here, is nearly \$1,000,000.

Definition 1 By a *risk quantifier*[2] we will mean a functional $\mathcal{R}$ that assigns to a random variable X a value $\mathcal{R}(X) \in (-\infty, \infty]$ standing as a surrogate for the "overall risk" deemed to be present in X.

Some immediate and commonly employed examples are

$$\begin{aligned} \mathcal{R}(X) &= E[X] = \ \text{expected loss ("average" loss)}, \\ \mathcal{R}(X) &= \sup X = \ \text{worst-case loss}, \\ \mathcal{R}(X) &= E[X] + \lambda\sigma^2(X) = \ \text{mean/variance loss with parameter } \ \lambda > 0, \\ \mathcal{R}(X) &= q_\alpha(X) = \ \text{quantile loss at probability level } \ \alpha \in (0, 1), \end{aligned}$$

where the α-quantile of X is defined by

[1] Often people speak of smaller instead of lower, but smaller means "closer to 0," while lower means "closer to $-\infty$."

[2] The usual terminology in finance is "measure of risk" or "risk measure," but here we experiment with trying to avoid conflict with the standard concept of a "measure" in mathematics.

$$q_\alpha(X) = \min\left\{\, \tau \;\middle|\; F_X(\tau) \geq \alpha \right\},$$

and characterized in probability by

$$q_\alpha(X) \leq \tau \quad \Longleftrightarrow \quad \text{prob}[X > \tau] < 1 - \alpha.$$

In finance, the quantile loss is called the *value-at-risk* at the probability level α and denoted by $\text{VaR}_\alpha(X)$. A less obvious concept, closely related to to the latter, is

$$\mathcal{R}(X) = \overline{q}_\alpha(X) = \text{ superquantile loss at level } \quad \alpha \in (0, 1),$$

which was developed as the *conditional value-at-risk*, $\text{CVaR}_\alpha(X)$, in [18, 19] in defining it to be the average of X in its upper α-tail[3] but can also be expressed by the averaging formula[4]

$$\overline{q}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 q_\beta(X)d\beta$$

and most valuably by

$$\overline{q}_\alpha(X) = \min_{C \in \mathbb{R}} \left\{ C + \frac{1}{1-\alpha} E\big[\max\{0, X - C\}\big] \right\}. \tag{3}$$

The last is a type of risk formula that will have an important role later in connecting risk to utility. An interesting feature of the expression in (3) is that the quantile $q_\alpha(X)$ is a value of C that attains the minimum,[5] so the same one-dimensional optimization problem serves to calculate both, and furthermore does so with no need of delving into conditional expectations.[6] The term “superquantile” as a substitute for “conditional value-at-risk” was proposed in [14] as being more suitable for applications outside of finance where “quantile” was already preferred to “value-at-risk.”

Of course, not every functional $\mathcal{R}$ would deserve to be called a risk quantifier. What properties might be required in general, and to what extent do the examples just given meet the test? This issue was forcefully raised in finance by Artzner et al. in [2] in a critique of the widespread use of value-at-risk in assessing the safety of portfolios and in particular banking reserves. They took for the first time an axiomatic

[3] This is the conditional expectation of X with respect to X exceeding its α-quantile as long as the probability of X actually equaling its α-quantile is 0. But when that probability is not 0 the definition involves a more careful interpretation of the α-tail distribution associated with X, as set forth in [19].

[4] On the basis of this formula, the concept has been called “average value-at-risk” by Föllmer and Schied [7].

[5] In general the argmin set, if not actually a singleton, is a closed interval having the quantile as its left endpoint.

[6] The intimate connection between this formula and the very concept of the cumulative distribution function for a random variable X has been explained in [15].

approach to risk and argued that a valid quantifier should be *coherent* in the sense of satisfying[7]

$$\begin{aligned}
&(r1)\quad \mathcal{R}(X+X') \le \mathcal{R}(X)+\mathcal{R}(X'),\\
&(r2)\quad \mathcal{R}(\lambda X) = \lambda\mathcal{R}(X) \quad\text{when}\quad \lambda > 0,\\
&(r3)\quad \mathcal{R}(X) \le \mathcal{R}(X') \quad\text{when}\quad X \le X',\\
&(r4)\quad \mathcal{R}(X+C) = \mathcal{R}(X)+C \quad\text{for constants}\quad C,
\end{aligned} \tag{4}$$

where $X \le X'$ means that prob$[X > X'] = 0$. For them, axiom (r1) was critical under the interpretation that $\mathcal{R}(X)$ stands for the amount of cash that should be held in reserve to cover possible losses in a portfolio X. If two portfolios X and X' had $\mathcal{R}(X+X') > \mathcal{R}(X)+\mathcal{R}(X')$, that would signify that the portfolio obtained by combining them required more cash in reserve than the uncombined portfolios and thus was actually riskier. That situation, counter to the virtues of diversification, is one of the big troubles they identified in value-at-risk, i.e., in using quantiles to quantify risk. It is not the only trouble, though: quantiles can behave discontinuously, even inevitably so in the case of a *finite* probability space, i.e., when Ω is a finite discrete set.

Axioms (r1) and (r2) together are equivalent to $\mathcal{R}$ being *sublinear* in the sense that

$$\mathcal{R}\Big(\sum\nolimits_{i=1}^{m} \lambda_i X_i\Big) \le \sum\nolimits_{i=1}^{m} \lambda_i \mathcal{R}(X_i) \quad\text{for}\quad \lambda_i > 0,$$

and then, in particular, $\mathcal{R}$ is *convex*. That is where convex analysis was first brought into risk theory in finance. Axiom (r3) seems utterly desirable, but in fact, it stands in the way of another practice long followed in portfolio optimization in finance, namely, concentrating on mean-variance loss criteria. With that, it is actually possible to have X deemed less risky than X' even though the outcomes of X are worse than the outcomes of X' with probability 1! Actually, mean-variance loss does not even satisfy (r1) or (r2), but that can be remedied by switching to *mean-devation* loss, obtained by putting the standard deviation $\sigma(X)$ in place of the variance σ^2. On the other hand, it can be questioned whether the linear scaling axiom (r2) is really appropriate, the suggestion being that the sublinearity (r1)+(r2) might better be replaced by the direct assumption of convexity.

The ground-breaking coherency axioms (4) were only articulated in [2] for finite-valued $\mathcal{R}$, no ∞ admitted, and for a *finite* probability space, in which case the $\mathcal{L}^p$ spaces of random variables are finite-dimensional and differ only in norm (not topology). Then the convexity of $\mathcal{R}$ coming from the sublinearity in (r1)+(r2) implies the *continuity* of $\mathcal{R}$. To cover general probability spaces, however, $\mathcal{R}(X)$ should be allowed to be ∞, with the case of $\mathcal{R}(X) = \sup X$ being a prime candidate, and then a topological assumption, namely lower semicontinuity, needs to brought in to serve in place of the no-longer-automatic continuity.

These reflections on the subject lead to a broader set of axioms:

[7] They couched their conditions in a context of gains instead of losses, as we express them here.

$$\begin{array}{ll} (R1) & \mathcal{R} \text{ is lower semicontinuous, possibly taking on } \infty, \\ (R2) & \mathcal{R}((1-\lambda)X+\lambda X') \leq (1-\lambda)\mathcal{R}(X)+\lambda\mathcal{R}(X') \quad \text{for} \quad \lambda \in (0,1), \\ (R3) & \mathcal{R}(X) \leq \mathcal{R}(X') \quad \text{when} \quad X \leq X', \\ (R4) & \mathcal{R}(C) = C \quad \text{for constants} \quad C, \end{array} \tag{5}$$

with (R2) being the explicit assumption of convexity. Under that convexity the simpler (and easier to defend) condition (R4) in place of (r4) is actually equivalent to (r4), cf. [20]. An additional condition to contemplate in the mix as a sharpening of (R4) is

$$(R4') \quad \mathcal{R}(X) > E[X] \quad \text{for nonconstant} \quad X, \tag{6}$$

which is called *aversity* and insists on a positive "risk premium" being tied to uncertainty. It was first brought up in [21] and turned out to be crucial for the scheme developed in [20].

Where does this leave us with examples? Mean/variance and mean/deviation fail (R3) as with (r3), and quantile risk fails (R2) as it previously failed (r1). The "risk neutral" quantifier $\mathcal{R}(X) = E[X]$ satisfies (R1)–(R4) but lacks the aversity in (R4′). But superquantile risk and worst-case risk (its limit as $\alpha \nearrow 1$) satisfy *all* the conditions. Another, quite different looking example satisfying all the conditions is

$$\mathcal{R}(X) = \log E\big[\exp X\big] = \text{ log-exponential loss,}$$

which, in contrast to the ones just mentioned is truly just convex and not also sublinear, i.e., does not have (r1)+(r2). Furthermore any sum

$$\mathcal{R}(X) = \lambda_1\mathcal{R}_1(X) + \cdots + \lambda_m\mathcal{R}_m(X) \quad \text{with} \quad \lambda_i > 0,\ \lambda_1 + \cdots + \lambda_m = 1,$$

in which each $\mathcal{R}_i$ satisfies (R1)–(R4) again satisfies (R1)–(R4), and if at least one $\mathcal{R}_i$ has the aversity in (R4′), then $\mathcal{R}$ has it as well.[8]

For the broad class of risk quantifiers characterized by (R1)–(R4) and maybe also (R4′), a fundamental issue of convex analysis can be raised, namely, duality. Whenever we have a convex functional $\mathcal{R} : \mathcal{L}^2 \to (-\infty, \infty]$ that is lower semicontinuous and $\not\equiv \infty$, the formula

$$\mathcal{R}^*(Q) = \sup_X \big\{ \langle X, Q\rangle - \mathcal{R}(X) \big\}$$

defines a *conjugate* convex functional $\mathcal{R}^* : \mathcal{L}^2 \to (-\infty, \infty]$ that likewise is lower semicontinuous and $\not\equiv \infty$, and then $\mathcal{R}^{**} = \mathcal{R}$, or in other words,

$$\mathcal{R}(X) = \sup_Q \big\{ \langle X, Q\rangle - \mathcal{R}^*(Q) \big\}, \tag{7}$$

[8] This rule carries over from discrete sums to "continuous sums" with respect to a "weighting measure," see [1, 17, 21].

For our class of risk quantifiers $\mathcal{R}$, what are the conjugates $\mathcal{R}^*$ and what do they tell us about risk?

The answer is beautifully informative and, in the domain of conjugate duality—with its extensive catalog of dualization of properties—is a routine exercise to obtain, although such background was not familiar to researchers in finance. It takes us back to considering alternative probability measures P to our nominal probability measure P_0 in the context of their densities $Q = dP/dP_0$ in the probability simplex $\mathcal{P}_0$ in (2).

Theorem 1 *(risk dualization) The class of functionals $\mathcal{J}$ on $\mathcal{L}^2$ that come up as conjugates $\mathcal{R}^*$ of risk quantifiers $\mathcal{R}$ satisfying (R1)–(R4) is characterized by*

$$\begin{aligned}
&(J1) \quad \mathcal{J} \quad \textit{is lower semicontinuous,}\\
&(J2) \quad \mathcal{J}((1-\lambda)X + \lambda X') \le (1-\lambda)\mathcal{J}(X) + \lambda\mathcal{J}(X') \quad \textit{for} \quad \lambda \in (0,1),\\
&(J3) \quad \mathcal{P}_0 \supset \mathcal{Q} = \operatorname{dom}\mathcal{J} = \left\{ Q \,\middle|\, \mathcal{J}(Q) < \infty \right\},\\
&(J4) \quad \inf_{Q\in\mathcal{Q}} \mathcal{J}(Q) = 0.
\end{aligned}$$

The extra condition (R4′) corresponds in this to

$$(J4') \quad \mathcal{J}(1) = 0 \quad \textit{(entailing} \quad 1 \in \mathcal{Q} \quad \textit{), but} \quad \partial\mathcal{J}(1) \quad \textit{contains no nonconstant} \quad X.$$

The subclass consisting of the conjugates $\mathcal{R}^*$ of functionals $\mathcal{R}$ which are sublinear instead of just convex as in (R2) is identified with $\mathcal{J}$ being $\equiv 0$ on $\operatorname{dom}\mathcal{J}$. This subclass thus consists of the functionals $\mathcal{J}$ of the form

$$\mathcal{J} = \delta_{\mathcal{Q}} \quad \textit{for some nonempty closed convex set} \quad \mathcal{Q} \subset \mathcal{P}_0, \tag{8}$$

where $\delta_{\mathcal{Q}}$ denotes, as usual, the indicator of $\mathcal{Q}$, having the value 0 on $\mathcal{Q}$ but ∞ outside.

Proof For the most part, these facts are contained in the Envelope Theorem of [20] and also known elsewhere, but the treatment of (J4) and (J4′) has not been seen in this form. The equivalence of (J4) with (R4) under conjugacy is obvious from having $\mathcal{R}(X) = \sup_Q \left\{ E[XQ] - \mathcal{J}(Q)\right\}$ in taking $X = C$ for a constant C, since $E[CQ] = C$. The “1” in (J4′) refers to the constant function 1 as an element of $\mathcal{L}^2$, which is the density dP_0/dP_0. Since $\mathcal{R}(X) = \sup_Q \left\{ E[XQ] - \mathcal{J}(Q)\right\}$, having $\mathcal{J}(1) = 0$ says in conjunction with (J3) that $\mathcal{R}(X) \ge E[X]$. Subgradients of $\mathcal{J}$ are defined by

$$X \in \partial\mathcal{J}(Q) \iff \mathcal{J}(Q') \ge \mathcal{J}(Q) + \langle X, Q' - Q\rangle = \mathcal{J}(Q) + E[XQ'] - E[XQ],$$

for all Q', or equivalently through conjugacy, $\mathcal{R}(X) = E[XQ] - \mathcal{J}(Q)$, so having $X \in \partial\mathcal{J}(1)$ with $\mathcal{J}(1) = 0$ amounts to having $\mathcal{R}(X) = E[X]$. Forbidding this for nonconstant X corresponds to (R4′). □

The main revelation here is that selecting a risk quantifier $\mathcal{R}$ corresponds in a unique way through duality to selecting a nonempty convex subset set $\mathcal{Q}$ of $\mathcal{P}_0$ along with a nonnegative convex expression $\mathcal{J}(Q)$ on $\mathcal{Q}$ such that (J4) holds and the level sets $\{\, Q \in \mathcal{Q} \,|\, \mathcal{J}(Q) \le c \,\}$ for $c \in \boldsymbol{R}$ are closed.[9] In this way, from (7), a representation of $\mathcal{R}$ is obtained in the form

$$\mathcal{R}(X) = \sup_{Q\in\mathcal{Q}} \big\{\, E[XQ] - \mathcal{J}(Q) \big\} \tag{9}$$

which has a rewarding interpretation in probability. As a subset of the probability simplex $\mathcal{P}_0$ in (2), $\mathcal{Q}$ corresponds to *a set* $\mathcal{P}$ *of probability measures* for which (9) can be written via (1) as

$$\mathcal{R}(X) = \sup_{P\in\mathcal{P}} \big\{\, E_P[X] - \mathcal{J}(dP/dP_0) \big\}. \tag{10}$$

When $\mathcal{R}$ is sublinear, we are in the case of (8) and the representation comes down to

$$\mathcal{R}(X) = \sup_{Q\in\mathcal{Q}} E[XQ] = \sup_{P\in\mathcal{P}} E_P[X]. \tag{11}$$

Having $1 \in \mathcal{Q}$ as in (J4′) translates to having $P_0 \in \mathcal{P}$, inasmuch as $dP/dP_0 = 1$ means $P = P_0$.

The formula $\mathcal{R}(X) = \sup_{Q\in\mathcal{Q}_0} E[XQ]$ for an arbitrary subset $\mathcal{Q}_0 \ne \emptyset$ of $\mathcal{P}_0$, not necessarily convex and possibly just finite, still yields a risk quantifier satisfying (R1)–(R4), and different choices of such $\mathcal{Q}_0$ can give the same $\mathcal{R}$. But then (11) holds also for $\mathcal{Q}$ being the closed convex hull of $\mathcal{Q}_0$, which is uniquely the largest set of densities Q that can serve in this manner.

Dualization in the form of (11) holding for a closed convex set $\mathcal{Q} \subset \mathcal{P}_0$ is illustrated by

$$\begin{aligned} \mathcal{R}(X) = E[X] \quad &\longleftrightarrow \quad \mathcal{Q} = \{1\}, \\ \mathcal{R}(X) = \sup X \quad &\longleftrightarrow \quad \mathcal{Q} = \mathcal{P}_0, \\ \mathcal{R}(X) = \overline{q}_\alpha(X) \quad &\longleftrightarrow \quad \mathcal{Q} = \big\{\, Q \in \mathcal{P}_0 \,\big|\, Q \le (1-\alpha)^{-1} \big\}. \end{aligned}$$

An example in the broader picture of $\mathcal{J}$ not just being an indicator function as in (8) is

$$\mathcal{R}(X) = \log E[\exp X] \quad \longleftrightarrow \quad \mathcal{Q} = \mathcal{P}_0,\ \ \mathcal{J}(Q) = E[Q \log Q] \ \text{ on } \ \mathcal{P}_0, \tag{12}$$

with $Q(\omega)\log Q(\omega)$ taken to be 0 when $Q(\omega) = 0$. This is striking because

$$E[Q \log Q] \quad \text{for} \quad Q = dP/dP_0, \tag{13}$$

[9] Extending $\mathcal{J}$ from $\mathcal{Q}$ to all of $\mathcal{L}^2$ by assigning it the value ∞ outside of $\mathcal{Q}$ results, under this closedness condition, in $\mathcal{J}$ being lower semicontinuous.

is the Kullback-Leibler distance of P from P_0, also known as the *relative entropy* of P with respect to P_0. Many more examples are available in [20]; see also [17]. Observe that in the general case of when $\mathcal{J}$ is not just an indicator, (J4′) makes $\mathcal{J}(dP/dP_0)$ give a sort of penalty related to how far P is from P_0; the Kullback-Leibler distance in (13) is just one illustration of that phenomenon.

The chief lesson in this is that the systematic approach to “risk” as axiomatized by (R1)–(R4) intrinsically leads through duality to the kind of worst-case analysis in (10), or more specially (11), in which some collection of probability measures is brought into action, not just P_0, to achieve “robustness.” This brings out a fundamental connection between risk quantifiers and “robust optimization,” which is a term that has become popular for problem formulations in which a worst-case expression is minimized or constrained [3, 4]. In recent years this idea has pursued more specifically under the heading of “distributionally robust optimization” in which special schemes for constructing collections of alternative probability measures are explored, cf. [8, 10, 23].[10] The term “stochastic ambiguity” is likewise often used in situations where multiple probability measures are under consideration. Both distributional robustness and stochastic ambiguity are therefore at the heart of risk theory. Interestingly, stochastic ambiguity can also enter on a higher level and then be coordinated with the stochastic ambiguity in risk, as will be seen in Section 4.

3 How Should Utility Be Understood?

Utility theory in economics is traditionally occupied with preferences among bundles of goods, but here we are focusing on preferences among scalar random variables such as arise in financial optimization or reliability engineering. We have been looking so far at loss-oriented random variables X with the idea that X will be preferable to X' when, with respect to a risk quantifier $\mathcal{R}$ that reflects our interests, $\mathcal{R}(X) < \mathcal{R}(X')$. In taking up utility, however, we switch to the opposite orientation, signaled here notationally (to reduce confusion) by Y as the symbol for a random variable for which we like outcomes to be higher rather than lower. The space of random variables is, as always, $\mathcal{L}^2 = \mathcal{L}^2(\Omega, \mathcal{F}, P_0)$.

Without yet pinning down any specifics, we can build our discussion around the notion that a *utility quantifier* is a possibly extended-real-valued functional $\mathcal{U}$ on $\mathcal{L}^2$ which is aimed at aiding decisions by marking Y as preferred to Y' when $\mathcal{U}(Y) > \mathcal{U}(Y')$. Examples of this already in use will lead us to identifying specific axioms to place on such $\mathcal{U}$, although an obvious one from the start ought to be monotonicity: $Y \geq Y'$ should imply $\mathcal{U}(Y) \geq \mathcal{U}(Y')$.

The prime example of $\mathcal{U}$ to begin with is simple *expected utility*:

[10] Researchers in that subject don’t seem aware that it is a branch of the general theory of coherent measures of risk initiated much earlier in [2].

$$\mathcal{U}(Y) = E[u(Y)] = \int_\Omega u(Y(\omega))dP_0(\omega), \tag{14}$$

for a function $u : (-\infty, \infty) \to [-\infty, \infty)$. Here $-\infty$ has been allowed as a value of u to cover cases where the natural domain of u is not the whole real line, as for instance if $u(y) = \log y$ or $u(y) = \sqrt{y}$; then u is extended by $-\infty$. Clearly we, should want the utility function u to be nondecreasing since that will lead to the expected utility being monotone in assigning preferences. What else? In line with utility theory more generally, we should likely want the set of Y' preferred to Y to be convex. That essentially dictates in (14) that u should be a concave function, and then $\mathcal{U}$ will be concave.

But expected utility as in (14) is not the only approach to utility preferences. We can turn to *stochastic ambiguity* to get examples of the kind

$$\mathcal{U}(Y) = \inf_{P \in \mathcal{P}} E_P[u(Y)] = \inf_{P \in \mathcal{P}} E\Big[u(Y)\frac{dP}{dP_0}\Big] \tag{15}$$

for a set $\mathcal{P}$ of probability measures P that is contemplated as potentially having to be faced instead of just P_0. Clearly this could tie in with the risk ideas at the end of Section 2, and indeed we will come back to that in Section 4. The point for now is that (15) presents candidates beyond those in (14) which are definitely worthy of including in our picture of utility quantifiers.

Still another line of thinking leads to "relative utility" in the sense of comparing outcomes to some benchmark.[11] Thus, there may be some random variable B with respect to which we are interested in

$$\begin{aligned} \mathcal{U}(Y) &= E[u(B+Y)] - E[u(B)] = \textstyle\int_\Omega u_B(Y(\omega), \omega)dP_0(\omega) \ \text{ for} \\ u_B(y, \omega) &= u(B(\omega) + y) - u(B(\omega)). \end{aligned} \tag{16}$$

This is not of course covered by (14), because we do not just have a function $u_B(y)$ unless $B(\omega)$ is actually a constant b independent of ω.

In (16) we have $\mathcal{U}(0) = 0$, and that brings up something further. Back in the case of (14), there is a utility function u, but there is some arbitrariness in choosing it for getting preferences. The same standards for when Y is preferred to Y' would be captured by selecting any b with $u(b) > -\infty$ and replacing $u(y)$ by $u_b(y) = u(b+y) - u(b)$. Thus, no loss of generality is incurred by supposing in (14) that $u(0) = 0$, in which case $\mathcal{U}(0) = 0$ there as well. Moreover this would transfer to the setting of stochastic ambiguity in (15).

In this vein of "normalizing" a utility function u, we can also contemplate rescaling, since that too would not affect the preferences coming from it. This can have an important effect on relating utility to expectation, in connection with already having $u(0) = 0$. For instance in the case of u being differentiable at 0, if $u'(0) > 0$, as would be natural from the standpoint of monotonicity, we can divide by u by $u'(0)$

[11] In finance, for example one might want to compare the gains Y from a portfolio to a well known stock market index.

to normalize to having $u'(0) = 1$. With the concavity then implying $u(y) \leq y$ for all y, we thus can arrange that $\mathcal{U}(Y) \leq E[Y]$. Even without differentiability at 0, we can anyway divide u by some value between its right derivative $u'_+(0)$ and its left derivative $u'_-(0)$ (which exist in consequence of concavity) to get the same result.

These examples and considerations suggest working, in the random variable context here, at least, with the following set of conditions on a utility quantifier $\mathcal{U} : \mathcal{L}^2 \to [-\infty, \infty)$:

$$\begin{aligned}
&(U1) \quad \mathcal{U} \quad \text{is upper semicontinuous, possibly taking on} \quad -\infty,\\
&(U2) \quad \mathcal{U}((1-\lambda)Y + \lambda Y') \geq (1-\lambda)\mathcal{U}(Y) + \lambda\mathcal{U}(Y') \quad \text{for} \quad \lambda \in (0,1),\\
&(U3) \quad \mathcal{U}(Y) \geq \mathcal{U}(Y') \quad \text{when} \quad Y \geq Y',\\
&(U4) \quad \mathcal{U}(0) = 0,
\end{aligned} \tag{17}$$

with the sometime addition of

$$(U4') \quad \mathcal{U}(Y) < E[Y] \quad \text{for all} \quad Y \neq 0. \tag{18}$$

This perspective on utility differs, of course, from the common one in finance that mainly focuses on technical features of utility functions, like HARA, which have more to do with mathematical simplications than expressing true preferences, cf. [7]. It aims rather at a relative form of utility preferences which, for instance, builds on (16) for a benchmark return B through dividing also by $u'(B(\omega))$ to achieve (U4').

Our next step is dualizing the conditions (U1)–(U4) and (U4'). This could be done while keeping to the concavity in (U2), but it will really be better to revert to convexity, which entails reverting to the loss-oriented context of random variables X in Section 2. To this end we set up a one-to-one correspondence between functionals $\mathcal{U}$ and $\mathcal{V}$ through the relations

$$\mathcal{V}(X) = \mathcal{U}_*(X) := -\mathcal{U}(-X), \qquad \mathcal{U}(Y) = \mathcal{V}_*(Y) = -\mathcal{V}(-Y), \tag{19}$$

calling $\mathcal{V}$ the *regret quantifier* associated with $\mathcal{U}$ as utility quantifier. The conditions on $\mathcal{U}$ in (17) and (18) translate for its counterpart $\mathcal{V} : \mathcal{L}^2 \to (-\infty, \infty]$ into:

$$\begin{aligned}
&(V1) \quad \mathcal{V} \quad \text{is lower semicontinuous, possibly taking on} \quad \infty,\\
&(V2) \quad \mathcal{V}((1-\lambda)X + \lambda X') \leq (1-\lambda)\mathcal{V}(X) + \lambda\mathcal{V}(X') \quad \text{for} \quad \lambda \in (0,1),\\
&(V3) \quad \mathcal{V}(X) \geq \mathcal{V}(X') \quad \text{when} \quad X \geq X',\\
&(V4) \quad \mathcal{V}(0) = 0,
\end{aligned} \tag{20}$$

with the sometime addition of

$$(V4') \quad \mathcal{V}(X) > E[X] \quad \text{for all} \quad X \neq 0. \tag{21}$$

Once more we investigate what happens with these conditions under the conjugacy relations

$$\mathcal{V}^*(Q) = \sup_X \big\{ \langle X, Q\rangle - \mathcal{V}(X)\big\}, \qquad \mathcal{V}(X) = \sup_Q \big\{ \langle X, Q\rangle - \mathcal{V}^*(Q)\big\}, \quad (22)$$

A distinction will be that instead of dual elements Q only in the simplex $\mathcal{P}_0$, we will have them in the nonnegative orthant

$$\mathcal{L}^2_+ = \big\{ Q \in \mathcal{L}^2 \,\big|\, Q \geq 0\big\}.$$

Theorem 2 *(regret dualization) The class of functionals $\mathcal{K}$ on $\mathcal{L}^2$ that come up as conjugates $\mathcal{V}^*$ of regret quantifiers $\mathcal{V}$ satisfying (V1)–(V4) is characterized by*

$$\begin{aligned}
&(K1)\quad \mathcal{K} \quad \textit{is lower semicontinuous,}\\
&(K2)\quad \mathcal{K}((1-\lambda)X + \lambda X') \leq (1-\lambda)\mathcal{K}(X) + \lambda\mathcal{K}(X') \quad \textit{for} \quad \lambda \in (0, 1),\\
&(K3)\quad \mathcal{L}^2_+ \supset \mathcal{M} = \operatorname{dom}\mathcal{K} = \big\{ Q \,\big|\, \mathcal{K}(Q) < \infty\big\},\\
&(K4)\quad \inf_{Q\in\mathcal{M}} \mathcal{K}(Q) = 0.
\end{aligned}$$

The extra condition (V4′) corresponds in this to

$$(K4') \quad \mathcal{K}(1) = 0 \quad \textit{(entailing} \quad 1 \in \mathcal{M} \quad \textit{), but} \quad \partial\mathcal{K}(1) \quad \textit{contains no nonzero} \quad X.$$

The subclass consisting of the conjugates $\mathcal{V}^*$ of functionals $\mathcal{V}$ which are sublinear instead of just convex as (V2) is identified with $\mathcal{K}$ being $\equiv 0$ on $\mathcal{M}$. This subclass thus consists of the functionals $\mathcal{K}$ of the form

$$\mathcal{K} = \delta_{\mathcal{M}} \quad \textit{for some nonempty closed convex set} \quad \mathcal{M} \subset \mathcal{L}^2_+.$$

Proof Conditions (K1) and (K2) merely echo (V1) and (V2) under the conjugacy between $\mathcal{V}$ and $\mathcal{K}$. To confirm that (V3) implies (K3), we note that if $Q \in \mathcal{M}$ then $\sup_X \big\{ E[XQ] - \mathcal{V}(X)\big\} < \infty$. Taking $c \in \boldsymbol{R}$ to be the supremum, we in particular have for any $Z \in \mathcal{L}^2_+$ that $\mathcal{V}(-Z) \geq -E[ZQ] - c$, but also from (V3) that $\mathcal{V}(-Z) \leq \mathcal{V}(0) = 0$. Therefore $E[ZQ] \geq -c$ for all $Z \in \mathcal{L}^2_+$, which requires $Q \in \mathcal{L}^2_+$. In the other direction, if $\mathcal{K}$ satisfies (K3), then from writing the second formula in (22) as

$$\mathcal{V}(X) = \sup_{Q\in\mathcal{M}} \big\{ E[XQ] - \mathcal{K}(Q)\big\} \quad (23)$$

we see that $\mathcal{V}(X + Z) \geq \mathcal{V}(X)$ when $Z \in \mathcal{L}^2_+$, which is (V3). From (23) it is obvious as well that (K4) corresponds to (V4). To understand (K4′), start by observing that $\mathcal{K}(1) = 0$ means under conjugacy that $\sup_X \big\{ E[X\,1] - \mathcal{V}(X)\big\} = 0$, which is the same as $\mathcal{V}(X) \geq E[X]$ for all X, with equality holding for $X = 0$ by (V3). The subgradient condition means that $\mathcal{V}(X) = E[X\,1] - \mathcal{K}(1)$ only when $X = 0$, which in view of $\mathcal{K}(1)$ being 0 ensures that $\mathcal{V}(X) > E[X]$ for all $X \neq 0$. □

A particular category to look at more closely in these relationships is expected utility as in (14) and its regret counterpart. What conditions on the utility function u on $(-\infty, \infty)$ make $\mathcal{U}(Y) = E[u(Y)]$ satisfy (U1)–(U4)? It is elementary that this is true when

$$
\begin{aligned}
&(u1) \quad u \quad \text{is upper semicontinuous, possibly taking on} \quad -\infty,\\
&(u2) \quad u((1-\lambda)y + \lambda y') \geq (1-\lambda)u(y) + \lambda u(y') \quad \text{for} \quad \lambda \in (0,1),\\
&(u3) \quad u(y) \geq u(y') \quad \text{when} \quad y \geq y',\\
&(u4) \quad u(0) = 0,
\end{aligned}
$$

and that (U4′) will be satisfied as well when

$$(u4') \quad u(y) < y \quad \text{for all} \quad y \neq 0.$$

Moreover these conditions are not just sufficient but necessary, as seen by considering constants $Y \equiv y$.

On the side of regret in place of utility, it all works the same way. Expected regret takes the form

$$\mathcal{V}(X) = E[v(X)] = \int_{\Omega} v(X(\omega))dP_0(\omega) \quad \text{for a function} \quad v : (-\infty, \infty) \to (-\infty, \infty].$$

The conditions on v that are both necessary and sufficient for $\mathcal{V}$ to satisfy (V1)–(V4) are

$$
\begin{aligned}
&(v1) \quad v \quad \text{is lower semicontinuous, possibly taking on} \quad \infty,\\
&(v2) \quad v((1-\lambda)x + \lambda x') \leq (1-\lambda)v(x) + \lambda v(x') \quad \text{for} \quad \lambda \in (0,1),\\
&(v3) \quad v(x) \leq v(x') \quad \text{when} \quad x \geq x',\\
&(v4) \quad v(0) = 0,
\end{aligned}
$$

and (V4′) is associated with

$$(v4') \quad v(x) > x \quad \text{for all} \quad x \neq 0.$$

On the other hand, the sublinearity of $\mathcal{V}$ corresponds to the sublinearity of v, which means in the one-dimensional setting and in coordination with (v1)–(v4) that

$$
\begin{aligned}
&\text{on} \quad (-\infty, 0), \quad v \quad \text{is linear with a slope} \quad a \geq 0, \quad \text{while}\\
&\text{on} \quad (0, \infty), \quad v \quad \text{is linear with a slope} \quad b \geq a \quad \text{or} \quad v \equiv \infty.
\end{aligned}
\tag{24}
$$

The pairing of $\mathcal{U}$ and $\mathcal{V}$ in (19) corresponds to the pairing of u and v by

$$v(x) = u_*(x) = -u(-x), \qquad u(y) = v_*(y) = -v(-y), \tag{25}$$

which is easy to picture because it just means that the graphs of u and v are reflections of each other across a 45-degree line between the axes of $\boldsymbol{R}^2$.

Duality between $\mathcal{V}$ and $\mathcal{K} = \mathcal{V}^*$ as in Theorem 2 is easy in this context via [12]:

$$\begin{array}{c} \mathcal{V}(X) = E[v(X)] \iff \mathcal{K}(Q) = E[k(Q)], \\ \text{where} \quad k(q) = v^*(q) = \sup_x \big\{ xq - v(x) \big\}. \end{array} \tag{26}$$

The conditions on k that correspond to (K1)–(K4) on $\mathcal{K}$ are

$$\begin{array}{ll} (k1) & k \quad \text{is lower semicontinuous,} \\ (k2) & k((1-\lambda)X + \lambda X') \leq (1-\lambda)k(X) + \lambda k(X') \quad \text{for} \quad \lambda \in (0,1), \\ (k3) & k(q) < \infty \implies q \geq 0, \\ (k4) & \inf_q k(q) = 0. \end{array}$$

The extra condition (v4′) corresponds in this to

$$(k4') \quad k \quad \text{is differentiable at} \quad 1 \quad \text{with} \quad k'(1) = 0,$$

which in combination with (k4) means that $k(1) = 0$ and $\partial k(1) = \{0\}$. The case of sublinear $\mathcal{V}$, in which v has the form in (24), corresponds to k being the indicator of the interval $[a, b] \subset [0, \infty)$, or as the case may be, $[a, \infty)$.

4 Risk Versus Utility

One of our chief goals in this article is to clarify how risk and utility might be related to each other in our context, and this passage from utility quantifiers $\mathcal{U}$ via (19) to regret quantifiers $\mathcal{V}$ and their dualization sheds a lot of light on that. The regret conditions in (20) and (21) have turned out to be almost identical to the risk conditions in (5) and (6)! The only difference is seen in having constant/nonconstant in (R4) and (R4′) versus zero/nonzero in (V4) and (V4′). In the dualizations that is paralleled by having $Q \in \mathcal{P}_0$ in the risk case but only $Q \in \mathcal{L}^2_+$ in the regret case. Those seemingly small distinctions are nevertheless significant and have interesting consequences which we will explore next.

Theorem 3 *(risk from utility or regret) Let $\mathcal{V}$ be a regret quantifier satisfying (V1)–(V4) + (V4′), thus being associated under* (19) *with a utility quantifier $\mathcal{U}$ satisfying (U1)–(U4) + (U4′). Let*

$$\mathcal{R}(X) = \inf_{C \in \boldsymbol{R}} \big\{ C + \mathcal{V}(X - C) \big\}, \qquad \mathcal{S}(X) = \operatorname*{argmin}_{C \in \boldsymbol{R}} \big\{ C + \mathcal{V}(X - C) \big\}. \tag{27}$$

Then $\mathcal{S} \neq \emptyset$ and $\mathcal{R}$ is a risk quantifier satisfying (R1)–(R4) + (R4′). In this relationship sublinearity for $\mathcal{V}$ produces sublinearity for $\mathcal{R}$. The dualizations $\mathcal{J} = \mathcal{R}^*$ and

$\mathcal{K} = \mathcal{V}^*$ *are moreover tied together by*

$$\mathcal{J}(Q) = \begin{cases} \mathcal{K}(Q) & \textit{if} \quad E[Q] = 1, \\ \infty & \textit{if} \quad E[Q] \neq 1. \end{cases} \tag{28}$$

Proof This was established in [20] under an additional assumption on $\mathcal{V}$, but that assumption was shown to be unnecessary in [16]. □

The rule in (27) provides a vast generalization of the formula (3) for the superquantile $\overline{q}_\alpha(X)$, in which

$$\overline{q}_\alpha(X) = \min_{C \in I\!R} \big\{ C + \mathcal{V}_\alpha(X - C) \big\} \quad \text{for} \quad \mathcal{V}_\alpha(X) = \frac{1}{1-\alpha} E\big[\max\{0, X\} \big]. \tag{29}$$

The random variable $\max\{0, X\}$ gives the *absolute* loss in X, unbalanced by the desirable outcomes, if any, in which $X(\omega) < 0$. As mentioned earlier, the argmin set in (29), if not consisting of the quantile $q_\alpha(X)$ alone, is a closed interval with that quantile as its left endpoint. That exemplifies the nonemptiness of the set $\mathcal{S}$ in (27). Note that (29) falls into the case of sublinearity, which on the dual side has $\mathcal{J}$ and $\mathcal{K}$ indicators of the sets

$$\mathcal{Q}_\alpha = \big\{ Q \in \mathcal{P}_0 \,\big|\, Q \leq (1-\alpha)^{-1} \big\}, \qquad \mathcal{M}_\alpha = \big\{ Q \in \mathcal{L}^2_+ \,\big|\, Q \leq (1-\alpha)^{-1} \big\},$$

which illustrate the rule in (28). An example beyond the sublinear case is furnished by the log-exponential risk quantifier:

$$\log E[\exp X] = \min_{C \in I\!R} \big\{ C + \mathcal{V}(X - C) \big\} \quad \text{for} \quad \mathcal{V}(X) = E[\exp X - 1],$$

where the regret quantifier $\mathcal{V}$ dualizes to

$$\mathcal{K}(Q) = \begin{cases} E[Q \log Q - Q] & \text{when} \quad Q \geq 0, \\ \infty \quad \text{otherwise.} \end{cases}$$

A remarkable feature of this choice of $\mathcal{V}$ is that it not only produces $\log E[\exp X]$ as the value of $\mathcal{R}(X)$ but also $C = \log E[\exp X]$ as the unique minimizing C in $\mathcal{S}(X)$. This is reminiscent of facts about the exponential function, like it being its own derivative.

The general formula in (27) can be given an appealing interpretation in terms of the risk of incurring a loss. Through (19) the regret $\mathcal{V}(X)$ is a kind of anti-utility which stands for the overall displeasure in being saddled with the potential losses in X (with negative losses acting as gains). These losses occur in the future, but it is possible to account for them to some extent in the present by writing off a selected amount C of loss as being certain. Then in place of $\mathcal{V}(X)$ we have, after the write-off, only the regret in the reduced random loss variable $X - C$. We can optimize by determining a value of C that makes the combination $C + \mathcal{V}(X - C)$ as low as possible, i.e., $C \in \mathcal{S}$. There does exist such C because $\mathcal{S} \neq \emptyset$. The minimizing C can be thought of as the amount of compensation that ought to be demanded for shouldering the obligations represented by X.

Everything in terms of a regret quantifier $\mathcal{V}$ can be restated in terms of a utility quantifier through (19). For instance, (27) can given the form

$$\mathcal{R}_*(Y) = \sup_{D \in \mathbb{R}} \left\{ D + U(Y - D) \right\}, \quad \text{where} \quad \mathcal{R}_*(Y) = -\mathcal{R}(-Y) \tag{30}$$

in which the C in (27) is replaced by $D = -C$ when X is replaced by $Y = -X$. The convexity of $\mathcal{R}$ turns into the concavity of $\mathcal{R}_*$ and leads in dualization to recasting the formulas (9), (10), and (11) in terms of minimization. Mathematically this is a trivial exercise, but the different view in (30) is actually the original one in this subject. It is how the thinking went with Ben-Tal and Teboulle [5], who called a D giving the maximum in (30) the *optimized certainty equivalent* for Y. Their pioneering efforts only targeted expected utility as in (14), not necessarily "normalized," but that was enough to cover the crucial connection with some of the coherent risk quantifiers of major importance. The widening of the picture to other versions of $\mathcal{U}$, as facilitated in treatment by the introduction of "regret" as a reoriented partner to utility, was carried out in [20].

Without laying out all the parallel details in $\mathcal{U}$-$\mathcal{V}$ notation and formulation, we can take advantage of this prospect to understand more about stochastic ambiguity as seen by economists in utility terms. This brings us back to examing more closely utility quantifiers like the one in (15), which concern worst-case expected utility with respect to some collection of probability measures. In order not to disrupt the notational scheme for risk and that we have so far put in place, which already is based through duality on a form of stochastic ambiguity, we shift the formula in (15) to avoid using $P \in \mathcal{P}$ and instead incorporate uncertainty in terms of a collection $\bar{\mathcal{P}}$ of probability measures $\bar{P}$ and the associated set $\bar{\mathcal{Q}}$ of densities $\bar{Q} = d\bar{P}/dP_0$:

$$\mathcal{U}(Y) = \inf_{\bar{P} \in \bar{\mathcal{P}}} E_{\bar{P}}[u(Y)] = \inf_{\bar{Q} \in \bar{\mathcal{Q}}} E[u(Y)\bar{Q}]. \tag{31}$$

With no loss of generality (passing to the closed convex hull if necessary), we can suppose in this that $\bar{\mathcal{Q}}$ is a nonempty closed convex subset of the probability simplex $\mathcal{P}_0$. We then have

$$\mathcal{U}(Y) = -\bar{\mathcal{R}}(-u(Y)) = \bar{\mathcal{R}}_*(u(Y)) \quad \text{for the risk quantifier} \quad \bar{\mathcal{R}}(X) = \sup_{\bar{Q} \in \bar{\mathcal{Q}}} E[X\bar{Q}]. \tag{32}$$

For complete rigor here we need to be sure that the expectations in (31) are well defined, of course, and that is true because u is concave.[12] Fancier versions of stochas-

[12] The concavity of u implies the existence of an affine function a that majorizes u, and then $u(Y)\bar{Q} \leq a(Y)\bar{Q}$. For $Y \in \mathcal{L}^2$ we have $a(Y) \in \mathcal{L}^2$. Since $\bar{Q} \in \mathcal{L}^2$ as well, we know that $a(Y)\bar{Q}$ is integrable, giving $\langle a(Y), \bar{Q} \rangle$. Having $u(Y)\bar{Q}$ bounded above by the integrable function $a(Y)\bar{Q}$ ensures that the expectation in (31) is well defined as either a real value or $-\infty$. The latter case

tic ambiguity in utility have been studied in which (31) is replaced by

$$\mathcal{U}(Y) = \inf_{\bar{P}\in\bar{\mathcal{P}}} \left\{ E_{\bar{P}}[u(Y)] + c(\bar{P}) \right\} \tag{33}$$

for some function c, or in our density-type formulation,

$$\mathcal{U}(Y) = \inf_{\bar{Q}\in\bar{\mathcal{Q}}} \left\{ E[u(Y)\bar{Q}] + \bar{\mathcal{J}}(\bar{Q}) \right\} \tag{34}$$

for some function $\bar{\mathcal{J}}$. This type of utility quantifier was introduced by Maccheroni, Marinacci, and Rustichini [11] for doing a better job at capturing the preferences of financial decision makers; in earlier work of Gilboa and Schmeidler [9] the extra term was not present. The main thing for us here is that a formula of type (34) fits perfectly with the ideas of risk and its dualization in Theorem 1 yielding the representations (9)–(10): we can rewrite (34) in those terms as

$$\mathcal{U}(Y) = \bar{\mathcal{R}}_*(u(Y)) = -\bar{\mathcal{R}}(-u(Y)) \quad \text{for} \quad \bar{\mathcal{R}}(\bar{X}) = \sup_{\bar{Q}\in\bar{\mathcal{Q}}} \left\{ E[\bar{X}\bar{Q}] - \bar{\mathcal{J}}(\bar{Q}) \right\} \tag{35}$$

under the assumption that $\bar{\mathcal{J}}$ satisfies (J1)–(J4) with $\operatorname{dom}\bar{\mathcal{J}} = \bar{\mathcal{Q}}$.

Strzalecki [24] has taken special interest in the case of (33) which, expressed as (34), is

$$\mathcal{U}(Y) = \inf_{\bar{Q}\in\mathcal{P}_0} \left\{ E[u(Y)\bar{Q}] + E[\bar{Q}\log\bar{Q}] \right\}.$$

The extra term then expresses, for the probability measure $\bar{P}$ having $d\bar{P}/dP_0 = \bar{Q}$, the Kullback-Leibler distance of $\bar{P}$ from P_0, or the relative entropy of $\bar{P}$ with respect to P_0. According to (12) this means that

$$\mathcal{U}(Y) = \bar{\mathcal{R}}_*(u(Y)) = -\bar{\mathcal{R}}(-u(Y)) \quad \text{for the risk quantifier} \quad \bar{\mathcal{R}}(X) = \log E[\exp X]. \tag{36}$$

Two general questions emerge. Do utility quantifiers $\mathcal{U}$ defined in the manner of (35), with (32) and (36) as special cases, satisfy the conditions (U1)–(U4), and possibly (U4′), that we came up with earlier? And when they do, what can be said about the risk quantifier $\mathcal{R}$ that they produce through the formula in Theorem 3 and its dualization to $\mathcal{J}$?

In answering these questions we will proceed for convenience in the equivalent mode of regret instead of utility. This will take advantage of the pairing of the utility funtion u with a regret function v in (25) and the corresponding pairing of a utility quantifier $\mathcal{U}$ as in (34) with a regret quantifier

$$\mathcal{V}(X) = \bar{\mathcal{R}}(v(X)) \quad \text{for} \quad \bar{\mathcal{R}}(\bar{X}) = \sup_{\bar{Q}\in\bar{\mathcal{Q}}} \left\{ E[\bar{X}\bar{Q}] - \bar{\mathcal{J}}(\bar{Q}) \right\}. \tag{37}$$

occurs if and only if the set $\left\{ \omega \,\middle|\, u(Y(\omega))\bar{Q}(\omega) = -\infty \right\}$ has positive measure with respect to P_0 (under the usual interpretation that the product of $-\infty$ and 0 is 0).

Theorem 4 *(basic properties of ambiguous expected utility/regret) A regret quantifier $\mathcal{V}$ of the form in* (37) *satisfies (V1)–(V4) as long as the risk quantifier $\bar{\mathcal{R}}$ satisfies (R1)–(R4) (or its dual counterpart $\mathcal{J}$ satisfies (J1)–(J4)) and the underlying regret function v satisfies (v1)–(v4). It further satisfies (V4′) when $\bar{\mathcal{R}}$ satisfies (R4′) (or $\mathcal{J}$ satisfies (J4′)) and v satisfies (v4′), and then there is associated with $\mathcal{V}$ a risk quantifier $\mathcal{R}$ given by* (27) *as*

$$\mathcal{R}(X) = \min_C \left\{ C + \bar{\mathcal{R}}(v(X - C)) \right\},$$

which in turn satisfies (R1)–(R4) + (R4′).

In addition, $\mathcal{V}$ is sublinear when $\bar{\mathcal{R}}$ and v are sublinear, and then $\mathcal{R}$ is sublinear as well.

Proof These claims are elementary to verify except for the one about (V4′). If v satisfies (v4′), we have $v(X) \geq X$ for any $X \in \mathcal{L}^2$. Then by (R3) we have $\bar{\mathcal{R}}(v(X)) \geq \bar{\mathcal{R}}(X)$. On the other hand, by (R4′) we have $\bar{\mathcal{R}}(X) \geq E[X]$, with equality holding only when X is constant. Thus, $\bar{\mathcal{R}}(v(X)) = E[X]$ cannot hold unless X is a constant C, in which case we are looking at $\bar{\mathcal{R}}(v(C)) = C$. Since $\bar{\mathcal{R}}(v(C)) = v(C)$ by (R4), but $v(C) > C$ for $C \neq 0$ by (v4′), this is only possible when $C = 0$. Thus, $\bar{\mathcal{R}}(v(X))$, which is $\mathcal{V}(X)$, cannot equal $E[X]$ unless 0, and we have (V4′). The properties about $\mathcal{R}$ come then from applying Theorem 3 in this setting. □

Working our way now toward dualization of the quantifiers $\mathcal{V}$ and $\mathcal{R}$ in Theorem 4, we must begin with close scrutiny of the special case of (37), where only a simple change of P_0 to a different probability measure $\bar{P}$ is involved, so that $\mathcal{V}$ just has the form:

$$\mathcal{V}_{\bar{Q}}(X) = E[v(X)\bar{Q}] = E_{\bar{P}}[v(X)] \quad \text{for fixed} \quad \bar{P} \quad \text{and} \quad \bar{Q} = d\bar{P}/dP_0. \tag{38}$$

Under the assumption that v satisfies (v1)–(v4), this is covered by Theorem 4 and therefore enjoys the dualization in Theorem 2. This would reduce to (25) for the function $k = v^*$ satisfying (26) if $\bar{P} = P_0$, corresponding to $\bar{Q} \equiv 1$, but here we want to allow $\bar{P}$ to be different from P_0.

Lemma 1 *For v satisfying (v1)–(v4) and its conjugate k satisfying (k1)–(k4), the convex functional $\mathcal{V}^*_{\bar{Q}}$ conjugate to $\mathcal{V}_{\bar{Q}}$ in* (38) *is*

$$\mathcal{V}^*_{\bar{Q}}(Q) = E[\bar{k}(\bar{Q}, Q)], \quad \textit{where} \quad \bar{k}(\bar{q}, q) = \begin{cases} \bar{q}k(\bar{q}^{-1}q) & \textit{if } \bar{q} > 0, \\ 0 & \textit{if } \bar{q} = 0,\ q = 0, \\ \infty & \textit{otherwise}, \end{cases} \tag{39}$$

and therefore

$$\sup_Q \left\{ E[XQ] - E[\bar{k}(\bar{Q}, Q)] \right\} = \mathcal{V}_{\bar{Q}}(X). \tag{40}$$

Here, $\bar{k}$ is a sublinear function on $\mathbb{R} \times \mathbb{R}$ for which the lower semicontinuous hull is

$$\mathrm{cl}\ \bar{k}(\bar{q}, q) = \begin{cases} \bar{q}k(\bar{q}^{-1}q) & \textit{if} \ \ \bar{q} > 0, \\ k^{\infty}(q) & \textit{if} \ \ \bar{q} = 0, \\ \infty \ \ \textit{if} \ \ \bar{q} < 0, \end{cases}$$

where

$$k^{\infty}(q) = \lim_{t \to \infty} \frac{k(tq)}{t} = \sup \left\{ xq \,\middle|\, v(x) < \infty \right\}.$$

Moreover,

$$\sup_{Q} \left\{ E[XQ] - E[\mathrm{cl}\ \bar{k}(\bar{Q}, Q)] \right\} = \mathcal{V}_{\bar{Q}}(X) + \delta(X \,|\, X \leq b), \tag{41}$$

where $\delta(X \,|\, X \leq b)$ is the indicator of X being $\leq b = \sup \left\{ x \,\middle|\, v(x) < \infty \right\}$ with probability 1.

Proof We are dealing in (38) with an integral functional on $\mathcal{L}^2$ having the form

$$X \mapsto \int_{\Omega} f(X(\omega), \omega) dP_0(\omega) \quad \text{for} \quad f(x, \omega) = v(x)\bar{Q}(\omega)$$

and can utilize the rule in [12], according to which the conjugate convex functional on $\mathcal{L}^2$ is

$$Q \mapsto \int_{\Omega} f^*(Q(\omega), \omega) dP_0(\omega), \quad \text{where} \quad f^*(\cdot, \omega) \quad \text{is conjugate to} \quad f(\cdot, \omega).$$

The calculus of conjugates in convex analysis gives us

$$f^*(q, \omega) = \sup_x \left\{ xq - f(x, \omega) \right\} = \sup_x \left\{ xq - \bar{Q}(\omega)v(x) \right\} = [\bar{Q}(\omega)v]^*(q).$$

A further rule about multiplication in [13] says that

$$[\bar{q}v]^*(q) = \begin{cases} \bar{q}v^*(\bar{q}^{-1}q) & \text{if} \ \ \bar{q} > 0, \\ 0 \ \ \text{if} \ \ \bar{q} = 0, \ q = 0, \\ \infty \ \ \text{otherwise.} \end{cases}$$

Since $v^* = k$, this confirms the claim in (39).

The formula then for $\mathrm{cl}\ \bar{k}$ is well known from convex analysis [13, Corollary 8.5.1]. The conjugate of $\mathrm{cl}\ \bar{k}(0, \cdot)$ is then the conjugate of k^{∞}, which is the indicator of the closure of $\mathrm{dom}\, v$, namely $\delta_{(-\infty, b]}$. (Recall that because v is nondecreasing and $v(0) = 0$, the interval comprising $\mathrm{dom}\, v$ is not bounded from below.) In repeating the earlier integral functional argument to get the conjugate of $Q \mapsto \int_\Omega g(Q(\omega),\omega)dP_0(\omega)$ for $g(q,\omega) = \ \text{cl}\ \bar{k}(\bar{Q}(\omega),q)$ we get the integral functional $X \to \int_\Omega g^*(X(\omega),\omega)dP_0(\omega)$ with

$$g^*(x,\omega) = \begin{cases} v(x)\bar{Q}(\omega) & \text{if} \quad \bar{Q}(\omega) > 0, \\ \delta_{(\infty,b]}(x) & \text{if} \quad \bar{Q}(\omega) = 0. \end{cases}$$

However, this is the same as $g^*(x,\omega) = v(x)\bar{Q}(\omega) + \delta_{(-\infty,b]}(x)$. That yields (41). □

With this at our disposal we are ready for the broader dualizations.

Theorem 5 *(dualizations from ambiguous regret/utility) Under the assumption that $\bar{R}$ satisfies (R1)–(R4) + (R4′) while v satisfies (v1)–(v4) + (v4′), the regret quantifier $\mathcal{V}$ in (37) has the dual representation*

$$\mathcal{V}(X) = \sup_{Q \in \mathcal{L}^2_+} \Big\{ E[XQ] - \mathcal{K}_0(Q) \Big\} \quad \textit{for} \quad \mathcal{K}_0(Q) = \inf_{\bar{Q} \in \bar{\mathcal{Q}}} \Big\{ E[\text{cl } \bar{\text{k}}(\bar{\text{Q}}, \text{Q})] + \bar{\mathcal{J}}(\bar{Q}) \Big\} \tag{42}$$

with $\bar{k}$ coming from (39), where $\mathcal{K}_0$ is a convex functional on $\mathcal{L}^2$, and consequently the conjugate functional $\mathcal{K} = \mathcal{V}^*$ is the lower semicontinuous hull cl $\mathcal{K}_0$ of $\mathcal{K}_0$. The risk quantifier $\mathcal{R}$ associated with $\mathcal{V}$ by Theorem 3 then has the dual representation

$$\mathcal{R}(X) = \sup_{Q \in \mathcal{P}_0} \Big\{ E[XQ] - \mathcal{K}(Q) \Big\},$$

which corresponds to the conjugate functional $\mathcal{J} = \mathcal{R}^*$ being $\mathcal{K}$ plus the indicator $\delta(Q \,|\, E[Q] = 1)$.

Proof From (37) and (38) we have

$$\mathcal{V}(X) = \sup_{\bar{Q} \in \bar{\mathcal{Q}}} \big\{ \mathcal{V}_{\bar{Q}}(Q) - \bar{\mathcal{J}}(\bar{\mathcal{Q}}) \big\} \tag{43}$$

where, by the Lemma, $\mathcal{V}_{\bar{Q}}(X)$ is given by (40). In fact, in this situation it makes no difference if we replace $\mathcal{V}_{\bar{Q}}(Q)$ here by $\mathcal{V}_{\bar{Q}}(X) + \delta(X \,|\, X \le b)$. The reason is that 1 is one of the elements in $\bar{\mathcal{Q}}$ in consequence of (R4′) for $\bar{\mathcal{R}}$ (through its counterpart (J4′) for $\bar{\mathcal{J}}$), and for that choice of $\bar{\mathcal{Q}}$ one has $\mathcal{V}_{\bar{Q}}(Q) = E[v(X)] = \infty$ unless $v(X) \in$ dom v almost surely, implying $v(X) \le b$ almost surely (for b as introduced in the Lemma). Thus, nothing is changed in (43) if we express $\mathcal{V}_{\bar{Q}}(Q)$ by the supremum in (41) instead of the one in (38). Therefore, in writing (37) as

$$\mathcal{V}(Q) = \sup_{\bar{Q} \in \bar{\mathcal{Q}}} \big\{ \mathcal{V}_{\bar{Q}}(Q) - \bar{\mathcal{J}}(\bar{Q}) \big\}$$

we can replace $\mathcal{V}_{\bar{Q}}(Q)$ by the supremum on the left side of (41) and obtain

$$\begin{aligned}\mathcal{V}(Q) &= \sup_{\bar{Q}\in\bar{\mathcal{Q}}}\Big\{\sup_{Q}\big\{E[XQ]-E[\text{ cl }\bar{k}(\bar{Q},Q)]\big\}-\bar{\mathcal{J}}(\bar{Q})\Big\}\\ &= \sup_{\bar{Q}\in\bar{\mathcal{Q}}}\sup_{Q}\Big\{E[XQ]-E[\text{ cl }\bar{k}(\bar{Q},Q)]\big\}-\bar{\mathcal{J}}(\bar{Q})\Big\}\\ &= \sup_{Q}\Big\{E[XQ]-\inf_{\bar{Q}\in\bar{\mathcal{Q}}}\big\{E[\text{ cl }\bar{k}(\bar{Q},Q)]+\bar{\mathcal{J}}(\bar{Q})\big\}\Big\},\end{aligned}$$

which is (42) since $\bar{k}(\bar{q},q)=\infty$ when $q>0$. This says that $\mathcal{K}_0^*=\mathcal{V}$, and therefore $\mathcal{V}^*=\mathcal{K}_0^{**}=\text{ cl }K_0$. The assertion about $\mathcal{R}$ follows then from (28) in Theorem 3. □

Perhaps further analysis, invoking additional assumptions, could do away with the closure operation, letting $\mathcal{K}=\mathcal{K}_0$, but we leave that for future work.

An interesting example of the relationships in Theorem 5 can be obtained by taking v to be the function underlying the α-superquantile,

$$v(x)=(1-\alpha)^{-1}\max\{0,x\},\qquad k(q)=\delta_{[0,(1-\alpha)^{-1}]}(q),\quad\text{where}\quad \alpha\in(0,1).$$

Then $k^{\infty}=\delta_0$, so that there is no difference between cl $\bar{k}$ and $\bar{k}$, with

$$k(\bar{q},q)=\begin{cases}0 & \text{if }\ \bar{q}>0,\ \bar{q}\geq(1-\alpha)q,\\ 0 & \text{if }\ \bar{q}=0,\ q=0,\\ \infty & \text{otherwise.}\end{cases}$$

This yields in (42) that

$$\mathcal{K}_0(Q)=\inf\big\{\bar{\mathcal{J}}(\bar{Q})\,\big|\,\bar{Q}\geq(1-\alpha)Q\big\},$$

where the condition $\bar{Q}\in\bar{\mathcal{Q}}$ has been omitted as superfluous since it just corresponds to $\bar{\mathcal{J}}(\bar{Q})<\infty$.

References

1. Acerbi, C.: Spectral measures of risk: a coherent reprsentation of subjective risk aversion. J. Bank. Financ. **26**, 1505–1518 (2002)
2. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D.: Coherent measures of risk. Math. Financ. **9**, 203–227 (1999)
3. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton (2009)
4. Ben-Tal, A., Nemirovski, A.: Robust convex optimization. Math. Oper. Res. **23**, 769–805 (1998)
5. Ben-Tal, A., Teboulle, M.: An old-new concept of convex risk measures: the optimized certainty equivalent. Math. Financ. **17**, 449–476 (2007)
6. Dencheva, D., Ruszczỳnski, A.: Common mathematical foundations of expected utility and dual utility theories. SIAM J. Optim. **24**, 381–405 (2013)

7. Föllmer, H., Schied, A.: Stochastic Finance, 2nd edn. de Gruyter, New York (2004)
8. Gao, R., Kleywegt, A.J.: Distributionally robust stochastic optimization with Wasserstein distance (2016). arXiv:1604.02199
9. Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. J. Math. Econ. **18**, 141–153 (1989)
10. Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. Oper. Res. **58**, 902–917 (2010)
11. Maccheroni, F., Marinacci, M., Rustichini, A.: Ambiguity aversion, robustness, and the variational representation of preferences. Econometrica **74**, 1147–1498 (2006)
12. Rockafellar, R.T.: Integrals which are convex functionals. Pac. J. Math. **24**, 525–539 (1968)
13. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
14. Rockafellar, R.T., Royset, J.O.: On buffered failure probability in design and optimization of structures. J. Reliab. Eng. Syst. Saf. **99**, 499–510 (2010)
15. Rockafellar, R.T., Royset, J.O.: Random variables, monotone relations and convex analysis. Math. Program. B **148**, 297–331 (2014)
16. Rockafellar, R.T., Royset, J.O.: Measures of residual risk with connections to regression, risk tracking, surrogate models and ambiguity. SIAM J. Optim. **28**, 1179–1208 (2015)
17. Rockafellar, R.T., Royset, J.O.: Superquantile/CVaR risk measures: second-order theory. Ann. Oper. Res. **262**, 3–29 (2018)
18. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**, 21–42 (2000)
19. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Financ. **26**, 1443–1471 (2002)
20. Rockafellar, R.T., Uryasev, S.: The fundamental risk quadrangle in risk management, optimization and statistical estimation. Surv. Oper. Res. Manag. Sci. **18**, 33–53 (2013)
21. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Generalized deviations in risk analysis. Financ. Stoch. **10**, 51–74 (2006)
22. Rockafellar, R.T., Wets, R.J-B: Variational Analysis. No. 317 in the series Grundlehren der Mathematischen Wissenschaften. Springer, Berlin (1997)
23. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Oper. Res. **62**, 1358–1376 (2014)
24. Strzalecki, T.: Axiomatic foundations of multiplier preferences. Econometrica **79**, 47–73 (2011)

Characterizations of Robust and Stable Duality for Linearly Perturbed Uncertain Optimization Problems

Nguyen Dinh, Miguel A. Goberna, Marco A. López and Michel Volle

We introduce a robust optimization model consisting in a family of perturbation functions giving rise to certain pairs of dual optimization problems in which the dual variable depends on the uncertainty parameter. The interest of our approach is illustrated by some examples, including uncertain conic optimization and infinite optimization via discretization. The main results characterize desirable robust duality relations (as robust zero-duality gap) by formulas involving the epsilon-minima or the epsilon-subdifferentials of the objective function. The two extreme cases, namely, the usual perturbational duality (without uncertainty), and the duality for the supremum of functions (duality parameter vanishing) are analyzed in detail.

1 Introduction

Duality theory was one of Jonathan Borwein's favorite research topics. Indeed, 14 of his papers include the term "duality" in their titles. The present article, dedicated to Jon's vast contribution to the subject, will refer only to four works of his, all of these related to optimization problems posed in locally convex Hausdorff topological vector spaces.

N. Dinh
International University, Vietnam National University - HCMC, Linh Trung ward,
Ho Chi Minh city, Thu Duc district, Vietnam
e-mail: ndinh@hcmiu.edu.vn

M. A. Goberna (✉) · M. A. López
Department of Mathematics, University of Alicante, Alicante, Spain
e-mail: mgoberna@ua.es

M. A. López
CIAO, Federation University, Ballarat, Australia
e-mail: marco.antonio@ua.es

M. Volle
LMA, Avignon University, EA 2151 Avignon, France
e-mail: michel.volle@univ-avignon.fr

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_4

Duality theorems were provided in [3] for the minimum of arbitrary families of convex programs; the quasi-relative interior constraint qualification was introduced in [6] in order to obtain duality theorems for various optimization problems where the standard Slater condition fails; the same CQ was immediately used, in [5], to obtain duality theorems for convex optimization problems with constraints given by linear operators having finite-dimensional range together with a conical convex constraint; finally, quite recently, in [4], duality theorems for the minimization of the finite sum of convex functions were established, using conditions which involve the ε-subdifferential of the given functions.

In this paper, we consider a *family of perturbation functions*

$$F_u : X \times Y_u \to \mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}, \text{ with } u \in U,$$

and where X and Y_u, $u \in U$, are given locally convex Hausdorff topological vector spaces (briefly, lcHtvs), the index set U is called the *uncertainty set* of the family, X is its *decision space*, and each Y_u is a *parameter space*. Note that our model includes a parameter space Y_u, depending on $u \in U$, which is a novelty with respect to the "classical" robust duality scheme (see [21] and references therein, where a unique parameter space Y is considered), allowing us to cover a wider range of applications including uncertain optimization problems under linear perturbations of the objective function. The significance of our approach is illustrated along the paper by relevant cases extracted from deterministic optimization with linear perturbations, uncertain optimization without perturbations, uncertain conic optimization and infinite optimization. The antecedents of the paper are described in the paragraphs devoted to the first two cases in Section 2.

We associate with each family $\{F_u : u \in U\}$ of perturbation functions corresponding optimization problems whose definitions involve continuous linear functionals on the decision and the parameter spaces. We denote by 0_X, 0_X^*, 0_u, and 0_u^*, the null vectors of X, its topological dual X^*, Y_u, and its topological dual Y_u^*, respectively. The optimal value of a minimization (maximization, respectively) problem (P) is denoted by inf (P) (sup (P)); in particular, we write min (P) (max (P)) whenever the optimal value of (P) is attained. We adopt the usual convention that inf (P) $= +\infty$ (sup (P) $= -\infty$) when the problem (P) has no feasible solution. The associated optimization problems are the following:

- *Linearly perturbed uncertain problems*: for each $(u, x^*) \in U \times X^*$,

$$(P_u)_{x^*} : \qquad \inf_{x \in X} \left\{ F_u(x, 0_u) - \langle x^*, x \rangle \right\}.$$

- *Robust counterpart of* $\{(P_u)_{x^*}\}_{u \in U}$:

$$(\mathrm{RP})_{x^*} : \qquad \inf_{x \in X} \left\{ \sup_{u \in U} F_u(x, 0_u) - \langle x^*, x \rangle \right\}.$$

Denoting by $F_u^* : X^* \times Y_u^* \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, the *Fenchel conjugate* of F_u, namely,

$$F_u^*(x^*, y_u^*) := \sup_{(x,y_u)\in X\times Y_u} \Big\{\langle x^*, x\rangle + \langle y_u^*, y_u\rangle - F_u(x, y_u)\Big\}, \quad (x^*, y_u^*) \in X^* \times Y_u^*,$$

we now introduce the corresponding dual problems:

- *Perturbational dual of* $(P_u)_{x^*}$:

$$(\mathrm{D}_u)_{x^*}: \qquad \sup_{y_u^*\in Y_u^*} -F_u^*(x^*, y_u^*).$$

  Obviously,

$$\sup\,(D_u)_{x^*} \leq \inf\,(P_u)_{x^*} \leq \inf\,(\mathrm{RP})_{x^*}, \forall u \in U.$$

- *Optimistic dual of* $(\mathrm{RP})_{x^*}$:

$$(\mathrm{ODP})_{x^*} \sup_{(u,y_u^*)\in\Delta} -F_u^*(x^*, y_u^*),$$

  where $\Delta := \big\{\big(u, y_u^*\big) : u \in U,\ y^* \in Y_u^*\big\}$ is the disjoint union of the spaces Y_u^*. We have

$$\sup\,(\mathrm{ODP})_{x^*} = \sup_{u\in U}(D_u)_{x^*} \leq \inf\,(\mathrm{RP})_{x^*}.$$

We are interested in the following desirable robust duality properties:

- *Robust duality* is said to hold at x^* if $\inf\,(\mathrm{RP})_{x^*} = \sup\,(\mathrm{ODP})_{x^*}$,
- *Strong robust duality* at x^* means $\inf\,(\mathrm{RP})_{x^*} = \max\,(\mathrm{ODP})_{x^*}$,
- *Reverse strong robust duality* at x^* means $\min\,(\mathrm{RP})_{x^*} = \sup\,(\mathrm{ODP})_{x^*}$,
- *Min-max robust duality* at x^* means $\min\,(\mathrm{RP})_{x^*} = \max\,(\mathrm{ODP})_{x^*}$.

Each of the above desirable properties is said to be *stable* when it holds for any $x^* \in X^*$. The main results of this paper characterize these properties in terms of formulas involving the ε-minimizers and ε-subdifferentials of the objective function of the robust counterpart problem $(\mathrm{RP})_{0_X^*}$, namely, the function

$$p := \sup_{u\in U} F_u(\cdot, 0_u).$$

Theorem 1 characterizes robust duality at a given point $x^* \in X^*$ as a formula for the inverse mapping of the ε-subdifferential at x^* without any convexity assumption. The same is done in Theorem 2 to characterize strong robust duality. In the case, when a primal optimal solution does exist we give a formula for the exact minimizers of $p - x^*$ to characterize dual strong (resp. min-max) robust duality at x^*, see Theorem 3 (resp. Theorem 4). We show that stable robust duality gives rise to a formula for the ε-subdifferential of p (Theorem 5, see also Theorem 1). The same is done for stable strong robust duality (Theorem 6). A formula for the exact subdifferential of p is provided in relation with robust duality at appropriate points (Theorem 7). The most simple possible formula for the exact subdifferential of p (the so-called *Basic*

Robust Qualification condition) is studied in detail in Theorem 8. All the results from Sections 1–8 are specified for the two extreme cases (the case with no uncertainty and the one in absence of perturbations), namely, Cases 1 and 2 in Section 2 (for the sake of brevity, we do not give the specifications for Cases 3 and 4). It is worth noticing the generality of the mentioned results (as they do not require any assumption on the involved functions) and the absolute self-containment of their proofs. The use of convexity in the data will be addressed in a forthcoming paper.

2 Special Cases and Applications

In this section, we make explicit the meaning of the robust duality of the general model introduced in Section 1, composed by a family of perturbation functions together with its corresponding optimization problems. We are doing this by exploring the extreme case with no uncertainty, the extreme case in absence of perturbations, and two other significant situations. In all these cases, we propose ad hoc families of perturbation functions allowing to apply the duality results to given optimization problems, either turning back to variants of well-known formulas for conjugate functions or proposing new ones.

Let us recall the robust duality formula, $\inf\,(\mathrm{RP})_{x^*} = \sup\,(\mathrm{ODP})_{x^*}$, i.e.,

$$\inf_{x\in X}\sup_{u\in U}\left\{F_u\left(x,0_u\right)-\left\langle x^*,x\right\rangle\right\}=\sup_{\left(u,y_u^*\right)\in\Delta}-F_u^*\left(x^*,y_u^*\right). \tag{1}$$

We firstly study the two extreme cases: the case with no uncertainty and the one with no perturbations.

Case 1. The case with no uncertainty: Deterministic optimization with linear perturbations deals with parametric problems of the form:

$$(\mathrm{P})_{x^*}:\qquad \inf_{x\in X}\left\{f(x)-\left\langle x^*,x\right\rangle\right\},$$

where $f : X \to \mathbb{R}_\infty$ (i.e., $f \in (\mathbb{R}_\infty)^X$) is the *nominal objective function* and the *parameter* is $x^* \in X^*$. Taking a singleton uncertainty set $U = \{u_0\}$, $Y_{u_0} = Y$ and $F_{u_0} = F$ such that $F\left(x, 0_Y\right) = f(x)$ for all $x \in X$, (1) reads

$$\inf_{x\in X}\left\{F\left(x,0_Y\right)-\left\langle x^*,x\right\rangle\right\}=\sup_{y^*\in Y^*}-F^*\left(x^*,y^*\right), \tag{2}$$

which is *the fundamental perturbational duality formula* [7, 24, 28]. Stable and strong robust duality theorems are given in [9] (see also [11] and [20] for infinite optimization problems).

Case 2. The case with no perturbations: Uncertain optimization without perturbations deals with families of problems of the form

$$(\mathrm{P})_{x^*}: \qquad \left\{\inf_{x\in X} f_u(x) - \langle x^*, x\rangle\right\}_{u\in U},$$

where $f_u \in (\mathbb{R}_\infty)^X$, $u \in U$, and $x^* \in X^*$. The absence of perturbation is realized by taking F_u such that $F_u(x, y_u) = f_u(x)$ for all $u \in U$, $x \in X$ and $y_u \in Y_u$. Assuming dom $f_u \neq \emptyset$ we have

$$F_u^*\left(x^*, y_u^*\right) = \begin{cases} f_u^*\left(x^*\right), & \text{if } \ y_u^* = 0_u^*, \\ +\infty, & \text{if } \ y_u^* \neq 0_u^*. \end{cases} \tag{3}$$

Then (1) writes

$$\left(\sup_{u\in U} f_u\right)^*(x^*) = \inf_{u\in U} f_u^*(x^*), \tag{4}$$

which amounts, for $x^* = 0_X^*$, to the inf − sup *duality in robust optimization,* also called *robust infimum* (recall that any constrained optimization problem can be reduced to an unconstrained one by summing up the indicator function of the feasible set to the objective function):

$$\inf_{x\in X}\sup_{u\in U} f_u(x) = \sup_{u\in U}\inf_{x\in X} f_u(x).$$

Robust duality theorems without perturbations are given in [27] for a special class of uncertain non-convex optimization problems while [11] provides robust strong duality theorems for uncertain convex optimization problems which are expressed in terms of the closedness of suitable sets regarding the vertical axis of $X^* \times \mathbb{R}$.

Case 3. Conic optimization problem with uncertain constraints: Consider the uncertain problem

$$(\mathrm{P}): \qquad \left\{\inf_{x\in X} f(x) \ \text{ s.t. } H_u(x) \in -S_u\right\}_{u\in U},$$

where, for each $u \in U$, S_u is an ordering convex cone in Y_u, $H_u\colon X \to Y_u$, and $f \in (\mathbb{R}_\infty)^X$. Denote by $S_u^+ := \left\{y_u^* \in Y_u^* : \left\langle y_u^*, y_u\right\rangle \geq 0, \forall y_u \in S_u\right\}$ the dual cone of S_u.

Problems of this type arise, for instance, in the production planning of firms producing n commodities from uncertain amounts of resources by means of technologies which depend on the available resources (e.g., the technology differs when the energy is supplied by either fuel gas or a liquid fuel). The problem associated with each parameter $u \in U$ consists of maximizing the cash-flow $c\,(x_1, ..., x_n)$ of the total production, with x_i denoting the production level of the i-th commodity, $i = 1, .., n$. The decision vector $x = (x_1, ..., x_n)$ must satisfy a linear inequality system $A_u x \leq b_u$, where the matrix of technical coefficients A_u is $m_u \times n$ and $b_u \in \mathbb{R}^{m_u}$, for some $m_u \in \mathbb{N}$. Denoting by $\mathrm{i}_{\mathbb{R}_+^n}$ the indicator function of $\mathbb{R}_+^n$ (i.e., $\mathrm{i}_{\mathbb{R}_+^n}(x) = 0$, when $x \in \mathbb{R}_+^n$, and $\mathrm{i}_{\mathbb{R}_+^n}(x) = +\infty$, otherwise), the uncertain production planning problem can be formulated as

$$\text{(P)}: \qquad \left\{ \inf_{x \in \mathbb{R}^n} f(x) = -c(x) + \mathrm{i}_{\mathbb{R}^n_+}(x) \ \text{ s.t. } A_u x - b_u \in -\mathbb{R}^{m_u}_+ \right\}_{u \in U},$$

with the space $Y_u = \mathbb{R}^{m_u}$ depending on the uncertain parameter u.

For each $u \in U$, define the perturbation function

$$F_u(x, y_u) = \begin{cases} f(x), \text{ if } H_u(x) + y_u \in -S_u, \\ +\infty, \text{ else.} \end{cases}$$

On the one hand, $(\mathrm{RP})_{0_X^*}$ collapses to the robust counterpart of (P) in the sense of robust conic optimization with uncertain constraints:

$$\text{(RP)}: \qquad \inf_{x \in X} f(x) \ \text{ s.t. } \ H_u(x) \in -S_u, \ \forall u \in U.$$

On the other hand, it is easy to check that

$$F_u^*(x^*, y_u^*) = \begin{cases} \left(f + y_u^* \circ H_u\right)^*(x^*), \text{ if } y_u^* \in S_u^+. \\ +\infty, \qquad\qquad\qquad\quad \text{else,} \end{cases}$$

$(\mathrm{ODP})_{0_X^*}$ is nothing else than the optimistic dual in the sense of uncertain conic optimization:

$$\text{(ODP)}: \qquad \sup_{u \in U, y_u^* \in S_u^+} \inf_{x \in X} \left\{ f(x) + \left\langle y_u^*, H_u(x) \right\rangle \right\}$$

(a special case when $Y_u = Y$, $S_u = S$ for all $u \in U$ is studied in [12, p. 1097] and [21]). Thus,

- *Robust duality holds at* 0_X^* means that inf (RP) = sup (ODP),
- *Strong robust duality holds at* 0_X^* means that

$$\inf\{f(x) : H_u(x) \in -S_u, \forall u \in U\} = \max_{\substack{u \in U \\ y_u^* \in S_u^+}} \inf_{x \in X} \left\{ f(x) + \left\langle y_u^*, H_u(x) \right\rangle \right\}.$$

Conditions for having such an equality are provided in [12, Theorem 6.3], [13, Corollaries 5, 6], for the particular case $Y_u = Y$ for all $u \in U$.

Strong robust duality and uncertain Farkas lemma: We focus again on the case where $Y_u = Y$ and $S_u = S$ for all $u \in U$. For a given $r \in \mathbb{R}$, let us consider the following statements:

(i) $H_u(x) \in -S, \ \forall u \in U \quad \Longrightarrow \quad f(x) \geq r,$

(ii) $\exists u \in U, \exists y_u^* \in S^+$ such that $f(x) + \left\langle y_u^*, H_u(x) \right\rangle \geq r, \ \forall x \in X.$

Then, it is true that the strong robust duality holds at 0_X^* if and only if $[(i) \Longleftrightarrow (ii)]$ for each $r \in \mathbb{R}$, which can be seen as an uncertain Farkas lemma. For details see [12, Theorem 3.2] (also [13, Corollary 5 and Theorem 1]).

It is worth noticing that when return to problem (P), a given robust feasible solution $\overline{x}$ is a minimizer if and only if $f(\overline{x}) \leq f(x)$ for any robust feasible solution x. So, a robust (uncertain) Farkas lemma (with $r = f(\bar{x})$) will lead automatically to an optimality test for (P). Robust conic optimization problems are studied in [2] and [25]. **Case 4. Discretizing infinite optimization problems:** Let $f \in (\mathbb{R}_\infty)^X$ and $g_t \in \mathbb{R}^X$ for all $t \in T$ (a possibly infinite index set). Consider the set U of non-empty finite subsets of T, interpreted as admissible perturbations of T, and the parametric optimization problem

$$(\mathrm{P}): \quad \left\{\inf_{x\in X} f(x) \text{ s.t. } \quad g_t(x) \leq 0, \quad \forall t \in S\right\}_{S\in U}.$$

Consider the parameter space $Y_s := \mathbb{R}^S$ (depending on S) and the perturbation function $F_S : X \times \mathbb{R}^S \to \mathbb{R}_\infty$ such that, for any $x \in X$ and $\mu := (\mu_s)_{s\in S} \in \mathbb{R}^S$,

$$F_S(x, \mu) = \begin{cases} f(x), \text{ if } g_s(x) \leq -\mu_s, \ \forall s \in S, \\ +\infty, \ \text{else.} \end{cases}$$

We now interpret the problems associated with the family of function perturbations $\{F_S : S \in U\}$. One has $Y_s^* = \mathbb{R}^S$ and

$$F_S^*\left(x^*, \lambda\right) = \begin{cases} \left(f + \sum\limits_{s\in S} \lambda_s g_s\right)^*(x^*), \text{ if } \lambda \in \mathbb{R}_+^S, \\ +\infty, \qquad\qquad\qquad\qquad \text{else.} \end{cases}$$

The robust counterpart at 0_X^*,

$$(\mathrm{RP})_{0_X^*}: \quad \inf f(x) \ \text{ s.t. } \quad g_t(x) \leq 0 \ \text{ for all } t \in T,$$

is a general infinite optimization problem while the optimistic dual at 0_X^* is

$$(\mathrm{ODP})_{0_X^*}: \quad \sup_{S\in U, \lambda\in\mathbb{R}_+^S} \left\{\inf_{x\in X} \left(f(x) + \sum_{s\in S} \lambda_s g_s(x)\right)\right\},$$

or, equivalently, the Lagrange dual of $(\mathrm{RP})_{0_X^*}$, i.e.,

$$(\mathrm{ODP})_{0_{X^*}}: \quad \sup_{\lambda\in\mathbb{R}_+^{(T)}} \left\{\inf_{x\in X} \left(f(x) + \sum_{t\in T} \lambda_t g_t(x)\right)\right\},$$

where, for each $\lambda = (\lambda_t)_{t\in T} \in \mathbb{R}_+^{(T)}$ (the subspace of $\mathbb{R}^T$ formed by the functions λ whose support, $\mathrm{supp}\lambda := \{t \in T : \lambda_t \neq 0\}$, is finite),

$$\sum_{t\in T}\lambda_t g_t(x) := \begin{cases} \sum\limits_{t\in \operatorname{supp}\lambda} \lambda_t g_t(x), & \text{if } \lambda \neq 0, \\ 0, & \text{if } \lambda = 0. \end{cases}$$

Following [14, Section 8.3], we say that $(\mathrm{RP})_{0_X^*}$ is *discretizable* if there exists a sequence $(S_r)_{r\in\mathbb{N}} \subset U$ such that

$$\inf\, (\mathrm{RP})_{0_X^*} = \lim_r \inf \left\{ f(x) : g_t(x) \leq 0,\ \forall t \in S_r \right\}, \tag{5}$$

and it is *reducible* if there exists $S \in U$ such that

$$\inf\, (\mathrm{RP})_{0_X^*} = \inf \left\{ f(x) : g_t(x) \leq 0,\ \forall t \in S \right\}.$$

Obviously, $\inf\, (\mathrm{RP})_{0_X^*} = -\infty$ entails that $(\mathrm{RP})_{0_X^*}$ is reducible which, in turn, implies that $(\mathrm{RP})_{0_X^*}$ is discretizable.

Discretizable and reducible problems are important in practice. Indeed, on the one hand, discretization methods generate sequences $(S_r)_{r\in\mathbb{N}} \subset U$ satisfying (5) when $(\mathrm{RP})_{0_X^*}$ is discretizable; discretization methods for linear and nonlinear semi-infinite programs have been reviewed in [15, Subsection 2.3] and [23], while a hard infinite optimization problem has been recently solved via discretization in [22]. On the other hand, replacing the robust counterpart (a hard semi-infinite program when the uncertainty set is infinite) of a given uncertainty optimization problem, when it is reducible, by a finite subproblem allows many times to get the desired tractable reformulation (see e.g., [1] and [8]).

Example 1 (Discretizing linear infinite optimization problems) Consider the problems introduced in Case 4 above, with $f(\cdot) := \langle c^*, \cdot\rangle$ and $g_t(x) := \left\langle a_t^*, \cdot\right\rangle - b_t$, where $c^*, a_t^* \in X^*$ and $b_t \in \mathbb{R}$, for all $t \in T$. Then, $(\mathrm{RP})_{0_X^*}$ collapses to the linear infinite programming problem

$$(\mathrm{RP})_{0_X^*} : \quad \inf \quad \left\langle c^*, x\right\rangle \ \text{ s.t. } \ \left\langle a_t^*, x\right\rangle \leq b_t,\ \forall t \in T,$$

whose feasible set we denote by A. So, $\inf\, (\mathrm{RP})_{0_X^*} = \inf_{x\in X} \left\{ \langle c^*, x\rangle + i_A(x) \right\}.$ We assume that $A \neq \emptyset$.

Given $S \in U$ and $\mu, \lambda \in \mathbb{R}^S$,

$$F_S(x, \mu) = \begin{cases} \langle c^*, x\rangle, & \text{if } \left\langle a_s^*, x\right\rangle \leq b_s - \mu_s,\ \forall s \in S, \\ +\infty, & \text{else,} \end{cases} \tag{6}$$

and

$$F_S^*\left(x^*, \lambda\right) = \begin{cases} \sum\limits_{s\in S} \lambda_s b_s, & \text{if } \sum\limits_{s\in S} \lambda_s a_s^* = x^* - c^* \ \text{ and } \ \lambda_s \geq 0,\ \forall s \in S, \\ +\infty, & \text{else.} \end{cases} \tag{7}$$

Hence, $(\mathrm{ODP})_{0_X^*}$ collapses to the so-called *Haar dual problem* [16] of $(\mathrm{RP})_{0_X^*}$,

$$(\mathrm{ODP})_{0_X^*}: \quad \sup\left\{-\sum_{t\in\operatorname{supp}\lambda}\lambda_t b_t : -\sum_{t\in\operatorname{supp}\lambda}\lambda_t a_t^* = c^*, \quad \lambda\in\mathbb{R}_+^{(T)}\right\},$$

i.e.,

$$\sup(\mathrm{ODP})_{0_X^*} = -\inf_{S\in U,\lambda\in\mathbb{R}_+^S}\left\{\sum_{s\in S}\lambda_s b_s : \sum_{s\in S}\lambda_s a_s^* = -c^*\right\}. \tag{8}$$

From (8), if $\inf(\mathrm{RP})_{0_X^*} = \max(\mathrm{ODP})_{0_X^*} \in \mathbb{R}$, then there exist $S\in U$ and $\lambda\in\mathbb{R}_+^S$ such that

$$\sum_{s\in S}\lambda_s\left(a_s^*, b_s\right) = -\left(c^*, \inf(\mathrm{RP})_{0_X^*}\right). \tag{9}$$

Let $A_S := \left\{x\in X : \left\langle a_s^*, x\right\rangle \le b_s, \ \forall s\in S\right\}$. Given $x\in A_S$, from (9),

$$0 \ge \sum_{s\in S}\lambda_s\left(\left\langle a_s^*, x\right\rangle - b_s\right) = -\left\langle c^*, x\right\rangle + \inf(\mathrm{RP})_{0_X^*}.$$

Since

$$\inf(\mathrm{RP})_{0_X^*} \le \left\langle c^*, x\right\rangle, \forall x\in A_S,$$

$$\inf(\mathrm{RP})_{0_X^*} = \inf\left\{\left\langle c^*, x\right\rangle : \left\langle a_s^*, x\right\rangle \le b_s, \ \forall s\in S\right\}, \tag{10}$$

so that $(\mathrm{RP})_{0_X^*}$ is reducible. Conversely, if (10) holds with $\inf(\mathrm{RP})_{0_X^*}\in\mathbb{R}$ and $\operatorname{cone}\left\{\left(a_t^*, b_t\right) : t\in T\right\} + \mathbb{R}_+\left(0_X^*, 1\right)$ is weak*-closed, since $\inf(\mathrm{RP})_{0_X^*} \le \langle c^*, x\rangle$ is consequence of $\left\{\left\langle a_s^*, x\right\rangle \le b_s, \ \forall s\in S\right\}$, by the nonhomogeneous Farkas lemma in lcHtvs [10] and the closedness assumption, there exist $\lambda\in\mathbb{R}_+^S$ and $\mu\in\mathbb{R}_+$ such that

$$-\left(c^*, \inf(\mathrm{RP})_{0_X^*}\right) = \sum_{s\in S}\lambda_s\left(a_s^*, b_s\right) + \mu\left(0_X^*, 1\right),$$

which implies that $\mu = 0$ and $\inf(\mathrm{RP})_{0_X^*} = \max(\mathrm{ODP})_{0_X^*}$. The closedness assumption holds when X is finite dimensional (guaranteeing that any finitely generated convex cone in $X^*\times\mathbb{R}$ is closed). So, as proved in [14, Theorem 8.3], a linear semi-infinite program $(\mathrm{RP})_{0_X^*}$ is reducible if and only if (10) holds if and only if $\inf(\mathrm{RP})_{0_X^*} = \max(\mathrm{ODP})_{0_X^*}$.

We now assume that $\inf(\mathrm{RP})_{0_X^*} = \sup(\mathrm{ODP})_{0_X^*}\in\mathbb{R}$. By (8), there exist sequences $(S_r)_{r\in\mathbb{N}} \subset U$ and $(\lambda_r)_{r\in\mathbb{N}}$, with $\lambda^r\in\mathbb{R}_+^{S_r}$ for all $r\in\mathbb{N}$, such that

$$\lim_r \inf_{\lambda^r\in\mathbb{R}_+^{S_r}}\left\{\sum_{s\in S_r}\lambda_s^r b_s : \sum_{s\in S_r}\lambda_s^r a_s^* = -c^*\right\} = -\sup(\mathrm{ODP})_{0_X^*}.$$

Denote $v_r := -\sum_{s\in S_r}\lambda_s^r b_s$. Then,

$$\sum_{s\in S_r}\lambda_s\left(a_s^*,b_s\right)=-\left(c^*,v_r\right), \tag{11}$$

with $\lim_r v_r=\inf\,(\mathrm{RP})_{0_X^*}$. Let $A_r:=\left\{x\in X:\ \left\langle a_s^*,x\right\rangle\le b_s,\ \forall s\in S_r\right\},\ \ r\in\mathbb{N}$. Given $x\in A_r$, from (11),

$$0\ge\sum_{s\in S_r}\lambda_s^r\left(\left\langle a_s^*,x\right\rangle-b_s\right)=-\left\langle c^*,x\right\rangle+v_r.$$

Since $v_r\le\langle c^*,x\rangle$ for all $x\in A_r$,

$$v_r\le\inf\left\{\left\langle c^*,x\right\rangle:\left\langle a_s^*,x\right\rangle\le b_s,\ \forall s\in S_r\right\}\le\inf\,(\mathrm{RP})_{0_X^*}.$$

Thus,

$$\lim_r\inf\left\{\left\langle c^*,x\right\rangle:\left\langle a_s^*,x\right\rangle\le b_s,\ \forall s\in S_r\right\}=\inf\,(\mathrm{RP})_{0_X^*},$$

i.e., $(\mathrm{RP})_{0_X^*}$ is discretizable. Once again, the converse is true in linear semi-infinite programming [14, Corollary 8.2.1], but not in linear infinite programming.

3 Robust Conjugate Duality

We now turn back to the general perturbation function $F_u\colon X\times Y_u\to\mathbb{R}_\infty$, $u\in U$, and let $\Delta:=\left\{(u,y_u^*):u\in U,\,y_u^*\in Y_u^*\right\}$ be the disjoint union of the spaces Y_u^*. Recall that

$$(\mathrm{RP})_{x^*}:\quad\inf_{x\in X}\left\{\sup_{u\in U}F_u\left(x,0_u\right)-\left\langle x^*,x\right\rangle\right\}, \tag{12}$$

$$(\mathrm{ODP})_{x^*}:\qquad\sup_{(u,y_u^*)\in\Delta}-F_u^*\left(x^*,y_u^*\right). \tag{13}$$

Define $p\in\overline{\mathbb{R}}^X$ and $q\in\overline{\mathbb{R}}^{X^*}$ such that

$$p:=\sup_{u\in U}F_u(\cdot,0_u)\quad\text{and}\quad q:=\inf_{(u,y_u^*)\in\Delta}F_u^*(\cdot,y_u^*). \tag{14}$$

One then has

$$\begin{cases}p^*(x^*)=-\inf\,(\mathrm{RP})_{x^*},\quad q(x^*)=-\sup\,(\mathrm{ODP})_{x^*}\\ q^*=\displaystyle\sup_{(u,y_u^*)\in\Delta}\left(F_u^*\left(\cdot,y_u^*\right)\right)^*=\sup_{u\in U}F_u^{**}\left(\cdot,0_u\right)\le p,\end{cases} \tag{15}$$

and hence,

- *Weak robust duality* always holds

$$p^*(x^*) \leq q^{**}(x^*) \leq q(x^*), \text{ for all } x^* \in X^*. \tag{16}$$

- *Robust duality at x^** means

$$p^*(x^*) = q(x^*). \tag{17}$$

Robust duality at x^* also holds when either $p^*(x^*) = +\infty$ or $q(x^*) = -\infty$.

As an illustration, consider Case 4 with linear data, as in Example 1. Then, $p\,(x) = \langle c^*, x\rangle + \mathrm{i}_A\,(x)$, dom $p = A$, and so

$$p^*\left(0_X^*\right) = \sup_{x\in\mathbb{R}^n}\,(-p\,(x)) = -\inf_{x\in\mathbb{R}^n}\left\{\left\langle c^*, x\right\rangle + \mathrm{i}_A\,(x)\right\} = -\inf\,(\mathrm{RP})_{0_X^*}.$$

Similarly, from (7),

$$q\left(x^*\right) = \inf_{S\in U, \lambda\in\mathbb{R}^S}\left\{\sum_{s\in S}\lambda_s b_s : \sum_{s\in S}\lambda_s a_s^* = x^* - c^*\right\},$$

dom $q = c^* + \operatorname{cone}\left\{a_t^* : t \in T\right\}$ and

$$q\left(0_X^*\right) = \inf_{S\in U, \lambda\in\mathbb{R}_+^S}\left\{\sum_{s\in S}\lambda_s b_s : \sum_{s\in S}\lambda_s a_s^* = -\, c^*\right\} = -\sup\,(\mathrm{ODP})_{0_X^*}. \tag{18}$$

3.1 Basic Lemmas

Let us introduce the necessary notations. Given a lcHtvs Z, an extended real-valued function $h \in \overline{\mathbb{R}}^Z$, and $\varepsilon \in \mathbb{R}_+$, the set of ε-minimizers of h is defined by

$$\varepsilon - \operatorname{argmin} h := \begin{cases} \{z \in Z \,:\, h(z) \leq \inf_Z h + \varepsilon\}, & \text{if } \inf\limits_Z h \in \mathbb{R}, \\ \emptyset, & \text{if } \inf\limits_Z h \notin \mathbb{R}, \end{cases}$$

or, equivalently,

$$\varepsilon - \operatorname{argmin} h = \{z \in h^{-1}(\mathbb{R}) \,:\, h(z) \leq \inf_Z h + \varepsilon\}.$$

Note that $\varepsilon-$argmin $h \neq \emptyset$ when $\inf_Z h \in \mathbb{R}$ and $\varepsilon > 0$. Various calculus rules involving $\varepsilon-$argmin have been given in [26].

The ε-subdifferential of h at a point $a \in Z$ is the set (see, for instance, [19])

$$\begin{aligned}\partial^\varepsilon h(a) &:= \begin{cases}\{z^* \in Z^* \,:\, h(z) \geq h(a) + \langle z^*, z-a\rangle - \varepsilon, \forall z \in Z\}, & \text{if } \ h(a) \in \mathbb{R},\\ \emptyset, & \text{if } \ h(a) \notin \mathbb{R},\end{cases}\\ &= \Big\{z^* \in (h^*)^{-1}(\mathbb{R}) \,:\, h^*(z^*) + h(a) \leq \langle z^*, a\rangle + \varepsilon\Big\}.\end{aligned}$$

It can be checked that if $h \in \overline{\mathbb{R}}^X$ is convex and $h(a) \in \mathbb{R}$, then $\partial^\varepsilon h(a) \neq \emptyset$ for all $\varepsilon > 0$ if and only if h is lower semi-continuous at a.

The inverse of the set-valued mapping $\partial^\varepsilon h : Z \rightrightarrows Z^*$ is denoted by $M^\varepsilon h : Z^* \rightrightarrows Z$. For each $(\varepsilon, z^*) \in \mathbb{R}_+ \times Z^*$, we have

$$\Big(\partial^\varepsilon h\Big)^{-1}(z^*) = \Big(M^\varepsilon h\Big)(z^*) = \varepsilon - \text{argmin}\ (h - z^*).$$

Denoting by $\partial^\varepsilon h^*(z^*)$ the ε-subdifferential of h^* at $z^* \in Z^*$, namely,

$$\partial^\varepsilon h^*(z^*) = \Big\{z \in (h^{**})^{-1}(\mathbb{R}) \,:\, h^{**}(z) + h^*(z^*) \leq \langle z^*, z\rangle + \varepsilon\Big\},$$

where $h^{**}(z) := \sup\limits_{z^* \in Z^*} \{\langle z^*, z\rangle - h^*(z^*)\}$ is the biconjugate of h, we have

$$(M^\varepsilon h)(z^*) \subset (\partial^\varepsilon h^*\big)(z^*), \ \forall (\varepsilon, z^*) \in \mathbb{R}_+ \times Z^*,$$

with equality if and only if $h = h^{**}$.

For each $\varepsilon \in \mathbb{R}_+$, we consider the set-valued mapping $S^\varepsilon : X^* \rightrightarrows X$ as follows:

$$S^\varepsilon(x^*) := \big\{x \in p^{-1}(\mathbb{R}) \,:\, p(x) - \langle x^*, x\rangle \leq -q(x^*) + \varepsilon\big\}. \tag{19}$$

If $q(x^*) = -\infty$, then $S^\varepsilon(x^*) = p^{-1}(\mathbb{R})$. If $q(x^*) = +\infty$, then $S^\varepsilon(x^*) = \emptyset$.

Since $p^* \leq q$, it is clear that

$$S^\varepsilon(x^*) \subset (M^\varepsilon p)(x^*), \ \forall \varepsilon \geq 0, \ \forall x^* \in X^*. \tag{20}$$

Lemma 1 *Assume that* dom $p \neq \emptyset$. *Then, for each $x^* \in X^*$, the next statements are equivalent:*

(i) *Robust duality holds at x^*, i.e., $p^*(x^*) = q(x^*)$,*
(ii) $(M^\varepsilon p)\,(x^*) = S^\varepsilon(x^*), \quad \forall \varepsilon \geq 0,$
(iii) $\exists \bar{\varepsilon} > 0 : (M^\varepsilon p)\,(x^*) = S^\varepsilon(x^*), \quad \forall \varepsilon \in]0, \bar{\varepsilon}[.$

Proof [(i) ⇒ (ii)] By definition

$$\begin{aligned}(M^\varepsilon p)\,(x^*) &= \varepsilon - \text{argmin}(p - x^*)\\ &= \big\{x \in p^{-1}(\mathbb{R}) \,:\, p(x) - \langle x^*, x\rangle \leq -p^*(x^*) + \varepsilon\big\}.\end{aligned}$$

By (i) we thus have $(M^\varepsilon p)\,(x^*) = S^\varepsilon(x^*)$.
[(ii) ⇒ (iii)] It is obviously true.

[(iii) ⇒ (i)] Since $p^*(x^*) \leq q(x^*)$, (i) holds if $p^*(x^*) = +\infty$. Moreover, since $\operatorname{dom} p \neq \emptyset$, one has $p^*(x^*) \neq -\infty$. Let now $p^*(x^*) \in \mathbb{R}$. In order to get a contradiction, assume that $p^*(x^*) \neq q(x^*)$. Then $p^*(x^*) < q(x^*)$ and there exists $\varepsilon \in]0, \bar{\varepsilon}[$ such that $p^*(x^*) + \varepsilon < q(x^*)$. Since $\inf_{x \in X} \{p(x) - \langle x^*, x\rangle\} = -p^*(x^*) \in \mathbb{R}$ and $\varepsilon > 0$, we have $\varepsilon - \operatorname{argmin}(p - x^*) \neq \emptyset$. Let us pick $x \in (M^\varepsilon p)(x^*) = \varepsilon - \operatorname{argmin}(p - x^*)$. By (iii), we have $x \in S^\varepsilon(x^*)$ and

$$-p^*(x^*) \leq p(x) - \langle x^*, x\rangle \leq -q(x^*) + \varepsilon,$$

which contradicts $p^*(x^*) + \varepsilon < q(x^*)$. □

For each $\varepsilon \in \mathbb{R}_+$, let us introduce now the following set-valued mapping $J^\varepsilon : U \rightrightarrows X$:

$$J^\varepsilon(u) := \left\{x \in p^{-1}(\mathbb{R}) \ : \ p(x) \leq F_u(x, 0_u) + \varepsilon\right\}, \tag{21}$$

with the aim of making explicit the set $S^\varepsilon(x^*)$. To this purpose, given $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_+$, $u \in U$, and $y_u^* \in Y_u^*$, let us introduce the set-valued mapping $A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)} : X^* \rightrightarrows X$ such that

$$A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)}(x^*) := \Big\{x \in J^{\varepsilon_1}(u) \ : \ (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)\Big\}.$$

Lemma 2 *For each $x^* \in X^*$, $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_+$, $u \in U$, and $y_u^* \in Y_u^*$, one has*

$$A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)}(x^*) \ \subset \ S^{\varepsilon_1+\varepsilon_2}(x^*).$$

Proof Let $x \in J^{\varepsilon_1}(u)$ be such that $(x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)$. Then we have $F_u^*(x^*, y_u^*) \in \mathbb{R}$ and $F_u(x, 0_u) \in \mathbb{R}$. Moreover

$$F_u(x, 0_u) + \varepsilon_1 \geq p(x) \geq F_u(x, 0_u) \in \mathbb{R},$$

implying $p(x) \in \mathbb{R}$ and, by (15),

$$\begin{aligned} p(x) - \langle x^*, x\rangle &\leq F_u(x, 0_u) - \langle x^*, x\rangle + \varepsilon_1 \leq -F_u^*(x^*, y_u^*) + \varepsilon_1 + \varepsilon_2 \\ &\leq -q(x^*) + \varepsilon_1 + \varepsilon_2, \end{aligned}$$

that means $x \in S^{\varepsilon_1+\varepsilon_2}(x^*)$. □

Lemma 3 *Assume that*

$$\operatorname{dom} F_u \neq \emptyset, \forall u \in U. \tag{22}$$

Then, for each $x^* \in X^*$, $\varepsilon \in \mathbb{R}_+$, $\eta > 0$, one has

$$S^\varepsilon(x^*) \ \subset \bigcup_{\substack{u \in U \\ y_u^* \in Y_u^*}} \ \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon+\eta \\ \varepsilon_1 \geq 0, \ \varepsilon_2 \geq 0}} A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)}(x^*).$$

Proof Let $x \in p^{-1}(\mathbb{R})$ be such that $x \in S^{\varepsilon}(x^*)$, i.e.,

$$p(x) - \langle x^*, x\rangle \leq -q(x^*) + \varepsilon.$$

We then have, for any $\eta > 0$,

$$q(x^*) < \langle x^*, x\rangle - p(x) + \varepsilon + \eta$$

and, by definition of q and p, there exist $u \in U$, $y_u^* \in Y_u^*$ such that

$$F_u^*(x^*, y_u^*) \leq \langle x^*, x\rangle - p(x) + \varepsilon + \eta \leq \langle x^*, x\rangle - F_u(x, 0_u) + \varepsilon + \eta. \tag{23}$$

Since $p(x) \in \mathbb{R}$, $F_u^*(x^*, y_u^*) \neq +\infty$. In fact, by (22), $F_u^*(x^*, y_u^*) \in \mathbb{R}$. Similarly, $F_u(x, 0_u) \in \mathbb{R}$. Setting

$$\alpha_1 := p(x) - F_u(x, 0_u),\ \alpha_2 := F_u^*(x^*, y_u^*) + F_u(x, 0_u) - \langle x^*, x\rangle,$$

we get $\alpha_1 \in \mathbb{R}_+$, $\alpha_2 \in \mathbb{R}$. Actually $\alpha_2 \geq 0$ since, by definition of conjugate,

$$F_u^*(x^*, y_u^*) = \sup_{z \in X, y_u \in Y_u} \left\{\langle x^*, z\rangle + \langle y_u^*, y_u\rangle - F_u(z, y_u)\right\},$$

i.e., if $z = x$ and $y_u = 0_u$,

$$F_u^*(x^*, y_u^*) \geq \langle x^*, x\rangle - F_u(x, 0_u),$$

so that

$$F_u^*(x^*, y_u^*) + F_u(x, 0_u) - \langle x^*, x\rangle \geq 0.$$

Then, by (23), $0 \leq \alpha_1 + \alpha_2 \leq \varepsilon + \eta$. Consequently, there exist $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_+$ such that $\alpha_1 \leq \varepsilon_1, \alpha_2 \leq \varepsilon_2, \varepsilon_1 + \varepsilon_2 = \varepsilon + \eta$. Now $\alpha_1 \leq \varepsilon_1$ means that $x \in J^{\varepsilon_1}(u)$ and $\alpha_2 \leq \varepsilon_2$ means that $(x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)$, and we have $x \in A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)}(x^*)$. □

For each $x^* \in X^*$, $\varepsilon \in \mathbb{R}_+$, let us define

$$\begin{aligned}\mathcal{A}^{\varepsilon}(x^*) &:= \bigcap_{\eta>0} \bigcup_{\substack{u\in U \\ y_u^*\in Y_u^*}} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon+\eta \\ \varepsilon_1\geq 0,\ \varepsilon_2\geq 0}} A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)}(x^*) \\ &= \bigcap_{\eta>0} \bigcup_{\substack{u\in U \\ y_u^*\in Y_u^*}} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon+\eta \\ \varepsilon_1\geq 0,\ \varepsilon_2\geq 0}} \left\{x \in J^{\varepsilon_1}(u)\ :\ (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)\right\}.\end{aligned}$$

3.2 Robust Duality

We now can state the main result on characterizations of the robust conjugate duality.

Theorem 1 (**Robust duality**) *Assume that* dom $p \neq \emptyset$. *Then for each* $x^* \in X^*$, *the next statements are equivalent:*
(i) $\inf(\mathrm{RP})_{x^*} = \sup(\mathrm{ODP})_{x^*}$,
(ii) $(M^\varepsilon p)(x^*) = \mathcal{A}^\varepsilon(x^*), \quad \forall \varepsilon \geq 0$,
(iii) $\exists \bar{\varepsilon} > 0 : \ (M^\varepsilon p)(x^*) = \mathcal{A}^\varepsilon(x^*), \quad \forall \varepsilon \in]0, \bar{\varepsilon}[$.

Proof We firstly claim that if dom $p \neq \emptyset$ then for each $x^* \in X^*$, $\varepsilon \in \mathbb{R}_+$, it holds:

$$S^\varepsilon(x^*) \ = \ \mathcal{A}^\varepsilon(x^*). \tag{24}$$

Indeed, as dom $p \neq \emptyset$, (22) holds. It then follows from Lemma 3, $S^\varepsilon(x^*) \ \subset \ \mathcal{A}^\varepsilon(x^*)$. On the other hand, for each $\eta > 0$, one has, by Lemma 2,

$$\bigcup_{\substack{u\in U\\ y_u^*\in Y_u^*}} \ \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon+\eta\\ \varepsilon_1\geq 0,\ \varepsilon_2\geq 0}} A_{(u,y_u^*)}^{(\varepsilon_1,\varepsilon_2)}(x^*) \ \subset \ S^{\varepsilon+\eta}(x^*).$$

Taking the intersection over all $\eta > 0$ we get

$$\mathcal{A}^\varepsilon(x^*) \subset \bigcap_{\eta>0} S^{\varepsilon+\eta}(x^*) = S^\varepsilon(x^*),$$

and (24) follows. Taking into account the fact that (i) means $p^*(x^*) = q(x^*)$, the conclusions now follows from (24) and Lemma 1. □

For the deterministic optimization problem with linear perturbations (i.e., non-uncertain case where U is a singleton), the next result is a direct consequence of Theorem 1.

Corollary 1 (**Robust duality for Case 1**) *Let* $F : X \times Y \to \mathbb{R}_\infty$ *be such that* dom $F(\cdot, 0_Y) \neq \emptyset$. *Then, for each* $x^* \in X^*$, *the fundamental duality formula (2) holds, i.e.,*

$$\inf_{x\in X} \left\{F(x, 0_Y) - \langle x^*, x\rangle\right\} = \sup_{y\in Y^*} -F^*(x^*, y^*),$$

if and only any if the (equivalent) conditions (ii) or (iii) in Theorem 1 holds, where

$$\mathcal{A}^\varepsilon(x^*) = \bigcap_{\eta>0} \bigcup_{y^*\in Y^*} \left\{x \in X \ : \ (x, 0_Y) \in \left(M^{\varepsilon+\eta} F\right)(x^*, y^*)\right\}. \tag{25}$$

Proof Let $F_u = F : X \times Y \to \mathbb{R}_\infty, \ p = F(\cdot, 0_Y)$. In this case, one has,

$$J^\varepsilon(u) = \{x \in X \ : \ F(x, 0_Y) \in \mathbb{R}\}, \ \forall \varepsilon \geq 0,$$

and $\mathcal{A}^{\varepsilon}(x^*)$ will take the form (25). The conclusion follows from Theorem 1. □

For uncertain optimization problem without perturbations, the following result is a consequence of Theorem 1.

Corollary 2 (**Robust duality for Case 2**) *Let* $(f_u)_{u\in U} \subset \mathbb{R}_\infty^X$ *be a family of extended real-valued functions,* $p = \sup_{u\in U} f_u$ *be such that* $\operatorname{dom} p \neq \emptyset$*. Then, for each* $x^* \in X^*$, *the* $\inf-\sup$ *duality in robust optimization (4) holds, i.e.,*

$$\left(\sup_{u\in U} f_u\right)^*(x^*) = \inf_{u\in U} f_u^*(x^*)$$

if and only any of the (equivalent) conditions (ii) or (iii) in Theorem 1 holds, where

$$\mathcal{A}^{\varepsilon}(x^*) = \bigcap_{\eta>0} \bigcup_{u\in U} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon+\eta \\ \varepsilon_1\geq 0,\ \varepsilon_2\geq 0}} \left\{ J^{\varepsilon_1}(u) \cap (M^{\varepsilon_2} f_u)(x^*) \right\}, \tag{26}$$

with

$$J^{\varepsilon_1}(u) = \{x \in p^{-1}(\mathbb{R}) : \ f_u(x) \geq p(x) - \varepsilon_1\}.$$

Proof Let $F_u(x, y_u) = f_u(x)$, for all $u \in U$ and let $p = \sup\limits_{u\in U} f_u$. Then, by (21),

$$J^{\varepsilon}(u) = \left\{ x \in p^{-1}(\mathbb{R}) \ : \ f_u(x) \geq p(x) - \varepsilon \right\}, \ \forall \varepsilon \geq 0. \tag{27}$$

Moreover, recalling (3), for each $u \in U$ such that $\operatorname{dom} f_u \neq \emptyset$, $(x^*, y_u^*) \in X^* \times Y_u^*$, and $\varepsilon \geq 0$,

$$(M^{\varepsilon} F_u)\left(x^*, y_u^*\right) = \begin{cases} (M^{\varepsilon} f_u)(x^*), & \text{if } \ y_u^* = 0_u^*, \\ \emptyset, & \text{else.} \end{cases} \tag{28}$$

Finally, for each $(x^*, \varepsilon) \in X^* \times \mathbb{R}_+$, $\mathcal{A}^{\varepsilon}(x^*)$ takes the form as in (26). The conclusion now follows from Theorem 1. □

4 Strong Robust Duality

We retain the notations in Section 3 and consider the robust problem $(\mathrm{RP})_{x^*}$ and its robust dual problem $(\mathrm{ODP})_{x^*}$ given in (12) and (13), respectively. Let p and q be the functions defined by (14) and recall the relations in (15), that is,

$$\begin{cases} p^*(x^*) = -\inf\,(\mathrm{RP})_{x^*}, \quad q(x^*) = -\sup\,(\mathrm{ODP})_{x^*} \\ q^* = \sup\limits_{(u, y_u^*)\in\Delta} \left(F_u^*\left(\cdot, y_u^*\right)\right)^* = \sup\limits_{u\in U} F_u^{**}\left(\cdot, 0_u\right) \leq p. \end{cases}$$

In this section we establish characterizations of *strong robust duality at* x^*. Recall that the strong robust duality holds at x^* means that $\inf\,(\mathrm{RP})_{x^*} = \max\,(\mathrm{ODP})_{x^*}$, which is the same as

$$\exists (u, y_u^*) \in \Delta : p^*(x^*) = F_u^*(x^*, y_u^*).$$

For this, we need a technical lemma, but firstly, given $x^* \in X^*$, $u \in U$, $y_u^* \in Y_u^*$, and $\varepsilon \geq 0$, let us introduce the set

$$\begin{aligned} B^{\varepsilon}_{(u,y_u^*)}(x^*) &= \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon \\ \varepsilon_1\geqslant 0, \varepsilon_2\geqslant 0}} A^{(\varepsilon_1,\varepsilon_2)}_{(u,y_u^*)}(x^*) \\ &= \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon \\ \varepsilon_1\geqslant 0, \varepsilon_2\geqslant 0}} \Big\{ x \in J^{\varepsilon_1}(u) \ : \ (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*) \Big\}. \end{aligned}$$

Lemma 4 *Assume that* dom $F_u \neq \emptyset$, *for all* $u \in U$, *holds and let* $x^* \in X^*$ *be such that*

$$q(x^*) = \min_{\substack{u \in U \\ y_u^* \in Y_u^*}} F_u^*(x^*, y_u^*).$$

Then there exist $u \in U$, $y_u^* \in Y_u^*$ *such that*

$$S^{\varepsilon}(x^*) = B^{\varepsilon}_{(u,y_u^*)}(x^*), \ \forall \varepsilon \geq 0.$$

Proof By Lemma 2 we have $B^{\varepsilon}_{(u,y_u^*)}(x^*) \subset S^{\varepsilon}(x^*)$. Conversely, let $x \in S^{\varepsilon}(x^*)$. By the exactness of q at x^*, there exist $u \in U$ and $y_u^* \in Y_u^*$ such that

$$p(x) - \langle x^*, x \rangle \leq -F_u^*(x^*, y_u^*) + \varepsilon.$$

Since $p(x) \in \mathbb{R}$ and dom $F_u \neq \emptyset$, for all $u \in U$, we have $F_u^*(x^*, y_u^*) \in \mathbb{R}$, $F_u(x, 0_u) \in \mathbb{R}$,

$$\Big(p(x) - F_u(x, 0_u)\Big) + \Big(F_u(x, 0_u) + F_u^*(x^*, y_u^*) - \langle x^*, x \rangle\Big) \leq \varepsilon.$$

Consequently, there exist $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$ such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$,

$$p(x) - F_u(x, 0_u) \leq \varepsilon_1 \text{ and } F_u(x, 0_u) + F_u^*(x^*, y_u^*) - \langle x^*, x \rangle \leq \varepsilon_2,$$

that is, $x \in J^{\varepsilon_1}(u)$ and $(x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)$. Thus, $x \in A^{(\varepsilon_1,\varepsilon_2)}_{(u,y_u^*)}(x^*) \subset B^{\varepsilon}_{(u,y_u^*)}(x^*)$, since $\varepsilon_1 + \varepsilon_2 = \varepsilon$. □

Theorem 2 (**Strong robust duality**) *Assume that* dom $p \neq \emptyset$ *and let* $x^* \in X^*$. *The next statements are equivalent:*

(i) $\inf\,(\mathrm{RP})_{x^*} = \max\,(\mathrm{ODP})_{x^*}$,
(ii) $\exists u \in U,\ \exists y_u^* \in Y_u^* : (M^\varepsilon p)\,(x^*) = B^\varepsilon_{(u,y_u^*)}(x^*), \forall \varepsilon \geq 0$,
(iii) $\exists \bar{\varepsilon} > 0,\ \exists u \in U, \exists y_u^* \in Y_u^* : (M^\varepsilon p)\,(x^*) = B^\varepsilon_{(u,y_u^*)}(x^*), \forall \varepsilon \in]0, \bar{\varepsilon}[$.

Proof Observe firstly that (i) means that

$$p^*(x^*) = q(x^*) = \min_{\substack{u \in U \\ y_u^* \in Y_u^*}} F_u^*(x^*, y_u^*).$$

As dom $p \neq \emptyset$, (22) holds, and then by Lemmas 1 and 4, (i) implies the remaining conditions, which are equivalent to each other, and also that (iii) implies $p^*(x^*) = q(x^*)$.

We now prove that (iii) implies $q(x^*) = F_u^*(x^*, y_u^*)$. Assume by contradiction that there exists $\varepsilon > 0$ such that $q(x^*) + \varepsilon < F_u^*(x^*, y_u^*)$, and without loss of generality one can take $\varepsilon \in\,]0, \bar{\varepsilon}[$, where $\bar{\varepsilon} > 0$ appeared in (iii). Then, by (iii), $(M^\varepsilon p)\,(x^*) = B^\varepsilon_{(u,y_u^*)}(x^*)$.

Pick $x \in (M^\varepsilon p)\,(x^*) = B^\varepsilon_{(u,y_u^*)}(x^*)$. Then, there are $\varepsilon_1 \geq 0, \varepsilon_2 \geq 0,\ \varepsilon_1 + \varepsilon_2 = \varepsilon$, $x \in J^{\varepsilon_1}(u)$ and $(x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)$. In other words,

$$p(x) \leq F_u(x, 0_u) + \varepsilon_1, \tag{29}$$

$$F^*((x^*, y_u^*) + F_u(x, 0_u) \leq \langle x^*, x\rangle + \varepsilon_2. \tag{30}$$

It now follows from (29)–(30) that

$$\begin{aligned} p^*\left(x^*\right) &\geq \langle x^*, x\rangle - p\,(x) \geq \langle x^*, x\rangle - F_u(x, 0_u) - \varepsilon_1 \\ &\geq \langle x^*, x\rangle + F_u^*(x^*, y_u^*) - \langle x^*, x\rangle - \varepsilon_2 - \varepsilon_1 = F_u^*(x^*, y_u^*) - \varepsilon > q(x^*), \end{aligned}$$

which contradicts the fact that $p^*(x^*) = q(x^*)$. □

In deterministic optimization with linear perturbations we get the next consequence from Theorem 2.

Corollary 3 (**Strong robust duality for Case 1**) *Let* $F : X \times Y \to \mathbb{R}_\infty$, $p = F(\cdot, 0_Y)$, *and assume that* dom $p \neq \emptyset$. *Then, for each* $x^* \in X^*$, *the strong duality for* $(\mathrm{P})_{x^*}$ *in Case 1 holds at* x^*, *i.e.,*

$$\inf_{x \in X} \left\{F\,(x, 0_Y) - \left\langle x^*, x\right\rangle\right\} = \max_{y^* \in Y^*} -F^*\left(x^*, y^*\right),$$

if and only if one of the (equivalent) conditions (ii) or (iii) in Theorem 2 holds with $B^\varepsilon_{(u,y_u^*)}(x^*)$ *being replaced by*

$$B^\varepsilon_{y^*}(x^*) := \left\{x \in X\ :\ (x, 0_Y) \in (M^\varepsilon F)(x^*, y^*)\right\}. \tag{31}$$

Proof It is worth observing that we are in the non-uncertainty case (i.e., U is a singleton), and the set $B^{\varepsilon}_{(u,y^*_u)}(x^*)$ writes as in (31) for each $(x^*, y^*) \in X^* \times Y^*$, $\varepsilon \geq 0$. The conclusion follows from Theorem 2. □

In the non-perturbation case, Theorem 2 gives rise to

Corollary 4 (**Strong robust duality for Case 2**) *Let* $(f_u)_{u\in U} \subset \mathbb{R}^X_\infty$, $x^* \in X^*$, *and* $p = \sup\limits_{u\in U} f_u$ *such that* dom $p \neq \emptyset$. *Then, the robust duality formula*

$$\Big(\sup_{u\in U} f_u\Big)^*(x^*) = \min_{u\in U} f^*_u(x^*)$$

holds if and only if one of the (equivalent) conditions (ii) or (iii) in Theorem 2 holds with $B^{\varepsilon}_{(u,y^*_u)}(x^*)$ *being replaced by*

$$B^{\varepsilon}_u(x^*) := \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon\\ \varepsilon_1\geqslant 0,\varepsilon_2\geqslant 0}} \Big(J^{\varepsilon_1}(u) \cap (M^{\varepsilon_2} f_u)(x^*)\Big). \tag{32}$$

Proof Let $F_u(x, y_u) = f_u(x)$, $p = \sup\limits_{u\in U} f_u$, and, from (27) and (28) (see the proof of Corollary 2),

$$B^{\varepsilon}_{(u,y^*_u)}(x^*) = \begin{cases} \bigcup\limits_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon\\ \varepsilon_1\geqslant 0,\varepsilon_2\geqslant 0}} \Big(J^{\varepsilon_1}(u) \cap (M^{\varepsilon_2} f_u)(x^*)\Big), & \text{if } \; y^*_u = 0^*_u,\\ \emptyset, & \text{else,} \end{cases}$$

which in our situation, collapses to the set $B^{\varepsilon}_u(x^*)$ defined by (32). The conclusion now follows from Theorem 2. □

5 Reverse Strong and Min-Max Robust Duality

Given $F_u : X \times Y_u \to (\mathbb{R}_\infty)^X$ for each $u \in U$, $p = \sup\limits_{u\in U} F_u(\cdot, 0_u)$, and $x^* \in X^*$, we assume in this section that the problem $(\mathrm{RP})_{x^*}$ is finite-valued and admits an optimal solution or, in other words, that $\operatorname{argmin}(p - x^*) = (M^0 p)(x^*) \neq \emptyset$. For convenience, we set

$$(Mp)(x^*) := (M^0 p)(x^*), \quad S(x^*) := S^0(x^*), \text{ and}$$

$$\mathcal{A}(x^*) := \mathcal{A}^0(x^*) = \bigcap_{\eta>0} \bigcup_{\substack{u\in U\\ y^*_u\in Y^*_u}} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\eta\\ \varepsilon_1\geq 0,\ \varepsilon_2\geq 0}} \Big\{x \in J^{\varepsilon_1}(u) \ : \ (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y^*_u)\Big\}. \tag{33}$$

Theorem 3 (**Reverse strong robust duality**) *Let $x^* \in X^*$ be such that $(Mp)(x^*) \neq \emptyset$ and let $\mathcal{A}(x^*)$ be as in (33). The next statements are equivalent:*
(i) $\min\,(\mathrm{RP})_{x^*} = \sup\,(\mathrm{ODP})_{x^*}$,
(ii) $(Mp)(x^*) = \mathcal{A}(x^*)$.

Proof Since $(Mp)(x^*)\neq\emptyset$, dom $p\neq\emptyset$. It follows from Theorem 1 that [(i) $\Longrightarrow$ (ii)]. For the converse, let us pick $x \in (Mp)(x^*)$. Then by (ii), for each $\eta > 0$ there exist $u \in U$, $y_u^* \in Y_u^*$, $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$ such that $\varepsilon_1 + \varepsilon_2 = \eta$, $x \in J^{\varepsilon_1}(u)$, $(x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)$ and we have

$$\begin{aligned} q(x^*) &\leq F_u^*(x^*, y_u^*) \leq \langle x^*, x\rangle - F_u(x, 0_u) + \varepsilon_2 \\ &\leq \langle x^*, x\rangle - p(x) + \varepsilon_1 + \varepsilon_2 \leq p^*(x^*) + \eta. \end{aligned}$$

Since $\eta > 0$ is arbitrary we get $q(x^*) \leq p^*(x^*)$, which, together with the weak duality (see (16)), yields $q(x^*) = \langle x^*, x\rangle - p(x) = p^*(x^*)$, i.e., (i) holds and we are done. □

In the deterministic case we obtain from Theorem 3:

Corollary 5 (**Reverse strong robust duality for Case 1**) *Let $F : X \times Y \to \mathbb{R}_\infty$, $x^* \in X^*$, $p = F(\cdot, 0_Y)$, and*

$$\mathcal{A}(x^*) = \bigcap_{\eta>0} \bigcup_{y^*\in Y^*} \Big\{x \in X \;:\; (x, 0_Y) \in (M^\eta F)(x^*, y^*)\Big\}.$$

Assume that $(Mp)(x^*) \neq \emptyset$. Then the next statements are equivalent:
(i) $\min\limits_{x\in X} \{F(x, 0_Y) - \langle x^*, x\rangle\} = \sup\limits_{y^*\in Y^*} -F^*(x^*, y^*)$,
(ii) $(Mp)(x^*) = \mathcal{A}(x^*)$.

Corollary 6 (**Reverse strong robust duality for Case 2**) *Let $(f_u)_{u\in U} \subset \mathbb{R}_\infty^X$, $p = \sup\limits_{u\in U} f_u$, $x^* \in X^*$, and*

$$\mathcal{A}(x^*) := \bigcap_{\eta>0} \bigcup_{u\in U} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\eta \\ \varepsilon_1\geq 0,\ \varepsilon_2\geq 0}} \Big(J^{\varepsilon_1}(u) \cap (M^{\varepsilon_2} f_u)(x^*)\Big),$$

where

$$J^{\varepsilon_1}(u) = \Big\{x \in p^{-1}(\mathbb{R}) \;:\; f_u(x) \geq p(x) - \varepsilon_1\Big\}.$$

Assume that $(Mp)(x^*) \neq \emptyset$. Then the next statements are equivalent:
(i) $\Big(\sup\limits_{u\in U} f_u\Big)^*(x^*) = \inf\limits_{u\in U} f_u^*(x^*)$, *with attainment at the first member,*
(ii) $(Mp)(x^*) = \mathcal{A}(x^*)$.

Now, for each $u \in U$, $y_u^* \in Y_u^*$, $x^* \in X^*$, we set

$$J(u) := J^0(u) = \Big\{x \in p^{-1}(\mathbb{R}) \ : \ F_u(x, 0_u) = p(x)\Big\},$$

$$(MF_u)(x^*, y_u^*) := (M^0 F_u)(x^*, y_u^*) = \operatorname{argmin}\Big(F_u - \langle x^*, \cdot\rangle - \langle y_u^*, \cdot\rangle\Big),$$

and

$$B_{(u,y_u^*)}(x^*) := B^0_{(u,y_u^*)}(x^*) = \Big\{x \in J(u) \ : \ (x, 0_u) \in (MF_u)(x^*, y_u^*)\Big\}. \tag{34}$$

Theorem 4 (**Min-max robust duality**) *Let $x^* \in X^*$ be such that $(Mp)(x^*) \neq \emptyset$. The next statements are equivalent:*

(i) $\min\,(\mathrm{RP})_{x^*} = \max\,(\mathrm{ODP})_{x^*}$,
(ii) $\exists u \in U,\ \exists y_u^* \in Y_u^* \ : \ (Mp)(x^*) = B_{(u,y_u^*)}(x^*)$,
where $B_{(u,y_u^*)}(x^*)$ is the set defined in (34).

Proof By Theorem 2 we know that [(i) $\Longrightarrow$ (ii)]. We now prove that [(ii) $\Longrightarrow$ (i)]. Pick $x \in (Mp)(x^*)$ which is non-empty by assumption. Then by (ii), $x \in B_{(u,y_u^*)}(x^*)$, which yields $x \in J(u)$ and $(x, 0_u) \in (MF_u)(x^*, y_u^*)$. Hence,

$$\begin{aligned} q(x^*) &\leq F_u^*(x^*, y_u^*) \leq \langle x^*, x\rangle - F_u(x, 0_u) \\ &\leq \langle x^*, x\rangle - p(x) \leq p^*(x^*) \leq q(x^*), \end{aligned}$$

which means that $q(x^*) = F_u^*(x^*, y_u^*) = \langle x^*, x\rangle - p(x) = p^*(x^*)$ and (i) follows. □

Corollary 7 (**Min-max robust duality for Case 1**) *Let $F : X \times Y \to \mathbb{R}_\infty$, $x^* \in X^*$, $p = F(\cdot, 0_Y)$, and for each $y^* \in Y^*$,*

$$B_{y^*}(x^*) := \Big\{x \in X \ : \ (x, 0_Y) \in (MF)(x^*, y^*)\Big\}.$$

Assume that $(Mp)(x^*) \neq \emptyset$. The next statements are equivalent:

(i) $\min\limits_{x\in X}\{F\,(x, 0_Y) - \langle x^*, x\rangle\} = \max\limits_{y^*\in Y^*} -F^*\,(x^*, y^*)$,
(ii) $\exists y^* \in Y^*$: $(Mp)(x^*) = B_{y^*}(x^*)$.

Corollary 8 (**Min-max robust duality for Case 2**) *Let $(f_u)_{u\in U} \subset \mathbb{R}_\infty^X$, $p = \sup\limits_{u\in U} f_u$, $x^* \in X^*$, and for each $u \in U$,*

$$B_u(x^*) := J(u) \cap (Mf_u)(x^*),$$

where $J(u) = \{x \in p^{-1}(\mathbb{R}) \ : \ f_u(x) = p(x)\}$. Assume that $(Mp)(x^*) \neq \emptyset$. Then the next statements are equivalent:

(i) $\Big(\sup\limits_{u\in U} f_u\Big)^*(x^*) = \min\limits_{u\in U} f_u^*(x^*)$, *with attainment at the first member,*
(ii) $\exists u \in U$: $(Mp)(x^*) = B_u(x^*)$.

6 Stable Robust Duality

Let us first recall some notations. Given $F_u : X \times Y_u \to \mathbb{R}_\infty$, $u \in U$, $p = \sup\limits_{u \in U} F_u(\cdot, 0_u)$ and $q = \inf\limits_{\substack{u \in U \\ y_u^* \in Y_u^*}} F_u^*(\cdot, y_u^*)$. Remember that $p^*(x^*) \leq q(x^*)$ for each $x^* \in X^*$. *Stable robust duality* means that $\inf\,(\mathrm{RP})_{x^*} = \sup\,(\mathrm{ODP})_{x^*}$ for all $x^* \in X^*$, or equivalently,

$$p^*(x^*) = q(x^*),\ \forall x^* \in X^*.$$

Theorem 1 says that, if $\operatorname{dom} p \neq \emptyset$, then stable robust duality holds if and only if for each $\varepsilon \geq 0$ the set-valued mappings $M^\varepsilon p,\ \mathcal{A}^\varepsilon : X^* \rightrightarrows X$ coincide, where, for each $x^* \in X^*$,

$$\begin{aligned}(M^\varepsilon p)(x^*) &:= \ \varepsilon - \operatorname{argmin}(p - x^*),\\ \mathcal{A}^\varepsilon(x^*) &:= \bigcap_{\eta > 0} \bigcup_{\substack{u \in U \\ y_u^* \in Y_u^*}} \bigcup_{\substack{\varepsilon_1 + \varepsilon_2 = \varepsilon + \eta \\ \varepsilon_1 \geq 0,\ \varepsilon_2 \geq 0}} \Big\{x \in J^{\varepsilon_1}(u)\ :\ (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)\Big\}.\end{aligned}$$

Consequently, stable robust duality holds if and only if for each $\varepsilon \geq 0$, the inverse set-valued mappings

$$(M^\varepsilon p)^{-1},\ (\mathcal{A}^\varepsilon)^{-1} : X \rightrightarrows X^*,$$

coincide. Recall that $(M^\varepsilon p)^{-1}$ is nothing but the ε-subdifferential of p at x.

Let us now make explicit $(\mathcal{A}^\varepsilon)^{-1}$. To this end we need to introduce for each $\varepsilon \geq 0$ the (ε-active indexes) set-valued mapping $I^\varepsilon : X \rightrightarrows U$ with

$$I^\varepsilon(x) = \begin{cases} \Big\{u \in U\ :\ F_u(x, 0_u) \geq p(x) - \varepsilon\Big\}, & \text{if}\quad p(x) \in \mathbb{R},\\ \emptyset, & \text{if}\quad p(x) \notin \mathbb{R}. \end{cases} \tag{35}$$

We observe that I^ε is nothing but the inverse of the set-valued mapping $J^\varepsilon : U \rightrightarrows X$ defined in (21).

Lemma 5 *For each $(\varepsilon, x) \in \mathbb{R}_+ \times X$ one has*

$$(\mathcal{A}^\varepsilon)^{-1}(x) = \bigcap_{\eta > 0} \bigcup_{\substack{\varepsilon_1 + \varepsilon_2 = \varepsilon + \eta \\ \varepsilon_1 \geqslant 0, \varepsilon_1 \geqslant 0}} \bigcup_{u \in I^{\varepsilon_1}(x)} \operatorname{proj}_{X^*}^u \partial^{\varepsilon_2} F_u(x, 0_u),$$

where $\operatorname{proj}_{X^*}^u : X^* \times Y_u^* \longrightarrow X^*$ *is the projection mapping* $\operatorname{proj}_{X^*}^u(x^*, y_u^*) = x^*$.

Proof Let $(\varepsilon, x, x^*) \in \mathbb{R}_+ \times X \times X^*$. One has

$$
\begin{aligned}
x^* \in (\mathcal{A}^\varepsilon)^{-1}(x) &\Leftrightarrow x \in \mathcal{A}^\varepsilon(x^*) \\
&\Leftrightarrow \begin{cases} \forall \eta > 0, \exists u \in U, \exists y_u^* \in Y_u^*, \exists(\varepsilon_1, \varepsilon_2) \in \mathbb{R}_+^2 \quad \text{such that} \\ \varepsilon_1 + \varepsilon_2 = \varepsilon + \eta, \quad x \in J^{\varepsilon_1}(u) \text{ and } (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*) \end{cases} \\
&\Leftrightarrow \begin{cases} \forall \eta > 0, \exists u \in U, \exists y_u^* \in Y_u^*, \exists(\varepsilon_1, \varepsilon_2) \in \mathbb{R}_+^2 \quad \text{such that} \\ \varepsilon_1 + \varepsilon_2 = \varepsilon + \eta, \quad u \in I^{\varepsilon_1}(x), \text{ and } (x^*, y_u^*) \in \Big(\partial^{\varepsilon_2} F_u\Big)(x, 0_u) \end{cases} \\
&\Leftrightarrow \begin{cases} \forall \eta > 0, \exists u \in U, \exists(\varepsilon_1, \varepsilon_2) \in \mathbb{R}_+^2 \quad \text{such that} \\ \varepsilon_1 + \varepsilon_2 = \varepsilon + \eta, \quad u \in I^{\varepsilon_1}(x), \text{ and } x^* \in \operatorname{proj}_{X^*}^u \Big(\partial^{\varepsilon_2} F_u\Big)(x, 0_u) \end{cases} \\
&\Leftrightarrow x^* \in \bigcap_{\eta > 0} \bigcup_{\substack{\varepsilon_1 + \varepsilon_2 = \varepsilon + \eta \\ \varepsilon_1 \geqslant 0, \varepsilon_1 \geqslant 0}} \bigcup_{u \in I^{\varepsilon_1}(x)} \operatorname{proj}_{X^*}^u (\partial^{\varepsilon_2} F_u)(x, 0_u).
\end{aligned}
$$

□

Now, for each $(\varepsilon, x) \in \mathbb{R}_+ \times X$, let us set

$$
C^\varepsilon(x) := \bigcap_{\eta > 0} \bigcup_{\substack{\varepsilon_1 + \varepsilon_2 = \varepsilon + \eta \\ \varepsilon_1 \geqslant 0, \varepsilon_1 \geqslant 0}} \bigcup_{u \in I^{\varepsilon_1}(x)} \operatorname{proj}_{X^*}^u (\partial^{\varepsilon_2} F_u)(x, 0_u). \tag{36}
$$

Applying Theorem 1 and Lemma 5 we obtain:

Theorem 5 (**Stable robust duality**) *Assume that* dom $p \neq \emptyset$. *The next statements are equivalent:*

(i) $\inf (\mathrm{RP})_{x^*} = \sup (\mathrm{ODP})_{x^*}$ *for all* $x^* \in X^*$,
(ii) $\partial^\varepsilon p(x) = C^\varepsilon(x), \ \forall (\varepsilon, x) \in \mathbb{R}_+ \times X$,
(iii) $\exists \bar{\varepsilon} > 0$: $\partial^\varepsilon p(x) = C^\varepsilon(x), \ \forall (\varepsilon, x) \in]0, \bar{\varepsilon}[\times X$.

Corollary 9 (**Stable robust duality for Case 1**) *Let* $F : X \times Y \to \mathbb{R}_\infty$ *be such that* dom $F(\cdot, 0_Y) \neq \emptyset$. *Let* $\operatorname{proj}_{X^*} : X^* \times Y^* \longrightarrow X^*$ *be the projection mapping* $\operatorname{proj}_{X^*}(x^*, y^*) = x^*$. *Then, the next statements are equivalent:*

(i) $\inf\limits_{x \in X} \Big\{ F(x, 0_Y) - \langle x^*, x \rangle \Big\} = \sup\limits_{y^* \in Y^*} -F^*(x^*, y^*), \ \forall x^* \in X^*$,
(ii) $(\partial^\varepsilon p)(x) = \bigcap_{\eta > 0} \operatorname{proj}_{X^*}(\partial^{\varepsilon + \eta} F)(x, 0_Y), \ \forall (\varepsilon, x) \in \mathbb{R}_+ \times X$,
(iii) $\exists \bar{\varepsilon} > 0$: $(\partial^\varepsilon p)(x) = \bigcap_{\eta > 0} \operatorname{proj}_{X^*}(\partial^{\varepsilon + \eta} F)(x, 0_Y), \ \forall (\varepsilon, x) \in]0, \bar{\varepsilon}[\times X$.

Proof Let $U = \{u_0\}$ and $F = F_{u_0} : X \times Y \to \mathbb{R}_\infty$, $Y = Y_{u_0}$, $p = F(\cdot, 0_Y)$. Then for each $(\varepsilon, x) \in \mathbb{R}_+ \times X$,

$$
I^\varepsilon(x) = \begin{cases} \{u_0\}, & \text{if} \quad p(x) \in \mathbb{R}, \\ \emptyset, & \text{if} \quad p(x) \notin \mathbb{R}, \end{cases}
$$

and

$$
\bigcup_{\substack{\varepsilon_1 + \varepsilon_2 = \varepsilon \\ \varepsilon_1 \geqslant 0, \varepsilon_1 \geqslant 0}} \bigcup_{u \in I^{\varepsilon_1}(x)} \operatorname{proj}_{X^*}^u (\partial^{\varepsilon_2} F_u)(x, 0_u) = \operatorname{proj}_{X^*} \Big(\partial^\varepsilon F\Big)(x, 0_Y). \tag{37}
$$

The conclusion now follows from (36)–(37) and Theorem 5. □

Remark 1 Condition (ii) in Corollary 9 was quoted in [17, Theorem 4.3] for all $(\varepsilon, x) \in]0, +\infty[\times \mathbb{R}$, which is equivalent.

Corollary 10 **(Stable robust duality for Case 2)** *Let* $(f_u)_{u\in U} \subset \mathbb{R}_\infty^X$, $p = \sup\limits_{u\in U} f_u$, *and assume that* $\operatorname{dom} p \neq \emptyset$. *The next statements are equivalent:*

(i) $\left(\sup\limits_{u\in U} f_u\right)^*(x^*) = \inf\limits_{u\in U} f_u^*(x^*),\ \forall x^* \in X^*,$
(ii) $(\partial^\varepsilon p)(x) = C^\varepsilon(x),\ \forall(\varepsilon, x) \in \mathbb{R}_+ \times X,$
(iii) $\exists \bar{\varepsilon} > 0$: $(\partial^\varepsilon p)(x) = C^\varepsilon(x),\ \forall(\varepsilon, x) \in]0, \bar{\varepsilon}[\times X,$
where $C^\varepsilon(x)$ *is the set*

$$C^\varepsilon(x) = \bigcap_{\eta>0} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon+\eta \\ \varepsilon_1\geqslant 0, \varepsilon_1\geqslant 0}} \bigcup_{u\in I^{\varepsilon_1}(x)} (\partial^{\varepsilon_2} f_u)(x),\ \ \forall(\varepsilon, x) \in \mathbb{R}_+ \times X. \tag{38}$$

Proof Let $F_u : X \times Y_u \to \mathbb{R}_\infty$ be such that $F_u(x, y_u) = f_u(x)$ for all $u \in U$. Then for any $(\varepsilon, x) \in \mathbb{R}_+ \times X$,

$$I^\varepsilon(x) = \begin{cases} \Big\{u \in U \ :\ f_u(x) \geq p(x) - \varepsilon\Big\}, & \text{if} \quad p(x) \in \mathbb{R}, \\ \emptyset, & \text{if} \quad p(x) \notin \mathbb{R}, \end{cases}$$

$$(\partial^\varepsilon F_u)(x, 0_u) = (\partial^\varepsilon f_u)(x) \times \{0_u^*\}, \quad \forall(u, \varepsilon, x) \in U \times \mathbb{R}_+ \times X,$$

and $C^\varepsilon(x)$ reads as in (38). The conclusion now follows from Theorem 5. □

7 Stable Strong Robust Duality

We retain all the notations used in the Sections 3–6. Given $(\varepsilon, u) \in \mathbb{R}_+ \times U$ and $y_u^* \in Y_U^*$ we have introduced in Section 4 the set-valued mapping $B^\varepsilon_{(u,y_u^*)} : X^* \rightrightarrows X$ defined by

$$B^\varepsilon_{(u,y_u^*)}(x^*) = \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon \\ \varepsilon_1\geqslant 0, \varepsilon_2\geqslant 0}} \Big\{x \in J^{\varepsilon_1}(u) \ :\ (x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)\Big\}.$$

Let us now define $B^\varepsilon : X^* \rightrightarrows X$ by setting

$$B^\varepsilon(x^*) := \bigcup_{\substack{u\in U \\ y_u^*\in Y_u^*}} B^\varepsilon_{(u,y_u^*)}(x^*),\ \ \forall x^* \in X^*.$$

Lemma 6 *For each* $(\varepsilon, x) \in \mathbb{R}_+ \times X$ *we have*

$$(B^\varepsilon)^{-1}(x) = \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon \\ \varepsilon_1\geqslant 0, \varepsilon_2\geqslant 0}} \bigcup_{u\in I^{\varepsilon_1}(x)} \operatorname{proj}^u_{X^*}(\partial^{\varepsilon_2} F_u)(x, 0_u).$$

Proof $x^* \in (B^\varepsilon)^{-1}(x)$ means that there exist $u \in U$, $y_u^* \in Y_u^*$ $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$, such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$, $x \in J^{\varepsilon_1}(u)$, and $(x, 0_u) \in (M^{\varepsilon_2} F_u)(x^*, y_u^*)$, or, equivalently, $u \in I^{\varepsilon_1}(x)$, and $(x^*, y_u^*) \in (\partial^{\varepsilon_2} F_u)(x, 0_u)$. In other words, there exist $u \in U$, $y_u^* \in Y_u^*$ such that $x \in B^\varepsilon_{(u,y_u^*)}(x^*)$, that is $x \in B^\varepsilon(x^*)$. □

For each $\varepsilon \geq 0$ let us introduce the set-valued mapping $D^\varepsilon := (B^\varepsilon)^{-1}$. Now Lemma 6 writes

$$D^\varepsilon(x) \; = \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon \\ \varepsilon_1\geqslant 0, \varepsilon_2\geqslant 0}} \bigcup_{u\in I^{\varepsilon_1}(x)} \operatorname{proj}^u_{X^*}(\partial^{\varepsilon_2} F_u)(x, 0_u), \; \forall (\varepsilon, x) \in \mathbb{R}_+ \times X. \tag{39}$$

Note that

$$C^\varepsilon(x) = \bigcap_{\eta>0} D^{\varepsilon+\eta}(x), \; \forall(\varepsilon, x) \in \mathbb{R}_+ \times X, \tag{40}$$

and that $D^\varepsilon(x) = \emptyset$ whenever $p(x) \notin \mathbb{R}$.

We now provide a characterization of stable strong robust duality in terms of ε-subdifferential formulas.

Theorem 6 (**Stable strong robust duality**) *Assume that* dom $p \neq \emptyset$, *and let* D^ε *as in* (39). *The next statements are equivalent:*
(i) $\inf\,(\mathrm{RP})_{x^*} = \max\,(\mathrm{ODP})_{x^*} = \max\limits_{\substack{u\in U \\ y_u^*\in Y_u^*}} -F_u^*(x^*, y_u^*), \;\; \forall x^* \in X^*,$
(ii) $\partial^\varepsilon p(x) = D^\varepsilon(x), \; \forall(\varepsilon, x) \in \mathbb{R}_+ \times X$.

Proof [(i) $\Longrightarrow$ (ii)] Let $x^* \in \partial^\varepsilon p(x)$. Then $x \in (M^\varepsilon p)(x^*)$. Since strong robust duality holds at x^*, Theorem 2 says that there exist $u \in U$, $y_u^* \in Y_u^*$ such that $x \in B^\varepsilon_{(u,y_u^*)}(x^*) \subset B^\varepsilon(x^*)$, and finally $x^* \in D^\varepsilon(x)$ by Lemma 6. Thus $\partial^\varepsilon p(x) \subset D^\varepsilon(x)$.

Now, let $x^* \in D^\varepsilon(x)$. By Lemma 6 we have $x \in B^\varepsilon(x^*)$ and there exist $u \in U$, $y_u^* \in Y_u^*$ such that $x \in B^\varepsilon_{(u,y_u^*)}(x^*)$. By Lemma 2 and the definition of $B^\varepsilon_{(u,y_u^*)}(x^*)$ we have $x \in S^\varepsilon(x^*)$, and, by (20), $x \in (M^\varepsilon p)(x^*)$ which means that $x^* \in \partial^\varepsilon p(x)$, and hence, $D^\varepsilon(x) \subset \partial^\varepsilon p(x)$. Thus (ii) follows.

[(ii) $\Longrightarrow$ (i)] If $p^*(x^*) = +\infty$ then $q(x^*) = +\infty$ and one has $p^*(x^*) = F_u^*(x^*, y_u^*) = +\infty$ for all $u \in U$, $y_u^* \in Y_u^*$, and (i) holds. Assume that $p^*(x^*) \in \mathbb{R}$ and pick $x \in p^{-1}(\mathbb{R})$ which is non-empty as dom $p \neq \emptyset$ and $p^*(x^*) \in \mathbb{R}$. Let $\varepsilon := p(x) + p^*(x^*) - \langle x^*, x\rangle$. Then $\varepsilon \geq 0$ and we have $x^* \in \partial^\varepsilon p(x)$. By (ii) $x \in D^\varepsilon(x)$ and hence, there exist $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$, $u \in U$, and $y_u^* \in Y_u^*$ such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$, $u \in I^{\varepsilon_1}(x)$, $(x^*, y_u^*) \in (\partial^{\varepsilon_2} F_u)(x, 0_u)$. We have

$$\begin{aligned} q(x^*) &\le F_u^*(x^*, y_u^*) \le \langle x^*, x\rangle - F_u(x, 0_u) + \varepsilon_2 \\ &\le \langle x^*, x\rangle - p(x) + \varepsilon_1 + \varepsilon_2 = p^*(x^*) \qquad \text{(by definition of } \varepsilon) \\ &\le q(x^*), \end{aligned}$$

and finally, $q(x^*) = F_u^*(x^*, y_u^*) = p^*(x^*)$, which is (i). □

Next, as usual, we give two consequences of Theorem 6 for the non-uncertainty and non-parametric cases.

Corollary 11 (**Stable strong duality for Case 1**) *Let* $F : X\times Y \to \mathbb{R}_\infty$, $p = F(\cdot, 0_Y)$, dom $p \neq \emptyset$. *The next statements are equivalent:*
(i) $\inf\limits_{x\in X} \Big\{F(x, 0_Y) - \langle x^*, x\rangle\Big\} = \max\limits_{y^*\in Y^*} -F^*(x^*, y^*),\ \forall x^* \in X^*,$
(ii) $\partial^\varepsilon p(x) = \mathrm{proj}_{X^*}(\partial^\varepsilon F)(x, 0_y),\ \forall(\varepsilon, x) \in \mathbb{R}_+ \times X.$

Proof This is the non-uncertainty case (i.e., the uncertainty set is a singleton) of the general problem $(\mathrm{RP})_{x^*}$, with $U = \{u_0\}$ and $F_{u_0} = F : X \times Y \to \mathbb{R}_\infty$. We have from (37),

$$D^\varepsilon(x) = \mathrm{proj}_{X^*}(\partial^\varepsilon F)(x, 0_Y),\ \forall(\varepsilon, x) \in \mathbb{R}_+ \times X. \tag{41}$$

The conclusion now follows from Theorem 6. □

Corollary 12 (**Stable strong duality for Case 2**) *Let* $(f_u)_{u\in U} \subset \mathbb{R}_\infty^X$, $p = \sup\limits_{u\in U} f_u$, *and* dom $p \neq \emptyset$. *The next statements are equivalent:*
(i) $(\sup\limits_{u\in U} f_u)^*(x^*) = \min\limits_{u\in U} f_u^*(x^*),\ \forall x^* \in X^*,$
(ii) $\partial^\varepsilon p(x) = D^\varepsilon(x),\ \forall(\varepsilon, x) \in \mathbb{R}_+ \times X$, *where*

$$D^\varepsilon(x) = \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\varepsilon \\ \varepsilon_1\geqslant 0,\varepsilon_2\geqslant 0}} \bigcup_{u\in I^{\varepsilon_1}(x)} (\partial^{\varepsilon_2} f_u)(x),\ \forall(\varepsilon, x) \in \mathbb{R}_+ \times X, \tag{42}$$

and

$$I^\varepsilon(x) = \begin{cases} \Big\{u \in U\ :\ f_u(x) \ge p(x) - \varepsilon\Big\} & \text{if}\quad p(x) \in \mathbb{R}, \\ \emptyset & \text{if}\quad p(x) \notin \mathbb{R}. \end{cases}$$

Proof In this non-parametric situation, let $F_u(x, y_u) = f_u(x)$. It is easy to see that in this case, the set $D^\varepsilon(x)$ can be expressed as in (42), and the conclusion follows from Theorem 6. □

8 Exact Subdifferential Formulas: Robust Basic Qualification Condition

Given $F_u : X \times Y_u \to \mathbb{R}_\infty,\ u \in U$, as usual, we let $p = \sup\limits_{u\in U} F_u(\cdot, 0_u)$, $q := \inf\limits_{(u,y_u^*)\in\Delta} F_u^*(\cdot, y_u^*)$. Again, we consider the robust problem $(\mathrm{RP})_{x^*}$ and its robust dual problem $(\mathrm{ODP})_{x^*}$ given in (12) and (13), respectively. Note that the reverse strong robust duality holds at x^* means that, for some $\bar{x} \in X$, it holds

$$\begin{aligned} - p^*(x^*) = \min\,(\mathrm{RP})_{x^*} &= \sup_{u\in U} F_u(\bar{x}, 0_u) - \langle x^*, \bar{x}\rangle \\ &= p(\bar{x}) - \langle x^*, \bar{x}\rangle = \sup\,(\mathrm{ODP})_{x^*} = -q(x^*). \end{aligned} \tag{43}$$

Now, let us set, for each $x \in X$,

$$D(x) := \ D^0(x) = \bigcup_{u\in I(x)} \mathrm{proj}^u_{X^*}(\partial F_u)(x, 0_u), \tag{44}$$

$$C(x) := \ C^0(x) = \bigcap_{\eta>0} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\eta \\ \varepsilon_1\geqslant 0,\varepsilon_2\geqslant 0}} \bigcup_{u\in I^{\varepsilon_1}(x)} \mathrm{proj}^u_{X^*}(\partial^{\varepsilon_2} F_u)(x, 0_u), \tag{45}$$

where $I^{\varepsilon_1}(x)$ is defined as in (35) and

$$I(x) := \begin{cases} \{u \in U : F_u(x, 0_u) = p(x)\}, & \text{if} \quad p(x) \in \mathbb{R}, \\ \emptyset, & \text{if} \quad p(x) \notin \mathbb{R}. \end{cases} \tag{46}$$

Lemma 7 *For each $x \in X$, it holds*

$$D(x) \subset C(x) \subset \partial p(x).$$

Proof The first inclusion is easy to check. Now let $x^* \in C(x)$. For each $\eta > 0$ there exist $(\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2_+$, $u \in I^{\varepsilon_1}(x)$, and $y_u^* \in Y_u^*$ such that $\varepsilon_1 + \varepsilon_2 = \eta$ and $(x^*, y_u^*) \in (\partial^{\varepsilon_2} F_u)(x, 0_u)$. We then have $F_u^*(x^*, y_u^*) + F_u(x, 0_u) - \langle x^*, x\rangle \leq \varepsilon_2$, $p(x) \leq F_u(x, 0_u) + \varepsilon_1$ (as $u \in I^{\varepsilon_1}(x)$), and $p^*(x^*) \leq q(x^*) \leq F_u^*(x^*, y_u^*)$. Consequently,

$$p^*(x^*) + p(x) - \langle x^*, x\rangle \leq F_u^*(x^*, y_u^*) + F_u(x, 0_u) + \varepsilon_1 - \langle x^*, x\rangle \leq \varepsilon_1 + \varepsilon_2 = \eta.$$

Since $\eta > 0$ is arbitrary we get $p^*(x^*) + p(x) - \langle x^*, x\rangle \leq 0$, which means that $x^* \in \partial p(x)$. The proof is complete. □

Theorem 7 *Let $x \in p^{-1}(\mathbb{R})$ and $C(x)$ be as in (45). The next statements are equivalent:*

(i) $\partial p(x) = C(x)$,

(ii) *Reverse strong robust duality holds at each $x^* \in \partial p(x)$,*

(iii) *Robust duality holds at each $x^* \in \partial p(x)$.*

Proof [(i) $\Longrightarrow$ (ii)] Let $x^* \in \partial p(x)$. We have $x^* \in C(x) = (\mathcal{A})^{-1}(x)$ (see Lemma 5 with $\varepsilon = 0$). Then $x \in \mathcal{A}(x^*) = S(x^*)$ (see (24) with $\varepsilon = 0$), and therefore,

$$-p^*(x^*) \leq p(x) - \langle x^*, x\rangle \leq -q(x^*) \leq -p^*(x^*),$$

$$-p^*(x^*) = \min_{z\in X}\{p(z) - \langle x^*, z\rangle\} = p(x) - \langle x^*, x\rangle = -q(x^*),$$

that means that reverse strong robust duality holds at x^* (see (43)).
[(ii) $\Longrightarrow$ (iii)] is obvious.
[(iii) $\Longrightarrow$ (i)] By Lemma 7 it suffices to check that the inclusion "$\subset$" holds. Let $x^* \in \partial p(x)$. We have $x \in (Mp)(x^*)$. Since robust duality holds at x^*, Theorem 1 (with $\varepsilon = 0$) says that $x \in \mathcal{A}(x^*)$. Thus, $x^* \in \mathcal{A}^{-1}(x)$, and, by Lemma 5, $x^* \in C(x)$. □

In the deterministic and the non-parametric cases, we get the next results from Theorem 7.

Corollary 13 *Let $F : X \times Y \to \mathbb{R}_\infty$, $p = F(\cdot, 0_Y)$, and $x \in p^{-1}(\mathbb{R})$. The next statements are equivalent:*
(i) $\partial p(x) = \bigcap\limits_{\eta>0} \operatorname{proj}_{X^*}(\partial^\eta F)(x, 0_Y)$,
(ii) $\min\limits_{z\in X}\left\{F(z, 0_Y) - \langle x^*, x\rangle\right\} = \sup\limits_{y^*\in Y^*} -F^*(x^*, y^*),\ \forall x^* \in \partial p(x)$,
(iii) $\inf\limits_{z\in X}\left\{F(z, 0_Y) - \langle x^*, x\rangle\right\} = \sup\limits_{y^*\in Y^*} -F^*(x^*, y^*),\ \forall x^* \in \partial p(x)$.

Proof Let $F_u = F : X \times Y \to \mathbb{R}_\infty$ and $p = F(\cdot, 0_Y)$. We then have

$$C(x) = \bigcap_{\eta>0} \operatorname{proj}_{X^*}(\partial^\eta F)(x, 0_Y),\ \forall x \in X,$$

(see Corollary 9) and the conclusion follows directly from Theorem 7. □

Corollary 14 *Let $(f_u)_{u\in U} \subset \mathbb{R}^X_\infty$, $p = \sup\limits_{u\in U} f_u$, $x \in p^{-1}(\mathbb{R})$. The next statements are equivalent:*
(i) $\partial\left(\sup\limits_{u\in U} f_u\right)(x) = C(x)$,
(ii) $\max\limits_{z\in X}\left\{\langle x^*, z\rangle - p(z)\right\} = \inf\limits_{u\in U} f_u^*(x^*),\ \forall x^* \in \partial p(x)$,
(iii) $\left(\sup\limits_{u\in U} f_u\right)^*(x^*) = \inf\limits_{u\in U} f_u^*(x^*),\ \forall x^* \in \partial p(x)$,
where

$$C(x) = \bigcap_{\eta>0} \bigcup_{\substack{\varepsilon_1+\varepsilon_2=\eta\\ \varepsilon_1\geqslant 0,\varepsilon_2\geqslant 0}} \bigcup_{u\in I^{\varepsilon_1}(x)} \operatorname{proj}^u_{X^*}(\partial^{\varepsilon_2} f_u)(x), \forall x \in X. \tag{47}$$

Proof Let $F_u(x, y_u) = f_u(x)$. Then it is easy to see that in this case, $C(x)$ can be expressed as in (47). The conclusion now follows from Theorem 7. □

Let us come back to the general case and consider the most simple subdifferential formula one can expect for the robust objective function $p = \sup\limits_{u\in U} F_u(\cdot, 0_u)$:

$$\partial p(x) = \bigcup_{u\in I(x)} \operatorname{proj}^u_{X^*}(\partial F_u)(x, 0_u), \tag{48}$$

where the set of active indexes at x, $I(x)$, is defined by (46).

In Case 3 we have

$$p(x) = \begin{cases} f(x), \text{ if } \quad H_u(x) \in -S_u, \forall u \in U, \\ +\infty, \text{ else,} \end{cases}$$

$I(x) = U$ for each $x \in p^{-1}(\mathbb{R})$, and (48) writes

$$\partial p(x) = \bigcup_{\substack{u\in U,\ z_u^*\in S_u^+ \\ \langle z_u^*, H_u(x)\rangle = 0}} \partial(f + z_u^* \circ H_u)(x),$$

which has been called *Basic Robust Subdifferential Condition* (BRSC) in [8] (see [18, page 307] for the deterministic case). More generally, let us introduce the following terminology:

Definition 1 Given $F_u : X \times Y_u \to \mathbb{R}_\infty$ for each $u \in U$, and $p = \sup\limits_{u\in U} F_u(\cdot, 0_u)$, we will say that **Basic Robust Subdifferential Condition** holds at a point $x \in p^{-1}(\mathbb{R})$ if (48) is satisfied, that is $\partial p(x) = D(x)$.

Recall that, in Example 1, $p\,(x) = \langle c^*, x\rangle + \mathrm{i}_A\,(x)$, where $A = p^{-1}(\mathbb{R})$ is the feasible set of the linear system. So, given $x \in A$, $\partial p(x)$ is the sum of c^* with the normal cone of A at x, i.e., Basic Robust Subdifferential Condition (at x) asserts that such a cone can be expressed in some prescribed way.

Theorem 8 *Let $x \in p^{-1}(\mathbb{R})$. The next statements are equivalent:*
(i) *Basic Robust Subdifferential Condition holds at x,*
(ii) *Min-max robust duality holds at each $x^* \in \partial p(x)$,*
(iii) *Strong robust duality holds at each $x^* \in \partial p(x)$.*

Proof [(i) $\Longrightarrow$ (ii)] Let $x^* \in \partial p(x)$. We have $x^* \in D(x)$ and, by (44), there exist $u \in I(x)$ (i.e., $p(x) = F_u(x, 0_u)$), $y_u^* \in Y_u^*$, such that $(x^*, y_u^*) \in \partial F_u)(x, 0_u)$. Then,

$$\begin{aligned} p^*(x^*) &\geq \langle x^*, x\rangle - p(x) = \langle x^*, x\rangle - F_u(x, 0_u) = F_u^*(x^*, y_u^*) \\ &\geq q(x^*) \geq p^*(x^*). \end{aligned}$$

It follows that

$$\max_{z\in X}\{\langle x^*, z\rangle - p(z)\} = \langle x^*, x\rangle - p(x) = F_u^*(x^*, y_u^*) = q(x^*),$$

and min-max robust duality holds at x^*.

[(ii) $\Longrightarrow$ (iii)] It is obvious.

[(iii) $\Longrightarrow$ (i)] By Lemma 7, it suffices to check that $\partial p(x) \subset D(x)$. Let $x^* \in \partial p(x)$. We have $x \in (Mp)(x^*)$. Since strong robust duality holds at x^*, Theorem 2 says that there exist $u \in U$, $y_u^* \in Y_u^*$ such that $x \in B^0_{(u,y_u^*)}(x^*)$, that means (see (34))

$$(x, 0_u) \in (MF_u)(x^*, y_u^*),\ (x^*, y_u^*) \in (\partial F_u)(x, 0_u),$$

and by (44), $x^* \in D(x)$. □

As usual, Theorem 8 gives us corresponding results for the two extreme cases: non-uncertainty and non-perturbation cases.

Corollary 15 *Let $F : X \times Y \to \mathbb{R}_\infty$, $p = F(\cdot, 0_Y)$, and $x \in p^{-1}(\mathbb{R})$. The next statements are equivalent:*

(i) $\partial p(x) = \text{proj}_{X^*}(\partial F)(x, 0_Y)$,

(ii) $\max\limits_{z\in X} \Big\{\langle x^*, z\rangle - F(z, 0_Y)\Big\} = \min\limits_{y^*\in Y^*} F^*(x^*, y^*),\ \forall x^* \in \partial p(x)$,

(iii) $p^*(x^*) = \min\limits_{y^*\in Y^*} F^*(x^*, y^*),\ \forall x^* \in \partial p(x)$.

Proof In this case we have, by (41), $D(x) = \text{proj}_{X^*}(\partial F)(x, 0_Y)$ and the conclusion is a direct consequence of Theorem 8. □

Corollary 16 *Let $(f_u)_{u\in U} \subset \mathbb{R}_\infty^X$, $p = \sup\limits_{u\in U} f_u$, $x \in p^{-1}(\mathbb{R})$. The next statements are equivalent:*

(i) $\partial p(x) = \bigcup\limits_{u\in I(x)} \partial f_u(x)$,

(ii) $\max\limits_{z\in X} \Big\{\langle x^*, z\rangle - p(z)\Big\} = \min\limits_{u\in U} f_u^*(x^*),\ \forall x^* \in \partial p(x)$,

(iii) $(\sup\limits_{u\in U} f_u)^*(x^*) = \min\limits_{y^*\in Y^*} f_u^*(x^*),\ \forall x^* \in \partial p(x)$.

Proof In this non-parametric case, let $F_u(x, y_u) = f_u(x)$, $p = \sup\limits_{u\in U} f_u$. We have

$$D(x) = \bigcup_{u \in I(x)} \partial f_u(x), \; I(x) = \{u \in U \; : \; f_u(x) = p(x) \in \mathbb{R}\}$$

and Theorem 8 applies. □

Acknowledgements This research was supported by the Vietnam National University - HCM city, Vietnam, project B2019-28-02, by PGC2018-097960-B-C22 of the Ministerio de Ciencia, Innovación y Universidades (MCIU), the Agencia Estatal de Investigación (AEI), and the European Regional Development Fund (ERDF), and by the Australian Research Council, Project DP180100602.

References

1. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton (2009)
2. Bertsimas, D., Sim, M.: Tractable approximations to robust conic optimization problems. Math. Program. **107B**, 5–36 (2006)
3. Borwein, J.M.: A strong duality theorem for the minimum of a family of convex programs. J. Optim. Theory Appl. **31**, 453–472 (1980)
4. Borwein, J.M., Burachik, R.S., Yao, L.: Conditions for zero duality gap in convex programming. J. Nonlinear Convex Anal. **15**, 167–190 (2014)
5. Borwein, J.M., Lewis, A.S.: Partially finite convex programming. I. Quasi relative interiors and duality theory. Math. Program. **57B**, 15-48 (1992)
6. Borwein, J.M., Lewis, A.S.: Practical conditions for Fenchel duality in infinite dimensions. In: Fixed Point Theory and Applications, pp. 83–89, Pitman Research Notes in Mathematics Series, vol. 252, Longman Scientific and Technology, Harlow (1991)
7. Bot, R.I.: Conjugate Duality in Convex Optimization. Springer, Berlin (2010)
8. Bot, R.I., Jeyakumar, V., Li, G.Y.: Robust duality in parametric convex optimization. Set-Valued Var. Anal. **21**, 177–189 (2013)
9. Burachick, R.S., Jeyakumar, V., Wu, Z.-Y.: Necessary and sufficient condition for stable conjugate duality. Nonlinear Anal. **64**, 1998–2006 (2006)
10. Chu, Y.C.: Generalization of some fundamental theorems on linear inequalities. Acta Math Sinica **16**, 25–40 (1966)
11. Dinh, N., Goberna, M.A., López, M.A., Volle, M.: A unifying approach to robust convex infinite optimization duality. J. Optim. Theory Appl. **174**, 650–685 (2017)
12. Dinh, N., Mo, T.H., Vallet, G., Volle, M.: A unified approach to robust Farkas-type results with applications to robust optimization problems. SIAM J. Optim. **27**, 1075–1101 (2017)
13. Dinh, N., Long, D.H.: Complete characterizations of robust strong duality for robust optimization problems. Vietnam J. Math. **46**, 293–328 (2018). https://doi.org/10.1007/s10013-018-0283-1
14. Goberna, M.A., López, M.A.: Linear Semi-Infinite Optimization. Wiley, Chichester (1998)
15. Goberna, M.A., López, M.A.: Recent contributions to linear semi-infinite optimization. 4OR-Q. J. Oper. Res. **15,** 221–264 (2017)
16. Goberna, M.A., López, M.A., Volle, M.: Primal attainment in convex infinite optimization duality. J. Convex Anal. **21**, 1043–1064 (2014)
17. Grad, S.-M.: Closedness type regularity conditions in convex optimization and beyond. Front. Appl. Math. Stat. **16**, September (2016). https://doi.org/10.3389/fams.2016.00014
18. Hiriart-Urruty, J.-B., Lemarechal, C.: Convex Analysis and Minimization Algoritms I. Springer, Berlin (1993)

19. Hiriart-Urruty, J.-B., Moussaoui, M., Seeger, A., Volle, M.: Subdifferential calculus without qualification conditions, using approximate subdifferentials: a survey. Nonlinear Anal. **24**, 1727–1754 (1995)
20. Jeyakumar, V., Li, G.Y.: Stable zero duality gaps in convex programming: complete dual characterisations with applications to semidefinite programs. J. Math. Anal. Appl. **360**, 156–167 (2009)
21. Li, G.Y., Jeyakuma, V., Lee, G.M.: Robust conjugate duality for convex optimization under uncertainty with application to data classification. Nonlinear Anal. **74**, 2327–2341 (2011)
22. Lindsey, M., Rubinstein, Y.A.: Optimal transport via a Monge-Ampère optimization problem. SIAM J. Math. Anal. **49**, 3073–3124 (2017)
23. López, M.A., Still, G.: Semi-infinite programming. European J. Oper. Res. **180**, 491–518 (2007)
24. Rockafellar, R.T.: Conjugate Duality and Optimization. CBMS Lecture Notes Series No. 162. SIAM, Philadelphia (1974)
25. Vera, J.R.: Geometric measures of convex sets and bounds on problem sensitivity and robustness for conic linear optimization. Math. Program. **147A**, 47–79 (2014)
26. Volle, M.: Caculus rules for global approximate minima and applications to approximate subdifferential calculus. J. Global Optim. **5**, 131–157 (1994)
27. Wang, Y., Shi, R., Shi, J.: Duality and robust duality for special nonconvex homogeneous quadratic programming under certainty and uncertainty environment. J. Global Optim. **62**, 643–659 (2015)
28. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific, River Edge (2002)

Comparing Averaged Relaxed Cutters and Projection Methods: Theory and Examples

Reinier Díaz Millán, Scott B. Lindstrom and Vera Roshchina

1 Introduction

Projection and reflection methods are used for solving the *feasibility* problem of finding a point in the intersection of a finite collection of closed, convex sets in a Hilbert space. Such problems have a wide range of applications in variational analysis, optimisation, physics and mathematics in general. One of the most successful methods from this class is the Douglas–Rachford method that uses a combination of reflections and averaging on each iteration. The idea first appeared in [35] as a numerical scheme for solving differential equations, and the convergence of a more general scheme for finding a zero of the sum of two maximally monotone operators was framed in [45] (also see [13, Chapter 26] for a modern treatment). More recently, a modification of the Douglas–Rachford method for finding closest feasible points has been introduced [7].

Convergence rates for such methods are the subject of extensive research; we provide a brief sampling. Under appropriate conditions, the Douglas–Rachford method converges in finitely many steps [13]. Convergence rates may frequently be obtained through analysis of regularity conditions [40]. Additionally, semi-algebraic structure admits further bounds on convergence rates for projection methods more generally

R. D. Millán
Federal Institute of Goiás, Goiás, Brazil
e-mail: rdiazmillan@gmail.com

S. B. Lindstrom (✉)
School of Mathematical and Physical Sciences, Centre for Computer-assisted Research Mathematics and its Applications (CARMA), The University of Newcastle, Callaghan, NSW, Australia
e-mail: scott.lindstrom@uon.edu.au

V. Roshchina
UNSW Sydney, Sydney, NSW, Australia
e-mail: v.roshchina@unsw.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_5

[21, 22, 36] and for the Douglas–Rachford method in particular [42]. For a recent survey on the Douglas–Rachford method, see [43].

The idea of replacing projections with their approximations, and specifically with the approximations constructed from the subdifferentials of the convex functions that describe the sets, was introduced in [38]. It has been used in various contexts recently, including the numerical solution of variational inequalities; see, for example, [17, 18]. In particular, relaxation parameters together with subgradient projections have been used in the construction of the *extrapolation method of parallel subgradient projections* (EMOPSP) algorithm for image recovery [29]. Of particular relevance are the following works: [1, 10, 25, 26, 30]. Books which contain useful information about general cutters (see Definition 1) include [25, 28, 48].

The subgradient projector, in particular, is quite well studied; early contributions include the foundational work on subgradient projections [48] and the analysis of the cyclic version [27]. Characterizations of finite convergence are provided in [15], and a systematic study of the subgradient projector in [14, 16]. This sampling of the literature on subgradient projections is far from exhaustive; the interested reader is referred to the literature referenced in the latter works.

We consider the two-set feasibility problem of finding

$$u \in A \cap B, \tag{1}$$

for closed, convex subsets A and B of a Hilbert space $\mathcal{H}$. In particular, we consider the behaviour of the dynamical systems which arise from iterated application of an operator T that is a weighted average of the identity map and the composition of two relaxed cutters for the two sets in question. The Douglas–Rachford method (reflect-reflect-average), the Peaceman–Rachford method, alternating projections and relaxed-reflect-reflect (RRR) are all special cases.

The *goal* of the present work is threefold:

1. We compare and contrast what is true of such operators in the special case where the cutters are projections (onto the constraint sets) with the more general case of cutters.
2. In particular, we discuss nonexpansivity in the former setting and quasi-nonexpansivity in the latter, analysing what may be shown through each.
3. We illustrate with examples, and we provide simple geometric arguments throughout the exposition.

We would also like to highlight the recent work of Jonathan Borwein and his collaborators, who successfully applied the Douglas–Rachford method to a range of large-scale *non-convex* problems and studied its convergence [3–5, 23, 24]. In Borwein's chapter of *Tools and Mathematics: Instruments for Learning* [20], he included, along with his own commentary, a quote he particularly liked:

> Long before current graphic, visualisation and geometric tools were available, John E. Littlewood, 1885-1977, wrote in his delightful Miscellany:

> A heavy warning used to be given [by lecturers] that pictures are not rigorous; this has never had its bluff called and has permanently frightened its victims into playing for safety. Some pictures, of course, are not rigorous, but I should say most are (and I use them whenever possible myself). [46, p. 53]

In this spirit, we present our results in a tutorial form, complete with many pictures and examples that highlight the geometric intuition underpinning them.

Outline

We introduce the main concepts in Section 2. In Section 3, we provide a simple proof of convergence to a feasible point for a method which averages the composition of two relaxed cutters with the identity. The parameterized method recovers alternating projections as one special case and the Douglas–Rachford method as a limiting, but not allowable, case. This comes as no surprise since projections onto constraint sets are a special case of cutters, and examples where the Douglas–Rachford method converges to fixed points which are not also feasible points are well known.

With projections onto the constraint sets, the fixed points of the Douglas–Rachford operator have the handy property that they may be used to find feasible points in a single step. In Examples 4 and 5, we show this may fail when projections onto the constraint sets are replaced with more general cutters. The elegance of this pairing is that the geometry illustrates why the proof fails if the limiting parameters are allowed, and the examples showcase what can then go wrong.

In Section 4, we provide several examples of implementations of the Douglas–Rachford method with cutters.

2 Background and preliminaries

Let A and B be two closed convex sets in a Hilbert space $\mathcal{H}$. Given a starting point $x_0 \in \mathcal{H}$, the classic method of alternating projections generates the sequence of points $\{x_n\}_{n\in\mathbb{N}}$, where

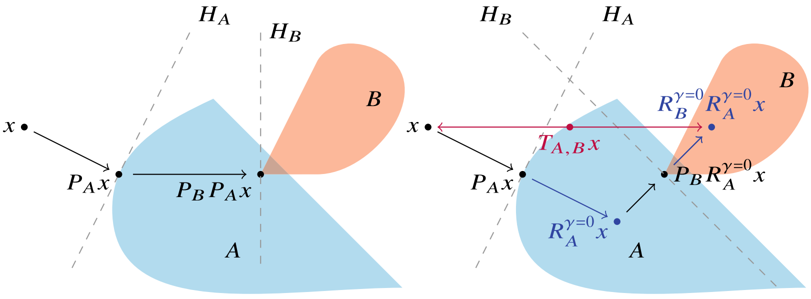
$$x_n := (P_B \circ P_A)^n x_0 \quad \forall n \in \mathbb{N}, \tag{2}$$

and by P_S we denote the Euclidean projection operator onto a closed convex set $S \subset \mathcal{H}$,

$$P_S(x) = \operatorname*{argmin}_{s\in S} \|s - x\|,$$

which is well defined (and single valued) for S closed, convex and non-empty. We assume these properties for all of our sets throughout. Observe (see Figure 1a) that each iteration of the method is the composition of projections onto the hyperplanes H_A and H_B that support the sets A and B at $P_A(x)$ and $P_B(x)$, respectively.

On each step of the classic Douglas–Rachford algorithm the previous iterate is first *reflected* through H_A, then reflected through H_B, and finally the resulting point is



(a) One step of alternating projections (b) One step of Douglas–Rachford method

Fig. 1 The operator $T^{\lambda}_{A^{\gamma},B^{\gamma}}$ for different values of γ, λ

averaged with the previous iterate; see Figure 1b. In this case, our iterative sequence $\{x_n\}_{n\in\mathbb{N}}$ is defined as

$$x_n := \left(\frac{1}{2}\left((2P_B - \mathrm{Id}) \circ (2P_A - \mathrm{Id})\right) + \frac{1}{2}\,\mathrm{Id}\right)^n x_0 \quad \forall n \in \mathbb{N}. \tag{3}$$

The reflection can be replaced by a *relaxed projection* which we denote by R_S^{γ}. For a fixed *reflection parameter* $\gamma \in [\,0, 2\,[$ we let

$$R_S^{\gamma} := (2-\gamma)(P_S - \mathrm{Id}) + \mathrm{Id}\,. \tag{4}$$

Observe that when $\gamma = 0$, the operator $R_S^{\gamma=0} = 2P_S - \mathrm{Id}$ is the standard *reflection* that we saw earlier, for $\gamma = 1$ we obtain the *projection*, $R_S^{\gamma=1} = P_S$. For $\gamma \in \,]\,1, 2\,[$ the operator R_S^{γ} can be called an *under-relaxed projection* following [34]. For $\gamma \in \,]\,0, 1\,[$ it may be called an *over-relaxed projection*.

In addition to using relaxed projections as in (4), the averaging step of the Douglas–Rachford iteration (3) can also be relaxed by choosing an arbitrary point on the interval between the second reflection and the initial iterate. This can be parameterized by some $\lambda \in \,]\,0, 1\,]$. We can hence define a λ-averaged relaxed sequence $\{x_n\}_{n\in\mathbb{N}}$ by

$$\begin{aligned} x_n &:= \left(T^{\lambda}_{A^{\gamma},B^{\mu}}\right)^n x_0, \\ \text{where} \quad T^{\lambda}_{A^{\gamma},B^{\mu}} &:= \lambda(R_B^{\mu} \circ R_A^{\gamma}) + (1-\lambda)\,\mathrm{Id}\,. \end{aligned} \tag{5}$$

When $\lambda = 1$ and $\gamma = \mu = 1$, this is the sequence generated by alternating projections (2). For $\gamma = \mu = 0$, this is the Douglas–Rachford method (3), and for $\lambda = 1$ the Peaceman–Rachford method. The case where $\gamma = \mu = 0$ and λ is flexible is often referred to as *relaxed-reflect-reflect* or RRR [37]. If $\gamma = \frac{2(\eta+1)}{2\eta+1}$, then

$$R_S^\gamma = \left(\frac{1}{2\eta+1} \,\mathrm{Id} + \frac{2\eta}{2\eta+1} P_S \right)$$

may be recognised as the form in which the relaxation was presented for the damped Douglas–Rachford variant in [21].

We note that the framework introduced here does not cover all possible projection methods. For example, one may want to vary the parameters γ, μ and λ on every step, or consider other variations of Douglas–Rachford-like operators (e.g. see [7]).

We recall the definition of a cutter (see [25, Definition 2.1.30]).

Definition 1 Let $x \in \mathcal{H}$, we say that y separates a subset $S \subseteq \mathcal{H}$ from x if $\langle x - y, z - y \rangle \leq 0$ for all $z \in S$. We say that an operator $T : \mathcal{H} \to \mathcal{H}$ is a *cutter* if for all $x \in \mathcal{H}$, Tx separates $\operatorname{Fix} T \neq \emptyset$ from x. In other words,

$$(\forall x \in \mathcal{H})\, (\forall z \in \operatorname{Fix} T) \quad \langle x - Tx, z - Tx \rangle \leq 0. \tag{6}$$

For the sake of simplicity, given a closed and convex set S, we denote by $\mathcal{P}_S$ a cutter if $S := \operatorname{Fix} \mathcal{P}_S$.

A cutter may be thought of as a map that assigns x to its projection onto a chosen separating hyperplane, as illustrated in Figure 2b. The Euclidean projection operator P_S for a closed, convex set S is an example of a cutter where the separating hyperplane is a supporting hyperplane to S, as illustrated in Figure 1 for alternating projections at left and the Douglas–Rachford method at right. We note that (6) is essential for cutter-based projection methods, and that for $x \in S$, $\mathcal{P}_S(x) = x$. We have the following elementary example that illustrates this.

Example 1 (Example of a cutter) In the one-dimensional real setting, assume that $S = \,]-\infty, 0\,]$ and

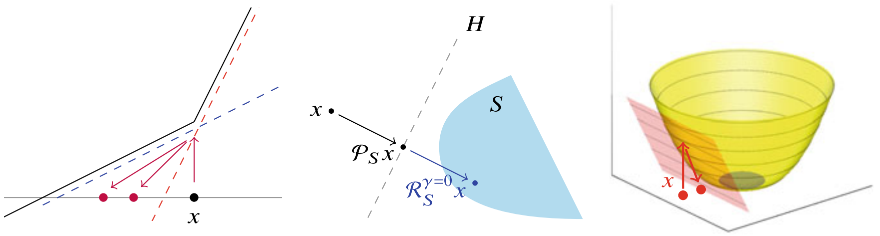
$$T(x) = \begin{cases} x, & x \in \,]-\infty, 0\,], \\ 0, & x \in \,]0, 1\,[\,, \\ 1, & x \in [\,1, +\infty\,[\,. \end{cases}$$

Observe that $y = T(x)$ is a separator, however, it is not a cutter: the point $x = 1 \notin S$ is a fixed point of T, and for $x \in \,]0, 1\,[$, the point $T(x) = 0$ does not separate the fixed points of T from x. ♢

A useful implementation of a cutter is the subgradient projection operator for a convex function f, which we recall in the following definition from [16, Definition 2.2], where ∂f denotes the usual Moreau–Rockafellar subdifferential of f.

Definition 2 Let $f : \mathcal{H} \to \mathbb{R}$ be lower semi-continuous and subdifferentiable. Let $s : \mathcal{H} \to \mathcal{H}$ be a selection for ∂f. Then the *subgradient projector* of f is

$$P_{\partial f} : \mathcal{H} \to \mathcal{H} : x \mapsto \begin{cases} x - \frac{f(x)}{\|s(x)\|^2} s(x) & \text{if } f(x) > 0; \\ x & \text{otherwise.} \end{cases} \tag{7}$$



(a) Subgradient projection (b) Cutter reflection (c) Subgradient projection

Fig. 2 Subgradient projections are cutters

The subgradient projection operator is a cutter with Fix $P_{\partial f} = \mathrm{lev}_{\leq 0} f$. We illustrate in Figure 2. In Figure 2c, we show the case where the selection operator s is uniquely determined since ∂f is single-valued everywhere. In Figure 2a, we show two possible values for the subgradient projection of x; we emphasise that the subgradient projector a is *single-valued operator*, and that the output depends on the chosen selection operator s in Definition 2.

In the case where projections onto the sets cannot be computed (or computing them exactly is undesirable), it makes sense to consider operators of the form (5) where the projections are replaced with subgradient projections or other kinds of cutters.

We will refer to all such discussed methods and their combination as *cutter methods* and use the notation

$$\mathcal{T}^{\lambda}_{A^{\gamma},B^{\mu}} := \lambda(\mathcal{R}^{\mu}_{B} \circ \mathcal{R}^{\gamma}_{A}) + (1-\lambda)\,\mathrm{Id},$$

where

$$\mathcal{R}^{\gamma}_{A} := (2-\gamma)(\mathcal{P}_A - \mathrm{Id}) + \mathrm{Id}, \quad \mathcal{R}^{\mu}_{B} := (2-\mu)(\mathcal{P}_B - \mathrm{Id}) + \mathrm{Id}$$

are the relaxed versions of the cutters $\mathcal{P}_A$ and $\mathcal{P}_B$, which may be projections onto the constraint sets or more general cutters, depending on the context.

In the case of subgradient projections, we will slightly abuse the notation and let

$$\mathcal{T}^{\lambda}_{f^{\gamma},g^{\mu}} := \mathcal{T}^{\lambda}_{(\mathrm{lev}_{\leq 0} f)^{\gamma},(\mathrm{lev}_{\leq 0} g)^{\mu}},$$

with cutters implemented via the subgradient projections (7).

Notice that if for some closed convex set S, we let $f := \mathrm{d}(\cdot, S)$ be the distance function for the set S given by

$$\mathrm{d}(x, S) = \min_{y \in S} \|x - y\|,$$

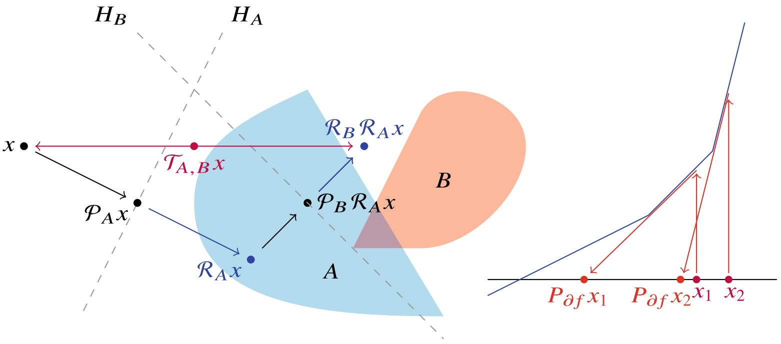


Fig. 3 An averaged relaxed cutter $\mathcal{T}^{\lambda}_{A^{\gamma},B^{\mu}}$ may not be nonexpansive

then P_S and P_f coincide. We will mainly focus on averaged cutter relaxations $\mathcal{T}^{\lambda}_{A^{\gamma},B^{\mu}}$, for which an example is shown in Figure 3, and will elaborate on the functional implementation in Section 4.

Let $\mathcal{A} = N_A$ and $\mathcal{B} = N_B$ be the normal cone operators for closed convex sets A and B. Then the resolvents $J^{\lambda}_{\mathcal{A}}$, $J^{\lambda}_{\mathcal{B}}$ (defined as $J^{\lambda}_{F} = (\mathrm{Id} + \lambda F)^{-1}$ for some set-valued mapping F) are the projection operators P_A, P_B, respectively, $T_{\mathcal{A},\mathcal{B}} = \frac{1}{2}R^{\gamma=0}_{B}R^{\gamma=0}_{A} + \frac{1}{2}\,\mathrm{Id}$ is what we recognise as the Douglas–Rachford method, and $J^{\lambda}_{\mathcal{A}}v = \mathrm{P}_A\, v \in A \cap B$ is a solution for the feasibility problem.

We quote the following key result from [45] that applies to a more general setting of maximal monotone operators.

Theorem 1 (**Lions & Mercier**) *Assume that $\mathcal{A}$, $\mathcal{B}$ are maximal monotone operators and $\mathcal{A}+\mathcal{B}$ is maximal monotone. Then for*

$$T_{\mathcal{A},\mathcal{B}} : \mathcal{H} \to \mathcal{H} \text{ by } x \mapsto J^{\lambda}_{\mathcal{B}}(2J^{\lambda}_{\mathcal{A}} - \mathrm{Id})x + (\mathrm{Id} - J^{\lambda}_{\mathcal{A}})x, \tag{8}$$

the sequence given by $x_{n+1} = T_{\mathcal{A},\mathcal{B}}x_n$ converges weakly to some $v \in \mathcal{H}$ as $n \to \infty$ such that $J^{\lambda}_{\mathcal{A}}v$ is a zero of $\mathcal{A}+\mathcal{B}$.

The authors of [12] showed that in the case of the feasibility problem (1) the requirement $\mathcal{A}+\mathcal{B}$ maximal monotone may be relaxed, a relaxation later made more general in [51]. See also [12, Theorem 26.11]. Both results rely on the firm nonexpansivity of $T_{\mathcal{A},\mathcal{B}}$, an immediate consequence of the fact that $R^{\gamma=0}_{\mathcal{B}}R^{\gamma=0}_{\mathcal{A}}$ is *nonexpansive* and so $T_{\mathcal{A},\mathcal{B}}$ is $1/2$-averaged. We define this term and several others which we summarise in the following definition (see [11, Def 4.1], [26, Def 2.2], and [25, Def 2.1.19] for more details).

Definition 3 (**Properties of operators**) Let $D \subset \mathcal{H}$ be non-empty and let $T : D \to \mathcal{H}$. Assume that $\mathrm{Fix}\, T := \{x \in \mathcal{H} \mid Tx = x\} \neq \emptyset$. Then T is

firmly nonexpansive if

$$\|T(x)-T(y)\|^2+\|(\mathrm{Id}-T)(x)-(\mathrm{Id}-T)(y)\|^2 \le \|x-y\|^2 \quad \forall x\in D,\quad \forall y\in D;$$

nonexpansive if it is Lipschitz continuous with constant 1,

$$\|T(x)-T(y)\| \le \|x-y\| \quad \forall x\in D,\quad \forall y\in D;$$

quasinonexpansive if $\|T(x)-y\| \le \|x-y\| \quad \forall x\in D, \quad \forall y\in \operatorname{Fix} T$ (an operator that is both quasinonexpansive and continuous is called paracontracting); **strictly quasinonexpansive** if

$$\|T(x)-y\| < \|x-y\| \quad \forall x\in D\setminus \operatorname{Fix} T,\quad \forall y\in \operatorname{Fix} T;$$

ρ**-strongly quasinonexpansive** for $\rho>0$ if

$$\|Tx-y\|^2 \le \|x-y\|^2-\rho\|Tx-x\|^2 \quad \forall x\in D\setminus \operatorname{Fix} T,\quad \forall y\in \operatorname{Fix} T.$$

We are focused on the feasible setting, so we can safely assume that for all operators T considered in the paper $\operatorname{Fix} T\neq\emptyset$. As soon as one moves from the setting of projections into the setting of more general cutters, the (firmly) nonexpansive property of $\mathcal{T}^{\lambda}_{A^{\gamma},B^{\mu}}$ may be lost, as illustrated in the following simple example.

Example 2 (Loss of nonexpansivity when using cutters) Define $f:\mathbb{R}\to\mathbb{R}$ by

$$f: x\mapsto \begin{cases} |x| & x\le 1,\\ 2x-1 & \text{otherwise.}\end{cases} \tag{9}$$

Then the subgradient cutter $P_{\partial f}:\mathbb{R}\to\mathbb{R}$ for the level set, $\operatorname{lev}_{\le 0} f$ is

$$P_{\partial f}: x\mapsto \begin{cases} 0 & x<1,\\ \frac{1}{2} & x>1,\\ \text{some } u\in[0,1/2] & x=1.\end{cases} \tag{10}$$

Observe that $P_{\partial f}$ is not nonexpansive for any choice of $x\in\,]0,1[\,,\, y\in\,]1,2[$ satisfying $|x-y|<\frac{1}{2}$. A similar polyhedral example is shown at right in Figure 3.
◊

Strong quasi-nonexpansivity is a less restrictive property that yields the desired convergence, though under a slightly more restrictive parameter scheme.

Definition 4 (**Fejér monotonicity**) A sequence $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to a closed convex set C if

$$\|x_{n+1} - x\| \leq \|x_n - x\| \qquad \forall x \in C, \quad \forall n \in \mathbb{N}.$$

A Fejér monotone sequence with respect to a closed convex set C may be thought of as a sequence defined by $x_n := T^n x_0$ where T is quasinonexpansive with respect to $C = \operatorname{Fix} T$. Note that a Fejér monotone sequence with respect to a non-empty set is always bounded.

We have the following well known convergence result (see [11, Theorem 5.11]).

Theorem 2 *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ and let C be a non-empty closed convex subset of $\mathcal{H}$. Suppose that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C. Then the following are equivalent:*

1. *the sequence $(x_n)_{n\in\mathbb{N}}$ converges strongly (i.e. in norm) to a point in C;*
2. *$(x_n)_{n\in\mathbb{N}}$ possesses a strong sequential cluster point in C; and*
3. *$\liminf\limits_{n\to\infty} \mathrm{d}(x_n, C) = 0$.*

3 Convergence of Projection Methods

In the following theorem, (i) is a known consequence of [25, Corollary 3.7.1(i)]. However, we provide a new proof which relies on simple geometry. We will then go on to analyse convergence for $\mathcal{T}^{\lambda}_{A^{\gamma},B^{\mu}}$, and the details of our proof will illustrate why for averaged cutter relaxation methods we may lose convergence in the case of $\gamma = 0$.

Theorem 3 *Let A be a closed convex set in a Hilbert space $\mathcal{H}$, and let $\mathcal{P}_A$ be a cutter. Then the following hold*

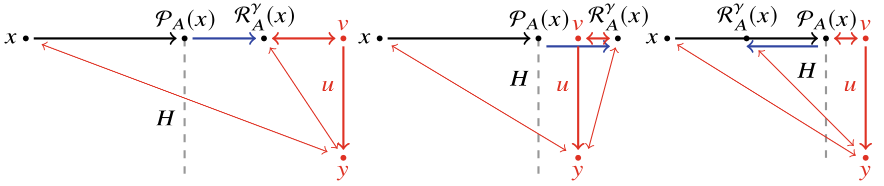
(i) $\|\mathcal{R}^{\gamma}_A(x) - y\|^2 \leq \gamma(\gamma - 2)\|x - \mathcal{P}_A(x)\|^2 + \|x - y\|^2 \qquad \forall y \in A \quad \forall x \in \mathcal{H}$*;*
(ii) *$\mathcal{R}^{\gamma}_A$ is $\gamma/(2-\gamma)$-strongly quasinonexpansive; and*
(iii) *$\mathcal{R}^{\gamma}_A$ is strictly quasinonexpansive for $\gamma \in \,]0, 2[$.*

Proof If $x \in A$, then $\mathcal{R}^{\gamma}_A(x) = \mathcal{P}_A(x) = x$ and the proof of (i) is trivial. Consider the case when $x \notin A$. Without loss of generality, we can assume that $x = 0$. Indeed, it is evident that for the affine change of variable $u' = u - x$ the induced mapping $\mathcal{P}_{A'}(u') = \mathcal{P}_{A-x}(u - x)$ is again a cutter for $A' = A - x$, and the relation (i) can be restated in terms of A' and $\mathcal{P}_{A'}$; this is also clear from the geometry illustrated in Figure 4.

Fix $y \in A$. We have $y := v + u$ where $v \in \operatorname{span}\{\mathcal{P}_A(x)\}$, $u \in \operatorname{span}\{\mathcal{P}_A(x)\}^{\perp}$. We will first show that

$$\|\mathcal{R}^{\gamma}_A(x) - v\| \leq \|v\| - \min\{\gamma, 2-\gamma\}\|\mathcal{P}_A(x)\|. \tag{11}$$

Here Figure 4 is most instructive, both for understanding this inequality and motivating its proof.



(a) Case 1

(b) Case 2

(c) Case 3

Fig. 4 Illustrations of the inequality (11) in the proof of Theorem 3

Since $\mathcal{P}_A$ is a cutter, we have

$$\langle y, \mathcal{P}_A(x)\rangle \geq \|\mathcal{P}_A(x)\|^2 \quad \forall\, y \in A.$$

Furthermore, we have $v = \beta \mathcal{P}_A(x)$, hence

$$\beta\|\mathcal{P}_A(x)\|^2 = \langle v, \mathcal{P}_A(x)\rangle = \langle y - u, \mathcal{P}_A(x)\rangle = \langle y, \mathcal{P}_A(x)\rangle \geq \|\mathcal{P}_A(x)\|^2,$$

which yields $\beta \geq 1$ (observe that $\|\mathcal{P}_A(x)\|^2 > 0$ since $x = 0 \notin A$). Now

$$\|\mathcal{R}_A^\gamma(x) - v\| = \|(2-\gamma)\mathcal{P}_A(x) - \beta\mathcal{P}_A(x)\| = |2-\gamma-\beta|\|\mathcal{P}_A(x)\|. \tag{12}$$

Observe that

$$\begin{aligned} |2-\gamma-\beta| &= \max\{2-\gamma-\beta, \beta+\gamma-2\} \\ &= \beta + \max\{2\underbrace{(1-\beta)}_{\leq 0}-\gamma, \gamma-2\} \\ &\leq \beta + \max\{-\gamma, \gamma-2\} \\ &= \beta - \min\{\gamma, 2-\gamma\}, \end{aligned}$$

hence we have (11). For convenience, let

$$\psi : [0,2) \to [0,1] \text{ defined by } \gamma \mapsto \min\{\gamma, 2-\gamma\} = \begin{cases} \gamma & \text{if } \gamma \in [0,1], \\ 2-\gamma & \text{if } \gamma \in (1,2). \end{cases} \tag{13}$$

Having shown that (11) is true, the Pythagorean theorem yields

$$\|\mathcal{R}_A^\gamma(x) - v\|^2 = \|y - \mathcal{R}_A^\gamma(x)\|^2 - \|u\|^2. \tag{14}$$

Together (14) and (11) yield

$$\|y - \mathcal{R}_A^\gamma(x)\|^2 \leq (\|v\| - \psi(\gamma)\|\mathcal{P}_A(x)\|)^2 + \|u\|^2. \tag{15}$$

Now the Pythagorean theorem also yields

$$\|u\|^2 = \|y\|^2 - \|v\|^2. \tag{16}$$

Equations (15) and (16) together yield

$$\begin{aligned}\|y - \mathcal{R}_A^\gamma(x)\|^2 &\leq (\|v\| - \psi(\gamma)\|\mathcal{P}_A(x)\|)^2 + \|y\|^2 - \|v\|^2\\ &= -2\psi(\gamma)\|\mathcal{P}_A(x)\| \cdot \|v\| + \psi(\gamma)^2\|\mathcal{P}_A(x)\|^2 + \|y\|^2.\end{aligned} \tag{17}$$

Now since $\|\mathcal{P}_A(x)\| \leq \|v\|$,

$$-2\psi(\gamma)\|\mathcal{P}_A(x)\|^2 \geq -2\psi(\gamma)\|\mathcal{P}_A(x)\| \cdot \|v\|. \tag{18}$$

Now (17) and (18) together yield

$$\begin{aligned}\|y - \mathcal{R}_A^\gamma(x)\|^2 &\leq -2\psi(\gamma)\|\mathcal{P}_A(x)\|^2 + \psi(\gamma)^2\|\mathcal{P}_A(x)\|^2 + \|y\|^2\\ &= \psi(\gamma)(\psi(\gamma) - 2)\|\mathcal{P}_A(x)\|^2 + \|y\|^2\\ &= \gamma(\gamma - 2)\|\mathcal{P}_A(x)\|^2 + \|y\|^2,\end{aligned} \tag{19}$$

where the final equality comes from the fact that $\psi(\gamma)(\psi(\gamma) - 2) = \gamma(\gamma - 2)$. This shows (i). Now since $\gamma \in [\,0, 2\,[$, we have that $\gamma(\gamma - 2) \leq 0$. Combining with the fact that $\|\mathcal{R}_A^\gamma(x)\| = (2 - \gamma)\|\mathcal{P}_A(x)\|$, we have from (19) that

$$\begin{aligned}\|y - \mathcal{R}_A^\gamma(x)\|^2 &\leq \gamma(\gamma - 2)\left(\frac{\|\mathcal{R}_A^\gamma(x)\|}{2 - \gamma}\right)^2 + \|y\|^2\\ &= -\frac{\gamma}{2 - \gamma}\|\mathcal{R}_A^\gamma(x)\|^2 + \|y\|^2,\end{aligned}$$

which shows (ii).

If we have $\gamma \in \,]0, 2\,[$, $x \notin A$, and $(\forall x \notin A)\, \mathcal{P}_A(x) \neq x$, then $\gamma(\gamma - 2)$ $\|\mathcal{P}_A(x)\|^2 < 0$ strictly and so

$$\|y - \mathcal{R}_A^\gamma(x)\|^2 \leq \gamma(\gamma - 2)\|\mathcal{P}_A(x)\|^2 + \|y\|^2 < \|y\|^2.$$

This shows (iii). □

Theorem 4 *The following hold*

(i) $\mathcal{T}^\lambda_{A^\gamma, B^\mu}$ *is quasinonexpansive and*
(ii) *if* $\mu, \gamma \in (0, 2)$ *then* $\mathcal{T}^\lambda_{A^\gamma, B^\mu}$ *is strictly quasinonexpansive and*

$$\lim_{n\to\infty} \|x_n - \mathcal{P}_A(x_n)\| = \lim_{n\to\infty} \|x_n - \mathcal{P}_B\mathcal{R}_A^\gamma(x_n)\| = 0.$$

Proof Fix $y \in A \cap B$. For any $x \in \mathcal{H}$, we have from Theorem 3:

$$\|\mathcal{R}_A^{\gamma}(x) - y\|^2 \leq \gamma(\gamma - 2)\|x - \mathcal{P}_A(x)\|^2 + \|x - y\|^2$$
$$\text{and} \quad \|\mathcal{R}_B^{\mu}\mathcal{R}_A^{\gamma}(x) - y\|^2 \leq \mu(\mu - 2)\|\mathcal{P}_B\mathcal{R}_A^{\gamma}(x) - \mathcal{R}_A^{\gamma}(x)\|^2 + \|\mathcal{R}_A^{\gamma}(x) - y\|^2.$$

Combining these two inequalities yields

$$\begin{aligned}&\|\mathcal{R}_B^{\mu}\mathcal{R}_A^{\gamma}(x) - y\|^2 \leq \theta(x) + \|x - y\|^2 \\ &\quad \text{where} \quad \theta(x) = \mu(\mu - 2)\|\mathcal{P}_B\mathcal{R}_A^{\gamma}(x) - \mathcal{R}_A^{\gamma}(x)\|^2 + \gamma(\gamma - 2)\|x - \mathcal{P}_A(x)\|^2.\end{aligned} \tag{20}$$

By convexity of $\| \cdot - y\|^2$,

$$\begin{aligned}\|\mathcal{T}_{A^{\gamma},B^{\mu}}^{\lambda}(x) - y\|^2 &= \| \left(\lambda\mathcal{R}_B^{\mu}\mathcal{R}_A^{\gamma}(x) + (1 - \lambda)x\right) - y\|^2 \\ &\leq \lambda\|\mathcal{R}_B^{\mu}\mathcal{R}_A^{\gamma}(x) - y\|^2 + (1 - \lambda)\|x - y\|^2.\end{aligned} \tag{21}$$

Combining (21) with (20) yields

$$\|\mathcal{T}_{A^{\gamma},B^{\mu}}(x) - y\|^2 \leq \lambda\left(\theta(x) + \|x - y\|^2\right) + (1 - \lambda)\|x - y\|^2 = \lambda\theta(x) + \|x - y\|^2. \tag{22}$$

Now notice that (22) implies the quasinonexpansiveness of $\mathcal{T}_{A^{\gamma},B^{\mu}}^{\lambda}$ since $\theta(x) \leq 0$ if $x \notin \operatorname{Fix}\mathcal{T}_{A^{\gamma},B^{\mu}}^{\lambda} \supset A \cap B$. If we additionally have $\lambda \in \,]0, 1\,]$ and $\mu, \gamma \in \,]0, 2\,[$, then $(x \notin \operatorname{Fix}\mathcal{T}_{A^{\gamma},B^{\mu}}^{\lambda}) \implies \theta(x) < 0$, which shows the strict quasi-nonexpansivity.

Now we have that

$$0 \leq \|x_{n+1} - y\|^2 \leq \|x_0 - y\|^2 + \lambda\sum_{j=0}^{n}\theta(x_j).$$

Since $\gamma(2 - \gamma) \leq 0$ and $\mu(2 - \mu) \leq 0$, we have $\theta(x_j) \leq 0\,\forall j$. Since $\sum_{j=0}^{\infty}\theta(x_j)$ is a sum of non-positive terms and is bounded from below, $\theta(x_j) \to 0$. In particular, let $\gamma, \mu \in \,]0, 2\,[$ and we have $\gamma(2 - \gamma) < 0$ and $\mu(2 - \mu) < 0$; combining this with the fact that $\theta(x_j) \to 0$, we obtain

$$\lim_{n\to\infty}\|x_n - \mathcal{P}_A(x_n)\| = 0, \tag{23}$$

$$\text{and} \lim_{n\to\infty}\|\mathcal{R}_A^{\gamma}(x_n) - \mathcal{P}_B\mathcal{R}_A^{\gamma}(x_n)\| = 0. \tag{24}$$

Now since $\|x_n - \mathcal{R}_A^{\gamma}(x_n)\| = (2 - \gamma)\|x_n - \mathcal{P}_A(x_n)\|$, (23) implies that

$$\lim_{n\to\infty}\|x_n - \mathcal{R}_A^{\gamma}(x_n)\| = 0. \tag{25}$$

Now the triangle inequality yields

$$\|x_n - \mathcal{P}_B\mathcal{R}_A^{\gamma}(x_n)\| \leq \|x_n - \mathcal{R}_A^{\gamma}(x_n)\| + \|\mathcal{R}_A^{\gamma}(x_n) - \mathcal{P}_B\mathcal{R}_A^{\gamma}(x_n)\|, \tag{26}$$

and so (25) and (26) together imply

$$\lim_{n\to\infty} \|x_n - \mathcal{P}_B\mathcal{R}_A^{\gamma}(x_n)\| = 0.$$

This completes the proof. □

From Theorem 4, we obtain a number of convergence results.

Theorem 5 *Let $\gamma, \mu \in \,]0, 2[$ and $\lambda \in \,]0, 1]$. Suppose that the following hold*

(I) $\lim_{n\to\infty} \|x_n - \mathcal{P}_A(x_n)\| = 0$ *implies* $\lim_{n\to\infty} \mathrm{d}(x_n, A) = 0$ *and*
(II) $\lim_{n\to\infty} \|x_n - \mathcal{P}_B\mathcal{R}_A^{\gamma}(x_n)\| = 0$ *implies* $\lim_{n\to\infty} \mathrm{d}(x_n, B) = 0.$

Then $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in $A \cap B$. Moreover, any one of the three conditions below guarantee that $(x_n)_{n\in\mathbb{N}}$ converges strongly to a point in $A \cap B$:

(i) $\mathcal{H}$ is finite dimensional;
(ii) One of A or B is compact; and
(iii) $\{A, B\}$ is regular. *That is,* $\max\{\mathrm{d}(x, A), \mathrm{d}(x, B)\} \to 0$ *implies that* $\mathrm{d}(x, A \cap B) \to 0$.

Proof First, we prove that the sequence is weakly convergent to $A \cap B$. Since Theorem 4 implies that $\lim_{n\to\infty} \|x_n - \mathcal{P}_A(x_n)\| = \lim_{n\to\infty} \|x_n - \mathcal{P}_B\mathcal{R}_A^{\gamma}(x_n)\| = 0$, by assumptions (I) and (II), we have that $\|x_n - P_A(x_n)\| \to 0$ and $\|x_n - P_B(x_n)\| \to 0$. Thus all weak clusters point of the sequence $(x_n)_{n\in\mathbb{N}}$ belong to A and B, and so all weak cluster points of the sequence belong to $A \cap B$. By Theorem 4, $\mathcal{T}^{\lambda}_{A^{\gamma}, B^{\mu}}$ is a quasi-nonexpansive operator, and $A \cap B \subseteq \operatorname{Fix} \mathcal{T}^{\lambda}_{A^{\gamma}, B^{\mu}}$, and so the sequence generated by (5) is Fejér monotone with respect to $A \cap B$. Since all weak cluster points belong to $A \cap B$, the whole sequence converges weakly to a point in $A \cap B$; see, for example [13, Theorem 5.5].

(i) This is obvious since weak convergence implies strong in finite-dimensional spaces.
(ii) Suppose, without loss of generality, that A is compact. Then, there exist a subsequence $(x_{k_n})_{k_n\in\mathbb{N}} \subseteq (x_n)_{n\in\mathbb{N}}$ such that $\left(P_A(x_{k_n})\right)_{k_n\in\mathbb{N}}$ is strongly convergent to a point in A. Now, let $\bar{x}$ be the weak limit of the sequence $(x_n)_{n\in\mathbb{N}}$. Since $\bar{x} \in A \cap B$, we must have $P_A(x_{k_n}) \to \bar{x}$. Now,

$$\|x_{k_n} - \bar{x}\| \leq \|x_{k_n} - P_A(x_{k_n})\| + \|P_A(x_{k_n}) - \bar{x}\| \to 0,$$

which proves that the sequence $(x_n)_{n\in\mathbb{N}}$ has a strong cluster point. By Theorem 2, we conclude the strong convergence.

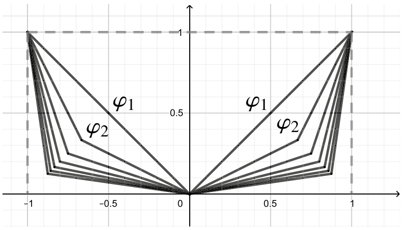


Fig. 5 The functions φ_k from (27)

(iii) Since $\max\{\mathrm{d}(x_n, A), \mathrm{d}(x_n, B)\} \to 0$, we have $\mathrm{d}(x_n, A \cap B) \to 0$; using the Fejér convergence of the sequence and Theorem 2, we obtain the strong convergence. □

Theorem 5 raises several natural questions. First, it is evident that the conditions (I) and (II) are satisfied in the case of projections onto the constraint sets. We will give examples of other cutter methods which satisfy them in Section 4.

Next we show that even for the very simple setting of a singleton set A, it is possible to construct the constraint function in such a way that condition (I) does not hold, highlighting that this condition is essential for the result.

Example 3 (On the importance of condition (I)) Let $\mathcal{H} = l_2$, and $A = \{0_{l_2}\}$. Note that A is the zero-level set of the function

$$f(x) = \sup_{k\in\mathbb{N}} \varphi_k(x^{(k)}),$$

where $x = (x^{(1)}, x^{(2)}, \dots, x^{(k)}, \dots)$ and

$$\varphi_k(t) = \max\left\{-\frac{1}{k}t, \frac{1}{k}t, kt + 1 - k, -kt + 1 - k\right\}. \tag{27}$$

These functions are shown in Figure 5 for $k \in \{1, \dots, 7\}$. Indeed, for any non-zero $x \in l_2$ we have $x^{(k)} \neq 0$ for at least one k, then $\varphi_k(x^{(k)}) > 0$, and hence $f(x) > 0$. At the same time, for $x = 0$ we have $\varphi_k(x^{(k)}) = 0$ for all $k \in \mathbb{N}$, so $f(0) = 0$.

Consider the sequence $\{x_n\}$, where x_n has all entries zero except for $x_n^{(n)} = 1$, so we have

$$x_1 = (1, 0, 0, 0, \dots), \quad x_2 = (0, 1, 0, 0, \dots), \quad x_3 = (0, 0, 1, 0, \dots), \dots.$$

For $|t| \le 1/2$, we have

$$|t|(k^2 - 1) - k^2 + k \le \frac{k^2 - 1}{2} - k^2 + k = -\frac{k^2 - 2k + 1}{2} = -\frac{(k-1)^2}{2} \le 0,$$

hence

$$\max\left\{\frac{1}{k}|t|, k|t|+1-k\right\} = \frac{|t|}{k} + \max\left\{0, \frac{|t|(k^2-1)-k^2+k}{k}\right\} = \frac{|t|}{k},$$

and

$$\varphi_k(t) = \frac{1}{k}|t| \quad \forall t, |t| \leq 1/2.$$

At the same time, for $|t| \geq \frac{k}{k+1}$ we have

$$k|t|+1-k = \frac{|t|}{k} + \frac{k^2-1}{k}|t| + 1 - k \geq \frac{|t|}{k},$$

hence,

$$\max\left\{\frac{1}{k}|t|, k|t|+1-k\right\} = k|t|+1-k \quad \forall t,\ |t| \geq \frac{k}{k+1},$$

so we have for $\|u\|_{l_2} \leq \frac{1}{2n}$ that

$$\varphi_k(u^{(k)}) = \frac{1}{k}|u^{(k)}| \leq \frac{1}{2} \quad \forall k \neq n, \quad \varphi_n(x^{(n)}+u^{(n)}) = \varphi_n(1+u^{(n)}) = nu^{(n)}+1 \geq \frac{1}{2},$$

hence

$$f(x_n+u) = \max\{\varphi_n(x_n^{(n)}+u^{(n)}), \sup_{k\neq n}\varphi_k(u^{(k)})\} = \varphi_n(x_n^{(n)}+u^{(n)}),$$

and so in a small neighbourhood of x_n we have

$$f(x) = \varphi_n(x^{(n)}) = nx^{(n)} + 1 - n.$$

The subgradient cutter then gives

$$\|x_n - P_{\partial f}(x_n)\| = \frac{1}{n} \to 0,$$

however

$$d(x_n, A) = \|x_n\| = 1,$$

so the condition (I) is violated. ◊

Due to an important example, we know that in infinite dimensions our algorithms may fail if we do not have subtransversality or compactness [39]. The above theorem also begs the question of what may go wrong in the case where we allow reflections $\gamma = 0$ or $\mu = 0$.

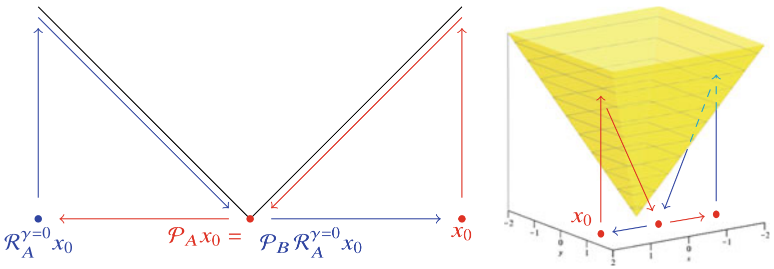

$\mathcal{R}_A^{\gamma=0}x_0$
$\mathcal{P}_A x_0 = \mathcal{P}_B\mathcal{R}_A^{\gamma=0}x_0$
x_0
x_0

Fig. 6 It is possible for every point to be a fixed point

Example 4 (Every point may be fixed) Letting $f, g : \mathbb{R} \to [\,0, \infty\,[$ by $f(x) := |x| =: g(x)$. Then every point in $\mathbb{R}$ is a fixed point of $T_{f^{\gamma=0}, g^{\mu=0}}$. This example is illustrated at left in Figure 6. ♢

For this example, all of the fixed points satisfy the property that $P_f(x) \in A \cap B$, which is analogous to the classical Douglas–Rachford fixed point result in Theorem 1. This property does not always hold, however, as illustrated in the next example.

Example 5 (Fixed points may not reveal much) Let $f, g : \mathbb{R}^2 \to [\,0, \infty\,[$ by $f, g : (x, y) \mapsto \max\{|x|, |y|\}$. This example is illustrated at right in Figure 6. Any point (x, y) satisfying $|x| \neq |y|$ is a fixed point of the operator $T_{f^{\gamma=0}, g^{\mu=0}}$; indeed, it is possible that every point is a fixed point, depending upon how the cutter is chosen when $|x| = |y|$. If additionally, $x \neq 0$, $y \neq 0$ and $|x| \neq |y|$, then (x, y) does not satisfy the property that $P_f(x) \in A \cap B$. ♢

Example 6 (Regularity conditions and convergence rates) One might also ask if the regularity conditions of Theorem 5 (iii) can be used to guarantee linear convergence rates, as is often the case with projection operators (see the many convergence results listed in [43]). However, Theorem 3 is for *very general* cutters, and so we can construct a counterexample.

Let $A, B := \{0\} \subset \mathbb{R}$. It is straightforward to verify that $\{A, B\}$ is regular. Let $C := \{1/n \mid n \in \mathbb{Z} \setminus \{0\}\}$. Now define

$$\mathcal{P} : x \mapsto \begin{cases} 0 & \text{if}\, x = 0 \\ 1 & \text{if}\, x > 1 \\ -1 & \text{if}\, x < -1 \\ 1/(n+1) & \text{if } 0 < x = 1/n \in C \\ 1/(n-1) & \text{if } 0 > x = 1/n \in C \\ 1/(n-1) \text{ for the unique } n \in \mathbb{Z} & \\ \qquad \text{satisfying } 1/n < x < 1/(n-1) & \text{if } \; -1 < x < 0 \,\text{and}\, x \notin C \\ 1/(n+1) \text{ for the unique } n \in \mathbb{Z} & \\ \qquad \text{satisfying } 1/(n+1) < x < 1/n & \text{if } 1 > x > 0 \,\text{and}\, x \notin C. \end{cases}$$

Clearly $\mathcal{P}$ is a cutter with respect to A and B. Set $\mathcal{P}_A := \mathcal{P}_B := \mathcal{P}$, $\gamma = \mu = 1$ and $\lambda = 1$. Then for $x_0 := 1$, we have $x_n := \mathcal{P}^2(x_{n-1}) = 1/(2n+1)$, so $x_n \to 0$ with a sublinear convergence rate. ◊

4 Implementations

For the classical implementation with projections onto the constraint sets, the assumptions of Theorem 5 are satisfied automatically, and hence we have the following result.

Corollary 1 *If $\mathcal{P}_A := P_A$ and $\mathcal{P}_B := P_B$ are projections onto the constraint sets, then assumptions (I) and (II) in Theorem 5 are satisfied, and we have the same weak convergence results.*

Proof Since in this case $\mathcal{P}_A = P_A$, we have

$$\|x_n - \mathcal{P}_A(x_n)\| = \|x_n - P_A(x_n)\| = \mathrm{d}(x_n, A) \xrightarrow[n\to\infty]{} 0, \tag{28}$$

hence (I) holds. Additionally, since $P_B R_A^{\gamma}(x_n) \in B$,

$$\mathrm{d}(x_n, B) \leq \|x_n - P_B R_A^{\gamma}(x_n)\| \xrightarrow[n\to\infty]{} 0. \tag{29}$$

□

Suppose that instead of two sets A and B, we are given a finite collection of closed convex sets $\Omega_i \subseteq \mathcal{H}$, where $i \in \{1, \dots, N\}$. The feasibility problem in this case consists of finding a point x such that

$$x \in \Omega := \bigcap_{i=1}^{N} \Omega_i.$$

Our two set formulation can be applied to this setting by working in the product space $\mathcal{H}^N$, and letting

$$A := \Omega_1 \times \cdots \times \Omega_N, \qquad B := \{x = (u_1, \dots, u_N) \,|\, u_1 = u_2 = \cdots = u_N\},$$

in which case the product space projections for $x = (u_1, \dots, u_N) \in \mathcal{H}^N$ are

$$P_A(x) = P_{\Omega_1, \times \cdots \times \Omega_N}(x) = (P_{\Omega_1}(u_1), \dots, P_{\Omega_N}(u_N)).$$

$$P_B(x) = \left(\frac{1}{N} \sum_{k=1}^{N} u_k, \dots, \frac{1}{N} \sum_{k=1}^{N} u_k \right). \tag{30}$$

This well known technique is used extensively in practical applications; see the important works [47, 50]. We note that even in the elementary case of alternating or cyclic projection method the convergence is much easier to study and understand in the case of two sets, and in fact there are some negative results in terms of the shape of limit sets for the infeasible case of the problem on more than two sets [9, 31].

We may use cutter methods together with the product space method to solve the *system of inequalities* expressed in feasibility form as

$$x \in \bigcap_{i=1}^{N} \operatorname{lev}_{\leq 0} f_i .$$

For example, one may employ *subgradient projections* with the cutter operators $P_{\partial f_i}$ defined by (7). From now on, we work in the Euclidean setting, letting $\mathbb{E}$ represent a finite-dimensional Hilbert space. We first prove the convergence for the special case of two convex functions.

Corollary 2 *Let $A := \operatorname{lev}_{\leq 0} f$ and $B := \operatorname{lev}_{\leq 0} g$ where $f : \mathbb{E} \to \mathbb{R}$ and $g : \mathbb{E} \to \mathbb{R}$ are convex functions with full domain. Suppose that $A \cap B \neq \emptyset$. Then the sequence $(x_n)_{n\in\mathbb{N}}$ generated by $x_{n+1} := T^{\lambda}_{f^{\gamma}, g^{\mu}}(x_n)$ with $\gamma, \mu \in \,]0, 2[$ converges strongly to a point $\bar{x} \in A \cap B$.*

Proof By Theorem 4(i), we have that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $A \cap B$. Thus $(x_n)_{n\in\mathbb{N}}$ is bounded.

First we will prove that the conditions (I) and (II) in Theorem 5 are satisfied. That is, $\|x_n - P_{\partial f}(x_n)\| \to 0$ implies that $\mathrm{d}(x_n, A) \to 0$, and $\|x_n - P_{\partial g} R^{\gamma}_{\partial f}(x_n)\| \to 0$ implies that $\mathrm{d}(x_n, B) \to 0$. Note that

$$\|x_n - P_{\partial f}(x_n)\| = \left\| x_n - \left(x_n - \frac{f(x_n)}{\|s(x_n)\|^2} s(x_n) \right) \right\| = \frac{|f(x_n)|}{\|s(x_n)\|},$$

where $s(x_n) \in \partial f(x_n)$. Since the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded, and f has full domain, we have that the sequence $\|s(x_n)\|_{n\in\mathbb{N}}$ is bounded (see, for example, [49, Theorem 24.7]). Since $\frac{f(x_n)}{\|s(x_n)\|} \to 0$, we have that $f(x_n) \to 0$.

We will show that $f(x_n) \to 0 \implies \mathrm{d}(x_n, A) \to 0$. Let

$$D := \mathbb{B}(P_{A\cap B}(x_0), \|x_0 - P_{A\cap B}(x_0)\|).$$

If $\|x_0 - P_{A\cap B}(x_0)\| = 0$, then we are done. Suppose then that $\|x_0 - P_{A\cap B}(x_0)\| > 0$. Let $\|x_0 - P_{A\cap B}(x_0)\| > \epsilon > 0$.

Since x_n is Fejér monotone with respect to $A \cap B$, we have that $x_n \in D$ for all n. Thus we may work with a restriction of f:

$$f|_D : x \mapsto \begin{cases} f(x) & \text{if } x \in D; \\ \infty & \text{otherwise,} \end{cases}$$

which is convex and coercive and satisfies $f|_D(x_n) = f(x_n)$ for all n, as well as $A' := A \cap D = \text{lev}_{\leq 0} f|_D$.

Without loss of generality let $0 \in A'$. Let $S := A' + \mathbb{B}(0, \epsilon)$. As A' is bounded, S is bounded. The condition $\|x_0 - P_{A \cap B}(x_0)\| > \epsilon$ ensures that $\text{bd}\, S \cap D \neq \emptyset$. As $f|_D$ is proper, lower semi-continuous, and bd S is closed and bounded with $\text{bd}\, S \cap D \neq \emptyset$, $f|_D$ attains a minimum on bd S. Let $\zeta := \min_{x \in \text{bd}\, S} f|_D(x)$. We will show $\text{lev}_{\leq \zeta} f|_D \subset S$. Suppose by way of contradiction that there exists $y \in \text{lev}_{\leq \zeta} f|_D \setminus S$. Then since $\mathbb{B}(0, \epsilon) \subset S$, we have that $y = \frac{1}{\lambda} u$ for some $u \in \text{bd}\, S$ and $\lambda \in [0, 1]$. Thus we have $u = \lambda y + (1 - \lambda)0$. This yields

$$\zeta \leq f|_D(u) = f|_D(\lambda y + (1 - \lambda)0) \leq \lambda f|_D(y) + (1 - \lambda) f|_D(0) \leq \lambda f|_D(y), \leq \lambda\zeta \tag{31}$$

where the first inequality is how we have defined ζ, the first equality is from how we have defined u, the second inequality is from convexity of $f|_D$, the third is because $0 \in A' = \text{lev}_{\leq 0} f|_D$, and the final inequality is because $0 < f|_D(y) \leq \zeta$. From (31), we have $\zeta \leq \lambda\zeta$, which is true only if $\lambda = 1$ or $\zeta = 0$. If $\zeta = 0$, then $y \in \text{lev}_{\leq 0} f|_D \subset S$, a contradiction. If $\lambda = 1$ then $y = u \in S$, a contradiction. Thus $\text{lev}_{\leq \zeta} f|_D \subset S$. Since $f|_D(x_n) \to 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $f|_D(x_n) < \zeta$ and so $x_n \in \text{lev}_{\leq \zeta} f|_D \subset S$ and so $\text{d}(A, x_n) \leq \text{d}(A', x_n) \leq \epsilon$. Thus $\text{d}(A, x_n) \to 0$.

Turning to the function g, with the same argument that is in Corollary 1 we have

$$\|x_n - P_{\partial g}(x_n)\| \leq \|x_n - P_{\partial g} R^{\gamma}_{\partial f}(x_n)\| + (2 - \gamma)\|P_{\partial f}(x_n) - x_n\| \to 0.$$

Thus, by the same arguments we used to show $f(x_n) \to 0$, we have that $g(x_n) \to 0$ and that $\text{d}(x_n, B) \to 0$.

Together with the fact that x_n is Fejér monotone with respect to $A \cap B$ and the fact that $\mathbb{E}$ satisfies condition (i) from Theorem 5, we conclude by Theorem 2 the convergence of the sequence to a point in $A \cap B$. □

Now we present a result for the case of more than two functions.

Corollary 3 *Let the* $(f_i)_{i \in \mathcal{I}}$ *where* $\mathcal{I} = \{1, 2, \cdots, N\}$, $N \in \mathbb{N}$, *are convex functions from* $\mathbb{E}$ *to* $\mathbb{R}$. *Consider for all* $i \in \mathcal{I}$, *the sets* $A_i := \{x \in \mathbb{E} : f_i(x) \leq 0\}$, *and suppose that* $C := \cap_{i \in \mathcal{I}} A_i \neq \emptyset$. *Consider the functions*

$$F : \mathbb{E}^N \to \mathbb{R} \;\; \text{defined by: } (x_1, x_2, \cdots, x_N) \to \sum_{i \in \mathcal{I}} \max\{f_i(x_i), 0\},$$

$$G : \mathbb{E}^N \to \mathbb{R} \;\; \text{defined by: } (x_1, x_2, \cdots, x_N) \to \sum_{i \in \mathcal{I}} \left\| x_i - \frac{1}{N} \sum_{j \in \mathcal{I}} x_j \right\|^2.$$

Let the sequence $(\mathbf{x}_n)_{n \in \mathbb{N}}$ *be as follows:*

$$\mathbf{x}_0 = (x_1^0, x_2^0, \cdots, x_N^0) \in \mathbb{E}^N,$$
$$\mathbf{x}_{n+1} = (x_1^{n+1}, x_2^{n+1}, \cdots, x_N^{n+1}) = \mathcal{T}^{\lambda}_{F^{\gamma}, G^{\mu}}(\mathbf{x}_n) = \mathcal{T}^{\lambda}_{F^{\gamma}, G^{\mu}}(x_1^n, x_2^n, \cdots, x_N^n).$$

Then $\mathbf{x}_n \to \bar{\mathbf{x}} = (\bar{x}, \bar{x}, \cdots, \bar{x}) \in D := \Pi_{i \in \mathcal{I}} A_i$ *with* $\bar{x} \in \mathbb{E}$*, which means that* $\bar{x} \in C$.

Proof The convexity of each f_i guarantees the convexity of F. Notice that $B := \text{lev}_{\leq 0} G = \{(x_1, x_2, \cdots, x_N) \in \mathbb{E}^N : x_1 = x_2 = \cdots = x_N\}$ is the linear subspace of agreement which we recognize from (30), and $G = \text{d}(B, \cdot)^2$ is actually the square of the distance function for B. As G is the square of the distance function for a convex set, G is convex. In fact, if one chooses to replace subgradient projection with respect to G by Euclidean projection directly onto its zero level set, the Euclidean projection is just as given in (30).

The algorithm is well defined because the domain of each f_i is the space $\mathbb{E}$. Finally notice that $D = \text{lev}_{\leq 0} F$. Applying Corollary 2, we have that $\mathbf{x}_n \to \bar{\mathbf{x}} := (\bar{x}, \bar{x}, \cdots, \bar{x}) \in D \cap B$ where $B = \{(x_1, x_2, \cdots, x_N) \in \mathbb{E}^N : x_1 = x_2 = \cdots = x_N\}$. □

As an immediate consequence, we also have strong convergence in the case where we work with projections onto the constraint sets and a finite number of sets $A_1, \ldots, A_N$; just let the N functions be given by $f_i := \text{d}_{A_i}(\cdot)$. See, for example, [16, Ex. 2.7].

Remark 1 (Sequences γ_n, μ_n*)* One may take sequences $(\gamma_n)_{n \in \mathbb{N}}, (\mu_n)_{n \in \mathbb{N}}$ and, provided that $\liminf \gamma_n(2 - \gamma_n) > 0$ and $\liminf \mu_n(2 - \mu_n) > 0$, all of the above convergence results will hold for sequence given by $x_n := \mathcal{T}^{\lambda}_{A^{\gamma_n}, B^{\mu_n}} x_{n-1}$. Indeed, this is the usual framework of [25] although we have avoided the use of these sequences for the simplicity of exposition.

5 Discussion

In the convex setting, when projections onto the constraint sets are replaced with cutters, the operator $\mathcal{T}^{\lambda}_{A^{\gamma}, B^{\mu}}$ loses firm nonexpansivity and yet retains many of its desirable convergence properties because of Fejér monotonicity. Subgradient projections are one useful context in which the firm nonexpansivity is lost while the Fejér monotonicity is retained. The similarities suggest avenues of further research: one in the convex setting and one outside of it.

5.1 Further Investigation

In the convex setting, the algorithmic differences corresponding to different choices of μ, γ, λ are a highly active area of investigation. See, for example, [8, 32, 33]. Figure 7 compares two variants of λ-averaged relaxed projection methods in the case of subgradient projections, and the behaviour differences are reminiscent of those known in the setting of projections onto the constraint sets. Further comparison of behaviour for choices of averaging and relaxation parameters invites experimental investigation.

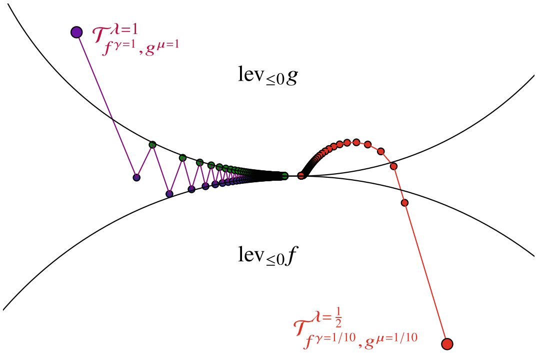

$\mathcal{T}^{\lambda=1}_{f^{\gamma=1},g^{\mu=1}}$
$\mathrm{lev}_{\leq 0}g$
$\mathrm{lev}_{\leq 0}f$
$\mathcal{T}^{\lambda=\frac{1}{2}}_{f^{\gamma=1/10},g^{\mu=1/10}}$

Fig. 7 Convergence for $\mathcal{T}^{\lambda=\frac{1}{2}}_{f^{\gamma=1/10},g^{\mu=1/10}}$ (parameters similar to Douglas–Rachford method) versus $\mathcal{T}^{\lambda=1}_{f^{\gamma=1},g^{\mu=1}}$ (parameters similar to alternating projections)

Even when the formulation of a problem allows for computations of projections onto the constraint sets, it may be undesirable (computationally expensive) to do so. Consider, for example, the projection onto an ellipse: $\mathcal{E} := \left\{(x, y) \in \mathbb{R}^2 \mid \frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1\right\}$ for given constants a, b, x_0, y_0. Computation of the exact projection for a point not in $\mathcal{E}$ requires solving a Lagrangian problem (see, for example [23] and [44]), while computation of the subgradient projection for the function $f : (x, y) \to \left(\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} - 1\right)^2$ does not. It is very natural to investigate the differences in behaviour induced by the choice of projection method.

Both the method of alternating projections and the Douglas–Rachford method have also been used to solve a variety of non-convex feasibility problems, with the latter generally the more robust. See, for example, [2–6, 19, 23, 41, 44]. It is reasonable to consider the behaviour of λ-averaged relaxed projection methods in the non-convex setting, and a very natural problem would be that of finding $x \in \mathrm{lev}_{\leq 0} f \cap \mathrm{lev}_{\leq 0} g$—using subgradient projections—where one or both of f, g are not convex. Indeed, any non-convex feasibility problem in $\mathbb{R}^N$ is an example of such a non-convex variational inequality problem where $f = \mathrm{d}_A(\cdot)$, $g = \mathrm{d}_B(\cdot)$, and so much investigation has already been done.

5.2 Conclusion

We learn much by comparing and contrasting what may be shown about λ-averaged relaxed cutter methods through the differing frameworks of firm nonexpansivity and quasi-nonexpansivity. That so many of the desirable properties carry over—from the more specific setting of projections onto the constraint sets to the more general setting of cutters—is especially useful. Splitting methods employing projections onto the constraint sets are an area of significant experimental research. We conclude by noting that those methods which employ other implementations of cutters merit further experimental investigation, and that the theory is elegant in its own regard.

Acknowledgements We are grateful to Heinz Bauschke for pointing out useful references. We also extend our gratitude to the organisers and participants of the workshop *Splitting Algorithms, Modern Operator Theory and Applications* held at Casa Matematica Oaxaca, Mexico, from 17–22 September 2017 and to the Australian Research Council for supporting our travel expenses via project DE150100240. We are also grateful for the helpful feedback of two anonymous referees, whose suggestions greatly improved this manuscript.
We dedicate this work to the memory of Jonathan M. Borwein, whose influence on the present authors and on the Australian mathematical community cannot be overstated.

References

1. Aleyner, A., Reich, S.: Block-iterative algorithms for solving convex feasibility problems in Hilbert and in Banach spaces. J. Math. Anal. Appl. **343**(1), 427–435 (2008). https://doi.org/10.1016/j.jmaa.2008.01.087
2. Aragón Artacho, F.J., Borwein, J.M.: Global convergence of a non-convex Douglas-Rachford iteration. J. Glob. Optim. **57**(3), 753–769 (2012)
3. Aragón Artacho, F.J., Borwein, J.M., Tam, M.K.: Douglas–Rachford feasibility methods for matrix completion problems. ANZIAM J. **55**(4), 299–326 (2014). https://doi.org/10.1017/S1446181114000145
4. Aragón Artacho, F.J., Borwein, J.M., Tam, M.K.: Recent results on Douglas–Rachford methods for combinatorial optimization problems. J. Optim. Theory Appl. **163**(1), 1–30 (2014). https://doi.org/10.1007/s10957-013-0488-0
5. Aragón Artacho, F.J., Borwein, J.M., Tam, M.K.: Global behavior of the Douglas–Rachford method for a nonconvex feasibility problem. J. Global Optim. **65**(2), 309–327 (2016). https://doi.org/10.1007/s10898-015-0380-6
6. Aragón Artacho, F.J., Campoy, R.: Solving graph coloring problems with the Douglas–Rachford algorithm. Set-Valued and Variational Analysis pp. 1–28 (2017)
7. Aragón Artacho, F.J., Campoy, R.: A new projection method for finding the closest point in the intersection of convex sets. Comput. Optim. Appl. **69**(1), 99–132 (2018). https://doi.org/10.1007/s10589-017-9942-5
8. Artacho, F.J.A., Campoy, R.: Optimal rates of linear convergence of the averaged alternating modified reflections method for two subspaces (2017). arXiv:1711.06521
9. Baillon, J.B., Combettes, P.L., Cominetti, R.: There is no variational characterization of the cycles in the method of periodic projections. J. Funct. Anal. **262**(1), 400–408 (2012). https://doi.org/10.1016/j.jfa.2011.09.002
10. Bauschke, H.H., Combettes, P.L.: A weak-to-strong convergence principle for fejér-monotone methods in hilbert spaces. Mathematics of Operations Research **26**(2), 248–264. http://www.jstor.org/stable/3690618 (2001)
11. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces, 2nd edn. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48311-5. With a foreword by Hédy Attouch
12. Bauschke, H.H., Combettes, P.L., Luke, D.R.: Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization. J. Opt. Soc. Am. A **19**(7), 1334–1345 (2002). https://doi.org/10.1364/JOSAA.19.001334
13. Bauschke, H.H., Dao, M.N., Noll, D., Phan, H.M.: On Slater's condition and finite convergence of the Douglas–Rachford algorithm for solving convex feasibility problems in Euclidean spaces. J. Global Optim. **65**(2), 329–349 (2016). https://doi.org/10.1007/s10898-015-0373-5
14. Bauschke, H.H., Wang, C., Wang, X., Xu, J.: On subgradient projectors. SIAM J. Optim. **25**(2), 1064–1082 (2015)
15. Bauschke, H.H., Wang, C., Wang, X., Xu, J.: On the finite convergence of a projected cutter method. J. Optim. Theory Appl. **165**(3), 901–916 (2015)

16. Bauschke, H.H., Wang, C., Wang, X., Xu, J.: Subgradient projectors: extensions, theory, and characterizations. Set-Valued Var. Anal. 1–70 (2017)
17. Bello Cruz, J.Y., Díaz Millán, R.: A relaxed-projection splitting algorithm for variational inequalities in Hilbert spaces. J. Global Optim. **65**(3), 597–614 (2016). https://doi.org/10.1007/s10898-015-0397-x
18. Bello Cruz, J.Y., Iusem, A.N.: An explicit algorithm for monotone variational inequalities. Optimization **61**(7), 855–871 (2012). https://doi.org/10.1080/02331934.2010.536232
19. Benoist, J.: The Douglas-Rachford algorithm for the case of the sphere and the line. J. Glob. Optim. **63**, 363–380 (2015)
20. Borwein, J.M.: The life of modern homo habilis mathematicus: experimental computation and visual theorems. Tools and Mathematics. Mathematics Education Library, vol. 347, pp. 23–90. Springer, Berlin (2016)
21. Borwein, J.M., Li, G., Tam, M.K.: Convergence rate analysis for averaged fixed point iterations in common fixed point problems. SIAM J. Optim. **27**(1), 1–33 (2017)
22. Borwein, J.M., Li, G., Yao, L.: Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets. SIAM J. Optim. **24**(1), 498–527 (2014). https://doi.org/10.1137/130919052
23. Borwein, J.M., Lindstrom, S.B., Sims, B., Schneider, A., Skerritt, M.P.: Dynamics of the Douglas-Rachford method for ellipses and p-spheres. Set-Valued Var. Anal. **26**(2), 385–403 (2018)
24. Borwein, J.M., Sims, B.: The Douglas–Rachford algorithm in the absence of convexity. Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and Its Applications, vol. 49, pp. 93–109. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-9569-8_6
25. Cegielski, A.: Iterative Methods for Fixed Point Problems in Hilbert Spaces. Lecture Notes in Mathematics, vol. 2057. Springer, Heidelberg (2012)
26. Cegielski, A., Reich, S., Zalas, R.: Regular sequences of quasi-nonexpansive operators and their applications (2017)
27. Censor, Y., Lent, A.: Cyclic subgradient projections. Math. Program. **24**(1), 233–235 (1982)
28. Censor, Y., Zenios, S.A.: Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press on Demand (1997)
29. Combettes, P.L.: Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections. IEEE Trans. Image Process. **6**(4), 493–506 (1997)
30. Combettes, P.L.: Quasi-Fejérian analysis of some optimization algorithms **8**, 115–152 http://www.sciencedirect.com/science/article/pii/S1570579X01800100 (2001)
31. Cominetti, R., Roshchina, V., Williamson, A.: A counterexample to De Pierro's conjecture on the convergence of under-relaxed cyclic projections (2018)
32. Dao, M.N., Phan, H.M.: Linear convergence of projection algorithms (2016). arXiv:1609.00341
33. Dao, M.N., Phan, H.M.: Linear convergence of the generalized Douglas–Rachford algorithm for feasibility problems (2017). arXiv:1710.09814
34. De Pierro, A.R.: From parallel to sequential projection methods and vice versa in convex feasibility: results and conjectures. Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications (Haifa, 2000). Studies in Computational Mathematics, vol. 8, pp. 187–201. North-Holland, Amsterdam (2001). https://doi.org/10.1016/S1570-579X(01)80012-4
35. Douglas Jr., J., Rachford Jr., H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**, 421–439 (1956). https://doi.org/10.2307/1993056
36. Drusvyatskiy, D., Li, G., Wolkowicz, H.: A note on alternating projections for ill-posed semidefinite feasibility problems. Math. Program. **162**(1-2, Ser. A), 537–548 (2017). https://doi.org/10.1007/s10107-016-1048-9
37. Elser, V.: Matrix product constraints by projection methods. J. Global Optim. **68**(2), 329–355 (2017)
38. Fukushima, M.: A relaxed projection method for variational inequalities. Math. Program. **35**(1), 58–70 (1986). https://doi.org/10.1007/BF01589441

39. Hundal, H.S.: An alternating projection that does not converge in norm. Nonlinear Analysis: Theory, Methods & Applications **57**(1), 35–61. http://www.sciencedirect.com/science/article/pii/S0362546X03004218 (2004). https://doi.org/10.1016/j.na.2003.11.004
40. Kruger, A.Y., Luke, D.R., Thao, N.H.: About subtransversality of collections of sets. Set-Valued Var. Anal. **25**(4), 701–729 (2017). https://doi.org/10.1007/s11228-017-0436-5
41. Lamichhane, B.P., Lindstrom, S.B., Sims, B.: Application of projection algorithms to differential equations: boundary value problems (2017). arXiv:1705.11032
42. Li, G., Pong, T.K.: Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Math. Program. **159**(1-2, Ser. A), 371–401 (2016). https://doi.org/10.1007/s10107-015-0963-5
43. Lindstrom, S.B., Sims, B.: Survey: sixty years of Douglas–Rachford (2018). arXiv:1809.07181
44. Lindstrom, S.B., Sims, B., Skerritt, M.P.: Computing intersections of implicitly specified plane curves. Nonlinear Conv. Anal. **18**(3), 347–359 (2017)
45. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979). https://doi.org/10.1137/0716071
46. Littlewood, J.E.: A Mathematician's Miscellany. Methuen London (1953)
47. Pierra, G.: Decomposition through formalization in a product space. Math. Program. **28**(1), 96–115 (1984)
48. Polyak, B.T.: Introduction to optimization. Translations Series in Mathematics and Engineering. Optimization Software (1987)
49. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
50. Spingarn, J.E.: Partial inverse of a monotone operator. Appl. Math. Optim. **10**(1), 247–265 (1983)
51. Svaiter, B.F.: On weak convergence of the Douglas-Rachford method. SIAM J. Control Optim. **49**(1), 280–287 (2011)

Part II
Education

Introduction

Naomi Simone Borwein

Jon was a passionate advocate for mathematics research and education. It is a sense of joviality and inquiry that sustained his dedication to both. His active engagement with the communication and growth of mathematics research and pedagogy crossed popular and academic lines, and spanned from primary to tertiary platforms. The papers contained in this volume showcase his dynamic set of research interests, which is equally mirrored in the Education-themed section of the September 2017 Jonathan Borwein Commemorative Conference, and its Satellite meetings on Indigenising mathematics curricula, entitled “Mathematics and Education: Spirit, Culture and Community”. As a somatic approach, haptics and heuristics became integral to Jon’s model of mathematical education, with long reaching ramifications, which this section explores through topical considerations.

Indeed, the papers contributed to the Education section of *From Analysis to Visualisation* encompass a broad spectrum of approaches, all of which in some way reflect Jon’s life, interests and legacy. An introductory paper by Naomi Simone Borwein and Judy-anne Heather Osborn (the section editors) functions as an inquiry “On the Educational Legacies of Jonathan M. Borwein”, exposing the extent of Jon’s engagement with teaching practice. The paper seeks to chronicle the Education section and its panel discussion, to undertake a biographical survey of Jon’s educational creed and to catalyse discourse on disciplinary conditions of mathematics and mathematical education that stand in the way of reform. The remaining seven contributions are thematically organised around three major topics (a, b, c). This first topic, a) **Historical—Biographical Focus**, contains a paper by Keith J. Devlin. In this piece, Devlin proffers recollections of himself, Jon, computation and the early American Mathematics Society (AMS) in “How Mathematicians Learned to Stop Worrying

N. S. Borwein (✉)
Faculty of Education, Western University, London, ON, Canada
e-mail: naomi.borwein@gmail.com

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_6

and Love the Computer". It is implicative of some of Jon's early roles in integrating computers and mathematics. The second topic b) provides exemplary **Educational Approaches** like Merrilyn Goos' careful and substantive paper, "Crossing Boundaries: Fostering Collaboration Between Mathematics Educators and Mathematicians in Initial Teacher Education Programs", and Kathryn Holmes' instructive "Mathematics Education in the Computational Age: Challenges and Opportunities", which continues the conversation on implementing technology to the apparatus of education in secondary public schools. Collin Grant Philips and Fu Ken Ly's description of a math workshop in "Mathematics Education for Indigenous Students in Preparation for Engineering and Information Technologies" explores this context in relation to Indigenous math education, highlighting challenges of teaching at the university level, and promoting within the broader community. In the final thematic grouping c) **Innovative Applications**, Michael Assis and Michael Donovan define and explicate "storigami", an operation and form that marries aboriginal storytelling practices and the art of origami folding as an effective teaching aide—an innovative mixed methodological approach to indigenising the curriculum. Taking the structure of classroom notes, the contribution by Damir Jungic and Veselin Jungic, entitled "Dynamic Visual Models: Ancient Ideas and New Technologies", presents animated visual proofs (dynamic visual models) juxtaposed to traditional handwritten formula proofs. Their inquiry into Jon's oft-misunderstood maxim "Sometimes it is easier to see than to say" through an apparently singular juxtaposition of animated visual proof in union with formal written proof becomes a way of inspecting improved learning and teaching tools, outcomes and approaches in standard university calculus classes. Their paper highlights the benefits of a dipole method. Robert Corless and Eunice Yu Sze Chan describe the creation of a first-year Experimental Mathematics course that utilises active learning strategies and devices in "A Random Walk Through Experimental Mathematics". The title itself is nicely reminiscent of Jon and Francisco J. Aragon Artacho's work on random walk visualisations.

From cultural history and biography to digital assistance and methodological innovation, together these papers are suggestive of diverse extensions of Jon's own work in myriad computational and experimental mathematics laboratories, his use of visual proof alongside written formulas, his desire to succour reform, development, and outreach, across educational and mathematical boundaries. The content matter of these contributions truly builds on a subset of Jon's interests and pursuits in mathematical education, and functions as a living textual memory.

On the Educational Legacies of Jonathan M. Borwein

Naomi Simone Borwein and Judy-anne Heather Osborn

Dedicated to the memory of Jonathan M. Borwein

1 Introduction

Jon Borwein was no stranger to being a catalytic agent between discourses and disciplines, at times both derided and esteemed. In life, he often had a central role in connecting people and ideas across diverse specialities and careers. Even in his passing, his role as *boundary broker* [88] and catalyst continues as the community reflects [5], including in this chapter.

During his life, Jon had many significant achievements touching multiple fields of mathematical research. Within each of these fields he will be remembered for his insights and contributions. This chapter discusses aspects of Jon's academic interests, which relate to all his mathematical research, and more broadly to his mathematical and educational philosophy. Out of this multifaceted view of Jonathan Borwein come multiple ideas and frameworks, which, if taken up and further developed by the community, may become lasting legacies within mathematical culture and education.

It is a contention of this chapter that the highly connected and somewhat radical philosophy and practice (praxis) of mathematics that Jon created and exemplified in his life has the potential to be richly fruitful in the future of mathematics itself, and particularly in mathematics education. That praxis was in development for his

N. S. Borwein
Faculty of Education, Western University, London, ON, Canada
e-mail: naomi.borwein@gmail.com

J. H. Osborn (✉)
School of Mathematical and Physical Sciences, The University of Newcastle, Newcastle, Australia
e-mail: Judy-anne.Osborn@newcastle.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_7

whole professional life, and may have been most fully elaborated in his conception of experimental mathematics and experimental “mathodology” [14, p 32], for both research and teaching.

Jon’s educational writing, such as in Proof and Proving in Mathematics Education [50, p 69–96], illustrates that praxis, as commented upon by the editors [50, p 4] when they wrote

> Jonathan Borwein, in his plenary chapter “Exploratory experimentation: Digitally assisted discovery and proof” argues that current computing-technologies offer revolutionary new scaffolding both to enhance mathematical reasoning and to restrain mathematical error. He shares Pólya’s view that intuition, enhanced by experimentation, mostly precedes deductive reasoning. He then gives and discusses some illustrative examples, which clearly show that the boundaries between mathematics and the natural sciences and between inductive and deductive reasoning, are blurred and getting more blurred.

Across Jon’s own work, he frequently ignored division lines that bifurcated mathematics and mathematics education, crossing seamlessly betwixt what are, today, commonly seen as distinct enterprises. More than this, his mathematics was entwined with art, music, history, science and an intense engagement with his fellow human beings. Further, his seemingly effortless boundary crossing within mathematics [13] is reflected in the depth and breadth of his mathematics as reflected in the five themes of the Jonathan Borwein Commemorative Conference (JBCC) [15], *from Analysis to Visualization*.

Throughout his life, Jon developed an understanding and philosophy of mathematics that was influenced by Mathematical Humanism as described by Hersh [52], seen in the light of experimental mathematics [4, 6, 12, 17]. This philosophy, one that embraces a holistic approach to disciplines, infused his practice. These ways of seeing, thinking, teaching and researching, as well as the digital impressions that Jon left behind, constitute a rich reservoir that we draw upon in this work, and encourage others to also draw upon.

In a sense, this chapter is about antitheticals: silos versus bridges (or sometimes chasm-sized leaps) and schisms versus schism-spanning stewardship. Explored in the context of mathematics culture and education, we represent this subject matter through two branches of scholarship, with different bodies of knowledge, nomenclature, style and interpretive sources. Methodologically, we draw upon very different kinds of resources in addition to our memories and archives of Jon. The first is the multifaceted community that Jon belonged to, especially as represented at the JBCC panel discussion on education, and the second is the theory of schisms, divisions and discipline formation.

Schisms and theory thereof are relevant, because mathematics and mathematics education matters to so many people in so many capacities, some of which are potentially conflicting. One such potentially problematic division, not universally experienced, but potent, is between mathematics (as seen by mathematics academics and practitioners in the world) and mathematics education (as seen by education academics and school teachers).

We are cognizant of the fact that, depending on where one comes from, mathematics and mathematics education are defined differently, sometimes as part of a whole enterprise, and sometimes as two distinct disciplines. Indeed, within the union of the two (whether defined as distinct or not) there are various subcultures (e.g. topologists, optimisers, educationalists, ...) with different semantics, approaches, literatures, styles and values. From the point of view of the aims and methodology of this chapter, the same infrastructure of paradigmatic function applies and much of the same broad patterns emerge independently of the particular divisions acknowledged or focused upon.

There are four sections in this chapter: *Introduction*, *Issues: Practitioner Voices*, *Schisms: a Theoretician's Voice* and *Jon's Coda*. In the introduction, we introduce ourselves, the aims of this work, and the people and context of the JBCC. In *Issues: Practitioner Voices*, we describe themes that arose in the JBCC panel discussion, and link them to the published literature written by mathematicians and mathematics educators on those same issues. In *Schisms: a Theoretician's Voice*, we describe theory of schisms, divisions and disciplinarity. In *Jon's Coda*, we use examples from his life to describe his philosophy and methodology, which throws fresh light on the content of the previous sections of the paper, and offers examples of practical ways forward.

We present various parts of what now follows as a dialogue between the authors. This was a style Jon appreciated and has used previously [22]. We use this form of presentation in an acknowledgement of the relationship between form and content since we want the chapter to foster ongoing conversations.

1.1 Introducing the Voices

The two authors of this chapter are amongst Jon's many collaborators, and in distinct capacities have worked closely with him on matters relating to mathematical culture and mathematics education. Both of us were, in different capacities, apprenticed to Jon, in the ways he saw, managed and wrote about educational spaces and mathematical spaces, and in the fluidity he saw and actively worked to create between the two.

Naomi Simone Borwein is a scholar of literary and cultural theory who had an intensely mathematical childhood, which from her early life included mathematical discussions with her father, Jonathan Borwein, and inclusion in the broader mathematical milieu that surrounded him. She has had a lifelong love of and interest in education, which has been actualised across a spectrum of disciplines throughout her life. Further, her engagement with education has occurred across a range of contexts, from popular outreach in the form of “Math in the Malls” [43], to teaching University classes in literature, to usability research for NIST's *Digital Library of Mathematical Functions*, to editorial collaboration with her father on various works, including *Tools and Mathematics: Instruments for Learning*. Her most recent work is in mathematical curriculum and pedagogy at Western University, Canada. Like her father, Naomi is an inveterate boundary crosser, with University training across Mathematics, Science, History and English. Naomi's focus throughout her work may be broadly categorised as cultural historiography but draws on this diverse body of knowledge.

The second author, Judy-anne Heather Osborn, was appointed in 2011 by Jon as the founding Educational Representative on the Executive Committee of his Priority Research Centre CARMA (Computer Assisted Research Mathematics and its Applications) at the University of Newcastle. In this capacity, Judy-anne Heather Osborn worked with Jon for the last 6 years, before his death, on common goals relating to mathematics education and educational research. Like Jon, Judy-anne Heather Osborn has interests in history and philosophy as well as mathematics and education. Indeed, at the conclusion of school Judy-anne Heather Osborn had wondered whether to pursue mathematics or history, so it was resonant for her to read in Jon's *Implications of Experimental Mathematics for the Philosophy of Mathematics* [19] that Jon had seriously contemplated the same choice, and chosen mathematics by a hair's breadth, at the last moment.

As part of the JBCC, the two authors organised a panel discussion called *Maths, Education, Research and Culture*, which took place from 11:00am to 12:30pm on Wednesday 27 September 2017. The intended purpose of the panel was to start a dynamic dialogue between mathematics educators and mathematicians within the conference, which extends to the broader conversation of which this paper is a part. The two talks immediately preceding the panel discussion were there by design, intended to spark discussion. They were "Investigating the Schism in Math(s)/Education: a transcultural perspective", by Naomi Simone Borwein, and "About Jon: Learner and Teacher", by Judy-anne Heather Osborn. Indeed, we had hoped that the panel would address the nexus between mathematics education and research practice, in light of Jon's influence of innovative models for working and being, and we assert that this hope was realised.

Panel members were chosen to represent the collaborative community both at the JBCC, and as a cross-section of Jon's collaborative world. The panel consisted of representatives from the five conference themes, plus three additional experts, as follows:

- Rob Corless standing in for Regina Burachik and Guoyin Li (chairs: Applied Analysis, Optimisation and Convex Functions).
- Kathryn Holmes standing in for Naomi Simone Borwein and Judy-anne Heather Osborn (chairs: Education).
- David H. Bailey (chair: Experimental Mathematics and Visualisation).
- Qiji J. Zhu (chair: Financial Mathematics).
- Richard Brent (chair: Number Theory, Special Functions and Pi).
- Veselin Jungic (keynote from preceding Satellite meeting: "Mathematics and Education: Spirit, Culture and Community", with expertise in Indigenous mathematics education in the Canadian context).
- Brailey Sims (expert at the cusp between the secondary and tertiary mathematics curriculum in Australia).
- Cyndi Garvan (interdisciplinary practitioner, Statistician in a College of Medicine collaborating on Educational Development in the United States).

The audience, who ultimately drove the conversation, represented a vast repository of knowledge across different specialised areas, in proportions roughly representative of Jon's own collaborations. These included many eminent researchers and people known for their advocacy and leadership within mathematics. The University-based research mathematicians and graduate students present spanned the five themes of the Conference and beyond. Non-University-based participants included school teachers and mathematicians in industry across a variety of roles including publishing. The following broad stimulatory questions were posed to the panel:

Q1. Can research practices be fruitfully incorporated in school math(s)?
Q2. Can mathematics be taught in such a way that the general population does not fear it?
Q3. How can we change mathematics research training to be more inclusive of a diversity of people and cultures?
Q4. How do the specific characters and needs of the research areas affect the above questions?

A free-flowing conversation emerged, of which Naomi took written notes as it unfolded. In the next section, we discuss this practitioner conversation in relation to both our own views as practitioners within Mathematics and Education, and in relation to the scholarly literature written by Mathematics and Education practitioners and theorists.

2 Issues: Practitioner Voices

2.1 Overview

In this section, we draw upon the conversation that emerged in the JBCC panel session. What transpired was an intellectually open forum in which a large number of well respected people openly discussed many collective concerns of Mathematics culture and Education. We have grouped the issues raised into three broad themes: *Diversity*, *Fear* and *Transformation.*

The starred entries in this section represent paraphrased commentary by audience and panel members, which we present anonymously here with permission of the speakers. Interspersed with starred entries are sections of dialogue between the authors: *NB* stands for Naomi Simone Borwein and *JO* for Judy-anne Heather Osborn. These vignettes give us the opportunity to extend and contextualise the JBCC participants' words; and more expansively are intended to provoke and promote further dialogue beyond these pages. Each subsection is supported by some relevant literature at its close.

2.2 Diversity

Diversity is a contested word with many possible shades of meaning. We use it here because it was used by the JBCC participants. We hope its meanings in context will emerge from the following quotes.

* We are not where we need to be with diversity.
* It is especially important to recruit talent where we find it, and those talented people come from diverse situations.

Most of the conversation around this topic focused on gender inequality.

* It is a very hard question, "how to move towards a warmer environment for women in mathematics?"
* The "gender schema" of the male doctor, male philosopher, etc. is a subtext that needs to be addressed.
* We are just not shifting that "male domain schema".
* As a socialised male (in both India and Canada), I was always told "you can do that". I was always two steps ahead, for several years.
* In our current society, girls and boys are socialised differently, so whether we like it or not "vroom, vroom fast cars" is not going to get many girls in. I do think that marketing needs to change.
* In Australia, all recent advertising of maths in terms of careers had female representation.
* There's this disparity in what turns people off.
* I want to pick up on something that Naomi said, as far as the schism in math research and math education goes. I was one of four men in attendance at a Women in Maths Conference with 150 delegates—very different to the usual numbers in which men dominate. The proportion of women in math education conferences is more than 50:50. That's part of the schism.
* There is an interesting story about an airline engineering project and an associated international high school competition. One year they surveyed the students and asked them about their confidence in their parent's advice. After the challenge, the boys still trusted their parent's advice, but the girls did not. A possible interpretation is that the competition led all participants to be more enthusiastic about maths-related careers, but this affected trust in parental advice differently for girls and boys due to girls potentially being more often told by their parents to avoid maths careers, and boys more often to pursue them. Parental advice to teenagers is important.
* In advanced high school maths courses nationally, there are twice as many boys as girls, and this is in part responsible for what is happening at University.
* Even at the tertiary level this is going wrong. In Science intakes females outperform males, but by Honours level the numbers of girls have dropped dramatically and it is very disproportionate.

* A word of warning: female participation at PhD level in Australian Universities is being driven by International Enrolment, which is masking a drop in Australian female participation.
* There's something that happens very early on in the education process which has ramifications right through, and we don't really understand it.

Ethnicity, race and culture were also part of the conversation that started around diversity. International comparisons led to conversation about the fact and implications of what is sometimes called *massification* [45] of tertiary education.

* In India we have this idea of learning from the West, but in the West there is no idea of learning from India.
* The general University rate in India is pretty high; there have been lots of changes in the last 20 years though graduate school in mathematics is declining.
* Many countries are experiencing change. Acceptance rates have gone from 5–6 percent of elite, to 50–60 percent.

There was concern that, worldwide, university mathematics education systems are not preparing people for the actual job market, and a desire to change this.

* In China there are around 4000 Universities and the government is asking that 600 of them be reversed into vocational schools.
* In the past 20 to 30 years, the Chinese education system has experienced a vast expanding of Universities—10 million students in any year. When the students graduate, they often cannot find a suitable job.
* There is a mismatch between the educational goals and the job market. From the literature, this problem is not unique to China.
* In the 1950s and 60s, there were different systems: Professional and State Universities, which taught in different ways.
* Should we consider one method of teaching mathematics for all, or different methods for different career orientations?

JO: What JBCC participants are saying is reminiscent of conversations that I often hear or participate in, in the corridors and tea rooms of Maths Departments.

NB: Also, there was an evident willingness and interest in addressing these issues and finding solutions.

JO: Yes, and a many-faceted approach to understanding causes, and on the large scale this does seem to be working, slowly.

NB: Yes. There is a deeper layer to the discussion, which can be understood as the social and political dimensions of math education. This topic has been the focus of an ICME survey [59]. The discussions in the panel are imbued by this theme, for instance, in the commentaries about gender, and about China and India versus "Western" math education and institutions. Both the ICME topical survey and the panel discussions point to possible negative implications of a lack of ability to incorporate diversity across several critical areas—see Subsection 3.2.

JO: Indeed, do we *celebrate diversity, seek diversity, cope with diversity?* In the discussion of women in maths, the sense seemed to be relatively straightforward: diversity is a good thing and we don't have enough of it—there are various things getting in the way including wider societal expectations so let's fix that. By and large that part of the conversation did not shine a critical lens upon mathematical culture itself. When the conversation moved towards consideration of ethnicity and race, there was some reflection that perhaps mathematical culture itself is not perfect?

NB: And not what we might conventionally consider it to be, as in Raju's manuscripts [73, 74].

JO: Exactly—this is where the decolonisation idea starts to occur in the literature, in Raju as you say, and more indirectly in books like those by Joseph [58].

NB: In fact, Raju outrightly calls it "academic censorship" in 2017 [72]. The third part of the conversation about the effects of massification, in this context is the commodification of mathematics education, and more broadly tertiary matriculation; these issues are raised indirectly in the discussion related to paradigmatic ideas in Subsection 3.1, where I unpack schisms through theory and critical examples, discussing works by Beecher and Trowler [8] and Clark [31].

JO: Yes, people are grappling with this and it's difficult. There are all kinds of issues about privilege and justice and opportunity. It is not obvious to me what the answer is in choosing and designing what to teach to whom. I value being respectful of differences in people's goals. Yet I also have an awareness of the sociological priming which biases who has which goals in the first place.

The literature reflects these concerns. For instance, in the gender context, Hill et al. [54, p 22] cite an extensive body of research showing that in general "women are more likely than men to prefer work with a clear social purpose" and that "most people do not view STEM occupations as directly benefiting society or individuals". Yet if women are directed towards more evidently "nurturing" courses because of this understanding, it can preclude their access to STEM careers [85, p 21], and thus become a self-fulfilling prophecy. See also [2]. The point is perhaps even more potent in the context of class.

Many studies show the seriousness of the disparities in terms of numbers, which tend to be similar worldwide. For instance regarding women, "In the United States, for example, 45% of undergraduate mathematics degrees are earned by women, 24% of mathematics PhDs go to women and women constitute 17% of tenured university mathematics faculty" [78], p 972. Comparing this more broadly with academia generally "The proportion of women among full-time faculty in US colleges and universities peaked at 36 percent in 1879, declined to 22 percent in the early 1960s (Bernard, 1964), and only surpassed its 1879 level in 2004 (AAUP, 2005)" [67]. Thus the under-representation of women in academic mathematics is part of, but more extreme than, a general under-representation in academia.

Every one of the JBCC speculations on causes is reflected in a large literature which also contains multiple strategies addressing these causes—for instance, see the 2017 UNESCO report [87]. As one example, in our allied disciplines of physics and computer science, Hill et al. [54, p 28] report “Research ... demonstrates how small improvements in the culture ... such as changing admissions requirements, presenting a broader overview of the field in introductory courses, and providing a student lounge, can add up to big gains in female student recruitment and retention”. Overall, from school maths through to mathematical careers, there is a narrowing gap between female and male participation and success [49, 54].

2.3 Fear

The theme of math fear was pervasive.

* I would like to speak to Question 2 (Can maths be taught so that the general population do not fear it?) because that is my favourite. I would like to relate two experiences, one about teaching math to my 9-year-old granddaughter, and another about teaching stats to women physicians in my course. For the students, I find that if they are successful then they are not afraid anymore. That’s what I try to figure out—how to teach so that they understand something, and can successfully do a data analysis or something. In the case of my granddaughter, she had missed something about what place value means, which meant she was struggling with rounding. I took her to lunch and she got a fortune cookie, crossed out the fortune and instead wrote, “I hate math” and said “See granny that is my fortune”. When I found out what her gap was and she understood, then she loved math and she didn’t fear it and didn’t hate it anymore.
* When students are successful and genuinely understand, the fear and hate go away.
* A professor of Maths Education and former maths teacher once told me that when school students plaintively asked her, “When am I ever going to use this?”, she learnt that her best response was “So when did I lose you?” She said it might have been two weeks back. Whenever it was, once the student understood again, they no longer seemed to care about the “When will I ever use this?” question.
* A lot of teachers themselves have math anxiety. We need to do a better job of making the fear go away for the teachers.
* Something like two-thirds of maths teachers in junior high school do not have mathematics training and are insecure in what they are doing.
* What is the cause of maths fear? I wonder if our practice at University, of always setting maths exams in which it is possible to get 100%, is partly to blame? A grade distribution that includes 100% is unusual, compared with the humanities. It could be responsible for the common societal view that some people are just really smart, and those people can do maths and get 100% every time. While the rest cannot. Such a dichotomy could be driving fear.
* There is something that I call the *Dark Side of Mathematics*. I think as a community, we need to at least be aware of this part of mathematics.

* It is important to look at adult numeracy in men and women and think about that in terms of levels of maths anxiety in people . . . there is a case to be made that there is significantly more "maths anxiety" for girls.

* "Maths anxiety" in girls in schools is something that has been happening since way back; it is leading to this issue, and I don't think that we have been able to crack that.

NB: So if you were to distill this commentary down to one overarching topic, theme or impulse, it would be *math in the community*: how math is conceived in a community, fear of math at a cultural level and a communal level, sub-stratified by many cultures of math and education, all of which encompasses individual, personal, family and community engagement.

JO: It is also very much about "school maths", and that has a very mixed reputation.

NB: Some children are naturally adept at math, but they get overly stressed by speed tests; some adults do too! It has this incredibly negative feedback loop even for advanced children. In fact, it can have just as much of an adverse effect on advanced children as it does on a child who might struggle. Obviously there is a place and time for this type of test, but judiciously used.

JO: The comment about the "dark side of mathematics" intrigued me, and since the speaker did not get a chance to elaborate on the day I made contact and asked for clarification afterwards. The speaker wrote: "For me the *dark side of mathematics* is mathematics as a source of fear and hate; mathematics used as a tool of judgement; and mathematics used as a tool of manipulation." Detailed elaboration from the speaker on all three aspects includes the following:

- "Almost in every class that I visit, and I visit Grades K-12, there is a group of students that act like they are somewhere far away from that math classroom. I understand this as a demonstration, in a passive way, of dislike and animosity towards mathematics."
- "In my view, the mathematical academic totem pole has a strange power to influence the sight of some of the mathematicians/math instructors on it: when those mathematicians/math instructors look up the pole, everyone seems very close, almost equal to them, but when they look down everyone seems very, very far away."
- "Supporting political or other arguments by lists of numbers, calculations, and tables that are incomprehensible for the general population is, in my view a *dark* way of using mathematics."

I find this view and these examples very insightful.

Expressions of a combination of fear and hate are strewn through decades of the literature [25–27, 41, 42, 44, 46, 55]. Hersh and John-Steiner [53, p 303] describe a survey in which 40% of adult participants said that they hated mathematics. By contrast, in the same survey, 25% said it was their favourite subject, a fact that shows the complexity of the issue. Swan [83] quotes an example in which an adult woman describes her childhood experience: *"I dreaded Friday mornings as this was the time* *for speed maths .... Whatever was not completed in this time had to be written out in full along with the answers ten times each. ... began to get me in tears before the weekly exercise task had even begun."*

Further, the comment that many teachers do not have proper maths training is only part of the story of fear—it is not just qualifications that count. The book by Liping Ma [65] describes how certain teachers of mathematics in China, with generally lower formal qualifications than their US counterparts, nonetheless had a generally deeper and more accurate understanding of the mathematics they taught, due to regular discussions with their colleagues about the meaning of the material and how best to teach it.

2.4 Transformation

This section is about change, both the kinds we can guide and choose, and the kinds that occur without our choice. The section is also about cause for hope, and the desire expressed by participants for renewed and active stewardship of mathematics.

* As Mathematicians we need to evolve our teaching. In a former life, I was teaching a tough crowd of Biology and Business Majors in America, whom I had for a class in Calculus. I really invested in those students and found ways to make the mathematics meaningful to them. They appreciated it. Then when I was teaching Integrals to Mathematics Majors, after I had done the official part of it I said, "You know what guys I am going to give you the same pitch I gave to my applied class. This is the picture of what it really means." The students were very appreciative. They said, "Why didn't we ever do this before?"
* There are so many examples in a History of Mathematics course or any Humanities course to explain things, but it is never done.
* Neither wants to talk to the other.
* In Physics courses there is an opportunity to help Mathematics by saying "Here's a real application of Mathematics". This often does not happen—the course is designed in such a way that the mathematics is not needed.
* We need to be incorporating IT with Math Education and the Mathematics Community, or is that a very good example of a schism?
* Everyone is guarding their own turf: its silos.
* The Mathematics Majors whom I was teaching: at the point I gave them the applications-based pitch, they already liked maths. It is not going to hurt to make them like it more!
* I have recently been teaching maths in schools on Practicum as part of teacher training, which I am doing alongside my current role teaching and researching mathematics at University. I can tell you that it is not enough to have a good grasp of maths and a love of maths, in order to teach it well in secondary schools. A big factor is, the school environment is so driven by external tests and speed.
* It's a time race.

* Our system encourages school students to rote learn.

* It is all about responding as an automaton, the students aren't really learning mathematics.

* Listening to people, it seems that there is a general agreement that there is too much testing and getting answers for exams, which turns students off. It would be better if we could get them using tools and exploring mathematics, as well as more mental arithmetic, which can be turned into a game—of which there is some discussion in the book "Surely you're joking, Mr Feynman". At present many students don't do maths because they are afraid that they won't get good marks.

* There's a move against it now. The HSC[1] is dead, it just hasn't fallen over yet.

* As a professional research mathematician, I find myself uniquely qualified or unqualified because I have never had a teaching appointment in my career, but I have seen at first hand through the experiences of my four daughters and what they were struggling with when they were learning mathematics at school.
I was shocked when they said that they had a homework problem to calculate the asymptotes of a tilted hyperbola. What in the world do you need this kind of thing for? That was part of the curriculum when I went to school and I was shocked that this had not changed. With the huge wave of technology, it *is* shocking how little has changed. I personally think that technology of the computer offers a compelling new platform for teaching mathematics.

* The whole process of discovery by the very best mathematicians—in our quest for purity, we have erased that experience. Why can't we resurrect that and make it part of the mathematical curriculum?

* That motivation and discovery and excitement, as Jon talked about—it would go a long way to getting us out of the stone ages of mathematics education.

* If you have not got good teacher training, then you will not have teachers who are able to embrace new areas of mathematics. There is a huge inertial force. It all comes down to having a very well-trained versatile teaching force.

* Somehow the mathematics profession has to recognise the need for ongoing math education for teachers.

* Teachers do not have time to "sit and reflect".

* At present, teachers do not have to be confident in numeracy or love maths to end up teaching maths in schools. How do we help?

* We have to recognise that teachers need more opportunities to improve as maths teachers. I think that they want to: it's a political issue.

* Teacher training has become over-regulated and beholden to external accreditation. Specific qualification courses "squeeze out" space.

* There is a place for more involvement from Maths Faculty in training teachers how to teach maths.

* We have to be good stewards of our field.

[1] Australian "Higher School Certificate" exam.

* It is a duty of a mathematician to be a custodian of the discipline, to do outreach and to get into schools.
* Outreach takes a huge commitment. One of the sad things is that the ability of research mathematicians to get into schools and do outreach is being squeezed out, because of the huge pressure to have success in research within one's own faculty.
* Most academics are academics because they love their discipline and they want to do outreach, but it is a role that has dwindled. It is bundled up in professional silos. It has become a business and whatever becomes a business gets filtered.
* Most people really underestimate the effect of portraying mathematicians and tech people in the movies, overwhelmingly caricatured in ridicule and stereotyped as overweight socially inept figures. People see this and think "this is definitely not my crowd".
* We have to get through to Hollywood.
* Helping with the "Life of Pi", which Jon did, and helping with outreach, has great power.

The closing anecdote of the Education-Led Panel Session resonated with hope, a sense of "not giving up on people", and the power of curiosity.

* I was visiting this outreach program in a small school on an island off the coast. The teacher said, "These are kind of slow kids so don't expect too much". This kid asked me if I had heard of *Graham's number*. I said "Yes I know *Graham's number*." I asked "What do you know about *Graham's number*?" The kid had found out about it from the Internet and said "it's the biggest number used in mathematics".

NB: What we're seeing here is really a whorl of ideas that embody the notion of transformation. We are seeing different angles from which math can be viewed, both theoretically and in a concrete context, in relation to neighbouring disciplines and fields. There is a migration of mathematics across disciplines, such as computer science.

The accompanying cordoning off and bifurcation of community is an example of what I term a "passive schism". These neighbouring disciplines have meaningful contributions to make to each other, and can be part of building a richer math education community together as a group, but the "passive schism" keeps them from enriching each other.

JO: Thinking about the computer science / mathematics schism leads me to a mathematical insight and related minor teaching epiphany: maybe we are not teaching the most generally useful algorithms for addition and multiplication at school? The point for me is that we are bionic people, extended by our tools. Thinking as a numerical analyst shows that the difference between the algorithms is the requirement for working memory. School algorithms are suitable when paper is ubiquitous since it extends working memory. But when the hands are busy, school algorithms are next to useless, which may explain why they are not retained.

Silos as a theme tie straight back to the previous discussion on fear, because it relates to meaning. I think that the pure versus applied division or schism is terribly important, and I think that teaching in a decontextualised way increases fear and mystification for many students. In a way it is part of maintaining an elite, closed-off, inaccessible mathematics (even though that may not be what we intend).

NB: Yes, some of these ideas about the nature of what an algorithm is and how that interfaces with our understanding of the body–mind synergy (or divide) in relation to math teaching and math tools has been explored in different capacities by Paul Drijvers, Ulrich Kortenkamp, Nathalie Sinclair, Richard Noss, Celia Hoyles, John Monaghan, Luc Trouche, and Jonathan M. Borwein. With the exception of Jonathan and Ulrich, much of this research has been done under the auspices of mathematics education—a silo. But, on fear, mystification and mythification, I will note if you go to a conference in mathematics, it is accepted that you will only understand a fraction of any topic that you hear because they are so specialised. There is an acceptance of that kind of esoteric culture.

JO: Indeed, and the trouble is that the language and notation can be exclusionary. That's part of what Jon was on about when he wrote about and used visualisation extensively. He said in one of his talks, "pictures are more democratic, but they come from formulas".

NB: Math has its own life and power. As many scholars note, it is a language with many dialects, much like any other language: for example, English, Spanish or French. There is a distinction between math and numbers, which in my mind I see as existing in nature, in a philosophical sense. If we think about it or apprehend it that way, then it should be something that everybody can access. The commentary about Jon and the Life of Pi reflect his unique philosophy of math education. He was trying to address a much broader spectrum of people than conventional math education as a discipline would expect or understand, in terms of disciplinary norms.

JO: I thought everyone was trying to teach everyone everything.

NB: But there is a very careful selection of who learns what.

JO: The school curriculum as documented is about universal access.

NB: None of those provisional doctrines or legislation actually make it happen.

JO: As the JBCC participants say, what is also critical is a well-trained flexible teaching force that is willing and able and empowered to introduce new mathematical thinking into the curriculum.

NB: When Jon wrote about *Homo Habilis Mathematicus*, he was advocating a new way of thinking about math education. Indeed, Jon labels the experimental mathematician *Modern Homo Habilis Mathematicus* where Australopithecus meets Homo Aestheticus with digital literacy and math tools. The term was playing on the extinct human ancestor, but it was also embodying the modern experimental mathematician as a math educator caught between two cultures [14]. For Jon, those cultures were old and new: Enlightenment era versus modern. But for me, and from conversations I had with him while he was writing

"The Life of Modern Homo Habilis Mathematicus: Experimental Computation and Visual Theorems", (Chapter 3 of *Tools and Math*), he was also playing with the idea of how experimental math sits between the mathematical research culture and the culture of math education. That's because experimental math is an education tool.

JO: Yes! It has been claimed by Davis et al. (2015, p. 61) that conditions are in place for transformation in mathematics education on the scale that accompanied the industrial revolution.

Transformation can disrupt divisions. The divisions noted in the JBCC conversation are reflected in the literature. For instance, even back in 1973 the renowned mathematician and educator Hans Freudenthal [47, p 72–73] discusses: a "dematematized physics", and how at a conference on teacher training the suggestion that computers be included was "met with stony silence". Yet he calls this a "pure accident. If a few computer people had been present at the conference, they would have hooked in on the subject and proposed a complete course in numerical mathematics".

The division or schism between pure and applied is part of the story [66]. Lave studied mathematics in "everyday life" and critiqued the hegemonic discourses which treat the "everyday" as less than the "scientific" [64, p 78]. Freudenthal [47, 69–73] writes about how mathematicians have a tendency to teach as though their students will all become copies of themselves, and how this can lead to a deification of beautiful isolated systems, with a likelihood that this will "stimulate an aversion to mathematics". In the same section, Freudenthal decries the associated omission of applications from swathes of mathematics teaching.

There is a disconnect between school maths and the kinds of maths that people use in their jobs or in making purchases in the supermarket, as was shown by studies by Lave [64] and others cited within in the 1970s and 80s. In these studies there was no correlation between (typically low) fluency in school maths-type tests, and (typically high) fluency in the maths they used in their lives. One difference turned out to be the algorithms. Participants were using techniques suitable for mental arithmetic, without the aid of paper (because they were working in circumstances that needed their hands for other things).

Lewis Carroll [29] wrote about the difference between algorithms suitable for pen and paper calculation versus mental arithmetic in his *Pillow Problems*. He says that for problems such as the "multiplying together of two numbers of 7 digits is no doubt best done, on paper, by beginning at the unit-end, and writing out 7 rows of figures, and adding up the columns in the usual way. But if would be very difficult indeed—to me quite impossible—to do such a thing in the *head*. The only chance seem to be to begin with the *millions*, and get *them* properly grouped; then the hundred-thousands, adding the results to the previous one; and so on. Very often it seems to happen, that the easiest *mental* process looks decidedly lengthy and round-about when committed to paper."

Major change is possible and the literature supports this. Davis et al. [36, p 61] wrote in 2015, "we close this chapter by observing once again that each of the major transitional moments in school mathematics was, firstly, triggered by significant socio-economic shifts and, secondly, enabled by new ways of thinking about knowledge and learning. We believe that both these conditions are currently being met in Western societies. On the former, few would dispute the suggestions that current socio-economic evolutions are on a par with those that accompanied the Industrial Revolution. On the latter ... theories of embodiment are emerging from cognitive science and other domains that challenge current conceptions of thinking and offer new insights into mathematical knowing and learning."

The literature also supports the fact that the philosophies of mathematics education are not what they were before. They have changed, hence have the capacity to change again. For instance, the work of Seymour Papert may have seemed radical 60 years ago, but it is now being embraced by eminent mathematics educators such as Celia Hoyles and Richard Noss, [56, 57]. Similarly, Jonathan Borwein and people he worked with and mentored such as Nathalie Sinclair were/are all swimming against the stream of the common philosophy that views mathematics as disembodied and universal [52]. Lave [64, p 78] is doing the same when he writes "The dichotomy between mind and body underlying Western epistemologies provide the framework for a similarly dichotomized sub-classification of rational and scientific modes of thought in opposition to primitive, non-rational, or irrational ones." Lakoff and Nũnez [63, p xiv] do similar when they write in 2000 of mathematics that "One of the great findings of cognitive science is that our ideas are shaped by our bodily experiences—not in any simpleminded one-to-one way but indirectly, through the grounding of our entire conceptual system in everyday life." Further, Lakoff and Nũnez cite Reuben Hersh as a major influence on their thinking, as he was for Jon [22].

The above synoptic review of various bodies of literature that bolsters and accompanies the major discussion points (or threads) by panel members. We are now going to explore these threads in terms of schisms, divisions and disciplinarity.

3 Schisms: A Theoretician's Voice

Thus far in this paper we have foregrounded "problems", with brief references to the theory of schisms, divisions and disciplinarity. Now we do the opposite and foreground the theory, with brief indications of where it may assist with mitigation of these issues. This approach is the result of the ongoing discussion in this paper on "problems in motion", which do not have fixed once-and-for-all solutions. Hence this work is at least as much about methodology for finding solutions, and resolutions, as it is about solutions themselves. In this section, we begin to more rigorously address the dynamic theoretic issues at work.

Lexicographically speaking, the *Oxford English Dictionary* [81, p. 1286] defines

> A *schism* is a division, separation, disagreement, discord or disharmony between two groups–often tinctured by political agendas of factions.

Schisms permeate Mathematics Education problems in ways that we cannot ignore if we want to solve them. The issues raised within the JBCC Panel Discussion are important and have been intransigently difficult. They have a long history and literature, throughout which is woven one or more schisms that impedes their resolution.

This section provides the groundwork for interpreting the final section, which then delves into the relationship between experimental mathematics methodology, humanistic philosophy of mathematics, and education.

3.1 Schisms, Theory, and Critical Examples

There are complicated factors to consider in critically examining schisms and divisions, including the vectors of culture and society at large, territories and terrains of academia, issues of interdisciplinarity, increasing sub-disciplines, conflicting epistemological frameworks and more. The literature is rife with examples of the math education schism. Each variant showcases a different permutation of the conflict, in different institutional systems, and across pedagogical, cultural, or individual levels.

- Mike Rose's 2012 article on the "Academic-Vocational Schism" [77] relies on the notion of practical versus conceptual value.
- Bharath Sriraman and Günter Törner examine the mathematician-didactician schism [24, p 662].
- Conversely in 2013, Peter Sullivan et al. [82, p 476] examine "this apparent schism"; they see visible traces in the Australian curriculum, and approaches to teaching in the classroom.
- Then, others scholars, like C. K. Raju and his article "Epistemic Divide in Mathematics" [74, p 1], extend the ideas of Martin L. Abbott et al. [1] in *Winning The Math Wars: No Teacher Left Behind* (on issues of the US curriculum, and conceptual knowledge for teachers). This is also seen in the patterns of disciplinarity in academic culture that burgeoned out of C. P. Snow's seminal *Two Cultures* [80], as briefly noted in our discussion section. The list goes on and on.

Broadly speaking, it is necessary to contemplate the idea of academic tribes and territories. Scholars have written extensively on the nature of academic schisms, outside math(s)/education. What can we learn about these divisions from critics like Collini [32], Clark [31], Becher and Trowler [8], Trowler et al. [86], Nealon [71], Monroe [69] and Davidson [35]? As Enzensberger intimates in *Drawbridge Up* [46, 70], such divides are rooted in long-standing dualities, sometimes attributed to enlightenment ideology and praxis. This has deeply influenced the twentieth-century critical idiom; consider Thomas Kuhn's [62] revolutionary work on scientific practice in the humanities to unseat the binary of science and arts in the era of C. P. Snow [80]. The underlying trends within academic schisms can be thought of first through this duality, and complicated by multi-, trans-, or inter- contemporary approaches.

To quote Collini's introduction to the 2012 fiftieth anniversary edition [80, p xiiii] of Snow's *Two Cultures*:

> At the heart of the concept of the "two culture" is a claim about academic disciplines. Other matters are obviously intimately involved—questions of educational structure, social attitudes, government policymaking and so on. But if the concept is to possess any continuing persuasiveness it must offer an illuminating characterisation of the divide between two sorts of intellectual enquiry.. . .The map of the disciplines [is constituted by] contradictory, or at least conflicting, forms. . .[complicated by]. . .the sprouting of ever more specialised sub-disciplines and growth of various forms of interdisciplinary endeavour.

Such a commentary touches on sweeping trends that impact schisms. Many of these concepts are important for understanding academic schisms, as pertinent to Math(s)/Education. Building on the work by Trowler et al. in 2012 [86], we see major shifts in the topology of academic knowledge, and the fields in which such knowledge lies. These are reconstituted by

1. changing landscapes, shifting territories,
2. levels of analysis,
3. "massification and marketization" (Trowler [86, p 14]).

In *Creating Entrepreneurial Universities*, Burton Clark [31] calls the growth in knowledge and the subsequent explosive growth in disciplines and their fragmentation into sub-disciplines the most important change affecting "massification and marketization" ([31]; [86, p 14]). In the shaping of fields, patterns of growth and fragmentation, alterity, politics, performative subjectivity, conflicting methodologies and the clash of subcultures (research, ecumenical and otherwise) come to bear.

3.2 *The Micro and Macro Level*

It is important to note that the divisions considered in the previous subsection are visible at both the micro and macro level. After considering the work of Collini, Snow, Trowler, Beecher and others, imagine how these patterns show up in relation to mathematics education and research practice at the micro level of the regional classroom and in the academic discourse at the transcultural/transnational level. "This apparent schism" in math/education is visible at the micro level of the Australian classroom. For example, take the implementation of the new Australian Curriculum Mathematics [3] with a focus on 1) cross-curriculum priorities (e.g., Aboriginal and Torres Strait Islander histories and cultures), and 2) general capabilities (e.g., numeracy, creative and critical thinking). Sullivan et al.'s 2013 article "Processes and priorities in planning mathematics teaching" [82, p 478] expands on this issue:

> These findings suggest that there is a disjuncture between teachers' conceptualisation and articulation of curriculum knowledge, and the ways in which it is represented in formal curriculum documentation. This apparent schism has import for both classroom practice and systems-level policy development and dissemination. Certainly, it suggests the need for further research on how teachers express and employ curriculum knowledge. There is also a need to work with teachers to develop ways of engaging with curriculum documents that assist them in articulating important ideas to their students.

Note that the above quote is the only example of the term “schism” appearing in the JRME database search, and it is in relation to the Australian curriculum. This rarity suggests the tenuous place discussions of such a schism have in established math education discourse.

Applied to the classroom and the translation of curriculum-based knowledge for teachers, it also reflects the same disjunct between or implementation of mathematics epistemology, but at the micro level of the Australian classroom. Issues that arise from the same basic dynamic disjunct are visible at the transnational level.

The politics of math education can be seen at the macro level of international culture. Here, divisions in math education take on vital transcultural considerations of identity, community, minority and disability—echoing concerns about schisms voiced at the JBCC panel discussion. Take the following example explored at Topical Survey Group 34 of the ICME 2013 [60]. The resultant 2016 study *Social and Political dimensions of mathematics education* by Murad Jurdak et al. [59] isolates and delineates the social and political impact of what they view as five critical areas in mathematics education.

The first two areas of [59, p 2] involve a) the equitable access and participation in quality mathematics education that focuses on ideology, policies and perspectives “in different contexts and from different ideological perspectives” and b) activism and the material conditions of inequality: the correlation between “achievement gaps”, and “theory gaps”.

The third area of [59, p 2] is representative of distributions of power and cultural regimes of truth: “It goes further to ascertain the critical role of mathematics education research in addressing key concepts such as mathematical literacy or modelling. It concludes that the contributions on the political nature of mathematics itself provide new insights into the political bias of the mathematics in the classroom”.

The fourth area of [59, p 2] revolves around mathematics identity, subjectivity and embodied dis/ability: “emphasis on language and discourse informs this research, and how new directions are being pursued to address the diverse material conditions that shape learning experiences in mathematics education”.

And finally, the fifth area of [59, p 2] pinpoints the importance of economic factors behind mathematics achievement: “the influence of national and global economic structures”. Fundamentally, a convergence of western and non-western educational systems with increased global mobility in research and social domains has led to an influx of diverse mathematics cultures and identities; in turn these have increased the necessity to renegotiate the terms of good practice. The survey by Murad Jurdak et al. [59] exposes the nexus of issues related to schisms at the macro level.

3.3 Theory and Interdisciplinarity

This section is a theoretical look at *issues of interdisciplinarity*.

Some basic issues arise when navigating interdisciplinary collaboration, teaching and research. Recently in a book on collaborative research, Marilyn Deegan and Willard McCarty [38] and their contributors describe key areas where different epistemological approaches are susceptible: visible in this is a gulf between disciplinary practice and personalities, lack of communication, perceived ownership of disciplinary material and practice, and divergent outcomes and expectations:

- "[T]hings have gone horrendously wrong, often between individuals who are supposed to be working closely together, but cannot bridge the personality—or discipline—clash to understand each other's approach" [84, p 222].
- Failures (the sustenance of schisms) "stem from a lack of communication. Perceived slights of status or disputed "ownership" of published outcomes and this would include ownership of content of the discipline [84, p 222].
- "Those in charge of managing either side of the research have no real understanding of the other discipline and require repeated correction of the same facts. . .which cumulatively" impact work outcomes content [84, p 223].
- Lack of identifiable outcomes, "huge differences in what are perceived as acceptable outcomes" [84, p 223].
- Understanding what a tool is, whether it is a practical "working tool" or juxtaposed to excogitation and theorising about a potential working tool [84, p 223].

This schema of issues relates to work by Cathy Davidson [35], Trowler and Becher [8] and others who all note that

- "[N]arratives of success, failure, compromise, change, and complication are, of course familiar to anyone pioneering interdisciplinary structures" [35, p 216].

Math education, as an inherently interdisciplinary field and practice, falls prey to these same basic division lines and complications. Likewise, clearly, transformation, integration, and mediation are central to advancing any multidisciplinary field. In *The Oxford Handbook to Interdisciplinarity* edited by Frodeman et al. [48], Cathy Davidson extrapolates on where conflicts and issues arise between teaching and research practice, building on the work of Rhoten and Pfrman (2007) [75]. These include intrapersonal cognitive connections (cross-fertilisation of ideas and methods between fields or disciplines), interpersonal collegial connections (team collaboration, teams or networks that span fields and disciplines), interdepartmental cross-field connections (field creation topics that sit at the intersection or edge of multiple fields or disciplines) and stakeholders with community connections (problem orientation multiple stakeholders missions outside of academia, such as those that service society) (p 389). Such areas of conjunction and multidisciplinarity affect math education in theory and practice. All of the aforementioned levels potentially give rise to conflicts or misunderstandings.

3.4 Standard Paradigms and Patterns

Of the myriad forms of math/education schisms, they all present the same standard paradigms and patterns that lead to difficulties. These include cross-cultural conflict, divergent desired outcomes, a clash of academic or bureaucratic or scholarly and mainstream needs and agendas, misapprehension of each other's epistemology and field, resistance to acknowledging disciplinary norms of other disciplines, an intellectual disparagement of the humanities side of the dynamic (see Davidson [35], Collini [32], Snow [80]) and combative response from the humanities. These standard patterns or areas of mistranslation are all complicated by what Trowler et al. [86] more broadly explores as changing academic landscapes, shifting territories, discursive patterns of growth and fragmentation, and "massification and marketization" of academic disciplines in general (Clark [31]; Trowler et al. [86, p 14]). Overwhelmingly, there is a call to understand the disciplinary practices and traditions of other fields, to be the other, as *structural and conceptual alterity* in math/education. Drawing on models posed by scholars like Cathy Davidson, Melissa Terras, Monaghan, Troche, and Borwein, this requires a requisite ability to understand and internalise the needs and viewpoints of other scholars, students, peers or groups, in the following six ways:

1. across disciplinary divides;
2. to create clearly defined "research expectations, publications, training, and project management" ([84, p 223]);
3. an increasing knowledge of tools employed by both sides of the divide: whether these are computational environments or abaci;
4. developing a communal nomenclature between mathematicians and mathematics educators, and on
5. devising practical ways to decrease tensions that accrue based on technical and non-technical issues as a way of alleviating resistance to adopting other disciplinary methods and tools; and
6. constructing real-world problems (applied) that integrate the conceptual and theoretical (pure), thus making interdisciplinary frameworks functional in working or teaching environments across institutional levels ([84, p 224]).

(See C. Davidson 2010 [35]; Deegan and McCarty 2012 [38]; Rockwell 2012 [76]; Terras 2012 [84, p 220–227]; Monaghan, Troche and Borwein 2016 [68].)

Jon's Coda, which follows, is a thoughtful extension of ideas Naomi first delved into in her Plenary Talk "From Lipschitz to Homo Habilis Mathematicus: a case study of Jon Borwein" delivered at the *Math and Tools: Instruments for learning* workshop in Sydney, Australia on 29 Tuesday, November 2016. The section gives biographical context to Jon's research, in order to illuminate and expand upon his philosophy, practice and many possible legacies.

4 Jon's Coda

Jon was a vocal advocate and disciplinary father of modern (computational and) Experimental Mathematics; the field is in fact both an educational and heuristic tool and a philosophical way of doing and experiencing mathematics. It sits on the boundary of pedagogy and research, and between proof and artefact. Jon's method and research has impacted math education pedagogy despite the many disputes over digital-assisted research and cross-disciplinary scholarship–hotbeds for “schisms”. Take the example of ICERM’s report based on results from an important experimental math conference held in 2011 [7]. The ICERM report on the conference states that perhaps the most succinct definition of what exactly “experimental mathematics” is, is given by Jon Borwein and Keith Devlin in *The Computer as Crucible* [39], which made Klaus Peters’ vision a reality.

An example of the deep and transformational collaboration that Jon was involved in is given by the outcomes of the ICERM conference [7]. A large number of people attended. They had working groups that approached all the different levels of transparency and methodology and these groups came up with suggestions on how to improve experimental math as a discipline. After these groups discussed their findings, one from each group presented theirs. They had a huge round-table discussion (itself a valuable education tool), and they came up with criteria for the report—and chose people to write the report, to fact check and finalise the document. The central thrust of the report was a demand for transparency in experimental methods for research and education. There was a call for data and methods used in published papers to be accessible online so they could be fact checked. Their basic mandate is linked to re-inventing mathematical education through technology: “Providing evidence-based rationales for experimental mathematics in the classroom,” using computer-based tools and direct, hands-on experimentation, to build a new generation of researchers: computer-based mathematics for a cadre of twenty-first century computer-savvy students eager to press forwards with these tools.

This vision for transformation is expressed in major reports. The ICERM report [7, p 4] concludes “. . .the entire process of mathematics education, from elementary to advanced levels, needs to be rethought. At the least, much ‘experimentation’ will be required to see which approaches really work”. In the ICMI report, the definition of “experimental mathematics is the use of a computer to run computations—sometimes no more than trial-and-error tests—to look for patterns, to identify particular numbers and sequences, to gather evidence in support of specific mathematical assertions that may themselves arise by computational means, including search” [50, p 1].

In a sense, more than some other fields, experimental mathematics is a perfect marriage of tools and math, research and education.

4.1 Experimental Mathematics in Education: Beauty and Philosophy

Jon was fond of the following expression from Hardy [51] and quoted it often:

> Beauty is the first test; there is no permanent place in the world for ugly mathematics. (G H Hardy, *A Mathematician’s Apology*)

Jon’s approach to research was in part a product of his rearing—his mother was an anatomist/botanist and his father was a summabilist. Jon also studied the Philosophy of Mathematics at Jesus College. However, two early pivotal research periods strongly shaped Jon's engagement with mathematical tools and education.

- In the mid-1980s:
  Jon took a sabbatical in Limoges, and Trinity College, Cambridge. The Cambridge stay provided work with mathematicians such as Preiss, the excellent old math library at Trinity College and an opportunity to meet Yasamasa Kanada of Japan. Limoges opened up many fruitful and long-standing interactions with mathematicians worldwide, facilitated by Michel Théra. Jon also worked on the *Collins Dictionary of Mathematics* with Ephraim Borowski [18], using the first Casio portable and file cards on the grounds of a Chateau near Rilhac-Rancon during his stay at Limoges.
- Moving to SFU and starting the *Centre for Experimental and Constructive Mathematics (CECM)*:
  The second pivotal moment coincides with Jon's move to Simon Fraser University (SFU), the founding and directorship of CECM [30] in 1992 where being awarded the Shrum Chair of Science allowed him to extend his ambit to the Science and Art of Math with proper funding. That is, it allowed him to develop a mathematics research methodology based on scientific discovery.

Jon's many other initiatives have been included, but were by no means limited to *The Organic Maths Project* [21], *The Dalhousie Distributed Research Institute and Virtual Environment (D-Drive) at Dalhousie* [34] and the *Canadian Coast to Coast Seminar Series (C2C)* at SFU [28], followed by the *Priority Research Centre for Computer Assisted Research Mathematics and its Applications (CARMA)* [33] at the University of Newcastle, Australia. In the translation from D-Drive to C2C then CARMA, Jon's use of Big Data, satellite technology and Big Images, continued to evolve. These periods and events time stamp important moments in the development of Jon's research aesthetic and tool set.

Jon's engagement with visualisation, and the nature of beauty, stem from a humanist philosophy of mathematics, which as a movement "has bloomed" in the last two decades and been a direct influence on "mathematics teaching" (Davis, Hersh and Marchisotto, [37, p xv]). "The idea of mathematics as a human creation has been advocated many times, by Aristotle, by the empiricists John Locke, David Hume, and John Stuart Mill, and by many others" (Hersh [52, p 182])—its relationship to sociocultural and historical "human experience" as a metaphysical concept. Such a dichotomist rationale is at the base of, or spans, the Math Wars divide, through perpetuation of old distinctions between science and arts. As Jon practiced it, his humanist philosophy of experimental mathematics was trussed to computer-assisted computational interfaces, exploratory scientific observation and approaches, and visualisation as beauty-proof-transcendent form. All of this amounted to "insight" and intuition.

This acknowledgement of the importance of "insight" in mathematics education has resulted in volumes being produced across decades that extend from teaching how to do "experimental mathematics" to lesson planning in works like "Dynamic Composition of Math Lessons" (Kellar, Borwein, Watters et al. [61]), "Digitally Activated Mathematics for a Brave New World Wide Web" [20], and *An Introduction* *to Modern Mathematical Computing* (Borwein and Skerritt 2012 [23]). The reach of his pedagogy was not lost on ICERM [7], which acknowledged the important pedagogical role of books like Devlin and Borwein's *Computer as Crucible* [39] in the development of current best practices for teaching (2012). Indeed, one of his later publications, co-authored with John Monaghan and Luc Troche, explores *Tools and Mathematics: Instruments for Learning* [68].

4.2 From Littlewood to Dawkins

Jon envisioned a mathematical world in which John Littlewood and Richard Dawkins should collide, reflective of an approach where visualisations meet scientific method. Long before current graphic, visualisation and geometric tools were available, John E. Littlewood (1885–1977) wrote in his *Miscellany* [9] another caution that Jon was fond of quoting [14, p 23]:

> A heavy warning used to be given [by lecturers] that pictures are not rigorous; this has never had its bluff called and has permanently frightened its victims into playing for safety. Some pictures, of course, are not rigorous, but I should say most are (and I use them whenever possible myself).

Jon notes "[a]esthetic criteria change: closed forms have yielded centre stage to 'recursion' much as biological and computational metaphors (even biological envy) have replaced Newtonian mental images with Richard Dawkins' 'the blind watchmake"', a chapter in a 2006 volume edited by Nathalie Sinclair, William Higginson, and David Pimm [11, p 23].

The idea of teaching through more romantic and humanist philosophy that takes into account history, science and arts influenced Jon's research sphere. For instance, Sinclair in *Mathematics and Beauty: Aesthetic Approaches to Teaching Children* [79] does just that; she partly bases her argument on E. Dissanayake's 1992 *Homo Aestheticus* [40], extending a rather Darwinian understanding of man's relationship to form.

Jonathan Borwein's work, like that of Nathalie Sinclair, highlights the blurring of boundaries between inductive and deductive reasoning, and natural sciences, arts and math: math as art, technology and science. And they continue to explore, as Jon did for decades, "what most working mathematicians experience" [16], a relationship to beauty, harmony and clarity [11, p 20]. It is more than an aesthetic or style problem; it becomes a methodological approach to visual aids in the classroom. (See also Borwein and Bailey's article "Why mathematics is beautiful and why that matters" [10, 16]).

Overarching facets of Jon's philosophy that are born out in his methodological approach to experimental mathematics in research and education include

i. a call for insight,
ii. intuition,
iii. experimentation,

iv. exploration
v. and accessibility.

Jon would add to these the ability to capture interest through intrigue and mystery, as well as promoting a strong sociocultural engagement with material.

These integral elements are strengthened by a Renaissance approach to education, emphasis on both science and the arts (be it Mathematics or Aesthetics) fortified by modern technology (science) and experimental mathematics. Through aesthetics and applications, the rigour and beauty of mathematical precision should, as Pólya thought and Littlewood less convincingly suggested, be brought into the classroom, hastened by a dose of intrigue, fun and social engagement. These elements were all a part of Jon and the way he presented mathematics, independent of platform, space or audience.

5 Conclusion

As Jon knew, a dialogue is sometimes the best place to start in order to catalyse change. Jon was actively trying to catalyse change in mathematics education, in many ways, which included but was by no means limited to publishing in the area. Everything Jon did was about trying to reach out and educate people about mathematics, and mathematics education was one of his primary focuses. He was constantly working on increasing stewardship and decreasing anathema amongst the mathematical community.

The three ultimate goals of this chapter are 1) to continue and promote an ongoing academic discourse or discussion about change, 2) to chronicle what transpired at the Education-themed session of the JBCC and 3) to analyse Jon's many possible legacies to mathematical education.

These legacies are deeply rooted in the heuristics and methodology Jon developed alongside his work in modern experimental mathematics—itself an education-research based praxis. Jon advocated experimental mathematics as a teaching tool.

The idea of a schism is in diametric opposition to the way Jon viewed mathematics education, the way he practiced research and the way he taught. This is a potential that the rest of the mathematical community has an opportunity to take up and develop.

The humanistic philosophy of mathematics that he engaged with is in fact part of a new cultural shift that has evolved along with the use of computational visualisation, digital-assisted learning and research. It is our understanding that Jon was a part of this cultural shift.

The animated discussion that took place during the Education-led Panel of the JBCC exposes the necessity of discussing diversity, hope and change, silos, meaning, but there would not have even been any discussion if it were not for the direct acknowledgement of underlying divides (or schisms) within math education, and a genuine interest in collectively and creatively addressing those divides.

Acknowledgements We thank CARMA and our co-organisers in the Jonathan Borwein Commemorative Conference (JBCC). Thanks to the many JBCC participants who gave their permission to be anonymously quoted. And thank you to Jon for a lifetime of boundary-breaking research and pedagogical exploration.

References

1. Abbott, M.L., Ferriso, B., Smith, K., Trzyna, T.: Winning the Math Wars: No Teacher Left Behind. University of Washington Press, Seattle (2015)
2. Annett, C.: Girls and Women in Sciences, Technology, Engineering and Mathematics. In: Hill-Notes: Research and Analysis from Canada's Library of Parliament by Clare Annett, Library of Parliament (2018). https://hillnotes.ca/2017/10/11/girls-and-women-in-science-technology-engineering-and-mathematics/. Accessed 28 Sep 2018
3. Australian Curriculum Assessment and Reporting Authority.: Cross Curriculum Priorities (2018). Retrieved from https://www.australiancurriculum.edu.au/f-10-curriculum/cross-curriculum-priorities/. Accessed 28 Sep 2018
4. Bailey, D.H., Borwein, J.M.: Experimental mathematics: examples, methods and implications. Not. AMS **52**(5), 502–513 (2005)
5. Bailey, D.H.: Jonathan Borwein dies at 65. In: Math Drudge: Two Mathematicians Contemplate the Cosmos (2016). http://experimentalmath.info/blog/2016/08/jonathan-borwein-dies-at-65/. Accessed 28 Sep 2018
6. Bailey, D.H., Borwein, J.M.: Mathematics by Experiment: Plausible Reasoning in the 21st Century. A K Peters/CRC Press, Wellesley and Baton Rouge (2004)
7. Bailey, D.H., Borwein, J.M., Martin, U., Salvy, B., Taufer, M.: Opportunities and Challenges in 21st Century Experimental Mathematical Computation: ICERM Workshop Report (2014). Retrieved from https://www.davidhbailey.com/dhbpapers/ICERM-2014.pdf. Accessed 28 Sep 2018
8. Becher, T., Trowler, P.: Academic Territories and Terrains, 2nd edn. Open University Press/SRHE, Buckingham (2006)
9. Bollobas, B., Littlewood, J.E.: Littlewood's Miscellany. Press Syndicate of the University of Cambridge, UK (1986)
10. Borwein, J.M.: Actually: Teaching and Research with Collaboration Tools and Technology (2011). https://carma.newcastle.edu.au/jon/aces11.pdf and https://www.austms.org.au/Publ/Gazette/2011/Jul11/CommsALTC.pdf. Accessed 28 Sep 2018
11. Borwein, J.M.: Aesthetics for the working mathematician. In: Sinclair, N., Pimm, D., Higginson, W. (eds.) Mathematics and the Aesthetic. CMS Books in Mathematics. Springer, New York (2006)
12. Borwein, J.M.: The experimental mathematician: the pleasure of discovery and the role of proof. Int. J. Comput. Math. Learn. **10**, 75–108 (2005)
13. Borwein, J.: Experimental mathematics leads to new insights. https://carma.newcastle.edu.au/jon/JMBinnov.pdf. Accessed 28 Sep 2018
14. Borwein, J.M.: The life of modern homo habilis mathematicus: experimental computation and visual theorems. In: Monaghan, J., Trouche, L., Borwein, J.M. (eds.) Tools and Mathematics: Instruments for Learning. Springer, Cham, Switzerland (2016)
15. Borwein, J.M.: Commemorative Conference, 2016. https://carma.newcastle.edu.au/meetings/jbcc/. Accessed 28 Sep 2018
16. Borwein, J.M., Bailey, D.H.: Why mathematics is beautiful and why it matters. In: Huffington Post (2017). https://www.huffingtonpost.com/david-h-bailey/why-mathematics-matters_b_4794617.html. Accessed 28 Sep 2018
17. Borwein, J., Bailey, D., Girgensohn, R.: Experimentation in Mathematics: Computational Paths to Discovery. CRC Press and Taylor and Francis Group, FL (2004)

18. Borwein, J.M., Borowski, E.: Collins Dictionary of Mathematics. HarperCollins, UK (2002)
19. Borwein J., Borwein P.: Implications of experimental mathematics for the philosophy of mathematics. In: Experimental and Computational Mathematics: Selected Writings. Perfectly Scientific Press, Portland, Oregon (2010)
20. Borwein, J.M., Borwein, P.B., Braham, S., Corless, R., Jorgenson, L.: Digitally activated mathematics for a brave new world wide web. Educ. Res. Perspect. **23**(2), 28–47 (1996)
21. Borwein, J.M., Jungic, V.: Organic mathematics: then and now. Not. Am. Math. Soc. **59**(3) (2012)
22. Borwein, J.M.: Osborn, J: review of ‘loving and hating mathematics’ by Reuben Hersh and Vera John-Steiner. Math. Intell. **33**, 63–69 (2011)
23. Borwein, J.M., Skerritt, M.P.: An Introduction to Modern Mathematical Computing: With Mathematica. Springer, New York (2012)
24. Bharath, S., Törner, G.: Political union/mathematical education disunion: building bridges in European didactic traditions. In: English, L. et al. (ed.) Handbook of International Research in Mathematics Education (2008)
25. Brown, M., Brown, P., Bibby, T.: I would rather die: reasons given by 16-year-olds for not continuing their study of mathematics. Res. Math. Educ. **10**, 3–18 (2008)
26. Burns, M. Illustrated by Weston, M.: The I Hate Mathematics! The Yolla Bolly Press, Covelo, California (1975)
27. Buxton, L.: Math Panic. The University of Michigan, Heinemann (1991)
28. Canadian Coast to Coast Seminar Series C2C. http://c2c.irmacs.sfu.ca. Accessed 28 Sep 2018
29. Carroll, L.: Mathematical Recreations of Lewis Carroll: Pillow Problems and a Tangled Tale. Dover Publications, Mineola, c1893 (1958)
30. Center for Experimental and Constructive Mathematics (CECM). http://www.cecm.sfu.ca. Accessed 28 Sep 2018
31. Clark, B.R.: Academic culture. Yale University, New Haven, Higher Education Research Group, Institution for Social and Policy Studies (1980)
32. Collini, S.: “Introduction” to Two Cultures by C.P. Snow. Cambridge University Press, Cambridge (2012)
33. Computer Assisted Research Mathematics and its Applications (CARMA). https://carma.newcastle.edu.au. Accessed 28 Sep 2018
34. Dalhousie University D-Drive Document Server. http://roar.eprints.org/281/. Accessed 28 Sep 2018
35. Davidson, C.: Humanities and technology in the Information age. In: Frodeman, R., Klein, J.T., Pacheco, R.C.D.S. (eds.) The Oxford Handbook of Interdisciplinarity, pp. 206–219. Oxford University Press, New York (2010)
36. Davis, B., The Spatial Reasoning Study Group.: Spatial Reasoning in the Early Years: Principles, Assertions, and Speculations. Routledge, New York (2015)
37. Davis, P., Hersh, R., Marchisotto, E.A.: The Mathematical Experience, Study Edition. Modern Birkhäser Classics, Boston (2012)
38. Deegan, M., McCarty, W. (eds.): Collaborative Research in the Digital Humanities. Ashgate Publishing Limited and Ashgate Publishing Company, England and USA (2012)
39. Devlin, K., Borwein, J.M.: The Computer as Crucible. A K Peters/CRC Press, Wellesley and Baton Rouge (2008)
40. Dissanayake, E.: Homo Aestheticus. Free Press, New York (1992)
41. Dowker, A., Sarkar, A., Looi, C. Y.: Mathematics anxiety: what have we learned in 60 years? In: Frontiers in Psychology **7**(42) (2016). https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00508. Accessed 17 Jan 2019
42. Dreger, R.M., Aiken, L.R.: The identification of number anxiety in a college population. J. Educ. Psychol. **48**(6), 344–351 (1957). https://doi.org/10.1037/h0045894
43. Dubiel, M., Heinrich, K.: https://cms.math.ca/Education/MallMath/. Accessed 28 Sep 2018
44. Dweck, C.: Is Math a gift: beliefs that put females at risk. In: Ceci, S.J., Williams, W. (eds.) Why Aren’t More Women in Science? Top Researchers Debate the Evidence. American Psychological Association, Washington (2006)

45. Engel, S., Halvorson, D.: Neoliberalism, massification and teaching transformative politics and international relations. Aust. J. Polit. Sci. **51**(3), 546–554 (2016). https://doi.org/10.1080/10361146.2016.1200706
46. Enzensberger, H.M.: Drawbridge Up. AK Peters, Natick (1999)
47. Freudenthal, H.: Mathematics as an Educational Task. D. Reidel Publishing Company, Dordrecht-Holland (1973)
48. Frodeman, R., Klein, J.T., Pacheco, R.C.D.S. (eds.): The Oxford Handbook of Interdisciplinarity, 2nd edn. Oxford University Press, Oxford (2017)
49. Hanna, G.: Reaching gender equity in mathematics education. Educ. Forum **67**(3), 204–214 (2003). https://doi.org/10.1080/00131720309335034
50. Hanna, G., de Villiers, M. (eds.): Proof and Proving in Mathematics Education: The 19th ICMI Study. International Commission on Mathematical Instruction. Springer, Dordrecht (2012)
51. Hardy, G.H. [1940].: A Mathematician's Apology. Cambridge University Press, Cambridge (2004). ISBN 978-0-521-42706-7
52. Hersh, R.: What is Mathematics, Really?. Oxford University Press, Oxford (1997)
53. Hersh, R., John-Steiner, V.: Loving and Hating Mathematics. Princeton University Press, Princeton (2011)
54. Hill, C., Corbett, C., and St. Rose., A.: Why So Few? Women in Science, Technology, Engineering and Mathematics. American Association of University Women, 1111 Sixteenth Street NW, Washington, DC 20036 (2010)
55. Holt, J.: How Children Fail. Pitman Publishing, Great Britain (1965)
56. Hoyles, C.: Changing the way people think, move and feel mathematically: the contribution of digital technologies. Panel (2016). https://carma.newcastle.edu.au/meetings/tools/#panel. Accessed 28 Sep 2018
57. Hoyles, C, Noss, R.: Visions for mathematical learning: the inspirational legacy of Seymour Papert (1928–2016). Educ. Forum, 34–36 (2017)
58. Joseph, G.G.: The Crest of the Peacock: Non-European Roots of Mathematics, 3rd edn. Princeton University Press, Princeton (2011)
59. Jurdak, M., de Freitas, R.V.E., Gates, P., Kollosche, D.: Social and Political Dimensions of Mathematics Education: Current Thinking. Springer Open, Switzerland (2016)
60. Kaiser, G. (Series ed.): ICME-13 Topical Surveys. Springer, Switzerland (2013). https://www.springer.com/series/14352. Accessed 28 Sep 2018
61. Kellar, M., MacKay, B., Zhang, R., Watters, C., Kaufman, D., Borwein, J.: Dynamic Composition of Math Lessons. J. Educ. Technol. Soc. Digit. Contents Educ. **6**(4), 100–111 (2003)
62. Kuhn, T.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago (1962)
63. Lakoff, G., Nũnez, R.E.: Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being. Basic Books, New York (2000)
64. Lave, J.: Cognition in Practice: Mind. Mathematics and Culture in Everyday Life. Cambridge University Press, Cambridge (1988)
65. Ma, L.: Knowing and Teaching Elementary Mathematics Teachers' Understanding of Fundamental Mathematics in China and the United States. Routledge, New York (2010)
66. Maddy, P.: How applied mathematics became pure. Rev. Symb. Logic. **1**(1), 16–41 (2008) . Retrieved from https://pdfs.semanticscholar.org/818f/f8e2b8d7619ee23940be5b7111a2badebf5f.pdf
67. Maranto, C., Griffin, A.: The antecedents of a 'chilly climate' for women faculty in higher education. Hum. Relat. **64**(2), 139–159 (2011). Retrieved from https://epublications.marquette.edu/cgi/viewcontent.cgi?referer=https://www.google.com.au/&httpsredir=1&article=1047&context=mgmt_fac
68. Monoghan, J., Trouche, L., Borwein, J.M.: Tools and Mathematics: Instruments for Learning. Springer, Cham (2016)
69. Monroe, J.: Introduction: the shapes of fields. In: Writing and Revising the Disciplines, pp. 1–12. Cornell University Press, Ithaca (2002)
70. Mumford, D.: Preface. In: Enzensberger, H.M. (ed.) Drawbridge Up. AK Peters, Natick (1999)

71. Nealon, J.: Alterity Politics: Ethics and Performative Subjectivity. Duke University Press, Durham (1998)
72. Raju, C.K.: Black thoughts matter: decolonized math, academic censorship, and the "Pythagorean" proposition. J. Black Stud. **48**(3), 256–278 (2017)
73. Raju, C.K.: Decolonising math and science education (2011a). http://ckraju.net/papers/decolonisation-paper.pdf. Accessed 18 Jan 2019
74. Raju, C.K.: Math wars and the epistemic divide in mathematics. Centre for Computer Science, MCRP University, Gyantantra Parisar (2011b). http://www.hbcse.tifr.res.in/episteme/episteme-1/themes/ckraju_finalpaper. Accessed 18 Jan 2019
75. Rhoten, D., Pfirman, S.: Women in interdisiciplinary science: exploring preferences and consequences. Res. Policy **36**, 56–75 (2007)
76. Rockwell, G.: Crowdsourcing the humanities: social research and collaboration. In: Deegan, M., McCarty, W. (eds.) Collaborative Research in the Digital Humanities, pp. 135–154. Ashgate Publishing, Farnham (2012)
77. Rose, M.: Healing the academic-vocational schism. The Higher Times Chronicle, 10 Sep 2012
78. Sengupta-Irving, T.: Gender equity and mathematics education. In: Banks, J.A. (ed.) Encyclopedia of Diversity in Education. SAGE Publications, University of California, Irvine (2012)
79. Sinclair, N.: Mathematics and Beauty: Aesthetic Approaches to Teaching Children. Teachers College Press. Teacher's College, Columbia University, New York and London (2006)
80. Snow, C.P.: The Two Cultures and a Second Look: An Expanded Version of the Two Cultures and the Scientific Revolution. Cambridge University Press, Cambridge (1969)
81. Stevenson, A., Waite, M. (eds.): Concise Oxford English Dictionary, Luxury edn. Oxford University Press, Oxford (2011)
82. Sullivan, P., Clarke, D.J., Clarke, D.M., Farrell, L., Gerrard, J.: Processes and priorities in planning mathematics teaching. Math. Educ. Res. J. **25**, 457–480 (2013)
83. Swan, P.: I hate Mathematics (2004). https://www.mav.vic.edu.au/files/conferences/2004/Swan.pdf. Accessed 18 Jan 2019
84. Terras, M.: Being the other. In: Mccarty, W., Deegan, M. (eds.) Collaborative Research in the Digital Humanities, pp. 213–230. Ashgate, UK (2012)
85. Tobias, S. [(1978)].: Overcoming math anxiety. Norton paperback edn. W. W. Norton., New York (1993)
86. Trowler, P., Saunders, M., Bamber, V. (eds.): Tribes and Territories in the 21st Century: Rethinking the Significance of Discipline in Higher Education, 1st edn. Routledge, London (2012)
87. UNESCO.: Cracking the Code: Girls' and Women's Education in Science, Technology, Engineering and Mathematics (STEM). United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France (2017). http://unesdoc.unesco.org/images/0025/002534/253479E.pdf. Accessed 18 Jan 2019
88. Wenger, E.: Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, Cambridge (1999)

How Mathematicians Learned to Stop Worrying and Love the Computer

Keith Devlin

How I learned to stop worrying and love the bomb,
subtitle to the 1964 movie Dr. Strangelove

The modern, programmable, digital computer grew from theoretical mathematical results obtained in the 1930s and 40s by mathematicians such as Alan Turing (1912–54) in the United Kingdom, John von Neumann (1903–57), and Alonzo Church (1903–95) in the United States. Indeed, both Turing and von Neumann were involved in the design and construction of early digital computing devices for military purposes in their respective countries during the Second World War.[1]

Yet, when digital computers became available for scientific work, starting in the 1950s, hardly any (pure) mathematicians made use of them. Indeed, that state of affairs continued through the 1960s and the 1970s, and well into the 1980s, a period in which the computer grew to be a ubiquitous tool in the natural sciences, in engineering, and the worlds of business and finance.

[1] While the results of Turing, von Neumann, and Church gave a theoretical underpinning to the subsequent developments of computers, it is clear that the technology would have been developed anyway, and indeed such advances were already underway. For example, Konrad Zuse took out patents for computing devices in 1936 and 1941. And the ENIAC, 1943–46, was designed by engineers Eckert and Mauchly, before von Neumann became involved in the project. Moreover, theoretical and practical work on computing devices was done much earlier by Pascal (1642), Leibniz (1674), and Babbage (1822).

K. Devlin (✉)
Stanford University, Stanford, CA 94305, USA
e-mail: kdevlin@stanford.edu
URL: https://web.stanford.edu/~kdevlin/

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_8

While mathematicians' seeming lack of interest in computers might have seemed strange—indeed paradoxical—to anyone outside the field, to those on the inside it was not at all surprising. The computer had little to offer to the vast majority of research mathematicians.

There was no paradox here. Society's widespread layperson's assumption that mathematics is essentially "higher arithmetic" was always completely off base: (pure) mathematicians study abstract patterns and relationships.[2] For the most part, they use logical reasoning rather than numerical computation. Indeed, the early mathematical work on computing by Turing, von Neumann, Church, and others focused entirely on the theoretical concept of computation. Were it not for the demands of the war effort at the time, it is highly unlikely that Turing or von Neumann would ever have become involved in the design and construction of physical computing devices (Pilot ACE and the Manchester Mark I for Turing, ENIAC for von Neumann) and the execution of actual computations (codebreaking in Turing's case and the calculation of artillery range tables and the design of the atomic bomb for von Neumann[3]).

To be sure, many applied mathematicians were quick to make use of the new technology, and specialized areas of mathematics such as numerical analysis grew considerably with the availability of computers. But the vast majority of mathematicians spent most of their time in day-to-day research activities that had remained largely unchanged for over two millennia. Their work progressed under a widespread, though unstated, assumption that computers could not possibly play a role in the construction of proofs of theorems.

That assumption was given a significant jolt in 1976, with the announcement by two mathematicians in the United States, the American Kenneth Appel and the German Wolfgang Haken, that they had made essential use of a computer to solve a famous, long-standing open problem in mathematics: the Four Color Problem. Dating back to 1852, the problem had all the hallmarks of a theoretical problem for which a computer might be of no help. It asked for a proof that any map drawn on a plane can be colored using at most four colors so no two countries that share a stretch of border are colored the same. The answer will be yes or no; it is not about calculating a number.

If you try a few cases, say with maps of countries, states, or counties, you quickly start to believe the answer might be yes. But what about fictitious maps with thousands of regions, designed to require five or more colors? How do you deal with that possibility? Since there are infinitely many possible map configurations, it is not possible to prove the answer is yes by trying to color every possible one, even with a fast computer.

Or maybe the answer is no. A computer could perhaps be used to solve the problem in the negative; you could let the computer generate map after map and try to color them until it finds one that cannot be colored with four or fewer colors. Since the number of possible coloring configurations for a given map is finite, that could work.

[2] The term "higher arithmetic" has acquired a special meaning in the mathematical world. That is not what I am referring to here.

[3] Making my title for this article a bit more than an irresistible play on words.

But if the answer to the problem is yes, that approach would never end. To solve the problem affirmatively, which is what the majority of mathematicians believed was the case, a logical argument would be required.

Attempts to solve the puzzle by many mathematicians over the years ended in failure, until Appel and Haken eventually came up with an approach that worked. Their approach did indeed involve a logical argument, but on its own that argument did not solve the problem. Rather, they were able to show that if every map in a specific collection of 1,476 particular map configurations could be colored with at most four colors, then the same would be true for all maps. The two researchers then wrote a computer program to examine all possible four-color coloring schemes for those 1,476 maps in turn to see if, in each case, it could find one that worked for that map. That task would have taken too long for a human to complete, even a team of humans, but their computer completed the task in a few months. (Today's computers could do it much faster.) The computer search proved successful, and the Four Color *Problem* became the Four Color *Theorem*. Mathematics had entered a new era.

Initially, the Appel and Haken proof generated a considerable amount of controversy among mathematicians, many of whom regarded the use of a computer to prove a theorem in the same way sports fans object to the use of performance-enhancing drugs. But when, over the ensuing years, a number of other theorems were proved using arguments that likewise required use of a computer, the objections gradually died down. The writing was clearly on the wall—or rather, on the computer screen. As in many other walks of life, for mathematics, the computer was here to stay.

Even so, for the vast majority of mathematicians, things remained the same. Proving a new result still required the construction of a suitable logical argument. The only new twist was that it became accepted that the argument might, on occasion, depend on the successful execution of a computation (often an exhaustive search through a large but finite sets of possibilities), and such arguments were accepted as legitimate proofs. Referees of papers submitted for publication would check the logic in the traditional way and either take the computation on trust or, if feasible, arrange for an independently written computer program running on a different computer to check that the computational part did as the authors claimed.

Notable examples of computer-assisted proofs, as they became known, are

- Proof of Feigenbaum's universality conjecture in nonlinear dynamics (1982),
- Proof of the nonexistence of a finite projective plane of order 10 (1989),
- Proof of the Robbins conjecture (1996), and
- Proof of Kepler's sphere packing conjecture (1998).

There were also cases where computers were used to establish negative results; for example, Odlyzko and te Riele's disproof of the Mertens Conjecture (1985). But since such examples were rare, mathematicians, by and large, continued doing business as usual. The only time they used a computer was for email, after it was introduced in the 1980s, and for typing manuscripts. As things turned out, that latter use of computers for manuscript preparation provided the final impetus that resulted in the American mathematical community embracing the new technology for teaching and research.

In 1978, the Stanford mathematician and computer scientist Donald Knuth released the first version of his mathematical typesetting system TeX, which enabled mathematicians to type their own books and papers using a regular computer keyboard. Special commands were used to product Greek letters and mathematical symbols, and the program took care of organizing the layout on the page so that it was both mathematically correct and esthetically pleasing. There was a fairly steep learning curve as a new user mastered the typesetting language, which was made somewhat easier with the appearance in the early 1980s of LaTeX, a more user-friendly front-end package for TeX, developed by Leslie Lamport of SRI.

So great and so obvious were the benefits of using LaTeX, that some mathematicians quickly adopted it, but even with LaTeX there was still a significant learning curve, and many were put off. They could still see the advantages of typing their own manuscripts, however, and so they went with one of a number of what-you-see-is-what-you-get, mathematical word-processing systems that offered drop-down menus of alternative alphabets and mathematical symbols, an approach that was much easier to learn but did not produce the elegant page layout you got from TeX.

In 1987, Richard Palais of Brandeis University wrote a series of articles for the American Mathematical Society *Notices*, surveying for mathematicians the various mathematical word processing systems that were available at the time. The interest in those articles was sufficiently strong for mathematicians at the AMS to start talking about the society taking a proactive role in helping the community take advantage of the new working possibilities that computers were starting to offer, not only in preparing manuscripts but also in teaching and research. That led to a decision for the *Notices*, which was sent to all members ten times a year, to introduce a regular section "Computers and Mathematics", that would serve both to provide inspiration for mathematicians to make greater use of computers, and to act as an information exchange for the various possibilities computers offered in their work.

That same year, 1987, was also when I moved from the United Kingdom to the United States, to spend a year as a Visiting Professor at Stanford. My host, Jon Barwise, was the mathematician the AMS asked to edit the new *Notices* section, and the two of us talked about the upcoming new column on a number of occasions.

As mathematicians, we both had spectators' interest in the use of computers within traditional mathematics—indeed, Jon attended the lavish event launching Steve Wolfram's new mathematical software system *Mathematica* on June 23, 1988—but our main interests took different forms. Jon's interest was primarily that of a logician, and he soon began working with his Stanford colleague John Etchemendy to develop instructional software to teach formal logic (*Turing's World*, *Tarski's World*, and *Hyperproof*). My focus was more as part of my growing interest in what would become known as mathematical cognition, where the focus was on studying mathematics as a mental tool, looking at how it arose, and how it related to, fitted in with, and complemented other forms of thinking. From that standpoint, the use of computers to assist in doing mathematics was but one component of what I would end up calling "mathematical thinking".

The "Computers and Mathematics" section launched in the May/June 1988 issue of the *Notices*, with Barwise leading off with an essay in which he declared that the goal was to reflect, both practically and philosophically, on cases where computers were affecting mathematicians and how they might do so in the future; to act as an information exchange into what software products were available; and to publish mathematicians' reviews of new software.

Barwise edited the section through to February 1991, after which the AMS asked me to take it over. I held the reins from the March 1991 issue until the AMS and I decided to end the special section in December 1994. The reason? That six-and-a-half-year run of the special section had achieved the intended goal. The computer had become a staple tool for mathematicians, both in teaching and research.

The general format of each column was to start with some form of editorial comment, then, frequently, a feature article solicited by the editor, and then a number of reviews of new mathematical software. In all, we published 59 feature articles, 19 editorial essays, and 115 reviews of mathematical software packages (31 features, 11 editorials, and 41 reviews under Barwise, 28 features, 8 editorials, and 74 reviews under me).

At around the same time the "Computers and Mathematics" section was starting up, a number of mathematicians were developing a new subfield of mathematics called "Experimental Mathematics". In this field, one of the primary goals in using computers was to formulate conjectures that could subsequently be proved by conventional means—which cast the computer as an additional weapon in the pure mathematician's armory rather than a completely separate technological endeavor. In 1992, a new journal with that name as its title was established by the American mathematicians David Epstein, Silvio Levy, and the German-American mathematics publisher Klaus Peters. And in the fall of that year, the Canadian mathematicians Jonathan and Peter Borwein sent me their article "Some Observations on Computer Aided Analysis", written to introduce their new field to the mathematical community at large, which I published in the October issue of "Computers and Mathematics".

At the same time as the computer was starting to change mathematics research and applications, various instructors brought it into their classrooms. Computer Algebra Systems such as *Mathematica* and *Maple* were used to teach calculus in a new way, and a number of new textbooks to support such teaching came onto the market. Some of the articles and product reviews in "Computers and Mathematics" were devoted to the increasing use of computers in the world of university mathematics education. Things were starting to move very quickly.

When he introduced the last section he edited, Barwise had written:

> Whether we like it or not, computers are changing the face of mathematics in radical ways, from research, to teaching, to writing, personal communication, and publication. Over the past couple of years we have seen numerous articles about these developments.
>
> Computers are even forcing us to expand our idea about what constitutes doing mathematics, by making us take much more seriously the role of experimentation in mathematics. (I draw attention to a new journal devoted to experimental mathematics below.)
>
> One view of the future is that mathematics will come to have (or already has) two distinct sides: experimentation, which can exploit the speed and graphics abilities of programs like

> Maple and Mathematica, to allow us to spot regularities and make conjectures, and proof, very much in the style of today's mathematics. . . .
>
> Whether we applaud or abhor all these changes in mathematics, there is no denying them by turning back the clock, anymore than there is in the rest of life. Computers are here to stay, just as writing is, and they are changing our subject.

It is surely obvious from those final remarks that the computer was seen as something of a threat by some mathematicians, and the "Computers and Mathematics" section was not without its detractors.

Taking over a month later, I began by saying that:

> This column is surely just a passing fad that will die away before long. Not because mathematics will cease to have much connection with computers, but rather, quite the reverse: the use of computers by mathematicians will become so commonplace that no one thinks to mention it any more.

When I wrapped up the section 4 years later, I wrote:

> With its midwifery role clearly coming to an end, the time was surely drawing near when "Computers and Mathematics" should come to an end. The change in the format of the *Notices*, which will take place at the end of this year, offered an obvious juncture to wind up the column. . . .
>
> The disappearance of this column does not mean that the *Notices* will stop publishing articles on the use of computers in mathematics. Rather, recognizing that the use of computer technology is now just one more aspect of mathematics, the new *Notices* will no longer single out computer use for special attention. I will drink to that.
>
> The child has come of age.

And so mathematics moved on. In 2004, Jon Borwein and David H. Bailey published (together with coauthors in two cases) the first of what would be three major research monographs on experimental mathematics, and in 2008, Jon and I published our expository text *The Computer as Crucible: An Introduction to Experimental Mathematics* [1]. A year later, Wolfram released *Wolfram Alpha*, an online computational tool that, among other things, was able to execute practically any mathematical method or procedure—faster and more accurately than any human, and with effectively no restrictions on data size.

The computer had, by then, completely revolutionized all of procedural mathematics. Only the pure mathematicians, who focus on finding proofs of precisely worded theorems, remained almost entirely unscathed by the revolution.

In late 2016, after I learned of Jon's tragic early passing, I looked back on my own mathematical work in the 20 years after I edited my last *Notices* "Computers and Mathematics" section, some of it with Jon. My reflections prompted me to pen—more accurately type (on a computer)—an opinion piece for the *Huffington Post*, which was published on January 1, 2017, with the startling, but absolutely accurate title: "All the Mathematical Methods I Learned in My University Math Degree Became Obsolete in My Lifetime" [3]. For the fact is, that over a period of just under a quarter-century, during which time I moved from working in pure mathematics (i.e., focusing on proofs) to making use of mathematics to solve large-scale, real-world problems, my daily experience of doing mathematics changed from using methods and executing procedures to putting problems into a form where I could apply a powerful computational tool such as (in my case) *Wolfram Alpha* or *Mathematica*.[4]

True, by then I was no longer a pure mathematician, so my experience here is not typical of pure math. But it is typical of the way doing math has changed for the vast majority of mathematicians in the world. Besides, no one can look at the computer-intensive work of Jonathan Borwein and David H. Bailey in Number Theory, where they also use *Mathematica*, and pretend it is anything other than pure math. To be sure, some pure mathematicians make hardly any professional use of computers aside from email and an occasional Google search. But for a great many, the computer is now an integral part of how they carry out their work.

That then is the story of how mathematicians learned to stop worrying and love the computer. I could go on, and dig much deeper into the details. But, given the ease with which, given a few key issues (and associated key words), we can now all dig down on our own, I'll let you get a sense of the mathematical computer revolution by browsing the image gallery associated with this short article. The gallery can be viewed as a browser-playable slide presentation at

https://www.icloud.com/keynote/0C-SmMChXRYKuvGeWFTwz8l3g#BorweinPaperIMAGES and as a browser-viewable or downloadable PDF file at

https://web.stanford.edu/%7Ekdevlin/BorweinPaperIMAGES.pdf

Further Reading

For a complete index to everything published in the "Computers and Mathematics" section of the AMS Notices, see [4].

Accessible books the reader may find helpful are: [1, 2, 5, 6]

References

1. Borwein, J.: Devlin, K,: The Computer as Crucible. AK Peters/CRC Press, Boca Raton (2008)
2. Ceruzzi, Paul E.: Computing: A Concise History. MIT Press, Cambridge (2012)
3. Devlin, K.: All the Mathematical Methods I Learned in My University Math Degree Became Obsolete in My Lifetime. Huffington Post (23 Jan 2017). https://www.huffingtonpost.com/entry/all-the-mathematical-methods-i-learned-in-my-university_us_58693ef9e4b014e7c72ee248
4. Devlin, K., Wilson, N.: Six-year index of "computers and mathematics". Not. Am. Math. Soc. **42**(2), 248–254 (1995). http://www.ams.org/journals/notices/199502/devlinsixyear.pdf
5. Dyson, G.: Turing's Cathedral: The Origins of the Digital Universe. Vintage, Visalia (2012)
6. Wilson, R.: Four Colors Suffice: How the Map Problem Was Solved. Princeton University Press, Princeton (2014)

[4] Full disclosure. I was a member of Wolfram's initial *Mathematica* Advisory Board in the products early years (we were all unpaid), so I naturally defaulted to using Wolfram products. But there were several CASs being developed around the same time, *Maple*, *Matlab*, *Magma*, *Sage*, etc.

Crossing Boundaries: Fostering Collaboration Between Mathematics Educators and Mathematicians in Initial Teacher Education Programmes

Merrilyn Goos

Dedicated to the memory of Jonathan M. Borwein, a visionary supporter of collaboration between mathematics educators and mathematicians

1 Introduction

There is great diversity in the structure of, and approaches to, mathematics teacher education across the world [1]. In many countries, however, secondary pre-service teacher education programmes are structured so that future teachers of mathematics learn the content they will teach by taking courses offered by the university's mathematics department, while they learn how to teach this content by taking content-specific pedagogy courses within the university's education department. Such program structures provide few opportunities to interweave content and pedagogy in ways that help develop professional knowledge of teaching. These structures also make it difficult for mathematicians and mathematics educators[1] to gain a mutual understanding of each other's roles in preparing future teachers or to generate a commitment to collaboration in addressing joint problems [2]. These are some of the challenges that were addressed by the Inspiring Mathematics and Science in Teacher Education (IMSITE) project, one of a suite of large-scale national projects funded by the Australian Government with the purpose of driving: a major improvement in the quality of mathematics and science teachers by supporting new pre-service

[1] We acknowledge the inadequacy of these categories for distinguishing between the two disciplinary fields and those who work in them (see Fried, 2014).

M. Goos (✉)
The University of Queensland, St Lucia Qld, Australia
e-mail: m.goos@uq.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_9

programmes in which faculties, schools or departments of science, mathematics and education collaborate on course design and delivery, combining content and pedagogy so that mathematics and science are taught as dynamic, forward-looking, and collaborative human endeavours [3].

The IMSITE project aimed to achieve the purposes outlined above by: (1) fostering genuine, lasting collaboration between mathematicians, scientists, and mathematics and science educators who prepare future teachers and (2) identifying and institutionalising new ways of integrating the content expertise of mathematicians and scientists with the pedagogical expertise of mathematics and science educators. The second of these aims provides the focus for this chapter, which discusses ways of integrating mathematics content and mathematics pedagogy in secondary pre-service teacher education programmes.

2 Project Context and Overview

The project involved a partnership between six Australian universities[2] that varied in terms of their institutional grouping, geographical location, pre-service programme structures, characteristics of the university student population, and characteristics of the students and schools to be experienced by graduating teachers. These variations ensured that the outcomes of the project were evaluated and embedded in a diverse range of institutional, geographical, and socioeconomic contexts.

The core project team comprised 23 university academics who were either education academics (mathematics and science teacher educators) or discipline academics (mathematicians and scientists). Each partner university's project team included at least one education academic and one discipline academic.

The project built on a diverse range of interdisciplinary strategies already piloted or envisioned in the partner universities. Thus, no single model of pre-service teacher education that privileges one structure for degree programmes, one way of combining content and pedagogy, or one form of collaboration between discipline and education academics was promoted. This approach generated a coherent suite of teacher education strategies that were shared and tested in new contexts, and fostered new collaborations between universities with common interests.

Previous publications have reported on processes of interdisciplinary collaboration between mathematicians and mathematics educators working on the IMSITE project [4, 5]. Conditions enabling collaboration included personal qualities such as open mindedness, trust, mutual respect, shared beliefs and values; and a common or shared problem, usually centred on supporting the professional learning of

[2] It is appropriate here to acknowledge the significant role played by Jon Borwein in advocating for the University of Newcastle to be part of the IMSITE project. As Director of CARMA Jon built bridges between mathematicians and mathematics educators, in ways that aligned with and foreshadowed the IMSITE project's aims. His support for the project was substantial, providing both mentoring and financial support for project staff. We are proud to claim IMSITE as part of Jon's legacy.

pre-service teachers. Conditions potentially hindering collaboration included the physical separation of the buildings where mathematicians and mathematics educators worked; university workload formulas and financial models that failed to recognise or reward interdisciplinary collaboration; and cultural differences between disciplines. This paper additionally examines the products of interdisciplinary collaborations. It identifies some characteristics of teacher education programmes that draw on the expertise of mathematicians and mathematics educators to integrate content with pedagogy in preparing future secondary school teachers.

3 Theoretical Background

The conceptual framework for the project drew on two major theoretical sources. The first is Wenger's [6] social theory of learning, and in particular the notions of community of practice, boundary practices, and brokering between disciplinary paradigms [7], to understand how the perspectives of mathematicians and mathematics educators can be coordinated and connected. The second concerns the construction of professional knowledge for teaching, to explore forms of knowledge that need to be addressed by teacher education programmes. When the project began, there were few known instances of productive interdisciplinary collaboration in the design and delivery of pre-service mathematics teacher education programmes in Australia, even though it has long been argued that both mathematicians and mathematics educators have an important role to play in the preparation of teachers [8].

Research on the role of subject matter knowledge and pedagogical content knowledge in mathematics provided the rationale for combining content and pedagogy in ways that aligned with the aims of the IMSITE project [9]. The design of the project rested on the assumption that the preparation of prospective mathematics teachers must include the development of subject matter knowledge as well as pedagogical content knowledge, that is, knowledge of how mathematics is learned; how to select, represent and sequence the big ideas, examples, and topics; how to deal with misunderstandings and conceptual blockages.

In mathematics education, numerous research studies have demonstrated that teachers need both content knowledge and pedagogical content knowledge (e.g., [10]), in particular, because both types of teacher knowledge predict students' mathematical achievement [11]. However, these types of knowledge are clearly related: content knowledge alone is insufficient for effective teaching, but it does form the basis for development of pedagogical content knowledge [11].

4 Research Methods

Each participating university prepared an initial implementation plan, building upon existing or envisioned collaborative initiatives and interdisciplinary strategies. The first task of the project was to identify and collate these strategies under headings that captured three phases in pre-service and beginning teachers' career trajectories: *(1) Recruitment strategies that promote teaching careers, (2) Innovative curriculum arrangements that combine content and pedagogy, and (3) Continuing professional learning that builds long term relationships with teacher education graduates.* Strategies identified at this stage included developing elective courses on mathematics education for inclusion in undergraduate mathematics programmes, new programmes that prepare specialist primary mathematics teachers (which differ from the usual preparation of generalist primary teachers in Australia), and online networks of graduates to support their continuing engagement with new mathematics content and mathematics pedagogy. In the first year of the project, each of the six universities implemented a selection of strategies so that each strategy was trialled by at least two universities. In the second year, participating universities undertook to implement the second set of strategies that had been successfully trialled in the previous year by other universities in the project team. In the third year, participating universities were encouraged to partner with another institution outside the project team that wished to adapt and implement some of the new teacher education strategies trialled in the first two years of the project. Through these processes, project approaches and outcomes were progressively adapted, tested, and transferred to new contexts.

Representatives of the full project team held regular teleconferences and face to face meetings to share practice and foster collaboration between universities. Individual university project team members held more frequent meetings—often at weekly intervals—to support interdisciplinary collaboration and design new recruitment, coursework, and continuing professional learning initiatives. Interviews were conducted in the first and third years of the project with the lead mathematician and mathematics educator at each of the project universities to investigate the processes and products of interdisciplinary collaboration [4, 5]. Each university also provided a written annual report that described activities and outcomes mapped against the two project aims of fostering interdisciplinary collaboration and integrating mathematics content and mathematics pedagogy. The interviews and annual reports provided evidence to address the aim of this paper, which is to characterise the products of interdisciplinary collaboration between mathematicians and mathematics educators in pre-service teacher education.

5 Towards Integrated Models of Teacher Education

Working between discipline communities and education communities requires the development of new practices that draw on the expertise of individuals from both communities. Evidence of these practices was seen in three ways: co-developed and co-taught courses, primary pre-service teacher education programmes designed specifically to integrate content and pedagogy, and approaches to building communities of pre-service mathematics teachers. As the focus of this paper is on preparing secondary teachers, it illustrates the first and third of these approaches.

5.1 Co-developed and Co-taught Courses Integrating Content and Pedagogy

Courses that integrate content and pedagogy for secondary pre-service teachers were either further developed or designed and implemented in four of the partner universities. Two examples are summarised below. Mathematical Content Knowledge for Lower Secondary Mathematics Teachers is offered in the Bachelor of Education (Secondary) programme at University A. It was co-designed and is co-taught by a mathematician and mathematics educator who had a prior history of collaboration. The mathematician described the motivation for developing the course:

> Well it's a subject specifically aimed for [...] pre-service maths teachers. The university has never had a subject like that and I'm not aware of many around the planet even. But it was a need that [mathematics educator] had expressed to me early on. [Her] opinion had been formed by her own students that they were getting a bunch of subjects in the maths department, that they felt as though didn't really prepare them for the maths they were going to teach in the classroom.

Student learning outcomes include developing deep content knowledge of lower secondary mathematics represented in the Australian Curriculum [12], understanding the links between these mathematics content areas, investigating and communicating mathematical ideas, and understanding the historical and socio-cultural development of mathematical ideas.

Reflective Communication in Mathematics was developed and is delivered collaboratively by a mathematician and a mathematics educator from University B to give non-education students an opportunity to explore teaching (e.g., students are enrolled in Bachelor of Engineering, Bachelor of Mathematics, Bachelor of Advanced Mathematics, Bachelor of Arts programmes.) The course was also made available to pre-service Bachelor of Mathematics Education students. In addition to coursework, students undertake private mathematics tutoring and participate in a range of mathematics outreach activities that bring secondary school students and their teachers to the university, for example, in 'Work like a mathematician' excursions. Intended learning outcomes include demonstrating the ability to analyse one's own understanding of mathematical concepts, demonstrating pedagogical content knowledge to explain mathematical concepts, and demonstrating technical and communication skills to explain mathematical ideas in creative ways. The intent of the course is to provide a 'risk free' experience in teaching to students who are not enrolled in a teaching degree, in order to encourage them to consider a future career in this field. However, the mathematician and mathematics educator who delivered the course also recognised unanticipated benefits for pre-service teacher education students who were taking the course as an elective. The mathematician commented:

> We both realised that [these students] had not made the connections between their maths subjects, their pedagogy subjects, and the maths that they were going to be teaching at school. This was the first subject that they had where we were talking about both at the same time, taking it further than anything had been taken—like take the syllabus from high school, push it into where it goes to university where they come back and talk about how might you teach it so that you get those outcomes.

The significance and innovation attached to these integrated courses needs to be understood in the context of institutional barriers to collaboration experienced by the mathematicians and mathematics educators in Australian universities. One mathematician commented on the difficulty of getting ‘things like what we do to be recognised in workload models’, because jointly taught courses are treated as invisible ‘extra work’. Another expressed frustration at financial models that discourage universities from sharing course income between different disciplines, even when these disciplines are contributing equally to course design and delivery as in the IMSITE project. One of the major achievements of the project was in overcoming some of these barriers so that workloads and funding were shared between the mathematics and education departments responsible for the new integrated courses.

5.2 Communities of Pre-service Mathematics Teachers

At the time the IMSITE project was conducted, pre-service teacher education programmes for secondary mathematics teachers typically involved either an undergraduate Education degree, a dual degree such as Bachelor of Science/Bachelor of Education (BSc/BEd), or an initial discipline-specific Bachelors degree followed by a one year Graduate Diploma in Education or a two year Master of Teaching. In all models, content and pedagogy are taught in separate courses. In dual degree programmes it is typical for mathematics content courses to be taught first, in the BSc component of the programme, and pedagogy courses some years later, in the BEd component. This means that pre-service mathematics teachers take their mathematics content courses together with a much larger group of BSc students who are not planning to become teachers, and they may not even be aware that there are other aspiring teachers in their content classes. The lack of a cohort experience in the early years of a pre-service teacher education programme makes it difficult to build a sense of community amongst prospective mathematics teachers, and could lead to unwanted attrition. This was a shared problem identified by the mathematician and mathematics educator at University C when they realised that they taught the same pre-service secondary teacher education students:

> Then I think you and I just started chatting one day ... and we thought, you know what? You teach the students maths and I teach them education. We should at least be sharing what we know about the students; starting to compare contrast, talk about issues, retention. We started talking about the fact that we would lose some of them. [mathematics educator]

University C offers a five year dual degree Bachelor of Science/Bachelor of Education programme, where overlap between mathematics and education courses does not occur until the third year of the programme. For this reason, the mathematician and mathematics educator participating in the IMSITE project collaborated to create early cohort experiences for pre-service mathematics teachers. For example, in a compulsory first year mathematics course, rather than randomly mixing mathematics education students in tutorial groups with non-education students, they are allocated together to special tutorials taught by former secondary school mathematics teachers. Regular lunches and social events are also held to bring together later year pre-service students with first year mathematics students who have not yet begun their education studies, for networking and sharing of experiences. An alumni conference has been held for the past three years where mathematics pre-service teachers nearing the end of their programme participate with recent graduates, mathematicians, and mathematics educators in a professional development day. The purpose of all these cohort-building activities is to create a strong sense of mathematics teacher identity and community from the earliest stages of the degree programme, and extending beyond graduation.

6 Concluding Comments

Internationally it is rare to find research or teaching collaborations between mathematicians and mathematics educators [2]. The IMSITE project has shed some light on how such collaborations can work, and what their outcomes might be. Within the universities participating in the project, there was evidence that these collaborations can contribute to curriculum development in teacher education courses and programmes that meet the needs of a diverse range of institutional contexts. There were also instances of collaboration between participating universities. This was not a simple process of transferring resources or courses from one institution to another; instead, each university had to recontextualise and transform the approaches originally developed elsewhere to suit its own circumstances. It could be argued that this process of appropriation and transformation was crucial to the embedding of strategies in new contexts because it required a mutually beneficial exchange of knowledge and understanding. There was less evidence of IMSITE strategies being taken up by universities outside the project, possibly because of the lack of existing or potential collaborations between academics from the two disciplines.

It will also be important to monitor the longer-term impact of the IMSITE project over the next five to ten years. In particular, a limitation in the project design was the lack of data from pre-service teachers. This limitation could be overcome in a future study by analysing enrolment, attrition, and graduation trends; and by administering longitudinal surveys of pre-service teacher attitudes, content knowledge, and pedagogical content knowledge.

The IMSITE project provides an example of how research might bring together mathematicians and mathematics educators with the aim of improving the preparation of future teachers of mathematics. Fried [2] notes that the fields of mathematics and mathematics education have been moving further apart for many years, as mathematics education research has become more aligned with the social sciences and mathematics research with the exact sciences. Not only do these two fields differ in the kinds of knowledge they generate, they also have different ways of pursuing knowledge. Fried [2] argues that the key to collaboration lies in acknowledging these differences, rather than members of each community trying to ‘convert’ each other to what they cannot be. Nevertheless, in the IMSITE project, although there were differences in the nature and extent of interdisciplinary collaboration, it was common for participants to recognise 'that each side is looking in the same direction but with very different, complementary eyes' [2, p. 15].

References

1. Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., ... Reckase, M.: Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Findings From the IEA Teacher Education and Development Study in Mathematics (TEDS-M). IEA, Amsterdam (2012)
2. Fried, M.: Mathematics and mathematics education: searching for common ground. In: Fried, M., Dreyfus, T. (eds.) Mathematics and Mathematics Education: Searching for Common Ground, pp. 3–22. Springer, New York (2014)
3. Department of Education and Training (DET): Enhancing the Training of Mathematics and Science Teachers Program. https://www.education.gov.au/enhancing-training-mathematics-and-science-teachers-program. Accessed 17 Nov 2017
4. Bennison, A., Goos, M.: Learning at the boundaries: collaboration between mathematicians and mathematics educators within and across institutions. In: White, B., Chinappan, M., Trenholm, S. (eds.) Opening Up Mathematics Education Research (Proceedings of the 39th Annual Conference of the Mathematics Education Research Group of Australasia), pp. 124–131. MERGA, Adelaide (2016)
5. Goos, M.: Learning at the boundaries. In: Marshman, M.,Geiger,V., Bennison, A. (eds.) Mathematics Education in the Margins (Proceedings of the 38th Annual Conference of the Mathematics Education Research Group of Australasia), pp. 269–276. MERGA, Adelaide (2015)
6. Wenger, E.: Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, Cambridge (1998)
7. Akkerman, S., Bakker, A.: Boundary crossing and boundary objects. Rev. Educ. Res. **81**, 132–169 (2011)
8. Hodgson, B.: The mathematical education of school teachers: role and responsibilities of university mathematicians. In: Holton, D. (ed.) The Teaching and Learning of Mathematics at University Level: An ICMI Study, pp. 501–518. Kluwer Academic Publishers, Dordrecht (2001)
9. Shulman, L.S.: Knowledge and teaching: foundations of the new reform. Harv. Educ. Rev. **57**(1), 1–22 (1987)
10. Goos, M.: Mathematical knowledge for teaching: what counts? Int. J. Math. Educ. Sci. Technol. **44**, 972–983 (2013)
11. Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y.-M.: Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. Am. Educ. Res. J. **47**, 133–180 (2010)
12. Australian Curriculum, Assessment, and Reporting Authority (ACARA): The Australian Curriculum (Version 8.3). http://www.australiancurriculum.edu.au/. Accessed 17 Nov 2017

Mathematics Education in the Computational Age: Challenges and Opportunities

Kathryn Holmes

Dedicated to the memory of my colleague Jonathan M. Borwein and his enthusiasm to challenge conventional thinking.

Jonathan Borwein was a pioneer in Experimental Mathematics; mathematics made possible through the use of computers. He wrote, with his colleague David H. Bailey, that this approach to mathematics had the potential to help mathematicians gain insight and intuition, visualise mathematics principles, discover new relationships, test and falsify conjectures, explore possible results to see if they merit formal proof, even to suggest approaches for formal proof, to replace lengthy hand derivations and to confirm analytically derived results [1]. While acknowledging the central role of formal proof in mathematics, they saw the potential for computational techniques to advance the discipline. And indeed they did. Interestingly, while Experimental Mathematics has been viewed as being at odds with the notion of proof as the defining feature of the discipline, we have recently witnessed the validation of a proof via computational techniques [2].

In facilitating Experimental Mathematics, computational techniques reveal insights that would otherwise remain hidden. It is fair to say, however, that the type of exploration and experimentation which is now supported by computers, has always taken place in mathematics. The mathematical greats of the past engaged in 'page after page of trial and error experimentation ... exploratory calculations, guesses formulated, hypotheses examined ...' [3]. Now, this type of exploratory work can be supercharged by computational techniques, accelerating both the pace of mathematical advances and changing the types of developments that are possible.

K. Holmes (✉)
Western Sydney University, Sydney, NSW, Australia
e-mail: K.Holmes@westernsydney.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_10

Given that the discipline of mathematics is changing in response to computational techniques, should there also be a corresponding change in the way mathematics is taught in schools? Conrad Wolfram, in his popular 2010 TED talk, argues that computers allow us to focus on the interesting parts of mathematics rather than computation. In contrast, he points out that the majority of school mathematics relies on pen and pencil computational techniques, which he labels as the drudgery of mathematics, and which can now be readily conducted with computers. He calls for mathematics education to be less procedural, focussing more on practical applications and conceptual understanding.

There are other reasons to consider changing how we teach mathematics in schools. There is ample evidence that students shy away from advanced mathematics in the senior years of schooling, preferring to take less demanding applied mathematics courses or no mathematics at all. While this trend applies to all students it is particularly acute for female students [4]. This is occurring at a time when demand for STEM skills are predicted to rise, leading to forecasts of skills shortages. There is also evidence that mathematics is widely disliked, causes anxiety amongst some students and that students don't see the relevance of the subject to their current or future lives [5]. These issues have persisted for decades with little evidence of any progress in countering the impact on student interest in mathematics [6]. Therefore, a compelling case can be made that continuing the status quo will not deliver the shifts desired in terms of student participation in and enjoyment of mathematics. So, while it is clear that approaches to mathematics teaching and learning need to change, there is markedly less clarity in relation to the nature of the required changes.

Debates over how best to learn mathematics have often focussed on the notions of procedural and conceptual knowledge. While most teachers would accept that it is desirable to develop both forms of mathematical knowledge, the best means by which to do so remains contentious [7]. Some would argue that one first needs to have a conceptual understanding in order to provide the learning of procedural mathematics skills with purpose and relevance [8]. Others contend that it is difficult to develop a deep conceptual knowledge without some fluency and automatic recall of underlying mathematics facts and procedures.

There is evidence that the development of procedural and conceptual knowledge can be thought of as being bi-directional and iterative, with learners experiencing gradual gains in both domains over time. Also, both types of knowledge have been linked to the development of 'procedural flexibility', a construct which relates to learners' knowledge of multiple procedures for solving mathematics problems and their capacity to apply these procedures adaptively in a variety of situations [9]. But how best might we develop procedural and conceptual knowledge and procedural flexibility in our school students, particularly in a time awash with new computational tools? Thinking about the most fruitful ways of approaching mathematics teaching in the future raises a number of significant questions:

- Is the learning of mathematical procedures necessary as a precursor to conceptual understanding?

- Does *fluency* in skills such as adding fractions, knowing times tables and fluency in algebraic manipulation, promote and support the development of conceptual understanding?
- Do procedural skills lead to an increased capacity to apply mathematics to new contexts or to solve novel problems, i.e. *procedural flexibility*?
- Is our focus on the automaticity of possibly redundant skills leading to our students' lack of interest in mathematics and reinforcing their perceptions of the *irrelevance* of mathematics to their lives?
- While it seems logical that possessing procedural fluency would not be a hindrance to the development of mathematical expertise, is it a *necessity*?
- Could we interest more students in mathematics by making available all of the *computational tools* that are currently available and that are regularly used in STEM careers?
- Are we wed to our current chronological approach to mathematics teaching purely because of *historical precedent*, thereby continuing a long tradition in the accepted 'order of operations'?

Perhaps the central question we should be looking to answer is: *'Which school level mathematical procedures/algorithms are useful for promoting the development of deep conceptual understanding?'* and then those that are not useful and can be completed with technological tools could be discarded.

On a personal note, in the late 1970's/early 1980's scientific calculators were readily available and were beginning to be permitted into senior school exams, removing the need for logarithmic and trigonometric tables. Around this time my father wanted to teach me his manual algorithm for finding the square root of any number to as many decimal points as desired. I resisted his attempts to teach me his method, as I had ready access to a calculator which would do the same thing at the press of a button. Despite not knowing how to carry out this calculation by hand, I did understand the concept of the square root operation and when it might be useful in solving a range of problems. I saw no need to learn a 'by hand' method and I was simply not prepared to put in the time and effort to do so. It strikes me that in persisting with teaching many of the algorithms and procedures that still remain in our school curricula, that we are asking students to spend their time on developing skills that have no place in their futures. They know, as I did with the square root algorithm, that these skills are irrelevant. As a result, many are turning away from mathematics, potentially limiting their future study and career options.

In addition to considering the salience of continuing to teach mathematical procedures which can more readily be conducted with technology, we should also consider how learning with technology can 'value-add' to school mathematics. The *ScratchMaths* project addresses the links that can be made between learning programming with a focus on mathematical concepts [10]. Such an approach embraces the technological world that students are now immersed in by capitalising on their inherent interest in digital technologies. Such an approach shows promise for increasing engagement levels in mathematics by allowing students to encounter mathematical ideas through the digital world which pervades most aspects of their lives [11]. The ideas behind *ScratchMaths* echo the approach of Jonathan Borwein to advancing and developing mathematical knowledge by leveraging the power of the computational age. By integrating the learning of computer science and mathematics in the crucial middle years of schooling, we may find new ways of sparking renewed student interest in mathematics.

The field of Experimental Mathematics came about through Jonathan Borwein and his colleagues' willingness to embrace new technological tools, thereby leading the advances not possible via traditional means. In doing so, he challenged traditional ways of thinking about and doing mathematics. In a similar vein, perhaps it is time to fully embrace the computational age in the teaching and learning of school mathematics, rather than persisting with traditional methods developed for the pre-digital age. Those concerned with the current state of mathematics education need to work to find a way to reform mathematics curricula for the computational age or we can expect the current student exodus to continue.

References

1. Bailey, D.H., Borwein, J.M.: Exploratory experimentation and computation. Not. AMS **58**(10), 1410–1419 (2012)
2. Hales, T., Adams, M., Bauer, G., Dang, T., Harrison, J., Hoang, L., Zumkeller, R.: A formal proof of the Kepler conjecture. Forum Math. Pi **5**, e2 (2017). https://doi.org/10.1017/fmp.2017.1
3. Borwein, J., Devlin, K.: The Computer As Crucible: An Introduction to Experimental Mathematics. AK Peters, Wellesley, MA (2008)
4. CESE: Why Aren't Students Studying Higher Level Maths? How ATAR Scaling may Affect Maths Uptake. Centre for Education Statistics and Evaluation, Sydney, NSW (2017)
5. Devine, A., Fawcett, K., Szucs, D., Dowker, A.: Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. Behav. Brain Funct. **8**, 33 (2012)
6. Rakes, C.: Challenging the status quo in mathematics: teaching for understanding. The Conversation. https://theconversation.com/challenging-the-status-quo-in-mathematics-teaching-for-understanding-78660. Accessed on 22 June 2016
7. Rittle-Johnson, B., Schneider, M.: Developing conceptual and procedural knowledge of mathematics. In: Oxford Handbook of Numerical Cognition, pp. 1102–1118. OUP Oxford, Oxford (2014)
8. Rittle-Johnson, B., Alibali, M.W.: Conceptual and procedural knowledge of mathematics: does one lead to the other? J. Educ. Psychol. **91**(1), 175 (1999)
9. Rittle-Johnson, B.: Developing mathematics knowledge. Child Dev. Perspect. **11**(3), 184–190 (2017)
10. Benton, L., Hoyles, C., Kalas, I., Noss, R.: Bridging primary programming and mathematics: some findings of design research in England. Digit. Exp. Math. Educ. **3**(2), 115–138 (2017)
11. Prieto, E., Hickmott, D., Holmes, K., Berger, N.: Exploring mathematics using computational thinking: the scratchmaths pilot project. Paper presented at the 2018 MERGA conference, Albany, New Zealand (2018)

Mathematics Education for Indigenous Students in Preparation for Engineering and Information Technologies

Collin Phillips and Fu Ken Ly

1 Introduction

It is well documented that there is a disparity in both presence and performance in Science, Technology, Engineering and Mathematics (STEM) subjects between Indigenous and non-Indigenous students. Indeed, a two-and-a-half year achievement lag has been observed in secondary schools for Indigenous students compared to their non-Indigenous peers [2, 8, 9], and this has resulted in a push in Australian education policy to close this disparity [6].

In September 2017, the Mathematics Learning Centre (MLC) at the University of Sydney conducted a Mathematics Workshop as part of an intensive week-long programme for Indigenous students organised by the Faculty of Engineering and Information Technologies (FEIT). The programme was called the 'STEM Spring Workshop' and one of its aims was to increase the Indigenous presence in STEM subjects, both at the University of Sydney, and more broadly.

Fourteen students attended the programme (seventeen invited) from year 10 and 11 students who had attended other programmes for Indigenous students at the University of Sydney (http://engineeringaid.org/sydney-university-iaess/ and http://sydney.edu.au/wpo/indigenous/summer-programme/index.shtml).

The selection was based on supporting documents (including references, academic records and personal statements), the recommendations of staff who had run the other programmes, consideration of socio-economic circumstances, and a priority for

C. Phillips (✉)
Mathematics Learning Centre, Education Portfolio, University of Sydney,
Sydney, NSW 2006, Australia
e-mail: collin.phillips@sydney.edu.au

F. K. Ly
Faculty of Science, School of Mathematics and Statistics, University of Sydney,
Sydney, NSW 2006, Australia
e-mail: ken.ly@sydney.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_11

including students who were taking or likely to take the required level of mathematics for entry into an engineering degree. There was no additional explicit benchmarking for ability and students came from both rural and urban communities. The students travelled considerable distances from their home states of New South Wales, the Australian Capital Territory, Queensland, South Australia and Tasmania to participate in the programme.

This paper describes the development, implementation and assessment of the Mathematics Workshop from its early stages through to its conclusion, in view of providing a useful resource for future initiatives of a similar kind. In this work, we shall refer to the Mathematics component of the STEM Spring Workshop as the *Mathematics Workshop* or just *Workshop* for brevity.

The philosophy of the MLC heavily influenced the Mathematics Workshop. As a Centre, we have particular expertise in educating underprepared students for and at University. Three principles inform our practice:

- Contribute to building a supportive, cooperative environment where all parties are valued in contributing their individual perspectives.
- Employ 'cultural plasticity', i.e. be receptive to, learn from and adapt to the cultural perspectives of others.
- Use feedback where students can have a genuine voice throughout the teaching and development process to adapt, modify and evolve teaching modes.

We are very aware of the ways in which students may have previously experienced mathematics. Before engaging in the ideas and concepts of mathematics, it can be critical to know the student's background, past experiences and culture, and to overcome past fears that may have arisen because standard teaching has failed them. For instance, students may have a different way of thinking and solving problems that need validation. Some students are not inclined to invest in the mathematical world unless or until there is a recognised real world application. Other students can be principally motivated and enthralled by patterns, connections and systems that have no direct, apparent practical application.

The ideas of mathematics itself also influence our practice and our philosophy. Learning mathematics can be challenging for a number of reasons. Firstly, to learn mathematics we need to engage with new concepts and ideas that may or may not have been encountered in any other experience of life. Examples of these include the formal mathematical concepts of a vector inner product, projections, exterior products, linear independence, complex numbers, number theory, category theory, etc. Regardless of how these new concepts are learnt—through real world applications, by mathematising real world problems or by engaging purely in the mathematical realm—new mathematical concepts must still be embraced. We are informed by the theory developed in [7] for these processes. Furthermore, there are many cultural practices and philosophies that are (for many people) common to both science and mathematics. These include reproducibility, refutability, proceeding from a set of axioms or postulates, internal consistency, and even being receptive to a new concept unless and until it is disproven. Thus, in order to learn, embrace and engage in the concepts and philosophies of the mathematical realm we need to extend and grow our world view and hence, in turn, extend our own cultural perspectives. As a result of this discussion we propose a broader question, which we do not answer directly, but provide some observations and insights:

> *Does engaging with mathematics requires an extension of one's culture?*

The article is organised as follows: Section 2 describes the development of the Mathematics Workshop through an assessment of the students' backgrounds and a questionnaire. We then describe our development of the mathematics programme and teaching material in response to the information gathered. Section 3 describes the implementation of the Workshop. Section 4 describes the feedback from students and the FEIT organisers. The conclusions are discussed in Section 5.

2 Development of the Curriculum and Schedule for the Mathematics Workshop

The decision was taken to create an entirely new course for this Workshop. Our primary purpose was to create a course that was more responsive and adaptive than any of our existing programmes, and that would allow us to change the modes of learning and even learning material according to continual feedback from the students and others involved. A secondary feature was the diversity of students' mathematical backgrounds in their school studies—as described below—which also led to a decision that one single stream of Mathematics would be inappropriate.

Our rationale for a highly adaptive course, in light of our principles, was driven by three empirical factors: our experience with the diverse cultural backgrounds of students including Indigenous students, the diverse backgrounds of these particular students (as described in Sections 2.1–2.2), and the responses to our questionnaire (as described in Sections 2.3–2.4). We believe that to advance the cooperation and collaboration of a group as a whole we need to respect the individuality of all participants of the group. This belief is consistent with the following observation:

> one approach to culturally responsive teaching that works with one Aboriginal student may not necessarily work with another. Consequently, any framework for culturally responsive teaching with Aboriginal and Torres Strait Islander students need to take into account the individuality that rests within all students. ([4], p.3–4)

2.1 Analysis of the Students' Backgrounds

As part of understanding student backgrounds, the MLC and FEIT surveyed the students about their home state and what level of mathematics subject (if any) the students were studying at school. The survey showed that all of the students were attending a mathematics course at school, but that the courses varied considerably. Because the students originated from many different states and territories and because the mathematics curricula vary considerably across these regions, then the students had studied some very different topics in mathematics at school. The level of mathematics ranged from year 10 pre-calculus courses to the non-calculus General Mathematics (year 11) through to Extension 1 Mathematics in New South Wales and Mathematics B & C in the Queensland state system. Specifically, this means that some of the students had studied calculus and others had not. Furthermore, all the different topics studied are taught using different techniques across the states and territories and even between local areas and schools.

2.2 Initial Consideration of Topics and Pedagogy, in Light of Background

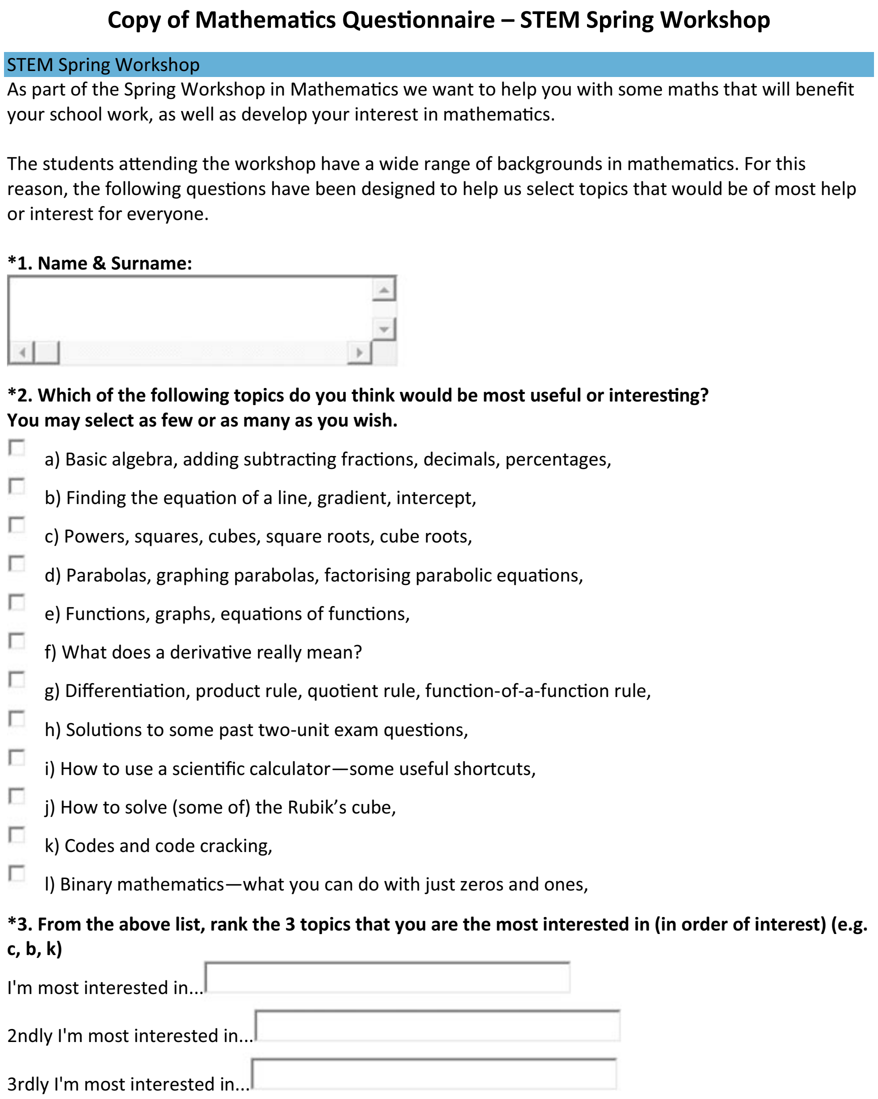
Several teaching decisions were made based on the students' backgrounds. In particular, it was decided to include calculus, but not as a mandatory component for all students. It was also decided to include group work components.

The decision to include calculus was made on the basis that calculus forms the most significant part of the Mathematics Higher School Certificate (HSC) curriculum needed for most university STEM courses. This decision was later bolstered by student input as reported below. However, it was considered critical that some of the students who had not studied calculus should not be forced into a class beyond their mathematical grounding at the time, because that could be detrimental. For instance, if a year 10 student was compelled to take a calculus class then that student may well be overwhelmed. This could lead the student to incorrectly conclude that the subject is beyond their capabilities when the subject may be well within their reach given a proper and timely grounding.

Group work was included in part because of its significance in promoting the worth of individuals' contributions. Once a student recognises that their individual input and perspective are valued, and possibly quite unique, this reinforces the value of their individual contribution to any collaborative work, which reinforces an understanding of the potential of collaboration being greater than the sum of the parts. Conversely, once a student sees that his or her individual contribution may advance the understanding of other members and improve the output of the group, this can in turn reinforce the value of each student's individual contribution. This is aligned with one of our principles; to respect, value and promote the individuality of each participant, and to provide an environment that can support and grow individual contribution to cooperation and collaboration.

2.3 Providing a Means of Feedback to Establish Workshop Topics, Levels and Teaching Modes

In order to design a course that would help all of the students advance their understanding of mathematics as much as possible, it was decided to ask the students what they thought they most needed help with.



Copy of Mathematics Questionnaire – STEM Spring Workshop

STEM Spring Workshop

As part of the Spring Workshop in Mathematics we want to help you with some maths that will benefit your school work, as well as develop your interest in mathematics.

The students attending the workshop have a wide range of backgrounds in mathematics. For this reason, the following questions have been designed to help us select topics that would be of most help or interest for everyone.

***1. Name & Surname:**

***2. Which of the following topics do you think would be most useful or interesting? You may select as few or as many as you wish.**

□ a) Basic algebra, adding subtracting fractions, decimals, percentages,

□ b) Finding the equation of a line, gradient, intercept,

□ c) Powers, squares, cubes, square roots, cube roots,

□ d) Parabolas, graphing parabolas, factorising parabolic equations,

□ e) Functions, graphs, equations of functions,

□ f) What does a derivative really mean?

□ g) Differentiation, product rule, quotient rule, function-of-a-function rule,

□ h) Solutions to some past two-unit exam questions,

□ i) How to use a scientific calculator—some useful shortcuts,

□ j) How to solve (some of) the Rubik's cube,

□ k) Codes and code cracking,

□ l) Binary mathematics—what you can do with just zeros and ones,

***3. From the above list, rank the 3 topics that you are the most interested in (in order of interest) (e.g. c, b, k)**

I'm most interested in...

2ndly I'm most interested in...

3rdly I'm most interested in...

Fig. 1 Introductory comments and questions 1–3 of the online questionnaire for students of the STEM Spring Workshop

A questionnaire was designed by four teachers of the Workshop, (the authors, Dr. Erwin Lobo and Mr. Alexander Majchrowski), and Mr. Collin Zheng. Questions 1–3 of the online questionnaire are given in Figure 1.

The remaining four questions asked were:

4. How difficult would you rate the following question: Find the equation of the straight line passing through the points (2,3) and $(-1,4)$
   □ easy □ somewhat easy □ somewhat difficult □ difficult
5. How do you rate your confidence in Mathematics
   □ not confident □ low confidence □ medium □ confident □ highly confident
6. Is there anything else you would like to let the organisers know?
7. Any other topics that you think would be useful or interesting to cover during the programme

The rationale behind the questionnaire design was as follows:

We provided options for topics to be covered in the Workshop firstly to stimulate thought about different topics in mathematics. However, we also provided a chance for the students to suggest any other topics that would be useful or interesting to cover during the Workshop. Some alternative, motivational topics were included that were not directly part of the HSC curriculum. These motivational topics that we call interest topics were included as a means of inspiring the students' interest and passion for mathematics beyond their high school curricula. The interest topics included aspects of the Rubik's cube, Cryptography and Binary Mathematics and anything else the students may suggest.

Students could share any cultural knowledge voluntarily through question 6, but were not required to answer direct questions about such cultural knowledge. This is consistent with methods established in the MLC. That is, rather than requiring a student to answer direct questions on cultural knowledge we would instead establish an environment of support and provide the opportunity for the student to then share their cultural knowledge voluntarily, if and when the student felt that this was appropriate, useful or desirable.

2.4 Results and Consequences of the Additional Questionnaire

The results of the questionnaire were informative, interesting and enlightening.

Of particular interest were the responses to Question 2; a histogram is given in Table 1. It is useful to distinguish between the types of topics that we asked about in Question 2. We call topics a–i in Figure 1 that are related to the HSC curriculum, *curriculum-topics*, and topics j–l in Figure 1 *interest-topics*.

There was a wide spread in preferences amongst the curriculum-topics. The highest ranked topic, covering basic algebra, fractions, decimals and percentages was originally included as a revision of elementary, but important concepts. The next two highest ranked curriculum-topics were 'finding the equation of a line …', which was included as a foundation topic for calculus, and 'product rule, quotient rule, …' of differential calculus, which is one of the most advanced topics among all of the topics in the HSC curriculum.

Table 1 Histogram of response to online questionnaire question 2, see Figure 1

Online questionnaire question	Response histogram
Basic algebra, adding subtracting fractions, decimals, percentages	X XX XXXXX X
Finding the equation of a line, gradient, intercept	X X X X X X X X
Powers, squares, cubes, square roots, cube roots	X X X X X X X
Parabolas, graphing parabolas, factorising parabolic equations	XXX X X X
Functions, graphs, equations of functions	X X X X X X X
What does a derivative really mean?	X X X
Differentiation, product rule, quotient rule, function-of-a-function rule	X X X X X X X X
Solutions to some past two-unit exam questions	X X X X X X X
How to use a scientific calculator–some useful shortcuts	X X X X
How to solve (some of) the Rubik's cube	X X X X X X X X
Codes and code cracking	X X X X X X X X
Binary Mathematics—what you can do with just zeros and ones	X X X X X

The interest topics were originally included to indicate to the students that they could choose 'extra curricular' topics other than core HSC subjects, partly because these may provide a different mode of learning and motivate interest in some different types of mathematics. Because these topics were not directly related to the students' HSC we were not sure how popular they would be. However, they proved popular with eight of the students choosing the 'Rubik's cube', eight choosing 'Codes and code cracking' (not all the same students) and four choosing 'Binary Mathematics'. This was encouraging because we thought that including these topics was a powerful way of cultivating and motivating further studies in STEM subjects, and allowed the teachers to further show a passion for mathematics; we thought that the students may share some fun with the teachers in exploring these ideas, and be inspired or motivated by the enthusiasm of the teachers.

This widespread of topics selected by the students, from foundational to more advanced topics, reinforced the need for different streams of classes. As a result of this information, it was decided that we should provide one stream that covered more advanced calculus topics and one stream that covered more fundamental foundation topics—primarily focused on the equation (and graphs) of a line and algebraic manipulations.

It was also decided that the two groups should form one group for the interest topics (being the Rubik's cube and Cryptography sessions). This was due both to the popularity of these topics, and because otherwise, students would not have the opportunity to interact with the whole group during the Workshop.

Beyond what we have described here, all student feedback was used in designing the Workshop lessons and teaching material, including responses to the questions on previous mathematical knowledge (question 4), confidence (question 5) and all other feedback (questions 6, 7). Furthermore, a spreadsheet of the students' backgrounds and responses were constructed and figured heavily in the course design; trying to ensure optimal help for the student group as a whole.

2.5 Resources Committed to Developing and Teaching the Workshop

Here we provide an indication of the resources required to develop such an initiative. Dr. Lobo and Mr. Majchrowski, who taught/moderated the algebra/calculus streams, contributed to the Workshop as casual employees. However, it is clear that they dedicated many extra hours beyond the standard preparation time to provide an exceptionally high-quality experience for all involved. The authors dedicated some weeks to organise and teach the Workshop. We dedicated time and effort beyond the normal duties as a recognition of the value of the Workshop. It is estimated that over 300 combined hours were devoted to produce the Workshop. This dedication by all of the teachers was at least in part a recognition of the importance of advancing the participation of Indigenous students in STEM studies for the future.

3 Teaching, Learning and Evolving the Workshop

It was imperative that the themes of knowing and responding to the students continued to play a significant role during the Workshop, and not be limited to the development phase alone. Here we recount how the Workshop evolved and how we attempted to remain responsive to the students' input and perspectives.

3.1 Setting Expectations that Students will Have a Genuine Voice

Our entire structure supported the principle that students have a respected and effective voice in the design of their learning. An example of this is provided by the first session. Given the information we had available from our questionnaire, we could have simply assigned students to one of two streams: calculus or algebra. Instead, we gave them a choice.

The way in which we provided choice was important. We first thanked them for travelling (in some cases considerable distances) during their holidays to attend the programme and for taking time to respond to our questionnaire. We then explained the format of what we had in mind for the Mathematics Workshop. The proposed content of material in the two streams–calculus and algebra–was outlined, and the students could then ask questions about the proposed subjects. We avoided associating the two streams with 'hard' or 'easy', in order to foster a sense of collegiality and cooperation rather than competition. Those who were not sure were offered the opportunity to speak to one of the teachers to discuss their concerns and then determine which option was most appropriate based on their individual circumstances. Throughout this process, it was emphasised that we would try to continuously and immediately respond to the students' input and change the sessions accordingly. Finally, the students could change classes so that they felt free to try a class even if they were not sure. In the end, one student opted to switch from the calculus to the algebra class. After this change, the classes stayed at eight for calculus and six for algebra.

3.2 Responding to Students Immediately and Continuously

It was important that all sessions contain significant student participation. For this reason, each class had a number of teachers helping at any one time. For instance, for both of the interest topics, all of the teachers helped with all of the student groups. However, each of the sessions had one person to moderate the learning. These were as follows: Dr. Lobo for algebra, Mr. Majchrowski for calculus, and Dr. Phillips for the interest topics. Because we intended the learning to be a cooperative venture, all of the teachers were considered as guides or moderators as instructors.

The teachers were specifically selected because, being members of the MLC, they had demonstrated an ability to be agile, responsive and adaptive to different perspectives including cultural perspectives, as described in the introduction. We also wanted to reassure the teachers that they could act on the input from students. To this end, it was made clear that they could at any stage change the learning mode of the class to suit the subject and individual feedback from the students for that topic. This could include utilising small groups, individual instruction, etc., and if necessary slowing down or speeding up the pace of the class. Moreover, the teachers were given scope to use their knowledge and experience to significantly modify the subject material in response to the input and needs of the students. In short, if something in the class was not working the teachers had complete autonomy to change any part of the sessions.

For example, the students in the calculus class needed and wanted specific practice at performing the mechanics associated with the product and chain rule, so for part of this session Mr. Majchrowski provided further practice examples and exercises. For the algebra session, however, students wanted a deeper understanding of how changing different elements in the equation of a line changed its graph. As a result, the students and Dr Lobo interacted cooperatively in question and answer feedback to advance the students understanding of this concept.

Hence, neither the method of learning nor subject matter were completely predetermined.

3.3 How a Culture of Inquiry was Shared Between Streams

By the end of the first hour on the first day, two different modes of operation had emerged in the two different streams. The students in the calculus class wanted to practice and hone mathematical techniques and were happy to work either individually or in small groups. This class was focused and diligent. The algebra class, however, evolved its own learning mode that suited the students and subject matter. As such, the algebra class engaged in strong interaction and much vigorous questioning.

The teachers collectively decided that the following Cryptography session would work well with the vigorous questioning, interactions and cooperation already evident in the algebra session, and so it was decided to transfer the calculus class into the algebra work room. This technique proved successful as the students from the calculus group soon adopted the questioning, interacting and cooperating nature of the algebra class. After this, the Workshop embraced a culture of supportive cooperation for all of the students, teachers and even the 'houseparents'. The houseparents (one for every three to four students) accompanied the students and helped as guides, mentors and chaperones. They were selected based on academic records, previous volunteering experiences, and willingness to inspire younger students to further study in STEM. The houseparents may themselves be Indigenous and or be past participants of similar workshops.

3.4 Incorporating Cooperation and Collaboration in the Cryptography Session

The Cryptography session started with a discussion about where and when codes have been used, and how they are being used now. With input from both the students and the teachers, it was recognised that coding is used extensively for encrypting information in internet communications. It was pointed out that buying or selling items on the internet or any other confidential internet transaction would be impossible without advances in Cryptography.

With input from both teachers and students, we continued to discuss how coding and code cracking played a critical role in the course and outcome of World War II. Some interesting details of the code cracking efforts at Bletchley Park, UK were used to help engage and motivate the students. It was interesting and inspiring for the teachers that some of the students already knew significant details that contributed to the breaking of the German enigma code.

The group then progressed to the process of encipheringVigenère cyphers. To describe the different modes of learning in the Cryptography session we briefly outline how to code and decode a Vigenère cypher. As an example, for a Vigenère cypher of key length 1 and shift 3, each letter in the original message is translated three steps in alphabetical order. Thus, A is translated to D and X is translated back to A etc. For a Vigenère cypher of key length n, we first assign each letter in the message a position p, and the message is then split into subgroups according to the value of $p \bmod n$. Each subgroup is then translated by a (possibly) different shift length.

In the Cryptography session, a set of exercises was developed over many iterations. The students had the opportunity to form different working groups of two students to work on these exercises. As an example of this, one of the questions in the exercises was to decode an encrypted message of key length 2. For a code of key length 2, the odd numbered letters in the original message will be translated by a given shift and the even numbered letters translated by a (possibly) different shift. For this exercise, some of the groups of students could crack the odd letters in the code and the other groups could crack the even letters in the code.

An effective method of deciphering any part of a Vigenère code is to compare the frequency distribution of the letters in the code to a 'standard' frequency distribution of letters in English text. Thus, it is possible to guess how far the letters in the original message has been shifted to produce the code. The students in each group worked together cooperatively to first sort the code into odd and even letters, construct the frequency distribution of their part of the code, attempt to guess the shift length and then finally crack the code by translating each of the letters backward in the alphabet by the shift length. Finally, as part of this exercise the groups who had cracked the odd numbered letters could combine their results with the groups who cracked the even numbered letters to decipher the whole message.

To promote curiosity and further investigation—either individually or as small groups—a more difficult (optional) question was given to the students to do after class. Some of the students spent considerable time and effort successfully deciphering this difficult code.

The students were generally engaged with all the Workshop topics, even though this could vary from topic to topic and from student to student. As Mr. Majchrowski observed, 'Some students wanted to move ahead more quickly than others...', while others 'wanted more problems or more challenging problems.' All of the participants of the Cryptography session seemed highly motivated and engaged. Indeed, even the houseparents wanted to try their hand at cracking some of the codes with the students. Perhaps this is because the process of cracking codes promoted a cooperative mode of learning. Possible improvements for a similar Cryptography session in the future include; setting more exercises after class because some of the students were keen to continue cracking more codes than were originally provided, spending more time describing the difference between a single key cypher and a 2-key cypher and providing more time motivating Cryptography through real world examples.

Generally, for all of the streams, we could provide help for the students through a number of different means. For example, we could provide individual help in answering such questions as; *Have I got this right?*, *Can I do this using an easier method* or *Am I right if I do it this way?*. Individual, tailored help was provided in the curriculum topics of algebra and calculus through to the interest topics of Cryptography and the Rubik's cube. The spectrum of help broadened for the interest topics with students more prepared to help other students, teachers helping students, students helping teachers and even the houseparents getting involved. These interactions grew as the Workshop progressed.

3.5 Promoting Inquisitive, Critical Thinking with the Rubik's Cube

The Rubik's cube can be regarded as a puzzle, a game or even a toy. However, there are some captivating and intriguing mathematical concepts embedded in the process of manipulating a cube. The first part of this session was used to motivate the idea that there are very different forms of mathematics including the mathematical subject of group theory. To this end, each student was given their own cube.

The students were given simple combinations of moves. As one example, the students were shown the combination of rotating; the front face clockwise, the right-hand face clockwise, the front face anti-clockwise, and then the right-hand face anti-clockwise. To be systematic this combination of moves was given the notation $FRF^{-1}R^{-1}$. The students were asked the question:

> *Can you find how many times we have to perform $C = FRF^{-1}R^{-1}$ before the cube returns to its original state?*

Most students were surprised or even amazed that the cube would progress through a seemingly 'random' pattern and return back to its original state after repeating this combination six times. This was used to motivate the idea of a cycle in group theory.

The teachers and students observed that if we firstly performed the combination two times called C^2 and secondly performed it another four times called C^4, then in a sense, the second action would 'undo' the first. That is, performing C^2 and then another C^4 is an identity operation, or $C^2C^4 = I$. In this sense C^4 can be regarded as an inverse of C^2.

Thus, by approaching this exercise as a game, the concepts of a group, a finite group, an identity operation and the idea of an inverse was introduced, invented or discovered by the students. The discussion then progressed to the multiple ways that we can identify inverses just from these observations.

It was also pointed out that C^3 only changed four corners on the cube and that this can be very useful in completing the Rubik's cube, such as, where one needs to manipulate four of the corner pieces and leave all other pieces unchanged. The students were engaged by these new mathematical concepts and wanted more time to play with these ideas.

The class then progressed to completing part of the cube. To complete a Rubik's cube some important basic concepts are needed. To (dramatically) demonstrate this idea a cube was dismantled in front of the students. By looking at the various pieces of the cube the students readily identified that the pieces at the centre of each face do not move to any other place. They also observed that edge pieces stay edge pieces and so do corner pieces. Thus, by asking students to use the colours at the centre of each face to decide 'what pieces must be moved to where' the students learned that it is possible to develop a systematic method of solving the cube. A series of operations for moving pieces into their identified places were then described and a series of 'how to' instructional handouts were given to the students. The students then progressed to trying to solve the cube individually or in small groups with the help of teachers and, as it turned out, from two enthusiastic students who could already solve the cube.

These students were keen to help the other students solve the puzzle. This session was designed to be an informal puzzle or play session to motivate, engage and inspire.

The interaction and enthusiasm demonstrated that the students possessed a capacity to work from the abstract mathematical realm and possibly relate this to their previous knowledge of mathematics.

4 Assessing the Mathematics Workshop

The Mathematics Workshop was evaluated using three different methods of assessment. The first method was an optional anonymous questionnaire, the second was a survey of the whole STEM Spring Workshop conducted by the FEIT, and the third was the continual and careful evaluation by teaching staff of student feedback and input during the Workshop.

The optional, anonymous questionnaire was completed by all 14 students, and included the questions:

1. As a result of attending the Workshop my understanding of Mathematics in the topics taught has:
   Decreased a lot Decreased Stayed the same Increased Increased a lot
2. As a result of attending the Workshop my confidence in Mathematics in the topics taught has:
   Decreased a lot Decreased Stayed the same Increased Increased a lot
3. Which parts of the Mathematics Workshops did you like (if any) and why?
4. Were there any parts of the Mathematics Workshops that you did not like? Why?
5. Do you think any parts of the course should be given more time? Which parts? Why?
6. Do you think any parts of the course should be given less time? Which parts? Why?
7. How do you think the Mathematics Workshops could be improved?

The primary aim of our evaluation questionnaire was to determine whether and how the Mathematics Workshop had affected the students' understanding and confidence in mathematics. The rationale behind the inclusion of both Questions 1 and 2 was our observation that an increase in confidence can be independent of (although associated with) an increase in understanding of mathematics. In Question 1, about understanding, all students indicated either 'increased' (57%) or 'increased a lot' (43%). In Question 2, about confidence, all students indicated one of: 'stayed the same' (22%), 'increased' (57%), or 'increased a lot' (21%). These results are heartening because many students do not normally consider mathematics the most inspirational or motivational of subjects, and in general in mathematics classes, it is possible for a student to leave a session feeling as if their understanding has not improved or may have even decreased. From our experience with the Workshop we believe the learning and teaching culture, along with the mode of learning, contributed the most to the increase in confidence.

In our evaluation questionnaire, we also wanted to know which parts of the Workshop may need to be either revised or removed and which parts needed to be expanded or improved. The overwhelming response was that the students mostly thought that there should be more time spent on Cryptography and the Rubik's cube. However, some students indicated that they wanted more time spent on calculus. No student indicated that there should be less time spent on any session.

The second method of assessing the Workshop was provided by the FEIT who conducted a survey of the entire STEM programme. The students were asked to score each of the different activities in the programme (including the Mathematics Workshop) out of 5. The Mathematics Workshop was rated at 63/70, which was one of the highest rankings of all the components of the whole programme.

The third method in which the Workshop was evaluated was the continual and careful evaluation of student feedback and input during the Workshop. A commonly unrecognised and yet critical component of this evaluation is the provision of an environment whereby the students' input is reinforced as being a valued contribution. The results and implications of this method of assessment are best understood through the description of the interactions during the Workshop itself, as described in Section 3. We emphasise that this method of continual evaluation can be significantly more important than just using an assessment after the course because unless the feedback from the students is genuinely valued and used continually to evolve the mode of learning, the engagement, enthusiasm and input from the students can be diminished. As expressed by Dr. Lobo: 'It is paramount to keep the students with you even if this is at a small cost to the material covered in the time.' For example, Dr. Lobo would ask the students if they wanted to contribute on the board: 'I would sit with the students to help them feel more relaxed and less intimidated. This was helpful.' Once the feedback cycle is broken it can be difficult, indeed impossible, to re-establish.

5 Conclusions

Our primary conclusion is that all the design elements in the Mathematics component of the STEM Spring Workshop was significant and important in its success in engaging and supporting the learning of the students.

The multiple efforts to provide students with a genuine voice (background survey, interests questionnaire and feedback throughout the Workshop) is part of our philosophy of cultural plasticity and led to innovations not present in any of our bridging courses. Unique innovations included the two differentiated streams, the inclusion of interest topics (for which the streams came together) and the process of empowering the teachers to freely adapt mode and content at the moment as needed. One of the reasons for such efforts arose from a heightened sense of responsibility to support the inclusion of Indigenous students in the fields of STEM education, and this is perhaps best expressed by Dr. Lobo: 'I put a lot of effort and care into thinking about the class in preparation for the Workshop because I care about Indigenous outcomes.'

As a consequence of this approach, the modes of engagement were collaborative and community focused throughout, consonant with the observation from [1] p.456, paraphrasing [3], that Aboriginal cultures '…emphasise relationship over task and cooperation over competition with a preference for cooperative and collaborative learning models in classrooms…'.

The results of adopting a practice of listening and responding, and employing cultural plasticity are reflected in observations from the teaching staff. Mr. Majchrowski noted, 'I was surprised how quickly the students knew each other, formed subgroups and how quickly the class warmed up as the Workshop went on.' Dr. Lobo kindly observed that:

> I was very pleased that the group coordinating this session asked for and valued our input in the preparation of these courses, organised times where we could discuss the material, talked at length about how best to approach the classes and received feedback from students about topics they were interested in. I felt this was all directed explicitly towards creating the best possible class we could for the students.

The multifaceted nature of the Workshop provided an excellent opportunity to connect many aspects of the three worlds of Mathematics, Science and Engineering as well as the students' whole-life experiences. The students exhibited an aptitude for embracing some of the broad principles of Science when, for instance, a simple operation was performed on the Rubik's cube over and over again. When the cube eventually returned to its original state the students were surprised, even astounded. This new idea that we could perform a complex operation until the cube returned to its original state demonstrated the idea of the order in a finite group. The students were also engaged with the concept that performing an operation until a system is returned to its original state can be abstracted and generalised to the idea of an inverse. The recognition that this was a completely new type of mathematical system was a revelation to many of the students. This demonstrated a willingness, indeed eagerness, to embrace the fact that a scientific theory can be advanced and valued even with no apparent (foreseeable) real world application.[1] Even though some students are motivated by real world applications to mathematics, many are also fascinated with the pure creativity, elegance and imagination of solving puzzles and problems within the mathematical realm.

The affinity that students showed for scientific ways of thinking, and their display of prior knowledge is consistent with the literature showing that Indigenous students may have slightly higher levels of interest in science vs non-Indigenous peers ([5] p. 2027), and that interest and engagement in science may be strongly correlated with science-related activities outside of school ([10] p.233). Furthermore, such activities may also be associated with higher levels of Science literacy:

[1] Some such examples include lasers, nuclear magnetic resonance, Galois theory, Hilbert spaces and differential geometry.

> High-performing Indigenous students, on average, report participating at higher levels in out-of-school science-related activities in comparison to all Indigenous students, and in comparison to all non-Indigenous students. ([9] p.vii)

Because of the unique nature of the selection criteria for the Workshop, which was different from our usual bridging courses that are open to general members of the public, an interesting research question arises. *Is this complete Workshop process—survey, questionnaire, design, response—equally valuable for other cohorts of students from different cultural backgrounds?* We hope that the knowledge gained in this project may help advance the learning of students from a broad cultural tapestry.

Acknowledgements We would like to graciously acknowledge the contributions of Erwin Lobo, Alexander Majchrowski and Collin Zheng for their conscientious and considerable contributions toward the development of the Mathematics Workshop programme and the teaching and learning during the Workshop. We would like to thank Jackie Nicholas for the valuable suggestions. We would like to acknowledge the stewardship and kind cooperation of Alberta-Mari Nortje and Keiran Passmore, two of the founders of the STEM Spring Workshop. We would like to acknowledge the suggestions of an anonymous referee and the insight and input of the editors. Finally, thanks to the inevitable wordsmith behind the scenes, Laurence Cantrill.

References

1. Boon, H.J., Lewthwaite, B.E.: Signatures of quality teaching for Indigenous students. Aust. Educ. Res. **43**(4), 453–471 (2016). https://doi.org/10.1007/s13384-016-0209-4
2. Dreise, T., Thomson, S.: Unfinished business: PISA shows Indigenous youth are being left behind. Australian Council for Educational Research, Melbourne (2014). https://www.acer.edu.au/occasional-essays/unfinished-business-pisa-shows-indigenous-youth-are-being-left-behind
3. Duchesne, S., McMaugh, A., Bochner, S., Krause, K.L.: Educational psychology: for learning and teaching, 4th edn. Cengage Learning, Australia (2015)
4. Krakouer, J.: Literature review relating to the current context and discourse on Indigenous cultural awareness in the teaching space: critical pedagogies and improving Indigenous learning outcomes through cultural responsiveness. Australian Council of Educational Research, Melbourne (2015). https://research.acer.edu.au/indigenous_education/42
5. McConney, A., Oliver, M., Woods-McConney, A., Schibeci, R.: Bridging the gap? A comparative, retrospective analysis of science literacy and interest in science for indigenous and non-indigenous Australian students. Int. J. Sci. Educ. **33**(14), 2017–2035 (2011). https://doi.org/10.1080/09500693.2010.529477
6. Perso, T.: Cultural responsiveness and school education: with particular focus on Australia's first peoples: a review and synthesis of the literature. Menzies School of Health Research, Centre for Child Development and Education, Darwin (2012)
7. Phillips, C.: An improved representation of mathematical modelling for teaching, learning and research. Int. J. Innov. Sci. Math. Educ. **23**(4), 51–63 (2015)
8. Song, S., Perry, L.B., McConney, A.: Explaining the achievement gap between Indigenous and non-Indigenous students: an analysis of PISA 2009 results for Australia and New Zealand. Educ. Res. Eval. **20**(3), 178–198 (2014). https://doi.org/10.1080/13803611.2014.892432

9. Woods-McConney, A., McConney, A.: Indigenous Student Success in Science. Murdoch University, School of Education, Perth (2014)
10. Woods-McConney, A., Oliver, M., McConney, A., Maor, D., Schibeci, R.: Science engagement and literacy: a retrospective analysis for indigenous and non-indigenous students in Aotearoa New Zealand and Australia. Res. Sci. Educ. **43**, 233–252 (2013). https://doi.org/10.1007/s11165-011-9265-y

Origami as a Teaching Tool for Indigenous Mathematics Education

Michael Assis and Michael Donovan

1 Introduction

Origami is an ancient art form, with deep roots in many cultures, having arisen independently multiple times throughout history. Since most cultures can relate to folded art in one medium or another, we posit that origami can be an ideal tool for use in Indigenous education, in particular mathematics. Due to the varied connections between origami and many areas of mathematics, e.g. [81], many educators are increasingly studying the use of origami as an educational tool in mathematics education, from primary school to university [82].

While origami may act as a tool for learning in the classroom, it can also readily be integrated with the Indigenous education methodology of narrative. The use of 'storigami', the telling of a progressive story throughout a folding procedure has been used [49, 50] in the teaching of origami to students and can readily be adapted to diverse cultures by the incorporation of Indigenous cultural knowledge and icons. As an art form, origami is very versatile in its ability to represent a wide range of objects and concepts [57]. For instance, we argue that origami is capable of representing Indigenous Australian dot and cross-hatch paintings using the origami style of 'tessellations' [29]. In particular, we demonstrate similarities of existing origami tessellations with Indigenous dot paintings. Furthermore, we also provide an original example of an origami representation of a woven Indigenous wakwak toy. Origami representations of Indigenous art are means through which mathematics can be taught and practised using Indigenous methodologies as well as artistic styles.

M. Assis (🖂)
School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, Australia
e-mail: assis.michael@gmail.com

M. Donovan
Academic Director Indigenous Learning and Teaching, Walanga Muru, Macquarie University, Sydney, NSW, Australia
e-mail: michael.j.donovan@mq.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_12

Translating the traditional paint medium or Indigenous woven articles into origami requires mathematical thinking at a level accessible to secondary students. Aside from the translation effort, the folding of origami models appears to have many benefits already at the Kindergarten level, e.g. [14].

The outline of the paper is as follows. We begin by defining origami, tracing the history of folded art and papermaking around the world in various Indigenous settings, and then consider the history of paper folding. The practice of folding and papermaking in Indigenous cultures and Australia in particular appears to have been little studied so far. We note that this review adds to and corrects existing origami histories of origami. Through this history, it is clear that both paper and folding have deep roots in many cultures around the world. We then give a brief history of the recent progress in the analysis of the mathematics found in origami, as well as the use of origami in education. Next, we discuss the importance of the Indigenous methodology of narrative and its effectiveness in Indigenous education, and the appearance of narrative in Indigenous artwork. We then show examples where origami can be used as a medium to represent Indigenous art, first through an original origami model representation of an Indigenous toy together with suggested mathematics explorations and then by comparing existing origami tessellations with Indigenous dot paintings. We then discuss recent original examples of Indigenous storigami and end with a summary.

2 Origami: Definition and History

The word origami typically evokes images of folded coloured paper, usually square, folded into recognizable shapes without the aid of cutting or glue. If origami often has the connotation of folding paper, it has more recently been recognized among origami artists that the art should more properly be defined as ‘the art of folding’, since some origami artists use media other than paper [66, 76]. The origami artist Chris Palmer, for example, has created many ‘origami’ works of folded fabric [89], and the name origami has been attached to other folded media as well [45, 101]. These examples provide the motivation for our working definition in this article of origami as ‘the art of folding’, regardless of medium.

From this perspective, the art of origami is much older and widespread than the history of papermaking. For example, the oldest documented employment of folding of cloth likely comes from ancient Egypt, where pleats were folded into the fabric of men’s skirts in the Old Kingdom era, that is, before 2130 BC [62]. However, the simplest example of folding materials employed in many cultures is flat leaves, such as banana leaves, used in cooking; the use of folding in cooking is likely very ancient and it is independently used across many cultures. Weaving, such as basket weaving, is also an ancient art form, although folding is not often employed in an essential manner. Paper is doubtless a very suitable material for folding, especially paper made using the modern papermaking techniques developed in China between the second century BC and the first century AD [97].

It is significant that traditional methods of papermaking have been independently discovered by a number of distinct Indigenous cultures, although it is not clear the extent to which these other types of papers could be folded beyond simple folds. Ancient Egyptians produced papyrus, which forms the root of the word 'paper' [97], and the ancient Mayans and Aztecs of Mesoamerica produced amate, a type of bark paper. Many indigenous cultures have used barkcloth throughout history, where the bark of a suitable tree is usually beaten or pressed and then used as clothing or in crafts. For example, Asia and the Pacific regions used tapa, which may have originated in China [97], Indigenous Australians used paperbark [25, 98], the Caribbeans used lacebark, which did not need to be beaten [12], the Pacific Northwest used cedar bark [80], and the Baganda people of Uganda made barkcloth from the Mutuba tree, which is listed as an UNESCO Intangible Cultural Heritage of Humanity [102]. Beyond simply the production of paper, these Indigenous papers were also folded. The oldest map known, drawn on papyrus and dated from around 1150 BC, is said to have been rolled up [33], the oldest known example of amate dated circa 74 AD is said to have been crumpled [6], and bark paper is mentioned as having wrapped poison in the second century BC in a historical account from China [97]. The rolling, the crumpling, and the wrapping of these ancient paper artefacts provide documented evidence of early folding of paper.

The knowledge of (modern) Chinese papermaking slowly spread around the world, bringing with it the possibility of independent paper folding traditions. It spread around Asia after a few centuries, and much later, through Arab traders, it arrived in Europe [35]. It appears that it was not until almost a millennium after its invention that folded paper is mentioned in China [97] in the Tang dynasty (618–907). Likewise, though Chinese papermaking arrived in Japan by the sixth century [94], the earliest reference to folding paper practices is a description in a later manuscript from 1850 of the start of the practice sometime around the twelfth century of the Helan period (794–1185), although possibly later. In Europe, which received the papermaking technique in the eleventh century [35], the earliest direct references to paper folding are from the fifteenth century. The famous Dutch manuscript dated circa 1440 'The Hours of Catherine of Cleves' [86] contains a marginal painting of three origami boxes, and a woodcut diagram in the 1498 Paris edition of the book Sphaera Mundi has a depiction of a folded paper boat [90]. Later, Europe developed a long tradition of folding cloth napkins, from at least the sixteenth century [63, 69, 94]. Spain in particular developed a long independent tradition in paper folding, with the earliest known reference possibly being from 1757 [94], and at some point early in the twentieth century this Spanish tradition migrated to Argentina, so that the continent of South America also developed its own independent tradition of paper folding [88]. By the mid-twentieth century , origami organizations and a flurry of books popularized origami internationally, and the Internet has further helped to transform it into a global art form.

Nevertheless, today origami is still often associated with Japanese culture. Japan was the first country to publish books dedicated to paper folding, starting in 1704 [94]. While paper folding in Europe appears to have been seen as mostly a hobby, there were deep cultural associations given to paper and folded patterns and their uses in

Japanese society [70, 72, 75]. Due the high price of manufacturing paper, initially it had only ceremonial uses [94]. As the price decreased and paper become more widely accessible, the tradition of paper folding became widespread, with models and patterns being passed down generally from mother to daughter [59]. Also, following the introduction to Japan in 1888 of Friedrich Fröbel's kindergarten curriculum [67, 73], Japan readily adopted the use of origami in school curricula, a practice which continues to this day. Equally notable is Japanese origami artist Akira Yoshizawa, who in the 1930s decided to devote his time to studying paper folding as an art form [64, 65]. Yoshizawa's artistic works were groundbreaking in their number, in his introduction of new techniques, and in their sensitivity and attention to detail. In 1951, his work received international recognition, sparking an international movement that eventually reached Lillian Oppenheimer in New York [68]. Lillian's enthusiasm was such that she decided to create the world's first international paper folding organization in 1958, to foster dialogue and collaboration of paper folding artists worldwide [71]. In honour of Yoshizawa's profound contributions to the art form, Oppenheimer decided to call the art form 'origami' in English [74], borrowing from the Japanese word which means 'folded paper', a name that has since been adopted in many countries and languages today.

It is perhaps due to Japan's extensive practice and popularization of origami that today origami still often mistakenly carries a Japanese connotation. However, it is clear that the art of folding has been independently discovered by many Indigenous cultures, spanning thousands of years, making origami a very suitable teaching tool in the Indigenous classroom.

3 Origami Mathematics

If the history of origami is very ancient as just discussed, the history of the analysis of the mathematics in origami is quite recent. Very few mathematical properties of origami were explored before the 1980s. In 1989 an international effort to compile known results and bring researchers together produced the first conference dedicated to origami science, mathematics and technology. This was the start of a dedicated movement to study origami mathematics and its scientific applications, continuing to present day. In the 1990s, several pioneering mathematical origami software was written by Robert Lang—using mathematical principles to design new models [60] and to find coordinate points constructed from folding procedures [58]. In this decade several important mathematical proofs appeared [8, 16] and the second conference on scientific origami occurred. With such a start, the next decade saw tremendous growth in origami mathematics, with two more conferences, the publishing of several books dedicated to origami mathematics [17, 31, 40, 56], and further mathematical origami software written [4, 95, 96]. Today origami mathematics and its applications are being developed at an ever faster rate, as seen by the number of published mathematical and scientific articles in recent years.

At a first glance, it is easy to recognize that origami can exemplify plane geometry through the shapes which appear after folding or the shapes which appear in the pattern of creases in the unfolded sheet of paper. Less obvious is the fact that constructions in origami can be treated using field and Galois theory [1, 2], that curved crease origami requires differential geometry for its analysis [37], and that number theory arises in the study of knotted strips of paper [78]. In fact, interesting mathematics seems to arise in every area of origami [39, 55, 81, 99], and much of it is suitable in mathematics education, from the concept of limits arising in the Fujimoto folding of the binary expansion of fractions [103] to the calculus used to optimize the area of curved folded surfaces [77].

4 Origami in Education

Origami has a long tradition of being used in education, starting with Friedrich Fröbel's Kindergarten movement in the nineteenth century. After his death in 1852, his supporters further developed the origami curriculum and incorporated traditional origami models [67, 73]. As the movement spread, so did the teaching of origami in schools, taking origami as a learning tool around the world by the end of the nineteenth century. The use of origami in Kindergarten progressively diminished in the twentieth century and its use has since shifted towards older students in the teaching of mathematics. We mention briefly that origami has also been used as an educational tool in areas beyond mathematics, such as in helping to teach English communication skills among non-native speakers [27, 36] and in various kinds of therapy [15, 93].

One early publication seeking to use origami to teach mathematics is the 1957 booklet ‘Paper folding for the mathematics class’ [46], which remained in print for several decades. Another similar book aimed at primary and secondary students is by Jones [47]. Due to the great progress in understanding origami mathematics in the 1990s, by the early 2000s, a number of textbooks appeared for the teaching of mathematics through origami. The first of its kind was the book Project Origami by Thomas Hull [40, 41], which contains a series of ready-made lesson plans and activities for mathematics classrooms, from primary to university level. The book serves the needs of a variety of teachers looking for a short mathematical activity to incorporate within a standard curriculum. In 2008 Kazuo Haga published a textbook with exercises, meant for high school level mathematics classes [31]; it is not clear to what extent Haga's book has been used as a course textbook. At a higher level, Erik Demaine published a university textbook with Joseph O'Rourke [17], which he uses as the main course textbook for a yearly course he teaches on algorithms at MIT.

Beyond these examples of published books and textbooks, many educators have written recently on the educational value of using origami in classrooms. There are many instances in the literature of its use in Kindergarten and primary schools [14, 26], middle schools [11], secondary schools [13], and universities [10, 28], for example. It appears that origami is useful for developing learning skills [30], spatial ability [10, 11], cognition and language reasoning [100], and problem-solving and creativity [87]. In fact, since 1991 regular conferences have been organized on the topic of origami's use in education. The proceedings from these conferences [15, 39, 55, 82, 93, 99] provide a useful compendium of the current and increasing knowledge on the use of origami in the modern mathematics classroom.

5 Narratives in Education: Theory and Evidence

Paralleling the growth of origami as an educational instrument is the increasing understanding of embracing cultural diversity through presenting diverse pedagogical practices. This can be seen with the engagement of narrative or the use of story to embed Indigenous and other cultures into mainstream educational institutions. The use of narrative is one of the most significant expressions of an Indigenous presence in educational research. This is shown when Indigenous voice is the central presentation of meaning; using Indigenous expression, tone and metaphor to paint the images that are portrayed through the narrative. The use of story is present through all aspects of the lives of Indigenous peoples, including Australian Aboriginal people, and has been used as a major learning and information tool [21, 34]. Using narrative as an educational tool emphasizes the importance of story in people's lives. Peacock notes this point in relation to his Ojibwe heritage and how this differs to many non-Indigenous interpretations of the use of narrative: 'Many Native people learn their way in life through stories. The use of stories reflects Ojibwe standpoint epistemology and differs from what many academic readers are used to' [7]. It allows the Aboriginal voices from the "past, present, future" (as cited in [24]) to be embedded within the understanding being presented in any current research questions. As Thomas King, an American Indian author and storyteller inform us, 'the truth about stories is that's all we are' [53]. King reinforces these comments by informing non-Indigenous readers of the depth of story and narrative within an Indigenous context. All stories from an Indigenous context can be seen as a paperbark tree, or as being multi-layered, where the learner can see the first layer but through reflection and exploration, other layers of information and understanding can present themselves as the learner gains a clearer footing in their examination.

Battiste et al. [5] highlight that societal institutions are normative, seeking conformity among the population as a homogenous group, leading to differences being ignored or pathologized. If teachers want to gain the best from their students these differences could be recognized and embraced in order to engage all students in the learning environment. This could include having the students' culture recognized within the students' learning environment such as through the use of a narrative or story from the students' cultural heritage. This is an intimate feature of Aboriginal culture and thus an appropriate tool to draw upon and focus the Aboriginal students towards the learning experiences through connected narratives. Story is such a culturally recognized Aboriginal learning practice in both traditional and contemporary Aboriginal society. The Aboriginal students are able to share their epistemological standpoint through the use of story, humour, and lived experiences, without being limited by the barriers that other methodological practices can have on personal expression such as questionnaires and surveys, particularly if these are designed without considering culturally diverse participants.

The ability of Aboriginal people to express who they are by positioning themselves within their community reinforces the importance of story as a social tool for Aboriginal people. It allows them to not only present their individual expressions but also place them into their social context. McLeod (as cited in [54]) reminds us that the cultural importance of story can carry different meaning and significance:

> The significance of story within Indigenous culture is less contested. Rather, it is the nature and structure of story that causes difficulties for non-tribal systems due to its divergence from the temporal narrative structure of Western culture. For tribal stories are not meant to be orientated within the linearity of time, but rather they transcend time and fasten themselves to places. (pp. 95–96)

6 Linking Story to Country and Culture

Story is an output of Aboriginal Australian literacy. All societies have some form of literacy that is based on the social practices and understandings of that community. The literacy of Aboriginal society is the reading of the Australian landscape from a local Aboriginal perspective, the landforms remind people of their local histories and stories that are embedded in how these landscapes were formed [32]. In traditional and contemporary Aboriginal communities, the re-telling of the Dreaming of their country reminds families of their values and culture. The 'Dreaming' for Aboriginal people is the notion of their 'past, present and future' being acknowledged in their lived experiences. It is the presentation of the Aboriginal worldview of Creating Spirits informing Aboriginal people of their existence and their relationship with the environment from the deepest roots in the land to the stars in the skies above. The Dreaming allows Aboriginal people to embrace their past to inform their present actions to allow for a future for their communities to maintain their existence [23]. This can be shown as a family passing through a place where an event occurred during the creation period of the Dreaming. The landscape features are ever-present markers to remind Aboriginal communities today of these past events thus reinforcing the values or cultural understandings that are implicit in that event and in the landscape, connecting communities to their Dreaming in country in their present time.

These actions of connecting culture to place will guide younger individuals towards correct behaviour in their future, thus maintaining the cyclic continuous nature of the Dreaming and highlighting the authority or rigours of behaviour that the Dreaming maintains in their lives. Whilst the land still exists and their spirit is in the land, the Dreaming (Aboriginal law) is present within Australia. Campbell, a senior Yarralin Elder, drew together these concepts of land, knowledge and belonging when he spoke with Bird-Rose [9]:

> You see that hill over there? Blackfella law like that hill. It never changes. Whitefella law goes this way, that way, all the time changing. Blackfella law is different. It never changes. Blackfella law hard—like a stone. Like that hill. The law is in the ground. (p. 56)

Hence story has authority within Aboriginal culture as a continuation of the Dreaming. By retelling the Dreaming, Aboriginal Australians are able to engage with their Creating Spirits' activities and the knowledge they passed down to all life in Australia, thus maintaining connection to country, including all living and inanimate features of Country.

7 Story, Culture and Artistic Expressions

Aboriginal knowledge are not separate aspects of Aboriginal people's lives, but they are embedded across all features of their behaviour. This has been consistently noted across Aboriginal artistic traditions where Aboriginal artists can use various art forms to re-tell their history, their life stories, their culture and their inter-relationships that they maintain with their country. This is noted in [42] when Arrernte and Pitjantjatjara communities highlight how the use of artistic techniques is important in the maintenance of culture. Through the process of making art the artists re-tell their oral histories passing on knowledge of law, values and appropriate behaviours. This educational information is embedded in the artistic practice that can help younger community members learn song, ceremony and place these understandings in context to their country and the landscape. As Ray Kelly, a prominent Aboriginal playwright and linguist states, 'Aboriginal people are people of story' [52]. Kelly [52] and Mundine [84] both reinforce this issue through highlighting that learning and understanding can be embedded within cultural practice and expression. This can be presented in story-telling, art or performance, but as with many Aboriginal representations of the arts, cultural validity is at the centre of the work.

The validity of cultural understandings is at the centre of Yunkaporta's 8-ways of Aboriginal Learning pedagogical framework (8-ways) [22, 104] highlighting that Aboriginal cultural understandings are not content driven but are developed through the Aboriginal epistemological processes that are ingrained with appropriate pedagogy, such as using the 8-ways framework. Narrative of origami can highlight mathematical principles whilst working through artistic processes. These processes engaging with narrative can be noted in many Aboriginal cultural practices such as weaving or tool making that connect to the practice and cultural relevance within the community. Yunkaporta [104] supports these principles in 8-ways, highlighting that the balancing of culturally relevant practices within educational processes can go beyond teaching through the use of simple worksheets in order to engage culture, story, community, history and land links through highlighting the relevance of the learning to the student.

8 Narrative in Origami

There are two different ways in which narrative has been utilized together with origami. The first is very old, where a folded origami model can assume several alternate forms and a story can be told relating each of the forms. Magicians have created stories to accompany the various transformations of the origami Troublewit model since at least the eighteenth century, if not earlier [94]. More recent examples include models published by Jeremy Shafer [91, 92]. In these examples, the origami model is being used as a prop for the story, and the audience does not interact with the model nor are they taught how to fold it.

A second type of narrative is increasingly being used in the teaching of origami, however, a technique called 'storigami', a term coined by Rachel Katz [49]. Storigami aims to tell a progressive story, so that each stage of the story roughly coincides with the state of the paper as it is being folded into its final shape [48, 79]. It is primarily used to help in the teaching of an origami model to an audience, such as a classroom, and it involves an interplay of the visual, auditory and tactile senses of the student [43, 44]. In a mathematics class, it can 'make mathematics visual and tangible' [79], as an instructional method based on constructivist theory. Many examples of ready-made storigami appropriate for teaching in classrooms are given in [49], with further examples found in [79, 91, 92]; also Iyer and Katz [43, 44] list known educators using it in schools. Of particular note is its use in the Cherokee nation in Oklahoma using traditional stories in the Cherokee language showing that is can readily be adapted to many different cultural contexts.

In 2005, Mastin conducted a survey of more than 1,000 pre-service and in-service educators who had previously taken part in her storigami workshops and subsequently used storigami in their own classrooms [79]. The educators reported an increase in confidence and motivation of their students for learning mathematics and an increase in their mathematical language and communication skills with other students and parents. The educators also noted the advantages of this approach, since it combines 'memory enrichment, small motor coordination, right brain-left brain interaction, and inventiveness' [79]. Using nothing but paper, the students were engaged directly with mathematics.

9 Origami and Indigenous Art

Origami has had a long history within various Indigenous cultures, as mentioned in section 2. These histories have been embedded within cultural stories to allow for Indigenous youth to engage with cultural practices but also to reinforce local story and extend their culture into every aspect of their lives [34, 52, 84]. Using the modern origami practices of origami, Indigenous Australian art forms can be reproduced using the different medium of paper whilst still allowing for the narrative behind the original art form to be engaged with. This allows an invitation to Indigenous students into the practice and learning of mathematics through the sharing of these complex cultural experiences.

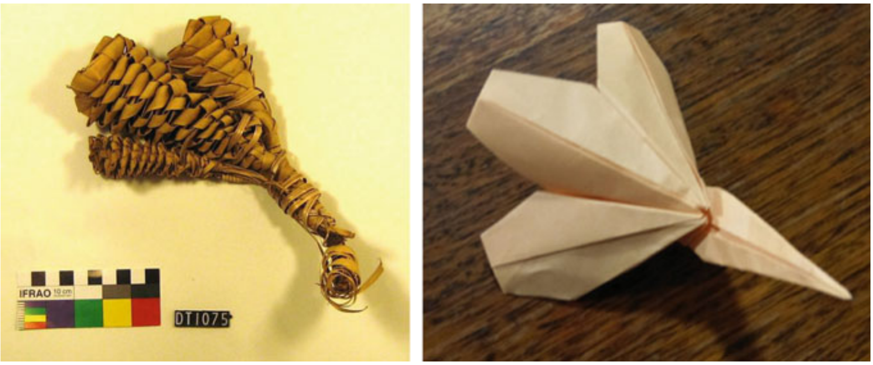


Fig. 1 On the left an example of a woven wakwak toy from 1937 of the Liyagalawumirr community in Glyde River, Eastern Arnhem Land, Northern Territory, Australia [20, 85]. On the right, an origami model representation of the wakwak toy, by the author M. A..

Origami as an art form uses a variety of techniques to achieve its goals of representational and abstract art, techniques which increasingly have been expounded in books, such as [19, 29, 57, 61, 83]. The available sophistication is such that any object of study could conceivably be represented using origami. Therefore, the representation of traditional Indigenous cultural heritage does not pose an artistic barrier and can therefore be utilized in the teaching of origami in the Indigenous mathematics classroom.

As an example, we show in Figure 1 an example of a woven wakwak toy from 1937 of the Liyagalawumirr community in the Northern Territory, Australia [20, 85]. It is the representation of the roots of a waterlily, woven using pandanus leaves. Next to it, we show a simplified origami representation, created by one of the authors (M.A.), appropriate for teaching in primary schools. As a tool in teaching, it can be used to teach plane geometry, for example. This simple model can be modified in various ways in order to better represent the original artwork.

In order to proceed from a basic model to increasingly more realistic models, an analysis of the unfolded paper's crease pattern is useful and illuminating, shown in Figure 2 on the left. We could, for example, wish the stem to be longer. We can accomplish that through the use of a border graft technique, described in [56, 57]. One choice of border graft is shown in the middle of Figure 2, with the result shown on the right of Figure 2.

From such the analysis of the crease pattern, questions of trigonometry arise naturally, appropriate for a secondary level class. If one wishes, calculus can be employed to optimize the extra stem length versus the resulting wasted paper area, shown in grey in Figure 2, appropriate for university level or perhaps for a secondary school student project. Furthermore, the design of similar models itself can be learned, for example through [56, 57], and in the process of design, a range of mathematical concepts will be acquired, including graph theory, trigonometry, circle packing and optimization. In engaging with the mathematics of an origami design, there is also

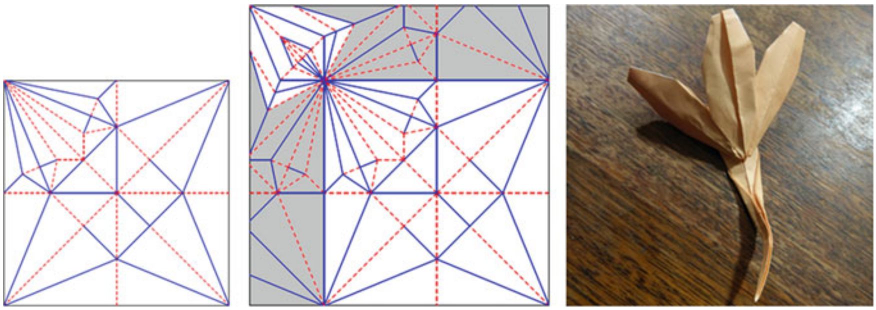


Fig. 2 On the left the unfolded simple origami models crease pattern, in the middle the crease pattern of the border graft modification, with 'wasted' paper shown in grey, on the right the result

a simultaneous engagement with the narrative traditions related to the Indigenous artwork.

As a second example, we show that the techniques available in origami tessellations can be used to represent patterns and motifs found in Indigenous Australian dot paintings. In Figure 3 we show two sketches of dot paintings that highlight some key elements of the paintings, paired with an existing origami tessellation, created and folded independently from the painting. Though the origami examples are not exact representations of the paintings, nor were they inspired by them, our goal is to show a sample of the range of suitable existing techniques available for further adaptation. These techniques can be learned through various sources [19, 29], so that geometric concepts can be acquired in an engaging and entertaining way while connecting to deeper Indigenous culture and heritage. In the process of learning origami tessellations, we posit that the student will naturally and playfully learn mathematical concepts such as plane geometry, trigonometry, and plane tiling.

As remarked in [42], the process of painting an Australian dot painting often begins in one area, progressing outward, and in the process, a parent may narrate a story to their child. Similarly, origami tessellations begin in one area of the paper and progress outward. Through the incorporation of traditional symbols and motifs into an origami tessellation, a story can be told in the process of folding the tessellation, so that the process of learning mathematics can be united with Indigenous culture through representation and narration.

In the process of teaching origami in an Indigenous classroom the technique of storigami can be usefully employed in a manner which is celebratory and respectful of the heritage of an Indigenous community. Indigenous storigami examples are very scarce in the literature. Recently both authors have participated in the Aboriginal and Torres Strait Islander Mathematics Alliance (ATSIMA) conference, where they presented two original examples of Indigenous storigami, developed by one of the authors (M.A.), which could be suitably employed in an Indigenous mathematics classroom [3]. One storigami told the Dreaming story about the Thukeri fish from the Ngarrindjeri people of Lake Alexandrina in South Australia. As the story was told, various intermediate states of the paper were used to represent distinct aspects of the

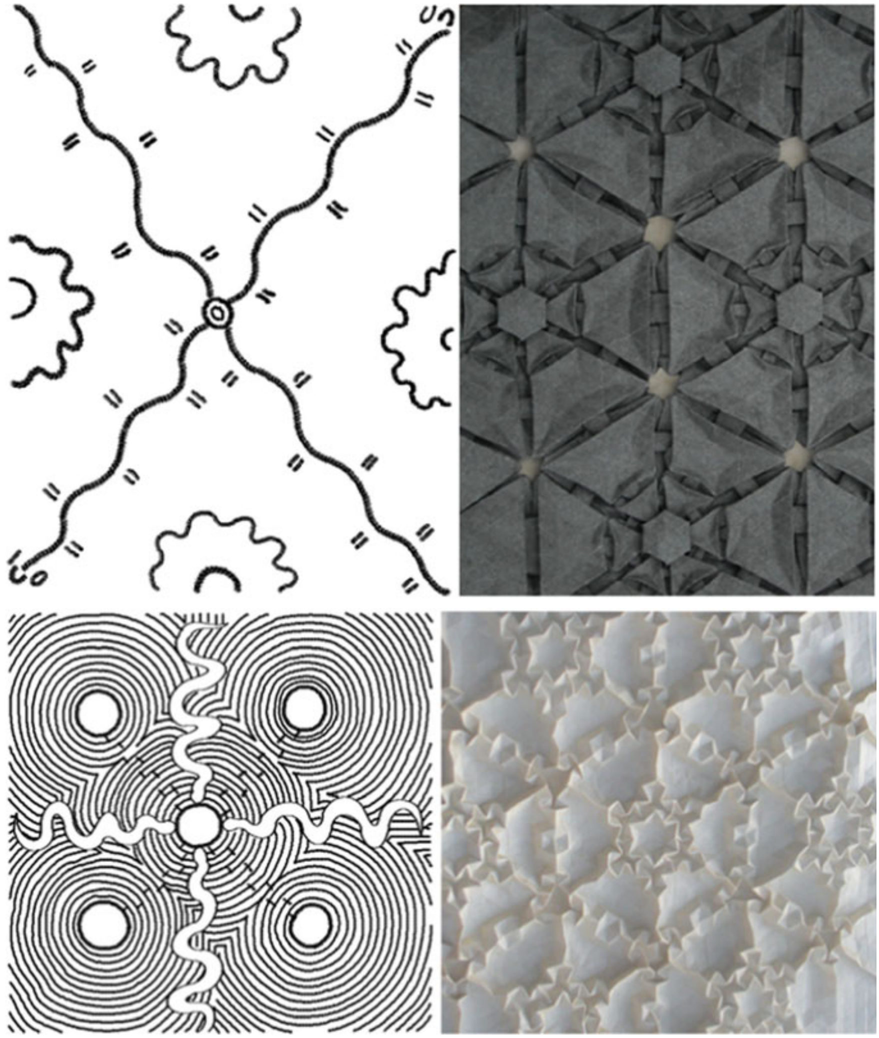


Fig. 3 On the left, sketches of original dot painting works; in the upper left a sketch of 'Four snakes at Piltadi' (1986) by Riley Major Tjangala and in the lower left a sketch of 'Yala dreaming' (1988) by Pauline Woods Nakamarra, in [42]. On the right, origami tessellation examples from Benjamin DiLeonardo-Parker [18]

story as it progressed, finally terminating in an origami fish. The second Dreaming story concerned how the water got to the plains, as well as the origin of the emu and the blue-tongued lizard, from the Butchulla people of Fraser Island, Queensland. Again, various points of the story were represented in the progression of folding, terminating first in an emu, then finally transformed into the blue-tongued lizard. We propose that as a teaching tool and method, storigami can be readily adapted to a wide variety of Indigenous classrooms. While our Indigenous storigami has not yet been presented in such a classroom setting, the wide-ranging survey results of [79] suggest that it will benefit Indigenous mathematics education.

As noted earlier origami has been recognized to support students' spatial learning, language reasoning and creativity. Hughes [38], one of the earliest Australian Indigenous pedagogical theorist has argued that spatial-visual learning is key learning preferences for Indigenous students. The use of practices that engage with these pedagogical principles would allow Indigenous students to connect with these mathematical skills. Spatial understanding was also highlighted by Kearins [51] who identified the sophisticated spatial understandings and positioning that Indigenous youth had in his work with remote South Australian children. Through engaging with origami practices, Indigenous students could be supported in their development of Standard Australian English (SAE) language acquisition through engaging with specific origami sequencing language [27, 36]. The development of origami spatial sequencing in response to Indigenous spatial understandings can help to engage and to support Indigenous creativity to develop complex and sophisticated mathematical origami art forms.

10 Summary

The story of origami development is an ancient and widespread practice that has existed within many Indigenous cultures and other old global communities. These skills have been shared internationally and altered to suit individual communities' needs and views on artistic beauty. In the late twentieth century, origami and its mathematical properties were noted and examined by a variety of mathematicians. These mathematicians together with mathematical educators drew together the beauty of the art form with the beauty of the mathematical understandings to help make mathematics education more accessible.

This paper highlights some mathematical educators in their approach via origami to engage their students in mathematical principles focused on various pedagogical practices. In order to support students in gaining greater understanding and fluidity of practice, the use of narrative and story allows for better student recall through their storigami practices. By consideration of Indigenous Australian culture and art, we provide examples of ways in which origami model representations, origami tessellations, and storigami could be adapted for use in an Indigenous Australian mathematics classroom, where the use of narrative can be highlighted and traditional artistic symbols and motifs can be represented. The central use of story in these pedagogical examples sits within the context of the global histories of paper and folding and the significant use of story that multiple Indigenous peoples use to inform and strengthen their cultural practices within our international and changing environments. Reinforcing the importance of story as a pedagogical tool, it can engage not only Indigenous students but a wide array of students to many complex mathematical principles by allowing the students' culture and worldviews to be easily adapted and incorporated within the students' thinking.

Despite the highlighted potential for origami and storigami to engage Indigenous students with mathematics, and in view of the large positive study conducted by Mastin in the United States on the use of storigami in mathematics classrooms [79], we are unaware of attempts to use origami as a tool in Indigenous Australian mathematics education. We posit that such a methodology will have a positive impact on Australian Indigenous mathematics students and hope this proposal will lead to its application in classrooms in Australia.

Acknowledgements M. D. gratefully acknowledges Walanga Muru at Macquarie University and the Wollotuka Institute at the University of Newcastle and the Aboriginal students and communities who have shared their stories and histories in the development the author's Aboriginal pedagogical theories. M. A. gratefully acknowledges origami artist Benjamin DiLeonardo-Parker for his help finding suitable origami tessellations and the use of his photographs, as well as Daniel Cohen for his help in finding the relevant page in [86].

We gratefully acknowledge the life of Jonathan M. Borwein who founded the CARMA research centre at the University of Newcastle, and in whose honour CARMA organized the Jonathan M. Borwein Commemorative Conference Satellite Meeting, Mathematics and Education: Spirit, Culture and Community which brought these two authors into collaboration. We also thank Judy-anne Osborn as an organizer of the conference and for suggesting the writing of this joint paper.

References

1. Alperin, R.C.: A mathematical theory of origami constructions and numbers. N. Y. J. Math. **6**, 119–133 (2000)
2. Alperin, R.C.: Trisections and totally real origami. Am. Math. Mon. **112**(3), 200–211 (2005)
3. Michael Assis, M., Donovan, M.: Storigami: using narrative and maths to embed an indigenous perspective. Paper presented at the Aboriginal and Torres Strait Islander Mathematics Alliance (ATSMIA) Conference, Melbourne (2018)
4. Bateman, A.: Tess, (2018). Software Retrieved http://www.papermosaics.co.uk/software.html
5. Battiste, M., Bell, L., Findlay, I.M., Findlay, L., Youngblood Henderson, J.S.: Thinking place: animating the indigenous humanities in education. Aust. J. Indig. Educ. **34**, 7–19 (2005)
6. Benz, B.F., López Mestas C.L., De La Vega, J.R.: Organic offerings, paper, and fibers from the Huitzilapa shaft tomb, Jalisco, Mexico. Anc. Mesoam. **17**(2), 283–296 (2006)
7. Bergstrom, A., Cleary, L.M., Peacock, T.D.: The seventh generation: native students speak about finding the good path. In: ERIC Clearinghouse on Rural Education and Small Schools, Charleston, WV (2003)
8. Bern, M., Hayes, B.: The complexity of flat origami. In: Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '96), pp. 175–183. Society for Industrial and Applied Mathematics, Harrisburg (1996)
9. Bird-Rose, D.: Dingo makes us human: life and land in an Australian Aboriginal culture. Cambridge University Press, Cambridge (1992)

10. Boakes, N.: Origami and spatial thinking of college-age students. In: Wang-Iverson, P., Lang, R.J., Yim, M. (eds.) Origami^5: Fifth International Meeting of Origami Science, Mathematics, and Education, p. 173. CRC Press, Boca Raton (2011)
11. Boakes, N.: The impact of origami-mathematics lessons on achievement and spatial ability of middle-school students. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, pp. 471–481. AK Peters Ltd., Natick (2009)
12. Brennan, E., Harris, L.-A., Nesbitt, M.: Object lesson Jamaican lace-bark: its history and uncertain future. Text. Hist. **44**(2), 235–253 (2013)
13. Cagle, M.: Modular origami in the secondary geometry classroom. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, pp. 497–506. AK Peters Ltd., Natick (2009)
14. Carter, J.A., Ferrucci, B.J.: Instances of Origami within Mathematics Content Texts for Preservice Elementary, p. 337. AK Peters Ltd., Natick (2002)
15. Cornelius, V.A.: COET95: Proceedings of the Second International Conference on Origami in Education and Therapy. Origami USA, New York (1995)
16. Demaine, E.D., Demaine, M.L.: Computing extreme origami bases. CS-97-22 (1997)
17. Demaine, E.D., O'Rourke, J.: Geometric Folding Algorithms: Linkages, Origami, Polyhedra. Cambridge University Press, Cambridge (2008)
18. DiLeonardo-Parker, B.: Best of, 2018. Retrieved https://www.flickr.com/photos/brdparker/sets/72157601694127026
19. DiLeonardo-Parker, B.: Six Simple Twists: The Pleat Pattern Approach to Origami Tessellation Design. CRC Press, Boca Raton (2016)
20. Donald Thomson Collection. On loan to Museums Victoria from The University of Melbourne. Item DT 1075 Toy, Wakwak, Liyagalawumirr, Glyde River, Eastern Arnhem Land, Northern Territory, Australia, (1937)
21. Donovan, M.J.: Aboriginal landscapes and their place in the classroom. Int. J. Sci. Soc. **2**(3), 243–252 (2011)
22. Donovan, M.J.: Aboriginal student stories, the missing voice to guide us towards change. Aust. Educ. Res. **42**(5), 613–625 (2015)
23. Donovan, M.J.: What form(s) of pedagogy are necessary for increasing the engagement of aboriginal school students? PhD thesis, University of Newcastle (2016)
24. Dunbar, C.: Critical Race Theory and Indigenous Methodologies, pp. 85–99. Sage Publications, Thousand Oaks (2008)
25. Elliot, W.R., Jones, D.L., et al.: Encyclopaedia of Australian Plants Suitable for Cultivation, vol. 6. K–M, p. 359. Lothian Publishing Company Pty Ltd., Melbourne (1993)
26. Fiol, M.L., Dasquens, N., Prat, M.: Student teachers introduce origami in kindergarten and primary schools: froebel revisited. In: Wang-Iverson, P., Lang, R.J., Yim, M. (eds.) Origami^5: Fifth International Meeting of Origami Science, Mathematics, and Education, pp. 151–165. CRC Press, Boca Raton (2011)
27. Foreman-Takano, D.: Applications of origami to the teaching of sophisticated communication techniques, p. 235. AK Peters Ltd., Natick (2002)
28. Frigerio, E.: In praise of the papercup: mathematics and origami at the university. In: Origami 3: International Meeting of Origami Science, Mathematics and Education, pp. 291–298. AK Peters, Ltd., Natick (2002)
29. Gjerde, E.: Origami tessellations: awe-inspiring geometric designs. AK Peters Ltd., Natick (2009)
30. Golan, M., Jackson, P.: Origametria: a program to teach geometry and to develop learning skills using the art of origami. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, pp. 459–469. AK Peters Ltd., Natick (2009)
31. Haga, K.: Origamics: Mathematical Explorations Through Paper Folding. World Scientific Publishing Company, Singapore (2008)
32. Hanlen, W.: Learning from the centre not the margin. Paper presented at the Ninth International Literacy and Education Research Network Conference on Learning, Beijing, China, July 16–21 (2002)

33. Harrell, J.A.: Maps and mapmaking in Egypt: turin papyrus map. In: Selin, H., (ed.) Encyclopaedia of the History of Science, Technology, and Medicine in Non-Western Cultures, pp. 1292–1301. Springer, New York (2008)
34. Heitmeyer, D.: It's Not a Race: Aboriginality and Education, pp. 220–248. Social Science Press, New Delhi, India (2004)
35. Hills, R.L.: Papermaking in Britain 1488–1988: A Short History. Bloomsbury Publishing, London (2015)
36. Ho, L.Y.: Origami and the Adult ESL Learner, p. 247. AK Peters Ltd., Natick (2002)
37. Huffman, D.A.: Curvature and creases: a primer on paper. IEEE Trans. Comput. **(C-25)**, 1010–1019 (1976)
38. Hughes, P., Moore, A.J.: Aboriginal ways of learning and learning styles. Paper presented at the Annual Conference of the Australian Association for Research in Education, December 1997, Brisbane, Australia
39. Hull, T.: Origami^3: Third International Meeting of Origami Science, Math, and Engineering. AK Peters Ltd., Natick (2002)
40. Hull, T.: Project Origami: Activities for Exploring Mathematics. AK Peters Ltd., Natick (2006)
41. Hull, T.: Project Origami: Activities for Exploring Mathematics, 2nd edn. CRC Press, Boca Raton (2012)
42. Isaacs, J.: Australian Aboriginal Paintings. Weldon Publishing, Richmond (2002)
43. Iyer, S., Katz, R.: Origami: an interdisciplinary approach using storigami. Paper presented to the Fifth International Meeting of Origami Science, Mathematics, and Education (5OSME), Singapore (2010)
44. Iyer, S., Katz, R.: Using storigami to teach origami (2011)
45. Jenkins, A.: The Lost Art of Towel Origami. Andrews McMeel Publishing, Kansas City (2005)
46. Johnson, D.A.: Paper Folding for the Mathematics Class. National Council of Teachers of Mathematics. National Education Association, Washington (1957)
47. Jones, R.: Paper Folding: A Fun and Effective Method for Learning Math. LWCD Incorporated, St. Louis (1995)
48. Kallevig, C.P., Draper, E.: Folding Stories: Storytelling and Origami Together as One. Storytime Ink International, Painesville (1991)
49. Katz, R.: Storigami: Storytelling and Origami by Rachel Katz (2018). Retrieved https://web.archive.org/web/20180818210230/http://origamiwithrachelkatz.com
50. Katz, R.: Personal communication
51. Kearins, J.: Visual spatial memory in aboriginal and white Australian children. Aust. J. Psychol. **38**(3), 203–214 (1986)
52. Kelly, R.: Remembered Words, pp. 64–72. Nelson Cengage Learning, Boston (2011)
53. King, T.: The Truth about Stories: A Native Narrative. University of Minnesota Press, Minneapolis (2003)
54. Kovach, M.: Indigenous Methodologies: Characteristics, Conversations, and Contexts. University of Toronto Press Incorporated, Toronto (2009)
55. Lang, R.J.: Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education. AK Peters Ltd., Natick (2009)
56. Lang, R.J.: Origami Design Secrets: Mathematical Methods for an Ancient Art. AK Peters Ltd., Natick (2003)
57. Lang, R.J.: Origami Design Secrets: Mathematical Methods for an Ancient Art, 2nd edn. CRC Press, Boca Raton (2011)
58. Lang, R.: Referencefinder (2018). Software Retrieved http://www.langorigami.com/article/referencefinder
59. Lang, R.J.: The Complete Book of Origami: Step-by-Step Instructions in Over 1000 Diagrams. Dover Publications Inc., Mineola (1988)
60. Lang, R.J.: Treemaker (2018). Software Retrieved http://www.langorigami.com/article/treemaker

61. Lang, R.J.: Twists, Tilings, and Tessellations: Mathematical Methods for Geometric Origami. CRC Press, Boca Raton (2018)
62. Laver, J., et al.: Dress (clothing): The History of Middle Eastern and Western Dress. Encyclopaedia Britannica, Inc., Chicago (2017). Retrieved https://www.britannica.com/topic/dress-clothing
63. Lister, D.: 16th century and later napkin folding (2018). Retrieved http://www.britishorigami.info/academic/lister/napkin_folding.php
64. Lister, D.: Akira Yoshizawa (2018). Retrieved http://www.britishorigami.info/academic/lister/yoshi1.php
65. Lister, D.: A tribute to Akira Yoshizawa 1911–2005 (2018). Retrieved http://www.britishorigami.info/academic/lister/yoshizawa_tribute.php
66. Lister, D.: Fold your arms: materials for folding (2018). Retrieved http://www.britishorigami.info/academic/lister/foldmat.php
67. Lister, D.: Friedrich Froebel (2018). Retrieved http://www.britishorigami.info/academic/lister/froebel.php
68. Lister, D.: The exhibition of paperfolding by Akira Yoshizawa in Amsterdam 1955 and its place in the origins of modern origami (2018). Retrieved http://www.britishorigami.info/academic/lister/yoshizawa_exhib1955.php
69. Lister, D.: The history of paperfolding : a German perspective, (2018). Retrieved http://www.britishorigami.info/academic/lister/german.php
70. Lister, D.: Japanese attitudes to origami (2018). Retrieved http://www.britishorigami.info/academic/lister/japanese_attitudes.php
71. Lister, D.: Lillian oppenheimer and her friends (2018). Retrieved http://www.britishorigami.info/academic/lister/oppenheimer.php
72. Lister, D.: Origami and the spiritual (2018). Retrieved http://www.britishorigami.info/academic/lister/origami_and_spirituality.php
73. Lister, D.: Origami in schools in japan and the west (2018). Retrieved http://www.britishorigami.info/academic/lister/oriinskools.php
74. Lister, D.: Origami v paper folding (2018). Retrieved http://www.britishorigami.info/academic/lister/ori_vs_paperfolding.php
75. Lister, D.: Paper and religion in Japan (2018). Retrieved http://www.britishorigami.info/academic/lister/religion.php
76. Lister, D.: What is origami? (2018). Retrieved http://www.britishorigami.info/academic/lister/what_is_ori.php
77. Maekawa, J.: A geometrical tree of fortune cookies. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, p. 449. AK Peters, Ltd., Natick (2009)
78. Maekawa, J.: Introduction to the study of tape knots. In: Wang-Iverson, P., Lang, R.J., Yim, M. (eds.) Origami^5: Fifth International Meeting of Origami Science, Mathematics, and Education, pp. 395–403. CRC Press, Boca Raton (2011)
79. Mastin, M.: Storytelling + origami = storigami mathematics. Teach. Child. Math. **14**(4), 206–212 (2007)
80. McColl, M.: Cedar bark weaving comes from deep roots, vol. 27. Windspeaker (2009). Retrieved http://www.ammsa.com/publications/windspeaker/cedar-bark-weaving-comes-deep-roots
81. Miura, K., Kawasaki, T., Tachi, T., Uehara, R., Lang, R.J., Iverson, P.W.: Origami^6: I. Mathematics. American Mathematical Society, Providence (2015)
82. Miura, K., Kawasaki, T., Tachi, T., Uehara, R., Lang, R.J., Iverson, P.W.: Origami^6: II. Education Technology, Art. American Mathematical Society, Providence (2015)
83. Montroll, J.: Origami Polyhedra Design. AK Peters Ltd., Natick (2009)
84. Mundine, D.: Aboriginal Art in New South Wales: Nothing is Lost—Nothing is Nothing—You are not Nothing, pp. 99–113. Nelson Cengage Learning, Boston (2011)
85. Museum of Victoria.: Women's Work : Aboriginal Women's Artefacts in the Museum of Victoria. Museum of Victoria, Melbourne (1995)

86. Plummer, J.: The Hours of Catherine of Cleves, pp. 306–307. George Braziller, New York (1966). MS M.917/945. http://www.themorgan.org/collection/hours-of-catherine-of-cleves/346
87. Pope, S., Lam, T.K.: Using origami to promote problem solving, creativity, and communication in mathematics education. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, pp. 517–524. AK Peters Ltd., Natick (2009)
88. Rozenberg, L.S.: Paper Life: The Story of Ligia Montoya. CreateSpace Independent Publishing Platform, Scotts Valley (2016)
89. Rutzky, J., Palmer, C.K.: Shadowfolds: Surprisingly Easy-to-make Geometric Designs in Fabric. Kodansha International, Toyko (2011)
90. de Sacro Bosco, J.: Uberrimum Joannis de Sacro Bosco Sphaere mundi commentum Petri Ciruelli, insertis etiam questionibus Petri de Aliaco. G. Mercatoris (1498)
91. Shafer, J.: Origami Ooh La La!: Action Origami for Performance and Play. CreateSpace Independent Publishing Platform, Scotts Valley (2010)
92. Shafer, J.: Origami to Astonish and Amuse. St. Martin's Griffin, New York (2001)
93. Smith, J.: Origami in Education and Therapy, Proceedings of the First International Conference on Origami Education and Therapy (COET '91). British Origami Society, Birmingham (1992)
94. Smith, J.: Notes on the history of origami, 3rd edn. British Origami Society, Birmingham (2014)
95. Tachi, T.: Origamizing polyhedral surfaces. IEEE Trans. Vis. Comput. Graph. **16**(2), 298–311 (2010)
96. Tachi, T.: Rigid origami simulator (2018). Software Retrieved http://origami.c.u-tokyo.ac.jp/~tachi/software/
97. Tsuen-Hsuin, T.: Science and Civilisation in China: Volume 5, Chemistry and Chemical Technology, Part 1, Paper and Printing. Cambridge University Press, Cambridge (1985)
98. Turnbull, D.: Maps and Mapmaking of the Australian Aboriginal People. Springer, New York (2008)
99. Wang-Iverson, P., Lang, R.J., Yim, M.: Origami^5: Fifth International Meeting of Origami Science, Mathematics, and Education. CRC Press, Boca Raton (2011)
100. Wilson, M., Flanagan, R., Gurkewitz, R., Skrip, L.: Understanding the effect of origami practice, cognition, and language on spatial reasoning. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, p. 483. AK Peters, Ltd., Natick (2009)
101. Wright, L.: Toilet Paper Origami: Delight Your Guests with Fancy Folds and Simple Surface Embellishments. Lindaloo Enterprises, Santa Barbara (2008)
102. United Nations Educational, Scientific and Cultural Organization (UNESCO): Decision of the intergovernmental committee: 3.com 1. Convention for the Safeguarding of the Intangible Cultural Heritage, Intergovernmental Committee for the Safeguarding of Intangible Cultural Heritage, Third session, Istanbul, Turkey, 4 to 8 November 2008 (2008). Retrieved https://ich.unesco.org/en/decisions/3.COM/1
103. Veenstra, T.B.: Fujimoto, number theory, and a new folding technique. In: Lang, R.J. (ed.) Origami^4: Fourth International Meeting of Origami Science, Mathematics, and Education, p. 405. AK Peters, Ltd., Natick (2009)
104. Yunkaporta, T.: Aboriginal pedagogies at the cultural interface. Ph.D. thesis, James Cook University (2009)

Dynamic Visual Models: Ancient Ideas and New Technologies

Damir Jungić and Veselin Jungić

1 Introduction

In this note, we discuss the use of a dynamic visual model as a surrogate for a proof when a quick justification of a mathematical fact is needed, but a demonstration of a formal proof is, for any reason, not convenient. We define a *dynamic visual model* as an animation that demonstrates a particular mathematical concept.

The choice of dynamic visual models in this work is motivated by the challenge that calculus instructors face when they need to make students go through examples that illustrate the definition of the definite integral as the limit of corresponding Riemann sums. Those examples mostly use the identities for the sum of the first n consecutive positive integers, or the sum of their squares, or cubes. Since, in general, students are not familiar with those identities the instructor needs to introduce them to the class in some way. In a topic-packed calculus course, even the most conscientious instructor would have difficulties to find time to do more than just flash identities in front of the class and to use the textbook as a reference.

One part of the challenge is, that standard calculus textbooks also, in the corresponding section, only list the identities. See, for example [18, 24]. Some textbooks offer their proofs in an Appendix at the end of the book or as exercises. Hence, the student is left to build their understanding of Riemann sums and the definition of the definite integral on trust that these identities, though seemingly coming from nowhere, are true. This issue has been noted by other math instructors. See, for example, [9].

D. Jungić · V. Jungić (✉)
Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
e-mail: vjungic@sfu.ca

D. Jungić
e-mail: dashj88@alumni.sfu.ca

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_13

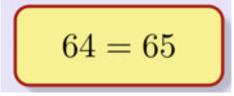


Fig. 1 Click on the image to link to a dynamic visual model of the "proof" that $64 = 65$. The url is http://people.math.sfu.ca/~vjungic/CLF/64-equals-65.html

In this note, we provide dynamic visual models[1] that could be a reasonable patch for the situation described above. Moreover, we wish to further promote dynamic visual models as a tool that instructors can use to communicate mathematical ideas. To underline the value of visualization in communicating mathematics, we accompany our dynamic visual models with the corresponding formal mathematical proofs.[2] We hope to convince the reader that sometimes, as Jonathan Borwein said, "It is easier to see than to say!" [7]

It is necessary for students to clearly understand the distinction between dynamic visual models and formal mathematical proofs. When using dynamic visual models or, so-called, *proofs without words*, the instructor should remind their audience that an important role of mathematics is to check if one's perception of an object obtained through human senses matches the established facts and standards. For example, the dynamic visual model of a Fibonacci Jigsaw Puzzle demonstrates a "proof" that $64 = 65$ (Fig. 1).

See [19] for more Fibonacci Jigsaw Puzzles and an extensive list of references related to this topic.

Therefore dynamic visual models should be used to produce

> the insight that will give the reasons why a property is true [12]

and to remind the student

> never to accept anything for true which [they] did not clearly know to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in [their] judgement than what was presented to [their] mind so clearly and distinctly as to exclude all ground of doubt [10].

Dynamic visual models play an important role in the presentation of mathematics in general. In the words of Mike Bostock:

> Visualization leverages the human visual system to augment human intellect: we can use it to better understand these important abstract processes, and perhaps other things, too [8].

All dynamic visual models that we have produced so far are based on what Alsina and Nelson [1] call *the Cantor principle* and give it as follows:

> If two sets are in one-to-one correspondence, then they have the same number of elements.

[1] Others have also created similar models. See for example [25–28].

[2] None of the proofs in this note uses mathematical induction.

2 Mathematical Proof in the Calculus Classroom

An obvious shortcoming of our dynamic visual models and visual proofs in general comes from the fact that such models demonstrate only a limited number of cases of the particular mathematical phenomenon. Ron Graham, Bruce Rothschild, and Joel Spencer stated the challenge of generalizing patterns recognized in a small number of cases in their *Law of Small Numbers*:

> Patterns discovered for small k disappear for k sufficiently large to make calculations difficult. [11]

Richard Gay reflected on the fact that there are patterns among small numbers that do not hold in the general case in his *Strong Law of Small Numbers*:

> There aren't enough small numbers to meet the many demands made of them. [13]

These *laws*, despite being lighthearted mathematical humour, point out at least two outcomes that the process of teaching and learning mathematics must achieve. One is the awareness that mathematics plays a role in keeping our intuition and hopeful thinking in check. Another is the realization that using not well-defined terms may produce vague statements.

We illustrate these two learning outcomes by the fact established by David H. Bailey et al. [2] that $\pi/8$ and the integral

$$\int_0^\infty \cos(2x) \prod_{n=1}^{\infty} \cos(x/n)\,dx$$

agree to 42 decimal digit accuracy, but are not equal. Only mathematics with its question "What about the 43rd digit?" prevents us from jumping to the conclusion that if the value of the integral agrees to 42 decimal digits with such a nice constant as $\pi/8$ is, then the integral *must* be equal to the constant. And, is 42 a *small number* of decimal places? In this case, apparently it is, even if something as small as 10^{-43} escapes human perception.

Hence, the existence of a formal mathematical proof is and will stay as the ultimate criterion in the process of establishing mathematical facts. But, going back to our calculus classroom, the question that the instructor faces is how to communicate this fundamental idea to a group of students that are often with no or a very limited experience with mathematical proofs?

In 2005 Jonathan Borwein wrote:

> There is a disconcerting pressure at all levels of the curriculum to derogate the role of proof. This is in part motivated by the aridity of some traditional teaching (e.g., of Euclid), by the alternatives now being offered by good software, by the difficulty of teaching and learning the tools of the traditional trade, and perhaps by laziness. [4]

Jonathan Borwein was not the only one to notice the "pressure (…) to derogate the role of proof" in teaching mathematics. For example, Gila Hanna examined "three specific factors that have lent imputes to the decline of proof in the curriculum" in the 1990s:

1. The position adopted in its *Standards* by the National Council of Teachers of Mathematics in the United States that proof needs to be explicitly taught only to those students who intend to pursue post-secondary education.
2. The view of some educators that deductive proof needs no longer be taught because heuristic techniques are more useful than proof in developing skills in reasoning and justification.
3. The idea, suggested and encouraged by the growing use of software with dynamic capabilities, that deductive proof might profitably be abandoned in the classroom in favour of a dynamical visual approach to mathematical justification. [14]

The content of the current curriculum for two high school precalculus courses, Principles of Mathematics 11 and 12, in British Columbia, Canada, supports both Borwein's observation and Hanna's examination: It mentions the notion of *proof* only a few times. The curriculum mandates that students taking Principles of Mathematics 11:

- verify and prove assertions in plane geometry, using coordinate geometry;
- investigate (…) circle properties using computers with dynamic geometry software, and prove them using established concepts and theorems.

The Principles of Mathematics 12 students need to

- use algebraic manipulation to simplify and prove given [trigonometric] identities for the general case;
- use algebraic manipulation to prove given identities involving sum and difference and double angle identities.

Hence, the precalculus curriculum in British Columbia does not demand that the general high school student population be exposed to mathematical proofs in a significant way. As a side note, we mention that the 2015 Programme for International Student Assessment (PISA) ranks the performance of Canadian high school students in mathematics as "Better than OECD average—decline since 2006" [23].

In the authors' experience the lack of exposure to mathematical proofs in high school has a ripple effect: a first-year university calculus course may be designed to reduce mathematical proofs to the level of a straightforward application of established concepts and theorems. We support this claim by a brief analysis of the current learning outcomes for Math 100—Differential Calculus with Applications to Physical Sciences and Engineering, a standard calculus course offered by the Department of Mathematics at the University of British Columbia (UBC), one of the leading Canadian universities. In the section entitled "Course-level Learning Goals" this 8-page long document contains the following:

> Students will also learn how to construct simple proofs. They will learn to show that a given mathematical statement is either true or false by constructing a logical explanation (proof) using appropriate Calculus theorems and properties of functions. In particular, when applying a theorem, students will recognize the importance of satisfying its hypotheses and drawing logical conclusions. [20]

This quote indicates that the authors of the document feel that they need to define to their students what a mathematical proof is ("a logical explanation") and to explicitly deconstruct the notion of a proof into its two key stages:

1. recognizing "the importance of satisfying its hypotheses,"
2. "drawing logical conclusions."

Implicitly, the above quote is suggesting that there is no expectation that the incoming students even know what a *mathematical proof* is.

In the section "Top-Level Learning Goals" the statement from the quote is elaborated in the context of the Intermediate Value theorem and the Mean Value theorem. For example,

> State the Mean Value theorem and its corollaries, use it to construct simple proofs about a given mathematical statement, specifically, recognize when the hypotheses of the theorem are satisfied, apply the theorem accordingly and draw logical conclusions based on it." [20]

The impression is that this *goal* resembles the Principles of Mathematics 11 curriculum, "prove [properties] using established concepts and theorems."

At this point, after realizing that a major Canadian university in its mainstream calculus course has a very modest place for proofs in its "Top-Level Learning Goals," it seems natural to ask the following questions:

Should students see proofs in a standard calculus course?
Should students do proofs in a standard calculus course?

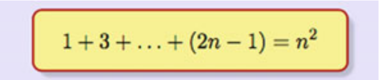
These are the exact questions that Thomas W. Tucker considered in [29]. In short, Tucker's answer was, "It depends which kind of proof we are talking about." Tucker distinguished two types of proofs. He described *Proof I* as the type of proof which "goal is to develop a formal language with which those results can be proved true." On the other hand, *Proof II* "is less formal, and is used to answer a question that is in doubt." While going through the list of "some commonly given reasons" why calculus students should see proofs, Tucker argued that Proof II should be preferred to Proof I:

> Seeing Proof II may help more [than Proof I] in later courses, but not nearly so much as doing Proof II. [29]

At the end of the paper, Tucker gave a few examples of Proof II and concluded:

> The point is that problems like these should be in a calculus course, and at present they are not. [29]

Even though Tucker's paper was written about 20 years ago, in our opinion, it is still relevant. Any conscientious math instructor will try to introduce to their students the notion of mathematical proof through examples and assignment and exam problems at the right level (this is the basic idea of Tucker's Proof II) and as much as possible. The challenges that this instructor will face include the limited capacity of their audience to engage in proving statements that require more than a simple application of the already established facts. This particular challenge should be patiently and persistently addressed throughout the learning process and with all available resources.



$1 + 3 + \ldots + (2n - 1) = n^2$

Fig. 2 A link to dynamic model of the proof of Theorem 1. The url is http://people.math.sfu.ca/~vjungic/CLF/sum-odd.html

In our view, dynamic visual models and technology in general may play a key role in helping the learner to grasp the essential idea of the mathematical proof.
We conclude this section with two observations.

A calculus instructor may find themselves in a seemingly contradictory situation. On one hand, in demonstrating proofs in the classroom there is no expectation to go beyond straightforward applications of already established facts. On the other hand, the instructor has the responsibility to introduce mathematical ideas, mathematical language and notation, and mathematical rigour as the essence of the mathematical proof at a level that will create the opportunity for all students to explore their mathematical talents and develop their mathematical skills to the best of their abilities. Simply, in the current university math learning environment, if calculus instructors do not do this, who will?

As we have demonstrated, the time and space for mathematical proofs in a calculus classroom are quite limited. On the other hand, as Jonathan Borwein's quote from above implicates, teaching and learning of mathematical proofs are hard and time-consuming. It is possible that a dynamic visual model or even a brief sketch on the board or screen will be the only "proofs" of the particular fact presented to the students. This may completely change the purpose of a dynamic visual model in the calculus classroom: instead of being a surrogate it may become *the proof.*

3 The Sum of the First *n* Consecutive Odd Positive Integers

Theorem 1 *For any positive integer n*

$$\sum_{i=1}^{n}(2i - 1) = n^2.$$

This fact was known to Pythagoras, c. 570–500 BCE, [15]. The first inductive proof of Theorem 1, has been attributed to Francesco Maurolico, 1494–1575, [17]. Here we offer a version of the proof attributed to Pythagoras. This version has been the base for a well-known *proof by picture* of Theorem 1. See, for example, [1, 22].

Click on the image on Fig. 2 to see an animation of the proof when $n = 6$.

Proof Let n be a positive integer and let $A = [0, n] \times [0, n]$ be the region in the coordinate plane. Clearly, A is a square and its area is $\mu(A) = n^2$. For all positive integers $i \in [1, n]$ we define a region A_i in the following way:

$$A_i = ([i-1, i] \times [0, i-1]) \cup ([0, i] \times [i-1, i]).$$

It follows that, for $i \neq j$, the interiors of A_i and A_j do not intersect and that $A = \cup_{i=1}^{n} A_i$. The region A_1 is a square and its area is $\mu(A_1) = 1$. For $i \geq 2$, the region A_i is the union of two rectangles with disjunct interiors, $[i-1, i] \times [0, i-1]$ and $[0, i] \times [i-1, i]$, which implies that the area of A_i is given by

$$\mu(A_i) = \mu([i-1, i] \times [0, i-1]) + \mu([0, i] \times [i-1, i]) = (i-1) + i = 2i - 1.$$

Therefore

$$n^2 = \mu(A) = \mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i) = \sum_{i=1}^{n} (2i-1).$$

4 The Sum of the First *n* Consecutive Positive Integers

Theorem 2 *For any positive integer n*

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

In the proof of Theorem 2 we follow the idea that has been attributed to Pythagoras, c. 570–500 BCE, [3, 15].

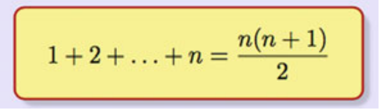
Click on the image on Fig. 3 to see an animation of the proof when $n = 6$.

Proof Let n be a positive integer and let $A = [0, n] \times [0, n+1]$ be the region in the coordinate plane. Clearly, A is a rectangle and its area is $\mu(A) = n(n+1)$. For all positive integers $i \in [1, n]$ we define regions A_i and A'_i in the following way:

$$A_i = [i-1, i] \times [0, i] \text{ and } A'_i = [i-1, i] \times [i, n+1].$$

We note that the interiors of A_i and A'_i are disjunct and that

$$A_i \cup A'_i = [i-1, i] \times [0, n+1].$$



$$1 + 2 + \ldots + n = \frac{n(n+1)}{2}$$

Fig. 3 A link to dynamic model of the proof of Theorem 2. The url is http://people.math.sfu.ca/~vjungic/CLF/sum-int.html

Hence, for $i \neq j$, the interiors of $A_i \cup A'_i$ and $A_j \cup A'_j$ do not intersect and $A = \cup_{i=1}^{n}(A_i \cup A'_i)$.

Also, since the area of the rectangle A_i is given by $\mu(A_i) = i$ and the area of the rectangle A'_i is $\mu(A'_i) = n + 1 - i$, we conclude that for any integer $i \in [1, n]$, $\mu(A'_i) = \mu(A_{n+1-i})$. Consequently

$$\sum_{i=1}^{n} \mu(A'_i) = \sum_{i=1}^{n} \mu(A_i) = \sum_{i=1}^{n} i.$$

Finally,

$$n(n+1) = \mu(A) = \mu\left(\bigcup_{i=1}^{n}(A_i \cup A'_i)\right) = \sum_{i=1}^{n} \mu(A_i \cup A'_i) = \sum_{i=1}^{n} \mu(A_i) + \sum_{i=1}^{n} \mu(A'_i) = 2\sum_{i=1}^{n} i.$$

5 The Sum of Squares

Theorem 3 *For any positive integer n*

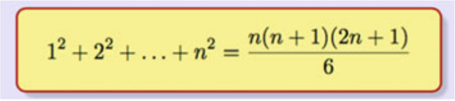
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(3n+1)}{6}.$$

The claim of Theorem 3 was known to Archimedes, c. 287–212 BCE, [16]. In the proof below we follow the idea that Nelsen [22] attributed to Martin Gardner and Dan Kalman. In [9] this method is called the *Greek rectangle method*.

Click on the image on Fig. 4 to see an animation of the proof when $n = 6$.

Proof Let n be a positive integer and let $A = [0, \frac{n(n+1)}{2}] \times [0, 2n + 1]$ be the region in the coordinate plane. Clearly, A is a rectangle and its area is $\mu(A) = \frac{n(n+1)(2n+1)}{2}$. For all positive integers $i \in [1, n]$ we define regions A_i, A'_i and A''_i in the following way:

$$A_i = \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] \times [0, i],\ A'_i = \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] \times [i, 2n - i + 1]$$



$$1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

Fig. 4 A link to dynamic model of the proof of Theorem 3. The url is http://people.math.sfu.ca/~vjungic/CLF/sum-square-int.html

and

$$A_i'' = \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] \times [2n-i+1, 2n+1].$$

We note that the interiors of A_i, A_i', and A_i'' are mutually disjunct and that

$$A_i \cup A_i' \cup A_i'' = \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] \times [0, 2n+1].$$

It follows that, for $i \neq j$, the interiors of $A_i \cup A_i' \cup A_i''$ and $A_j \cup A_j' \cup A_j''$ do not intersect.

From

$$\frac{i(i+1)}{2} - \frac{i(i-1)}{2} = i,$$

by Theorem 2, it follows that

$$\bigcup_{i=1}^{n} \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] = \left[0, \frac{n(n+1)}{2}\right]$$

and consequently

$$A = \bigcup_{i=1}^{n} \left(\left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] \times [0, 2n+1]\right) = \bigcup_{i=1}^{n} (A_i \cup A_i' \cup A_i'').$$

Hence the area of the rectangle A is

$$\begin{aligned} \frac{n(n+1)(2n+1)}{2} &= \mu(A) = \mu\left(\bigcup_{i=1}^{n} (A_i \cup A_i' \cup A_i'')\right) \\ &= \sum_{i=1}^{n} \mu(A_i \cup A_i' \cup A_i'') = \sum_{i=1}^{n} (\mu(A_i) + \mu(A_i') + \mu(A_i'')) \\ &= \sum_{i=1}^{n} \mu(A_i) + \sum_{i=1}^{n} \mu(A_i') + \sum_{i=1}^{n} \mu(A_i''). \end{aligned}$$

We observe that each A_i and A_i'' is a square with the side of the length i, which means that $\mu(A_i) = \mu(A_i'') = i^2$. Thus

$$\sum_{i=1}^{n} \mu(A_i) = \sum_{i=1}^{n} \mu(A_i'') = \sum_{i=0}^{n} i^2.$$

Since, for $1 \leq i \leq n$, $\mu(A'_i) = i \cdot (2n - 2i + 1) = i \cdot (2(n - i + 1) - 1)$, it follows that

$$\sum_{i=1}^{n} \mu(A'_i) = \sum_{i=1}^{n} i \cdot (2(n - i + 1) - 1) = \sum_{i=1}^{n} (n - i + 1) \cdot (2i - 1).$$

Thus, each odd number between 1 and $2n - 1$ that is written in the form $2k - 1$, $1 \leq k \leq n$, appears in the sum above exactly $n - k + 1$ times. Observe that, for $1 \leq i, j, k \leq n$, the odd number $2k - 1$ appears in the sum $\sum_{j=1}^{i} (2j - 1)$ if and only if the number i takes the value of one of those $n - k + 1$ integers i between 1 and n for which $k \leq i$. Hence

$$\sum_{i=1}^{n} \mu(A'_i) = \sum_{i=1}^{n} \sum_{j=1}^{i} (2j - 1).$$

and, by Theorem 1,

$$\sum_{i=1}^{n} \mu(A'_i) = \sum_{i=1}^{n} i^2.$$

Therefore

$$\frac{n(n+1)(2n+1)}{2} = \sum_{i=1}^{n} \mu(A_i) + \sum_{i=1}^{n} \mu(A'_i) + \sum_{i=1}^{n} \mu(A''_i) = 3 \cdot \sum_{i-1}^{n} i^2.$$

6 The Sum of Cubes

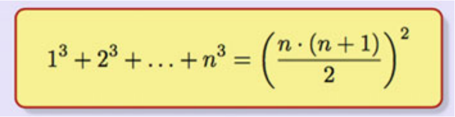
Theorem 4 *For any positive integer n*

$$\sum_{i=1}^{n} i^3 = \left(\frac{n(n+1)}{2} \right)^2.$$

Theorem 3 is attributed to Nicomachus, 60–120 CE [15]. Our dynamic visual model and the formal proof below follow the proof by Abu Bakr al-Karaji, 953–1029 CE [3, 15]. On the origin of al-Karaji's proof, Heath writes:

> Two alternatives are possible. Al-Karaji may have devised the proof himself in the Greek manner, following the hint supplied by Nicomachus's theorem. Or he may found the whole proof set out in some Greek treatise now lost and reproduced it. Whichever alternative is the true one, we can hardly doubt the Greek origin of the summation of the series of cubes [15, p. 110].

Click on the image on Fig. 5 to see an animation of the proof when $n = 4$.



$$1^3 + 2^3 + \ldots + n^3 = \left(\frac{n \cdot (n+1)}{2}\right)^2$$

Fig. 5 A link to dynamic model of the proof of Theorem 4. The url is http://people.math.sfu.ca/~vjungic/CLF/sum-cube-int.html

Proof Let n be a positive integer and let A be the region in the coordinate plane given by $A = [0, \frac{n(n+1)}{2}] \times [0, \frac{n(n+1)}{2}]$. Clearly, A is a square and its area is $\mu(A) = \left(\frac{n(n+1)}{2}\right)^2$. For all positive integers $i \in [1, n]$ we define regions A_i in the following way:

$$A_i = \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right] \times \left[0, \frac{i(i+1)}{2}\right] \bigcup \left[0, \frac{i(i-1)}{2}\right] \times \left[\frac{i(i-1)}{2}, \frac{i(i+1)}{2}\right].$$

It follows that the area of each A_i is given by

$$\begin{aligned}\mu(A_i) &= \left(\frac{i(i+1)}{2} - \frac{i(i-1)}{2}\right) \cdot \frac{i(i+1)}{2} + \frac{i(i-1)}{2} \cdot \left(\frac{i(i+1)}{2} - \frac{i(i-1)}{2}\right)\\ &= \left(\frac{i(i+1)}{2} - \frac{i(i-1)}{2}\right) \cdot \left(\frac{i(i+1)}{2} + \frac{i(i-1)}{2}\right)\\ &= i \cdot i^2 = i^3.\end{aligned}$$

We note that the interiors of A_i's are mutually disjunct. Therefore $A = \bigcup_{i=1}^n A_i$ and consequently

$$\left(\frac{n(n+1)}{2}\right)^2 = \mu(A) = \sum_{i=1}^{n} \mu(A_i) = \sum_{i=1}^{n} i^3.$$

7 Conclusion

In our view, the role of dynamic visual models in the math classrooms may be described by (at least) four of the "Eight Roles for Computation" stated by David H. Bailey and Jonathan Borwein in [6]:

> Gaining insight and intuition or just knowledge.
> Discovering new facts, patterns, and relationships.
> Graphing to expose mathematical facts, structures or principles.
> Suggesting [to the viewer] approaches for formal proof.

It should be recognized that dynamic visual models are only one of the enhancements that modern technology brings to the process of teaching and learning mathematics.

In a larger picture, dynamic visual models used in communicating ideas as part of the learning process fit in what Jonathan Borwein defined as a *visual theorem*:

> By a visual theorem I mean a picture or animation which gives one confidence that a desired result is true in Gianqunto's sense that it represents coming to believe it in an independent, reliable, and rational way [5].

It should be mentioned that the dynamic visual models presented in this note are based on mathematical facts that *originated* as visual theorems, in the sense of the above definition, in the time of ancient Greek mathematicians [15]. In addition, the names of the numbers that appear in this note are based on terms that one can easily visualize: the sum of the first n consecutive odd integers is a *perfect square*; the sum of the first n consecutive integers is a *triangular number*; the sum of the first n consecutive perfect squares is a *square pyramidal number*; and the sum of the first n consecutive cubes is a *squared triangular number*.

A generalization in another direction puts dynamic visual models in the category of *tools in mathematics*. For an extensive discussion about what the tools in mathematics are seen [21, pp. 5–8]. This generalization brings us back to the starting point and to the reasons why we created dynamic visual models presented in this note. To use Jonathan Borwein's words:

> Tools can help democratize appreciation of and ability in mathematics. [5]

References

1. Alsina, C., Neslon, R.B.: An invitation to proofs without words. Eur. J. Pure Appl. Math. **3**(1), 118–127 (2010)
2. Bailey, D.H., Borwein, J.M., Kapoor, V., Weisstein, E.: Ten problems in experimental mathematics. Am. Math. Mon. **113**, 409–481 (2006)
3. Beery, J.: Sums of powers of positive integers. In: Loci (2009). https://doi.org/10.4169/loci003284
4. Borwein, J.M.: The experimental mathematician: the pleasure of discovery and the role of proof. Int. J. Comput. Math. Learn. **10**(2), 75–108 (2005)
5. Borwein, J.M.: The life of modern homo habilis mathematicus: experimental computation and visual theorems. In: Monaghan, J., Trouche, L., Borwein, J. M.: Tools and Mathematics: Instruments for Learning. Mathematics Education Library, vol. 110. Springer, Berlin (2016)
6. Borwein, J. M., Bailey, D.H.: Mathematics by Experiment: Plausible Reasoning in the 21st Century, 2nd edn. A. K. Peters, Natick (2008)
7. Borwein, J., Jungić, V.: Organic mathematics: then and now. Not. Am. Math. Soc. **59**(3), 416–419 (2012)
8. Bostock, M.: Visualizing algorithms, posted on https://bost.ocks.org/mike/algorithms/. Accessed 9 Nov 2017
9. Cox, J.A.: Proofs of the sum of squares formula, posted on http://www.fredonia.edu/faculty/math/JonathanCox/math/SumOfSquares/SumOfSquares.html. Accessed 9 Nov 2017
10. Descartes, R.: A Discourse on Method. The Project Gutenberg EBook #59, posted on http://www.gutenberg.org/files/59/59-h/59-h.htm. Accessed 12 Nov 2017
11. Graham, R., Rothschild, B., Spencer, J.: Ramsey Theory. John Wiley and Sons, New York (1990)
12. Gravina, M. A.: Dynamical visual proof: what does it mean?. In: Santos, M., Shimizu, Y. (eds.) Proceedings of the 11th International Congress on Mathematical Education, Monterrey, Mexico (2008)

13. Guy, R.: The strong law of small numbers. Am. Math. Mon. **95**(8), 697–712 (1988)
14. Hanna, G.: A critical examination of three factors in decline of proof. Interchange **31**(1), 21–33 (2000)
15. Heath, T.L.: A History of Greek Mathematics, vol. 1. The Clarendon Press, Oxford (1921)
16. Heath, T.L.: A History of Greek Mathematics, vol. 2. The Clarendon Press, Oxford (1921)
17. O'Connor, J. J., Robertson, E. F.: Francesco Maurolico, posted on http://www-history.mcs.st-and.ac.uk/Biographies/Maurolico.html. Accessed 12 Nov 2017
18. Edwards, C., Penney, J.: Calculus, Early Transcendentals. 7th edn. Pearson, London (2012)
19. Knott, R.: Fibonacci puzzles, posted on http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/fibpuzzles2.html. Accessed on 12 Nov 2017
20. Math 100 — Learning objectives, posted on https://www.math.ubc.ca/~yhkim/yhkim-home/teaching/Math100-2017/pdfs/maths100_180_objectives.pdf. Accessed 22 Feb 2018
21. Monaghan, J., Trouche, L., Borwein, J. M.: Tools and Mathematics. Mathematics Education Library, vol. 110. Springer, Cham (2016)
22. Nelsen, R.: Proofs Without Words: Exercises in Visual Thinking. Mathematical Association of America, Washington (1993)
23. PISA 2015 — Canada. http://www.compareyourcountry.org/pisa/country/can?lg=en. Accessed 22 Feb 2018
24. Stewart, J.: Calculus, Early Transcendentals. 7th edn. Brook/Cole (2012)
25. Sum of n Integers, at GeoGebra. https://geogebra.org/m/ufxG9eHn. Accessed 22 Feb 2018
26. Sum of Odd Numbers, Wolfram Demonstration Project. http://demonstrations.wolfram.com/SumOfOddNumbers/. Accessed 22 Feb 2018
27. Sum of Squares, on GeoGebra. https://geogebra.org/m/JBnrZdn7. Accessed 22 Feb 2018
28. Sum of the Cubes of the First n Natural Numbers, on GeoGebra. https://geogebra.org/m/Z8tq2Usw. Accessed 22 Feb 2018
29. Tucker, T.W.: On the role of proof in calculus courses. In: Gavosto, Krantz, McCallum (eds.) Contemporary Issues in Mathematics Education, pp. 31–35. Cambridge University Press, Cambridge (1999)

A Random Walk Through Experimental Mathematics

Eunice Y. S. Chan and Robert M. Corless

1 Introduction

It has been known at least since the nineties that active learning is the most effective method for teaching science [13]. It is also clear to many people that the corpus of mathematical knowledge has changed dramatically since the introduction of computers. It is less widely acknowledged that this ought to change not only *how* we teach mathematics, but *what* mathematics we should teach in the first place.

Of course, there *have* been vocal advocates of just this, including for instance Gilbert Strang [21] and, very notably, Jon Borwein and coworkers. One of the present authors has held this view for decades and put it into practice on several occasions, see e.g. [6, 9, 10].

A major problem, pointed out by Hamming in his iconoclastic Calculus textbook [12], is that making only minor changes to our practice can only make things worse, if we're at a local optimum. To really do better, we have to make very large changes, basically all at once.

The difficulties with *that* are probably obvious, but we'll list some of them anyway: first, resources (it takes time and effort to design from scratch, not to mention to learn enough to use best pedagogical practices); second, inertia (the standard calculus, linear algebra, higher math sequence has enormous installed infrastructure including expectations); and finally outright hostility to change, amounting to denial.

In this paper, we'll talk about how we took advantage of an unusual confluence of opportunities to make a serious attempt, serious enough that we feel that this description will be useful for the next attempt.

E. Y. S. Chan (✉) · R. M. Corless
Ontario Research Centre for Computer Algebra, School of Mathematics and Statistical Sciences,
The Rotman Institute of Philosophy, Western University, London N6A 5B7, Canada
e-mail: echan295@uwo.ca

R. M. Corless
e-mail: rcorless@uwo.ca

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_14

1.1 *The Confluence of Opportunities*

In May 2014, David Jeffrey (then Chair of Applied Mathematics at Western) was lamenting the invisibility of our program to students in their first year of study. A decade previously, you see, the first-year calculus and linear algebra offerings of the separate Mathematics Department and of ours had been merged, so students only saw "Calculus" or "Linear Algebra." The Applied Math undergraduate program slowly declined, afterward. In response to the Chair's lament, RMC simply wrote "Introduction to Experimental Mathematics"—just that—on the whiteboard. The Chair and subsequently surveyed students were very enthusiastic, and so, with a verbal promise of a two-year trial both from the Chair and from the Dean of Science, RMC designed the course over June, July, and August, making (with undergraduate student help) the recruiting video https://vimeo.com/itrc/review/99140780/68995faea3 and taught the newly-christened AM 1999 September–December, of that same year, 2014.

With such short notice, enrolment was going to be an issue—Western does not give teaching credit for undergraduate courses with 9 or fewer students, or graduate courses with 4 or fewer students. Therefore, RMC invented and designed the sister course AM 9619 intended for graduate students and senior undergraduate students (such as EYSC, then—she's now a PhD student of RMC's at this time of writing). Again a verbal promise of a two-year trial was given by the Chair and Dean. These were to be taught simultaneously, and credit given for one course taught.

> "[on seeing the video] This is a wonderful course! It should be profiled!"—Charmaine Dean, Dean of Science

The major resources available for this were first RMC's time—he used a (northern) summer, normally used for research, to design the course, make the video, socialize it with students, design the grad course, attend the ICERM conference on Experimental Mathematics organized by Jon Borwein and David H. Bailey to seek advice there—Neil Calkin was especially helpful—explain the course to academic counselors, make a poster for Fall Preview Day, etc. Another task was to formally write up the course design and get it approved by Senate. Later, the Course Learning Outcomes had to be drafted for the undergraduate program description and harmonized with the Program Learning Outcomes. Second, RMC had some discretionary money he had available to pay the first TA, Steven Thornton, and an assistant, Torin Viger, to construct course materials (and to travel to the ICERM conference). The third resource was the \$400,000 Western Active Learning Space, which had just been built (RMC gave the very first class in it). We'll talk more about this space and how we used it later. Details of the space can be found at http://www.uwo.ca/wals/. The fourth resource was RMC's record in Experimental Mathematics and in the use of technology for teaching mathematics, which helped sell the course (both to the students and to the higher administration).

The members of the Undergraduate Society of Applied Mathematics (USAM) constituted a crucial fifth resource. They were consulted frequently in the "summer of design" and many of them helped to make the recruitment video linked above. Their enthusiastic support for the course was very heartening.

1.2 The Western Active Learning Space

The advanced features of WALS that we utilized included, first, the "Air Media" by which students or the TA could share their computer screens with their 7-member pods or with the whole class. This feature is not so expensive and one could imagine classes less richly furnished that still had it. WALS had multiple screens (one per pod) that could be used independently, and this was more expensive (but still useful). We also used the document camera frequently. However, we did not utilize the classroom's sound system or SMART Board pens.

We did use the low-tech whiteboards—there were three sizes: main boards, portable boards in A-frames and small "huddle boards" that could be placed at will on the main boards.

WALS came with eight recent vintage laptops each equipped with Maple. We used these each class; they were intended to be one per pod, plus the instructor's station, but we didn't always use them that way. The senior students had Maple on their personal laptops but the undergraduates did not. Therefore, priority of the classroom laptops was given to the undergraduate students.

1.3 Active Learning Techniques

Active learning techniques run from the obvious (get students to choose their own examples, and share) through the eccentric (interrupt students while programming similar but different programs and have them trade computers and problems) to the flaky (get them to do an interpretive dance or improvisational skit about their question). We tried to avoid the extremely flaky, but we did mention them, so that these introverted science students knew that this was within the realm of possibility.

The simplest activity was typing Maple programs that were handwritten on a whiteboard into a computer: this was simple but helpful because students learned the importance of precision, and had *immediate* help from their fellow students and from the TA.

Next in complexity was interactive programming exercises (integrated into the problems). Mathematicians tend to under-value the difficulty of learning syntax and semantics simultaneously.

We describe our one foray into eccentricity. The paper Strange Series and High Precision Fraud by Borwein and Borwein [2] has six similar sums. We had six teams program each sum, at a stage in their learning where this was difficult (five weeks into the course). After letting the teams work for twenty minutes, we forced one member of each team to join a new team; each team had to explain their program (none were working at this stage) to the new member. This exercise was most instructive. The lessons learned included:

- people approach similar problems very differently
- explaining what you are doing is as hard as doing it (maybe harder)
- basic software engineering (good variable names, clear structure, economy of thought) is important
- designing on paper first might be a good idea (nobody believed this, really, even after)
- social skills matter (including listening skills).

Perhaps the most important "active learning" technique used, and the hardest to describe accurately, was a shift in perspective: the instructor attempted to listen carefully to students' wishes, and alter the activity, discussion, or topic depending on what their questions were. Waiting long enough for students to actually ask questions was often uncomfortable for everyone. The students were, at least initially, reluctant to use their freedom.

1.4 Active Learning in Mathematics

It is widely accepted that "true" or "deep" learning only happens when students are actively engaged. The old saying goes, "I hear, and I forget. I see, and I remember. I do, and I understand." There is a lot of argument as to how best to make the students active, engaged, and make them do the work. There is a lot of discussion of the value of "discovery" versus the apparent efficiency of simple delivery of a lecture, or reading. We won't settle those arguments here because in our course, a mix was employed: some actual lectures (most often short, usually explaining how to program something in Maple), some individual activities, some partner work, some games, some peer assessment, a structured project, and giving them choices. RMC found it very hard (sometimes) to let the students make mistakes. There was an improvisational element: every class had a plan, and a goal (a "planned learning outcome") even if only a modest one, but at any moment tangential discussions could be seized as opportunities.

Another precious resource for this course was that it was not a prerequisite for anything—it had no required specific topics to cover—and thus the learning outcomes could be quite general, such as achieving the precision of thought necessary to program computers. Therefore, the class could afford to sail off on tangents, chosen by members of the class.

The students found this freedom ("lack of structure") frightening at first, but exhilarating at the same time. They learned to trust the instructor, and the instructor learned to trust them.

This is not to say that there was *no* structure: there certainly was. The main theme of the course was that of discrete dynamical systems, normally taught long after calculus and analysis. But one of the major achievements of computers for mathematics research (and education) is in making deep questions immediately accessible to students, questions such as "which initial guesses for Newton's method converge to which roots?" This question immediately gives access to fractals and their compelling images.

It is less appreciated that computers, via symbolic computation programs such as Maple or Mathematica, give students the same (or greater!) computational power to perform mathematical experiments that the giants of mathematical history had; Lagrange, Stirling, or even Euler, Gauss, and Newton [3]. The students can explore classical topics such as continued fractions experimentally and make deep discoveries themselves.

We ended the course with the Chaos Game Representation of DNA sequences. This has recently been used to generate an objective "Map of Life" [17]. We point out that at the time the course was first taught, this was new (not yet published) frontier research. The students understood it. We got them from entering university to frontline research results. The rest of this paper gives some examples of what we did, to do so.

One final introductory point: what did the students say they needed from our program, that we weren't delivering? The number one request (in our survey results, prior to designing the course) was "more programming." The number two consideration was "more say in the syllabus," that is, student-centerd curriculum design. We aimed to address both needs with this course.

2 Choices of Topics

As previously stated, the theme of the course was "discrete dynamical systems." This was chosen because it benefits massively from computer support, it's outside the normal curriculum (relieving pressure to teach to a specific goal), and it's accessible, beautiful, interesting, has applications, and because it has deep connections to classical topics in number theory (itself outside our standard curriculum). Before starting to teach, RMC drew up a list of potential vignettes. The top of the list was the classical theory of simple continued fractions. This topic is extremely accessible (we will give the first vignette in Section 2.1 below) and at the same time interesting, deep, and demanding of introductory programming skills (iteration, conditionals, and induction for proof of correctness). The next was the solution of nonlinear equations by Newton's method and its generalizations. Other topics included the Thue–Morse sequence, the Online Encyclopedia of integer sequences, computation of π, the game of life, the numerical solution of differential equations by Euler's method, visualization of complex functions by phase plots, and chaos game representation of DNA sequences. We sample from these below.

The graduate course AM 9619 had an extra lecture per week (the senior students attended the AM 1999 activities together with the first-year students, but they got some advanced material every week that the AM 1999 students did not). We will give one sample below.

2.1 Continued Fractions and Rational Approximations of $\sqrt{2}$

How does one give control over the pedagogy to the students? RMC could only think to do it gradually, by doing an example and then asking students to *choose* their own examples which we would (first) do together and then (for other choices) they would do themselves. The students were presented with the sequence

$$1\,,\ \frac{3}{2}\,,\ \frac{17}{12}\,,\ \frac{577}{408}\,,\ \frac{665857}{470832}\,,\ \cdots\,, \tag{1}$$

which at that point was plucked from thin air. Each term x_n is generated from its predecessor[1] by the rule $x_n = {}^{(x_{n-1}+{}^{2}\!/_{x_{n-1}})}\!/_{2}$. What means the same thing, if we label the numerators and denominators by $x_n = {}^{p_n}\!/_{q_n}$, in other words

$$\begin{aligned} p_n &= p_{n-1}^2 + 2q_{n-1}^2 \\ q_n &= 2p_{n-1}q_{n-1}\,. \end{aligned} \tag{2}$$

> "At first glance, nothing seems simpler or less significant than writing a number, say ${}^{9}\!/_{7}$, in the form
>
> $$\frac{9}{7} = 1+\frac{2}{7} = 1+\frac{1}{{}^{7}\!/_{2}} = 1+\frac{1}{3+{}^{1}\!/_{2}} = 1+\cfrac{1}{3+\cfrac{1}{1+{}^{1}\!/_{1}}}\,. \tag{3}$$
>
> It turns out, however, that fractions of this form, called *continued fractions*, provide must insight..."—from p. 3 of C. D. Olds, "Continued Fractions" [19].

Carl Douglas Olds won the 1973 Chauvenet Prize, the highest award for mathematical exposition, for his paper "The Simple Continued Fraction for *e*." The book cited above is likewise a model of lucidity and reads very well today. A new book [4] is similarly valuable to students.

What follows is a simulation of a whiteboard discussion.

What's happening [in Olds' example]? Let's do the same thing with each x_n. First, we take out the integer part. For our first two numbers, nothing much happens

$$x_0 = 1 \tag{4}$$

$$x_1 = \frac{3}{2} = 1+\frac{1}{2} = 1+\frac{1}{1+{}^{1}\!/_{1}}\,, \tag{5}$$

but this last isn't much use.

The next number is more interesting

$$x_2 = \frac{17}{12} = \frac{12+5}{12} = 1+\frac{5}{12}$$

[1] Here, $x_0 = 1$, $x_1 = {}^{3}\!/_{2}$, so on and so forth.

$$= 1 + \frac{1}{{}^{12}\!/_{5}} = 1 + \frac{1}{2 + {}^{2}\!/_{5}} = 1 + \cfrac{1}{2 + \cfrac{1}{{}^{5}\!/_{2}}}$$

$$= 1 + \cfrac{1}{2 + \cfrac{1}{2 + {}^{1}\!/_{2}}}. \tag{6}$$

The crucial step in this process is writing the fractional part that we get, after taking out the integer part, as a reciprocal of another fraction, i.e.,

$$\frac{5}{12} = \frac{1}{{}^{12}\!/_{5}}. \tag{7}$$

Now a longer example:

$$x_3 = \frac{577}{408} = \frac{408 + 169}{408} \qquad = 1 + \frac{169}{408} \qquad = 1 + \frac{1}{{}^{408}\!/_{169}}$$

$$= 1 + \frac{1}{2 + {}^{70}\!/_{169}} \qquad = 1 + \cfrac{1}{2 + \cfrac{1}{{}^{169}\!/_{70}}}$$

$$= 1 + \cfrac{1}{2 + \cfrac{1}{2 + {}^{29}\!/_{70}}} \qquad = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{{}^{70}\!/_{29}}}}$$

$$= 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + {}^{12}\!/_{19}}}} \qquad = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{{}^{29}\!/_{12}}}}}$$

$$= 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + {}^{5}\!/_{12}}}}} \qquad = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{{}^{12}\!/_{5}}}}}}$$

$$= 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + {}^{2}\!/_{5}}}}}} \qquad = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + {}^{1}\!/_{2}}}}}}}$$

$$= 1 + [2\,, 2\,, 2\,, 2\,, 2\,, 2\,, 2] \quad \text{for short.} \tag{8}$$

At this point, you may feel like sticking out your tongue and giving us a raspberry for such obvious cheating. Think of it like “television wrestling” and give the entertainment a chance!

When you think about it, it *is* a bit mysterious that the simple rule

$$x_n = \frac{(x_{n-1} + {}^{2}/_{x_{n-1}})}{2} \tag{9}$$

can generate the continued fractions

$$1\,,\,1+[2]\,,\,1+[2\,,2\,,2]\,,\text{ and }1+[2\,,2\,,2\,,2\,,2\,,2\,,2]\,. \tag{10}$$

The next one

$$x_4 = \frac{665857}{470832} = 1+[2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2] \tag{11}$$

has fifteen 2’s in it. (By the way; don’t worry, we’ll check that by computer later.) That’s one, three, seven, and fifteen twos. What’s next? We’ll leave that for now and go back to the first question, about $x_n^2 = {}^{p_n^2}/_{q_n^2}$.

The squares of our sequence are

$$\begin{gathered}1\,,\,\frac{9}{4} = 2\frac{1}{4}\,,\,\left(\frac{17}{12}\right)^2 = \frac{289}{144} = \frac{288+1}{144} = 2+\frac{1}{144} = 2+\frac{1}{12^2}\,,\\ \left(\frac{577}{408}\right)^2 = \frac{332929}{166464} = \frac{332928+1}{166464} = 2+\frac{1}{166464} = 2+\frac{1}{408^2}\end{gathered} \tag{12}$$

and at this point, we might be prepared to bet that

$$x_4^2 = \left(\frac{665857}{470832}\right)^2 = 2+\frac{1}{470832^2} \doteq 2+4.5\times 10^{-12}\,. \tag{13}$$

Checking using RMC’s phone (a Nexus 5), we see that this is, in fact, true. But what does it mean?

One thing it means is that our sequence can be written as

$$\sqrt{2-\frac{1}{1^2}}\,,\sqrt{2+\frac{1}{2^2}}\,,\sqrt{2+\frac{1}{12^2}}\,,\sqrt{2+\frac{1}{408^2}}\,,\sqrt{2+\frac{1}{470832^2}} \doteq \sqrt{2+4.5\times 10^{-12}} \tag{14}$$

that is a sequence of square roots of numbers that rapidly approach 2. The denominator of x_5 is

$$q_5 = 2p_4q_4 = 2\cdot 470832\cdot 665857 \doteq 6.5\times 10^{11}\,; \tag{15}$$

the next

$$\left(\frac{p_5}{q_5}\right)^2 = 2 + \frac{1}{q_5^2} \doteq 2 + 2 \times 10^{-24} \tag{16}$$

about as close to 2 as one molecule in a mole.[2]

Here ends our simulated whiteboard discussion. Some more questions present themselves. Does this continue? Is $x_5 = 1 + [2\,,2\,,\ldots,2]$ with thirty-one 2's in the continued fraction? Does x_6 have sixty-three 2's in it? Is $x_n^2 = 2 + {}^1\!/\!_{q_n^2}$ always? Does this mean that $x_n \doteq \sqrt{2}$?

Here's another question. What is

$$1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ddots}}}}, \tag{17}$$

where the 2's continue forever? Does this make sense? At this point, many students were surprised at the perfect predictability, and repeating nature, of this continued fraction, because it is indeed true that with quite natural definitions, this infinite continued fraction can only be $\sqrt{2}$.

But "everybody knows" that the decimal expansion for $\sqrt{2}$ does not repeat, because $\sqrt{2}$ is irrational! Why is this different? Is it something special about $\sqrt{2}$? (Of course a continued fraction is not a decimal expansion.)

The students were then asked to choose their own examples. Predictably, they chose only other square roots ($\sqrt{3}$, $\sqrt{5}$, etc.) until prodded.

After enough hand calculation, Maple was introduced as a calculator. Some hand calculation is necessary, at first, because the students need to *feel* that they are connected to the mathematics; they need to own it. Once they feel ownership (and once the computations get tedious) then computer assistance is welcomed.

Over the next three classes, the pace gradually increased until the Gauss map $G : x \mapsto \mathrm{frac}({}^1\!/\!_x)$ mapping $(0, 1)$ to $[0, 1)$ seemed natural; patterns were identified by the students, such as termination implies rationality, ultimate periodicity implies that the number is a quadratic irrational. This is Lagrange's classical result, and not easy to prove. No proofs are given in this section, only computation; Galois' result that purely periodic continued fractions arise from *reduced* quadratic irrationals is mentioned but not emphasized—we concentrate on the interval $[0, 1)$.

By this time the students are engaged, talking to each other, choosing different numbers. The number e causes amazement with its pattern: $e = 2 + [1, 2, 1, 1, 4, 1, 1,$

[2] Avogadro's number is $6 \cdot 10^{23}$, about.

6, 1, 1, 8, . . .]. The fact that the continued fraction for π is not known and has no apparent pattern, causes astonishment: in the first two weeks, we have arrived at a deep open problem.

By now the students are hooked. They've learned some Maple; they've learned some mathematics. They've learned about fixed points (equal points of period 1, including the golden ratio $\phi = 1 + [1, 1, 1, \ldots])$ and about periodic points. The journey has begun.

2.2 Newton's Method, $\sqrt{2}$, and Solving Nonlinear Equations

In the first vignette, we met with the sequence

$$1\,, \frac{3}{2}\,, \frac{17}{12}\,, \frac{577}{408}\,, \frac{665857}{470832}\,, \ldots \tag{18}$$

which was generated by

$$x_{n+1} = \left(x_n + \frac{2}{x_n}\right)/2$$

in words, the average of the number and two divided by the number. This vignette explores where that sequence came from, and its relationship to $\sqrt{2}$. We approached this algebraically, as Newton did. Consider the equation

$$x^2 - 2 = 0\,. \tag{19}$$

Clearly the solutions to this equation are $x = \sqrt{2}$ and $x = -\sqrt{2}$. Let us *shift the origin* by putting $x = 1 + s$; so $s = 0$ corresponds to $x = 1$. Then

$$(1 + s)^2 - 2 = 1 + 2s + s^2 - 2 = -1 + 2s + s^2 = 0\,. \tag{20}$$

We now make the surprising assumption that s is so small that we may ignore s^2 in comparison to $2s$. If it turned out that $s = 10^{-6}$, then $s^2 = 10^{-12}$, very much smaller than $2s = 2 \cdot 10^{-6}$; so there are small numbers s for which this is true; but we don't know that this is true, here. We just hope.

Then if s^2 can be ignored, our equation becomes

$$-1 + 2s = 0 \tag{21}$$

or $s = {}^1\!/\!_2$. This means $x = 1 + s = 1 + {}^1\!/\!_2 = {}^3\!/\!_2$.

We now repeat the process: shift the origin to ${}^3\!/\!_2$, not 1: put now

$$x = {}^3\!/\!_2 + s \tag{22}$$

so

$$({}^3\!/\!_2 + s)^2 = {}^9\!/\!_4 + 3s + s^2 - 2 = 0\,. \tag{23}$$

This gives $3s + s^2 + {}^1\!/\!_4 = 0$ and again we ignore s^2 and hope it's smaller than $3s$. This gives

$$3s + {}^1\!/\!_4 = 0 \tag{24}$$

or $s = -{}^1\!/\!_{12}$. This means $x = {}^3\!/\!_2 - {}^1\!/\!_{12}$ or $x = {}^{17}\!/\!_{12}$. Now we see the process. Again, shift the origin: $x = {}^{17}\!/\!_{12} + s$. Now

$$\left(\frac{17}{12} + s\right)^2 = \frac{289}{144} + \frac{17}{6}s + s^2 - 2 = 0\,. \tag{25}$$

Ignoring s^2,

$$\frac{17}{6}s + \frac{1}{144} = 0 \tag{26}$$

or

$$s = \frac{-6}{17 \cdot 144} = \frac{-1}{17 \cdot 24} = \frac{-1}{408}\,. \tag{27}$$

Thus,

$$x = \frac{17}{12} - \frac{1}{408} = \frac{577}{408}\,. \tag{28}$$

As we saw in the previous vignette, there are the exact square roots of numbers ever more close to 2. For instance,

$$\frac{577}{408} = \sqrt{2 + \frac{1}{408^2}}\,. \tag{29}$$

It was Euler who took Newton's "shift the origin" strategy and made a general method—that we call Newton's method—out of it. In modern notation, Euler considered solving $f(x) = 0$ for differentiable function $f(x)$, and used the tangent line approximation near an initial guess x_0: if $x = x_0 + s$ then, using $f'(x_0)$ to denote the slope at x_0, $0 = f(x) = f(x_0 + s) \doteq f(x_0) + f'(x_0)s$ ignoring terms of order s^2 or higher. Then

$$s = -\frac{f(x_0)}{f'(x_0)} \tag{30}$$

so

$$x \doteq x_0 + s = x - \frac{f(x_0)}{f'(x_0)}\,. \tag{31}$$

The fundamental idea of Newton's method is that, if it worked once, we can do it again: pass the parcel! Put

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{32}$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \tag{33}$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} \tag{34}$$

and keep going, until $f(x_k)$ is so small that you're happy.

Notice that each x_k solves

$$f(x) - f(x_k) = 0 \tag{35}$$

not $f(x) = 0$. But if $f(x_k)$ is really small, you've solved "almost as good" an equation, like finding $\sqrt{2 + {}^1/_{408^2}}$ instead of $\sqrt{2}$. So where did ${}^{(x_n + {}^2/_{s_n})}/_2$ come from?

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{\left(x_n^2 - 2\right)}{2x_n} \tag{36}$$

because if $f(x) = x^2 - 2$, $f'(x) = 2x - 0 = 2x$. Therefore,

$$\begin{aligned} x_{n+1} = x_n - \frac{\left(x_n^2 - 2\right)}{2x_n} &= \frac{2x_n^2 - x_n^2 + 2}{2x_n} \\ = \frac{x_n^2 + 2}{2x_n} \qquad &= \frac{1}{2}\left(x_n + \frac{2}{x_n}\right) \end{aligned} \tag{37}$$

as claimed.[3]

Executing this process in decimals, using a calculator (our handy HP48G+ again), with $x_0 = 1$, we get

$$\begin{aligned} x_0 &= 1 \\ x_1 &= \underline{1}.5 \\ x_2 &= \underline{1.4}1666\ldots \\ x_3 &= \underline{1.41421}568628 \\ x_4 &= \underline{1.41421356238} \\ x_5 &= x_4 \ \text{ to all 11 places in the calculator.} \end{aligned} \tag{38}$$

Now $\sqrt{2} = 1.41421356237$ on this calculator. We see (approximately) 1, 2, 5 then 10 correct digits. The convergence behavior is clearer in the continued fraction representation:

[3] For more complicated functions one *shouldn't* simplify for numerical stability reasons. But for $x^2 - a$, it's okay.

$$1\,,\,1+[2]\,\,,1+[2\,,2\,,2]\,\,,1+[2\,,2\,,2\,,2\,,2\,,2\,,2]\,\,,\\ 1+[2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2\,,2] \tag{39}$$

with 0, 1, 3, 7, 15 twos in the fraction part: each time doubling the previous plus 1, giving $2^0-1, 2^1-1, 2^2-1, 2^3-1, 2^4-1$ correct entries. This "almost doubling the number of correct digits with each iteration" is quite characteristic of Newton's method. This clearly demonstrates quadratic convergence.

We then look at the secant method and Halley's method, which in continued fractions notation generate a Fibonacci number of convergents (by the secant method) and $\mathcal{O}(3^n)$ convergents showing cubic convergents (by Halley's method). This is not part of any standard numerical analysis curriculum, by the way, that we know of.

We have a fine opportunity for a digression on Fibonacci numbers; another for a digression on Schröder iteration [20].

Remark 1 We avoid discussion of "convergence" and note that each iterate solves $f(x) - f(x_n) = 0$; we *interpret* residuals in context.

2.3 Backward Error

The reader will have noticed that we used $^{17}/_{12}$ as the exact square root of $2\ ^{1}/_{144}$; similarly $^{577}/_{408}$ is the exact square root of $2\ ^{1}/_{408^2}$. The is an example of what is called "Backward Error Analysis." Instead of an approximate answer to the reference problem, we have an exact answer to a nearby problem.

This is a very powerful old idea, used by von Neumann, by Turing, and others. It was popularized in numerical analysis by Wilkinson. The book [7] uses it in every chapter.

It is a useful idea pedagogically, as well due to its ease of understanding for most students. Students do understand it. If the course you're teaching the idea in has a formative assessment, the sly question comes up "Is it all right if I give an answer to a slightly different question on the midterm?" They get the idea.

Of course some problems are sensitive to changes in their formulation. This is also not quite mathematics: mathematics gains leverage by abstracting away context, and backward error needs problem context to be directly useful, but it can be done. If forward error (e.g. $\left|x_n - \sqrt{2}\right|$) is needed, then one can introduce the notion of derivative and relative derivative (also known as logarithmic derivative or "condition number"—von Neumann and Goldstine called it a "figure of merit" [22]).

But in this first-year course, a direct interpretation allows one to avoid infinity and to avoid the calculus. If $x = {}^{p_5}/_{q_5}$, then the difference between x^2 and 2 is less than $2 \cdot 10^{-24}$ light-years or $\sim 10^{-6}$mm; that is, we've found the exact length of triangle whose hypotenuse is not 2 light-years, but 2 light-years + less than a millionth of a millimeter. Students get this.

The next section, on modified equations, applies this idea in an advanced context for graduate students.

2.4 Modified Equations

As an example lecture for the senior students, consider the construction of a so-called *modified equation*. This was done in lecture format. In an attempt to understand a very simple numerical method (fixed time-step Euler's method) on a very simple but nonlinear initial value problem, namely, $\dot{y} = y^2$, $y(0) = y_0$, one might attempt to use a technique called "the method of modified equations". This leads to the series

$$1 - v + \frac{3}{2}v^2 - \frac{8}{3}v^3 + \cdots - \frac{9427}{210}v^7 + \cdots . \tag{40}$$

Computation of the series coefficients is somewhat involved but automatable. Consulting the Online Encyclopedia of Integer Sequences leads us to the work of Labelle [18], in combinatorics; this connection is powerful and unexpected.

The idea of this lecture is that we begin with "brute force"; then consult the OEIS or some other resource to try to identify our results and find faster/better ways, and to make connections to other works. Then we extract other useful materials from the results, proving what we can. We give some more details of the lecture, below.

The method of modified equations is a neat idea: if we solve a DE, say $\dot{y} = f(y)$, by a numerical method, say $y_{n+1} = y_n + h\dot{y}_n$ (Euler's method), then in order to *explain* the numerics we look for a modified equation

$$\dot{Y} = F(Y; h) \tag{41}$$

which (more nearly) has $Y(t_n) \doteq y_n$. References include [7, pp. 662–671], and [11] as well as [5, 8]. It's best explained by example. Consider $f(y) = y^2$. Then $\dot{y} = y^2$, $y(t_0) = y_0$ is actually easy to solve: $\frac{\dot{y}}{y^2} = 1$ so $\int_{t_0}^{t} \frac{\dot{y}\mathrm{d}t}{y} = \int_{t_0}^{t} \mathrm{d}t = t - t_0$ or $-\frac{1}{y}|_{t_0}^{t} = \frac{1}{y(t_0)} - \frac{1}{y(t)} = t - t_0$ or

$$y(t) = \frac{y_0}{1 - y_0 \cdot (t - t_0)} . \tag{42}$$

This is singular at $t = t_0 + \frac{1}{y_0}$.

Let's pretend we don't know this, and instead solve the problem using Euler's method with a small fixed time-step n:

$$y_{n+1} = y_n + hy_n^2 \quad n = 0, 1, 2, \ldots . \tag{43}$$

The key step in the method of modified equations is to replace this discrete recurrence relation with a functional equation: $Y(t)$ interpolating the data (t_n, y_n) so $Y(t_n) = y_n$; since $t_{n+1} = t_n + h$ we have necessarily $Y(t_n + h) = Y(t_n) + hY(t_n)^2$.

We now insist that

$$Y(t+h) = Y(t) + hY^2(t) \tag{44}$$

everywhere, not just at $t = t_n$. This is a weird thing to do. Even weirder, we now look for a differential equation for $Y(t)$, to replace the (harder) functional equation. Using Taylor expansion we have

$$Y(t+h) = Y(t) + \dot{Y}(t)h + \ddot{Y}(t)\frac{h^2}{2} + \dddot{Y}(t)\frac{h^3}{6} + \cdots \tag{45}$$

so we have (in a statement with a breathtaking lack of rigor)

$$Y(t) + h\dot{Y}(t) + \frac{h^2}{2}\ddot{Y}(t) + \frac{h^3}{6}\dddot{Y}(t) + \cdots = Y(t) + hY^2(t) \tag{46}$$

or

$$\dot{Y}(t) + \frac{h}{2}\ddot{Y}(t) + \frac{h^2}{3!}\dddot{Y}(t) + \cdots = Y^2(t)\,. \tag{47}$$

This is a singular perturbation of the original. It's not very helpful. We can truncate:

$$\sum_{k=1}^{N} Y^{(k)}(t)\frac{h^{k-1}}{k!} = Y^2(t) + \mathcal{O}(h^N)\,. \tag{48}$$

We can differentiate (no breath left!)

$$\ddot{Y} + \frac{h}{2}\dddot{Y} + \frac{h^2}{3!}Y^{\mathrm{IV}} + \cdots = 2Y\dot{Y} \tag{49}$$

and again

$$\dddot{Y} + \frac{h}{2}Y^{\mathrm{IV}} + \mathcal{O}(h^2) = 2(\dot{Y})^2 + 2Y\ddot{Y} \tag{50}$$

as many times as we need to.

Noticing that we need one less order in h with each derivative we have

$$\dot{Y} + \frac{h}{2}\ddot{Y} + \frac{h^2}{6}\dddot{Y} = Y^2 + \mathcal{O}(h^3) \tag{51}$$

$$\ddot{Y} + \frac{h}{2}\dddot{Y} = 2Y\dot{Y} + \mathcal{O}(h^2) \tag{52}$$

$$\dddot{Y} = 2(\dot{Y})^2 + 2Y\ddot{Y} + \mathcal{O}(h)\,. \tag{53}$$

Since the second equation implies $\ddot{Y} = 2Y\dot{Y} + \mathcal{O}(h)$, and the first equation implies $\dot{Y} = Y^2 + \mathcal{O}(h)$, the third equation can be simplified to

$$\begin{aligned}\dddot{Y} &= 2Y^2\dot{Y} + 2Y(2Y\dot{Y}) + \mathcal{O}(h) \\ &= 6Y^2\dot{Y} + \mathcal{O}(h). \end{aligned} \tag{54}$$

Using this in the second equation gives

$$\ddot{Y} + \frac{h}{2}(6Y^2\dot{Y}) = 2Y\dot{Y} + \mathcal{O}(h^2) \tag{55}$$

and using both of these in the first equation gives

$$\dot{Y} + \frac{h}{2}(2Y\dot{Y} - 3hY^2\dot{Y}) + \frac{h^2}{6}(6Y^2\dot{Y}) = Y^2 + \mathcal{O}(h^3) \tag{56}$$

or

$$\dot{Y} + hY\dot{Y} - \frac{3}{2}h^2Y^2\dot{Y} + h^2Y^2\dot{Y} = Y^2 + \mathcal{O}(h^3) \tag{57}$$

or

$$(1 + hY - \frac{1}{2}h^2Y^2)\dot{Y} = Y^2 + \mathcal{O}(h^3) \tag{58}$$

or

$$\dot{Y} = (1 + hY - \frac{1}{2}h^2Y^2)^{-1}Y^2 + \mathcal{O}(h^3) \tag{59}$$

or

$$\dot{Y} = (1 - hY + \frac{3}{2}h^2Y^2)Y^2 + \mathcal{O}(h^3)\,. \tag{60}$$

What does this mean? It means that Euler's method applied to $\dot{y} = y^2$ gives a better solution to

$$\dot{y} = (1 - hy + \frac{3}{2}h^2y^2)y^2 \tag{61}$$

(which wasn't intended); indeed it's $\mathcal{O}(h^3)$ accurate on *that* one.

Why is this our first senior lecture on Open Problems for Experimental Maths? Because it was among RMC's first forays into the subject. He wrote a Maple program to compute more terms; then tried to use "gfun" in Maple (which failed) and the Online Encyclopedia of Integer Sequences (which worked) to identify $B(v)$ where

$$\dot{y} = B(hy)y^2 \tag{62}$$

with $B(v) = 1 - v + \frac{3}{2}v^2 + \cdots$ is the modification. From the OEIS we are led to [18] which gives (in another context)

$$B(v) = \sum_{n\geq 0} c_n v^n \tag{63}$$

with $c_0 = 1$ and $c_n = -\frac{1}{n}\sum_{i=1}^{n}\binom{n-i+2}{i+1}c_{n-i}$. This series *diverges*.

> "Therefore, we may *do* something with it"—O. Heaviside [14].

For very small v,

$$B(v) \sim 1 - v + \frac{3}{2}v^2 - \frac{8}{3}v^3 + \mathcal{O}(v^4)\,. \tag{64}$$

Note: $y_{n+1} = y_n + hy_n^2 \;\rightarrow\; hy_{n+1} = hy_n + (hy_n)^2$ so putting $v(t) = hy(t)$ gives (with $\tau = \frac{t-t_0}{h}$)

$$v(\tau + 1) = v(\tau) + v^2(\tau) \tag{65}$$

$$\frac{\mathrm{d}v}{\mathrm{d}\tau} = B(v)v^2 \tag{66}$$

$$\therefore \quad \frac{\mathrm{d}v(\tau + 1)}{\mathrm{d}\tau} = B(v(\tau + 1) \cdot v(\tau + 1)^2 \tag{67}$$

$$\text{and} \quad \frac{\mathrm{d}v(\tau + 1)}{\mathrm{d}\tau} = \frac{\mathrm{d}v}{\mathrm{d}\tau} + 2v\frac{\mathrm{d}v}{\mathrm{d}\tau} = (1 + 2v) \cdot B(v) \cdot v^2 \tag{68}$$

$$\therefore \quad B(v + v^2) \cdot (v + v^2)^2 = (1 + 2v)B(v)v^2 \tag{69}$$

$$\text{or} \quad B(v + v^2) \cdot (1 + v)^2 = (1 + 2v)B(v) \quad \text{when } v \equiv 0\,. \tag{70}$$

Therefore, we have a functional equation for B.

$$B(v) = \frac{(1 + v)^2}{(1 + 2v)} \cdot B(v + v^2) \tag{71}$$

or

$$B(v + v^2) = \frac{(1 + 2v)}{(1 + v)^2}B(v) \tag{72}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = B(v)v^2 \tag{73}$$

$$\frac{\mathrm{d}v/\mathrm{d}t}{B(v)v^2} = 1 \tag{74}$$

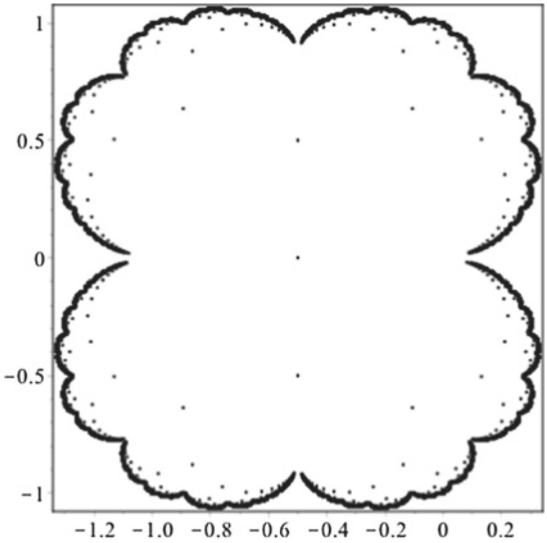
$$\int_{t_0}^{t} \frac{\mathrm{d}v/\mathrm{d}\tau}{B(v)v^2}\mathrm{d}\tau = t - t_0 \tag{75}$$

if we can evaluate $B(v)$, we can numerically integrate this to get $F(v(t)) - F(v(\tau_0)) = \tau - \tau_0$ which will allow us to plot $v(\tau)$.

What happens in the iteration $v_{n+1} = v_n + v_n^2$?

- $v_n \rightarrow \infty$ as $n \rightarrow \infty$,
- $v_n \rightarrow 0$ as $n \rightarrow \infty$,
- sometimes neither.

We can see this *experimentally*. Notice that B has a *pole* if $v = -\frac{1}{2}$ or if $v + v^2 = -\frac{1}{2}$ and so on. Therefore, the pre-images of $v_n = -\frac{1}{2}$ are *all* poles. Therefore, poles approach 0 arbitrarily closely. Therefore, the series at 0 cannot converge.



The Maple code to plot the pre-images is as follows:

```
> Digits := 15;
```

$$Digits := 15$$

```
> Wanted := 100000;
```

$$Wanted := 100000$$

```
> preimages := Array(0 ..Wanted);
```

$$preimages := \left[\begin{array}{c} \mathit{0\ ..\ 100000\ Array} \\ \mathit{Data\ Type:\ anything} \\ \mathit{Storage:\ rectangular} \\ \mathit{Order:\ Fortran\_order} \end{array}\right]$$

```
> iworking := 0;
```

$$iworking := 0$$

```
> ifree := 1;
```

$$ifree := 1$$

```
> preimages[0] := 0.5;
```

$$preimages_0 := -0.5$$

```
> while ifree < Wanted do
      p := preimages[iworking];
      preimages[ifree] := $-\frac{(1+\sqrt{1+4\cdot p})}{2}$;
      ifree := ifree + 1;
```

```
                                         p
        preimages[ifree] := - ────────────────────── ;
                              preimages[ifree - 1]
        ifree := ifree + 1;
        iworking := iworking + 1;
end do:
>plot(map(t→[Re(t), Im(t)],[seq(preimages[k], k = 0 ..Wanted)]),
style=point, color=black, symbol=point, symbolsize=1)
```

Remarks

- The pre-images of $-\frac{1}{2}$ are a subset of the Julia set. $B(v)$ is singular on *all* of those!

$$B(p_N) = \prod_{j=0}^{N-1} \frac{(1+v_j)^2}{(1+2v_j)} B(v_N) = B(-\frac{1}{2}) = \infty. \tag{76}$$

- The set of pre-images of $-\frac{1}{2}$ is an infinite set and is a *fractal* (looks like a cauliflower) and approaches 0 arbitrarily closely. Therefore, there are poles of $B(v)$ arbitrarily close to 0 and thus, the series $B(v) = 1 - v + \cdots$ *cannot* converge.
- The pre-images of -1 are *zeros* of B and they're also dense in the Julia set. Therefore, the Julia set forms a *natural boundary* for $B(v)$ and $\frac{\mathrm{d}v}{\mathrm{d}t} = B(v)v^2$ is *nasty*.
- The pre-images for the Mandelbrot derivation are computed similarly but more simply.
- We reused the pre-image code for the first-year course, for general Julia sets.

2.5 *Chaos Game Representation of DNA Sequences*

This final vignette shows that these dynamical ideas are not just toys or idle puzzles but useful tools for science. As mentioned in the previous section, when the course was first taught, this research area was new and not yet published. Since then, [15–17] have been published, in which the most recent paper [15] was published in a top journal, Bioinformatics (Figure 1).

It is general knowledge that DNA is a double helix and is made of four bases: adenine (A), guanine (G), cytosine (C), and thymine (T). However, for DNA sequencing, we only need to look at one of the two strands of the DNA since the bases are paired: A with G, and C with T. This means that if we know one strand, we automatically would know both strands (as long as there are no mutations). DNA sequences are unique to each organism; therefore, using chaos game representation, the goal is to be able to quantitatively classify the organisms by their classes (in this case, animal classes; the five most well-known classes of vertebrates are mammals, birds, fish, reptiles, and amphibians).

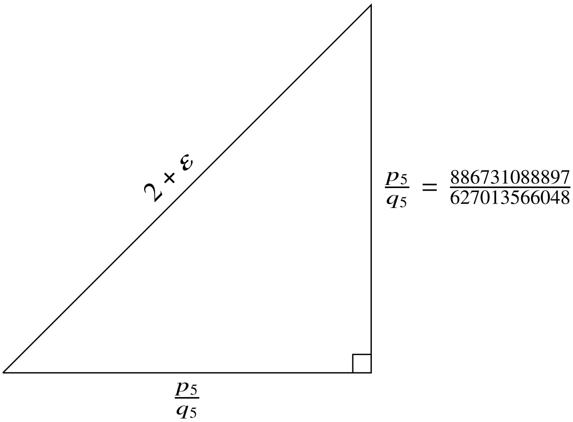


Fig. 1 Example for backward error: instead of a 2 light year hypotenuse, we have a $2 + \varepsilon$ one, where $\varepsilon < 10^{-24}$ light-years or 10^{-6} mm

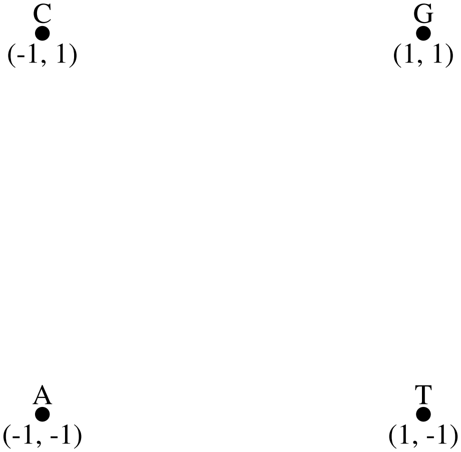


Fig. 2 Chaos game representation setup

We start with a plot with four corners, each of which represents one of the four bases in DNA, as shown in Figure 2. The algorithm for chaos game representation is outlined in the following steps:

1. Put a dot in the center; this is the "current dot" at the beginning.
2. Pick the next letter in the sequence (if none, stop), and draw an invisible line from the current dot toward the corner representing the base.
3. Put a dot halfway on the line; this becomes the new "current dot".
4. Return to Step 2

The short Maple code below explains this precisely.

However, for the course, we did not use any DNA sequences; instead, we generated a sequence of (not so) random numbers, in which each number represents an index of our variable `ACGT`, which corresponds to a coordinate point (representing a base). The following is the Maple procedure that was used to generate chaos game representation plots.

```
> CGR := proc(s::list)
     local ACGT, k, n, p;
     ACGT := table();
     ACGT[0] := [-1, -1];
     ACGT[1] := [-1, 1];
     ACGT[2] := [1, 1];
     ACGT[3] := [1, -1];
     n := nops(s);
     p := Array(0 ..n);
     p[0] := [0, 0];
     for k to n do
         p[k] := (ACGT[s[k]] + p[k-1])/2.0;
     end do:
     plots[pointplot]([seq(p[k], k=1..n)],color=black,axes=none);
end proc:
```

We can see that in the code, instead of using letters for the bases, we have used the numbers 0, 1, 2, 3 instead.

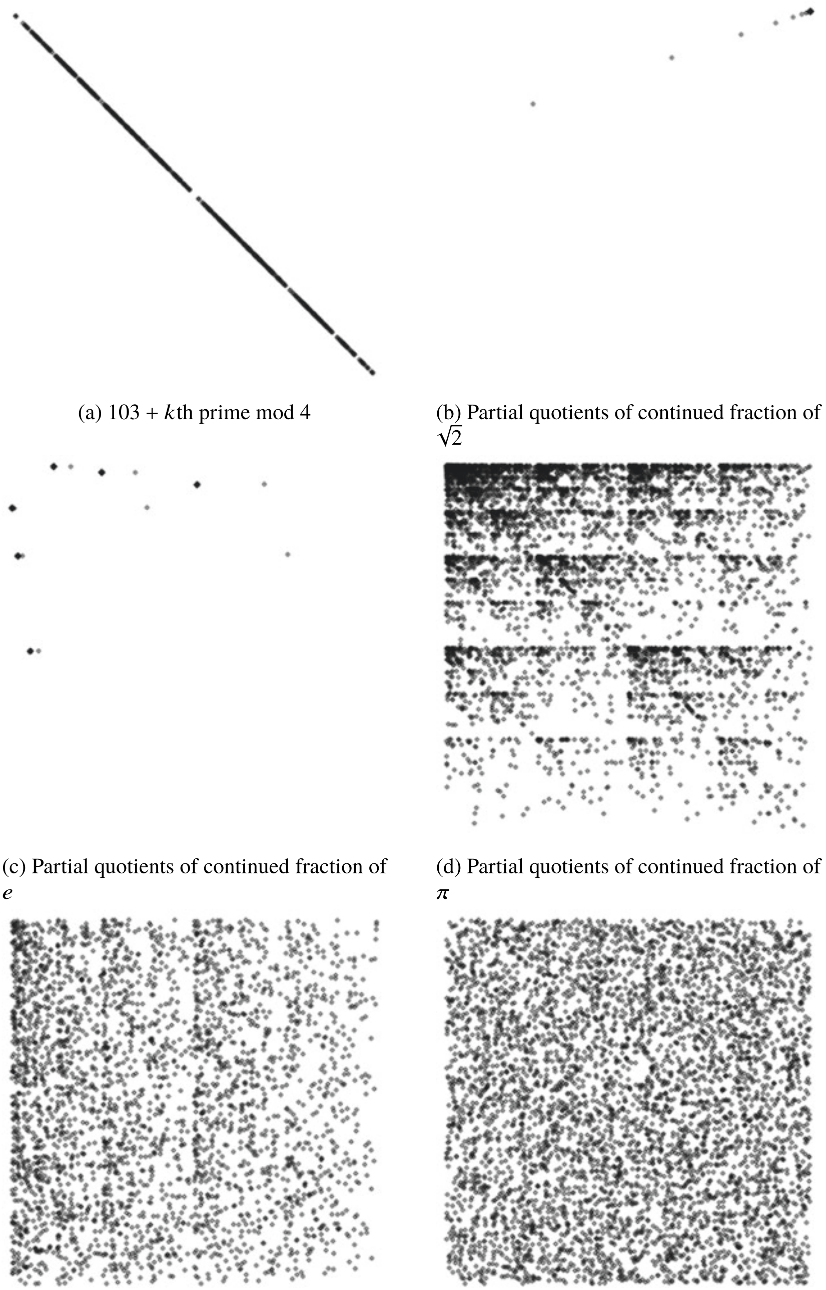
The first sequence the class looked at was the ith prime number (actually, we took the $103 + k$th prime). The result, a diagonal line (Figure 3a), was surprising at first. But actually, it is not *that* surprising in hindsight. As we all know, prime numbers bigger than 2 will never be even, so the coordinate points (in this case $(1, 1)$ and $(-1, -1)$) that represented "even" values would never occur, thus creating a straight diagonal line. If we had reassigned the values the coordinate points represent, the plot would turn out different; instead of a diagonal line, it could possibly be a horizontal or vertical line. One needs to be mindful of how the coordinate points are assigned.

The students were then encouraged to try other sequences. Some started with something familiar: one recurring theme of this course was $\sqrt{2}$, so with this in mind, some students jumped on the opportunity to plot the chaos game representation of the quotients of the continued fraction of $\sqrt{2}$. As we have seen earlier in the course, the sequence goes like

$$1 + [2, 2, 2, 2, \ldots, 2]\,. \tag{77}$$

Because all elements of the sequence (apart from the first one) are 2's, it is not surprising to see a (faint) diagonal line with most of the points in the upper right corner, shown in Figure 3b. The students also tried the partial quotient of e, shown in Figure 3c. What is seen here makes sense as the sequence mostly contains 1's and these alternates with even values, so in this case, either 0 or 2. Therefore, it is clear that there are no points in the lower right corner since that represents the value 3. Unfortunately, these two plots do not look all that impressive; in fact, it is pretty underwhelming. Disappointed with this result, the students decided to experiment with other sequences in which all four coordinates occur.

The students then thought, "Why not take the partial quotients of the continued fraction of π? The results *must* be random." As shown in Figure 3d, unexpectedly, there is indeed a pattern, which shows that the sequence of partial quotients of the



(a) 103 + *k*th prime mod 4

(b) Partial quotients of continued fraction of $\sqrt{2}$

(c) Partial quotients of continued fraction of *e*

(d) Partial quotients of continued fraction of π

(e) Digits of π

(f) Random numbers

Fig. 3 Examples of chaos game representations (all examples are mod 4)

continued fraction of π is not as random as we thought at the beginning of the course. Although the continued fraction quotients of π may not actually be random, it is known that the digits of π themselves are. The chaos game representation plot (Figure 3e) of this shows some randomness. This can be compared to the chaos game representation plot of randomly generated values mod 4 (Figure 3f).

Additionally, this brings up the thought of whether this idea can be extended to other figures—base 6, base 8, base 3. This could possibly be explored in a future version of the course. This connects interestingly with the random walk on the digits of π [1].

3 Concluding Remarks

There were many successful outcomes from this course. The most political achievement was meeting the Hon. Kathleen Wynne, Premier of Ontario, at the time of publication as part of a WALS demo (she said that she was frightened of mathematics). Secondly, two undergraduate students who participated in the first year of the course (Alex Wu and Tiam Koukpari) founded Mustang Capital, initially known as The Algorithm Trading Club, which has over 100 members. They use the pedagogical principles of this course to teach themselves about algorithmic trading. There were also many academic achievements that came from this course: beautiful and prize-winning posters (in ISSAC 2016) were created, and two students (Yang Wang and Ao Li) published papers from their projects. In addition to this, RMC recruited at least one PhD student (possibly more are pending). Lastly, interest in the course has been generated for The Digital Humanities Program.

A Less Happy Outcome

The course AM 1999 was taught only once, in spite of the promises from the outgoing Chair and the Dean to run it for two years. RMC had failed to get the Dean's promise in writing, and the incoming acting Chair canceled the second offering, owing to an important misunderstanding on the part of some colleagues. Jon Borwein was scathing about this decision—any new program needs time for growth of awareness, and this was especially acute here given the short period from conception (May 2014) to first delivery (Sept 2014). The senior course AM 9619 *was* offered a second time, which would have made the second offering of AM 1999 "free"; this makes the lost opportunity even more sharp.

This bit of data is not included here as a lament, but rather for clarity and as a recommendation: before undertaking such a serious undertaking, get your promised support in writing, and be sure to tell your colleagues what you are doing.

Acknowledgements We thank David Jeffrey for his early encouragement, Steven Thornton and Torin Viger for their help with the course material, and the membership of USAM: Julia Jankowski, Andy Wilmot, and Anna Zhu. Gavan Watson, Stephanie Oliver, and Wendy Crocker were very helpful with active learning and the use of WALS. We thank Lila Kari for introducing us to the chaos game representation. We also thank the Rotman Institute of Philosophy and the Fields Institute for Research in the Mathematical Sciences for their sponsorship of the Computational Discovery conference (www.acmes.org), at which some of these ideas presented here were refined. We also thank the Center for Teaching and Learning for a Fellowship in Teaching Innovation Award 2018–2019.

References

1. Artacho, F.J.A., Bailey, D.H., Borwein, J.M., Borwein, P.B.: Walking on real numbers. Math. Intell. **35**(1), 42–60 (2013)
2. Borwein, J.M., Borwein, P.B.: Strange series and high precision fraud. Am. Math. Mon. **99**(7), 622–640 (1992)
3. Borwein, J., Devlin, K.: The computer as crucible: an introduction to experimental mathematics. Aust. Math. Soc. 208 (2009)
4. Borwein, J., van der Poorten, A., Shallit, J., Zudilin, W.: Neverending Fractions: An Introduction to Continued Fractions, vol. 23. Cambridge University Press, Cambridge (2014)
5. Corless, R.M.: Error backward. Contemporary Mathematics, vol. 172, pp. 31–31 (1994)
6. Corless, R.M.: Computer-mediated thinking. In: Proceedings of Technology in Mathematics Education (2004). http://www.apmaths.uwo.ca/~rcorless/frames/PAPERS/EDUC/CMTpaper.pdf
7. Corless, R.M., Fillion, N.: A graduate introduction to numerical methods. AMC **10**, 12 (2013)
8. Corless, R.M., Jankowski, J.E.: Variations on a theme of Euler. SIAM Rev. **58**(4), 775–792 (2016)
9. Corless, R.M., Jeffrey, D.J.: Scientific computing: one part of the revolution. J. Symb. Comput. **23**(5), 485–495 (1997)
10. Corless, R.M., Essex, C., Sullivan, P.J.: First year engineering mathematics using supercalculators, 2nd edn. SciTex, The University of Western Ontario, London (1995)
11. Griffiths, D.F., Sanz-Serna, J.M.: On the scope of the method of modified equations. SIAM J. Sci. Stat. Comput. **7**(3), 994–1008 (1986)
12. Hamming, R.W.: Methods of mathematics applied to calculus, probability, and statistics. Courier Corporation, Chelmsford (2012)
13. Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S.M., et al.: Scientific teaching. Science **304**(5670), 521–522 (2004)
14. Hardy, G.: Divergent series. 1949. Theorem **65**, 122 (1949)
15. Karamichalis, R., Kari, L.: MoDMaps3D: an interactive webtool for the quantification and 3D visualization of interrelationships in a dataset of DNA sequences. Bioinformatics (2017)
16. Karamichalis, R., Kari, L., Konstantinidis, S., Kopecki, S.: An investigation into inter-and intragenomic variations of graphic genomic signatures. BMC Bioinform. **16**(1), 246 (2015)
17. Kari, L., Hill, K.A., Sayem, A.S., Karamichalis, R., Bryans, N., Davis, K., Dattani, N.S.: Mapping the space of genomic signatures. PloS One **10**(5), e0119815 (2015)
18. Labelle, G.: Sur l'inversion et l'itération continue des séries formelles. Eur. J. Comb. **1**(2), 113–138 (1980)
19. Olds, C.D.: Continued Fractions, vol. 18. Random House, New York (1963)
20. Schröder, E.: On Infinitely Many Algorithms for Solving Equations (1993)
21. Strang, G.: Too much calculus (2001). http://www-math.mit.edu/~gs/papers/essay.pdf
22. Von Neumann, J., Goldstine, H.H.: Numerical inverting of matrices of high order. Bull. Am. Math. Soc. **53**(11), 1021–1099 (1947)

Part III
Financial Mathematics

Introduction

David H. Bailey and Qiji J. Zhu

It is well known that Jonathan Borwein had numerous wide-ranging interests and intellectual persuasions. Moreover, he never stopped only at purely musing a subject, but often concerned himself with its impact to science and society. Jon's involvement in financial mathematics research is an excellent illustration of these multidisciplinary interests.

Jon's most important contribution in the area of financial mathematics was his joint paper with David H. Bailey, Marcos Lopez de Prado and Qiji J. Zhu entitled "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of sample performance" [4]. The paper grew out of the authors' concern that although mathematics has become a standard language to quantify financial phenomena, it is also often used in a misguided or substandard fashion, lending a patina of rigor to the topic at hand, but also masking some serious deficiencies. This paper unequivocally demonstrated that many financial strategies and fund designs, claiming to be backed up by extensive "backtests" (analyses based on historical market data), are nothing more than illusory artifacts resulting from statistical overfitting. Indeed, this paper concluded that backtest overfitting is the most likely reason why so many financial strategies and fund designs that look great on paper often fall flat when actually fielded. This paper was followed by a more detailed analysis on how to estimate the probability of backtest overfitting [5].

Seeing the potential harm of this abusive use of mathematics to the general investors and society, Borwein and some collaborators established the blog site

D. H. Bailey (✉)
Lawrence Berkeley National Laboratory, Berkeley, CA, USA
e-mail: david@davidhbailey.com

Q. J. Zhu
Department of Mathematics, Western Michigan University, Kalamazoo, MI 49009, USA
e-mail: zhu@wmich.edu

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_15

"The Mathematical Investor: Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA)" [6]. From 2013 to 2017 the blog site published more than 65 blogs to explain to the general public, in accessible language, the many pitfalls from misusing mathematics in finance, and the bankruptcy of many popularly promoted investment methods. In the wake of Borwein's death, Bailey has continued to pursue this cause in his new Mathematical Investor blog, which includes material from the earlier blog, and the MAFFIA.org site [2].

At the Jonathan Borwein Commemorative Conference (JBCC), the talk [7] addressed the related issue of objectively measuring the reliability of predictions by financial forecasters. The talk was a concise summary of the joint paper of Jonathan Borwein with David H. Bailey, Amir Salehipour and Marcos Lopez de Prado [8]. Jon also co-authored another related paper, summarized in a separate talk at the conference, that specifically addressed overfitting in stock portfolio design [3].

Another line of Jon's work in the area of financial mathematics was reflected in the talk [10] at JBCC. In this work, Jon and his co-author Qiji J. Zhu observed that most important results in financial mathematics can be derived using a unified framework of entropy maximization. This is an interesting cross-disciplinary study because entropy maximization is a physical principle. The fact many fundamental financial results can be derived from such a physical principle begs the question: Do financial markets behave like physical systems? What can be learned from this apparent correspondence?

Of course, the abuse and misuse of mathematics are not limited to the area of financial mathematics. The talk given by M. Altman [1] provides an example in the area of economics. Jon and his collaborators also discussed another example in the area of scientific computing in [9].

The talks in the financial mathematics session at the Jonathan Borwein Commemorative Conference provides us a glimpse at depth and breadth of Jon's contribution to the financial arena. The references below hopefully will give the reader a better picture of the scope of Jon's work.

References

1. Altman, M.: A Holistic Approach to Empirical Analysis: The Insignificance of P, Hypothesis Testing, and Statistical Significance
2. Bailey, D.H.: Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA). https://www.maffia.org/; blog is at https://www.mathinvestor.org
3. Bailey, D.H., Borwein, J.M., de Prado, M.L.: Stock portfolio design and backtest overfitting. J. Invest. Manag. **17**(1) (2017)
4. Bailey, D.H., Borwein, J.M., de Prado, M.L., Zhu, Q.J.: Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. Notices Am. Math. Soc. 458–471 (2014)
5. Bailey, D.H., Borwein, J.M., de Prado, M.L., Zhu, Q.J.: The probability of backtest overfitting. J. Comput. Financ. **20**, 1–31 (2016)

6. Bailey, D.H., Borwein, J.M., de Prado, M.L., Zhu, Q.J.: The Mathematical Investor: Mathematicians Against Fraudulent Financial and Investment Advice (MAFFIA)," Blog (2013–2017). This blog's contents and newer material are now at [2]
7. Bailey, D.H., Borwein, J.M., Salehipour, A., de Prado, M.L.: Evaluation and ranking of market forecasters
8. Bailey, D.H., Borwein, J.M., Salehipour, A., de Prado, M.L.: Evaluation and ranking of market forecasters. J. Invest. Manag. **16**(2), 47–64 (2018)
9. Bailey, D.H., Borwein, J.M., Stodden, V.: Facilitating reproducibility in scientific computing: principles and practice. In: Atmanspacher, H., Maasen, S. (eds.) Reproducibility: Principles, Problems, Practices and Prospects, pp. 205–232. Wiley, New York (2016)
10. Borwein, J.M., Zhu, Q.J.: Entropy maximization in finance

A Holistic Approach to Empirical Analysis: The Insignificance of P, Hypothesis Testing and Statistical Significance*

Morris Altman

Jon was one of the most creative, critical, energetic and open-minded academics I ever met and we hit it off almost immediately. Four years ago we started our collaboration on workshops and a book, on underlying fundamental problems to applied statistics, and math from a multidisciplinary perspective. What a severe loss his passing was to the community of critical thinkers and thought leaders. And, I lost a friend.

1 Introduction

This chapter sets out to address the question of what are some of the key underlying necessary conditions or foundational requirements for robust scientific statistical practice, especially in the social sciences, but with a strong bearing on other areas of applied scientific discourse, including the so-called hard sciences. There is currently an excessive focus on technique, especially on statistical approaches to hypothesis testing which, in turn, emphasizes P tests and, relatedly, tests of statistical significance. There is also a focus on correlation analysis, with little or no emphasis on causality nor on the theoretical basis for the applied modelling structure, apart from the underlying statistical theories. The focus on technique without much relationship to how the data is collected and collated and cleaned, the representativeness of the data (issues of external validity), what data is collected and why (relates to the chosen models-theories being employed), the focus on correlations without a

M. Altman (✉)
Dean, University of Dundee School of Business (UDSB), Dundee, Scotland, UK
e-mail: MAltman001@dundee.ac.uk; morris.altman@usask.ca

Chair Professor of Behavioural and Institutional Economics, and Co-operatives, Dundee, Scotland, UK

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_16

theoretical context, and a single-minded focus on statistical hypothesis testing can generate highly misleading analyses under the guise of robust scientific procedures. This can generate highly misguided policy which can be touted as being scientifically robust because of its statistical significance and the relatively high correlation between particular variables.

We discuss how to best use statistics to better understand socio-economic and behavioural phenomena. And, a first step in so doing is to appreciate the statistical input that goes into statistical analysis, the assumptions underlying the statistical theories that are applied, and the fundamental importance of causal theories (non-statistical theories), for applied analysis. In order of analytical importance, I argue that the use of P, statistical hypothesis testing and, relatedly tests of statistical significance, are not of high order importance and should be deemed analytically insignificant. The focus on P, statistical hypothesis testing and tests of statistical significance detract from the much more important exercises of data collection, construction and causal modelling. I provide examples to illustrate the importance of re-focusing analysis to theory informed applied analysis and the importance of better understanding data collection and construction. This also serves to pinpoint critical weaknesses in current approaches to applied research. I also briefly discuss why sub-optimal approaches to applied research that yield incorrect scientific results and deleterious policy outcomes can persist over time. I conclude that this is largely a product of rational behaviour in a perverse decision-making environment enveloped in a world of imperfect and asymmetric information and asymmetric power relationship in determining what gets published and who receives research grants, for example. This decision-making environment is further polluted by incorrect mental models about what is the most appropriate and most scientifically robust approach to applied research.

2 What is the Problem

The first order of business is to address the question of why the concern about how applied research is and has been undertaken. Both inside and outside of the hard sciences it has been well documented that there has been a focus upon P, statistical hypothesis testing and tests of statistical significance and that statistical significance is often used as a proxy for substantive significance [4, 5, 9, 10, 14, 20, 25–27, 29, 29, 33, 35, 36]. Moreover, scholars and practitioners tend to assume that the conditions for using these tests appropriately are met, with little effort to test this hypothesis. Also, hypothesis testing is limited to the testing of the null hypothesis in terms of the statistical significance of the estimated outcome(s). This is in contrast with hypothesis testing that relates to addressing questions of causality and specifying necessary and sufficient conditions. The latter lies outside of the domain of statistical analysis but is critical to causal analysis. Distinguishing between the two types of hypothesis testing is of vital importance, where currently the statistical approach to hypothesis testing in terms of statistical significance now dominates. Additionally, correlation analysis is used to suggest causation without adequate regard to the possibility of spurious correlation. This relates to lack of theoretical focus informing statistical analysis and inadequate regard to alternative theories to explain causation given particular circumstances. The main critiques of statistical practice do not pay much attention to the importance of the robustness of the data used in one's statistical analysis and the importance of non-statistical theory driving empirical research. Relatedly little attention is paid to the extent to which data and sample data construction are transparent. A necessary condition for robust statistical analysis is the quality of the data being analysed. This quality needs to be demonstrated, not assumed.

Overall, the main critique of statistical practice in the literature relates to the abuse of tests of statistical significance. This is a point of commonality across economics, finance, psychology, and medical/pharma critiques of applied research. P is used as a proxy for both statistical significance and substantive significance. Relatedly, statistical significance translates into rejecting the null hypothesis (there is a real (not a fluke) difference between the empirical result and the null) and accepting the alternative hypothesis. But a statistically insignificant result translates into accepting the null since there is no real (fluke) difference between the derived coefficient and the null. Such statistical hypothesis testing (the statistically flukiness of one's results) is used to determine analytical, substantive, clinical significance. So serious is the malpractice involved in applied research that the American Statistical Association (ASA) released a statement and a related document trying to nudge researchers taken into a more scientific approach to applied analysis. Here are snippets of the ASA statement [9, 34]:

- Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.
- The p-value was never intended to be a substitute for scientific reasoning. Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a post $p < 0.05$ era.
- Over time it appears the p-value has become a gatekeeper for whether work is publishable, at least in some fields This apparent editorial bias leads to the file-drawer effect, in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as p-hacking and data dredging that emphasize the search for small p-values over other statistical and scientific reasoning.

There is this persistent problem in this application of tests of statistical significance, variously critiqued and condemned. But what needs further emphasis are the limitations of P and statistical hypothesis testing in general, what are the proper scientific substitutes for these tests, what is the theoretical context of statistical analysis, and what is the validity and the limitations of the data being interrogated using statistical methodologies. Decreeing over and over again the limitations of P and its abuse, without robust and clear alternatives being put in place will not result in a mass transformation in statistical practice. This is especially the case, given the many obstacles to change discussed below.

3 P and Statistical Hypothesis Testing in Context

Hypothesis testing as defined in statistics is not the same thing as testing for cause and effect, determining necessary or sufficient conditions, or testing for the impactfulness of the independent variables upon the dependent variables. To reiterate, practitioners test for the validity of the null hypothesis against the alternative, thereby testing for the statistical significance of a chosen variable (but the test is the flukiness of the result and nothing more). The null hypothesis is the benchmark for the variables (typically the independent variables) being tested, being tested for whether or not they are different for the benchmark or the null. Not only are they being tested for differences from the null, they are tested in terms of a selected level of confidence. One can end up rejecting the null for a dependent variable at high degree of confidence (the norm, based on nothing scientific) which is, of course, 95 percent. In this case, one's result is statistically significant. One can end up accepting the null and then one's result is statistically insignificant. As long as the estimated P-value is equal or less than the level of confidence, the null is rejected and we have statistical significance. But hypothesis testing here is simply, very specifically and narrowly related to whether or not one's estimates are reliable, whether or not they are statistical flukes, whether or not repeating one's analysis with a different sample is likely to generate similar results in terms of accepting or rejecting the null hypothesis. This is, to reiterate, a very special type of hypothesis testing. No causality test here, no test for the importance or impactfullness (size effect), no effort to determine necessary or sufficient conditions, no determination of which variables should be estimated based on a scientifically determined reasonable model.

But even given this rather narrow conceptual framework, that is accepting or rejecting the null hypothesis, having one's empirical results deemed statistically insignificant or significant, is based on the presumption of that one's confidence interval is scientifically determined. Otherwise, one's analytical determination in terms of statistical significance is ephemeral, with benchmarks (in this case confidence intervals and relatedly, P-values), shifting like sand in a storm.

A directly related contextual point is that what is significant at one level of confidence (95 percent) might not be at another higher level of confidence (99 percent). Alternatively, what is insignificant at a 95 percent level of significance becomes significant at 94.5 percent, 94 percent, 90 percent, 89.5 percent, and so on. Taking hypothesis testing and, therefore, tests of statistical significance at face value, one's empirically determined truth statement–"my estimated coefficient is significant" (statistically significant)–critically depends on the chosen level of confidence. But the scientific value of rejecting or accepting the null hypothesis is based on a leap of faith, often a blind leap of faith, as analysts blindly accept a particular confidence interval as God-given, is highly tenuous. Analysts somehow assume that 95 percent is just right, like the law of gravity or the world is not flat, level of significance.

This assumption is typically made implicitly, based on accepted practice (this is what other people do, this is what my peers do, this what my and other folk's software dishes out). No questions asked. Analysts typically assume that this accepted practice is most probably scientifically based. Were this practice not scientifically based, then wouldn't this practice be abandoned or forced out of the market of statistical analysis procedures by better practice standards? We discuss below why good practices or procedures need not chase out our bad or worse procedures (below the known optimal procedures). Imagine concluding that a cancer medication is statistically insignificant (accept the null), at a 95 percent level of confidence, where it is significant at 94.5 level of confidence or at a 90 percent level of confidence. This treatment is rejected as not effective, not different from the benchmark null, even though the confidence level norm is not scientifically determined. And, based on such hypothesis testing an empirical result not being significant at 95 percent, is sufficient to reject the null, even though the result might be insignificant at a 95.5 percent level of confidence. The same can be said for the impact of minimum wages on employment, regulation on banking stability, immigration on domestic wages, government intervention and business cycle volatility, and the frequency and extent of deep recessions. Hypothesis testing here becomes tenuous, even if one accepted the proposition that these tests (statistical hypothesis testing) could be of some scientific value.

This problem can't be overcome. All that one can do is to be transparent about the meaning of statistical hypothesis testing, providing information on statistical significance across an array of confidence intervals. But then there is no mechanical and simplistic determination of what is or what is not statistically significant. Therefore, even at its very foundation, statistical hypothesis testing rests on very shaky foundations. At best, testing for the reliability of one's estimates can't be precise, contains a lot of noise, and is very much assumption-based (on the specified level of confidence).

Apart from how confidence intervals are applied to hypothesis testing, there is the critically important issue of the robustness of the data being used in one's statistical analysis. This also raises the issue of sample size, the representativeness of one's sample and biases in the sample as well as possible errors in data construction. The latter speaks to the importance of transparency on the methods used to construct data and to construct one's sample. Transparency in data construction also makes it easier to replicate studies using different samples where replication is becoming increasingly important to the empirical enterprise [12]. One can also better check for data falsification, which speaks against the robustness of the data used to test hypotheses [17]. And, one should also check for data that is simply a product of modelling misspecification, generating data that yield meaningless estimates [24]. Whichever way we test one's hypothesis, a necessary condition for the robustness of one's results is the robustness and representativeness of the data used in the analysis as well as the modelling that drives the search for the data to be tested. Statistical analysis is not all about sophisticated, complex and mathematical analytical tools applied to the data. But this is what the focus has been increasingly about, with economics and finance being good examples of this perspective. Data construction does not appear to have the lustre and sophistication of the statistical models employed and their underlying mathematical proofs. But poor data results in impoverished analysis whether you apply narrow statistical hypothesis testing or the broader hypothesis testing discussed throughout this paper, which relates to analytical or substantive significance.

Statistical hypothesis testing is meaningful (testing for the reliability of results; probability of the outcome being a fluke given a specified confidence interval) only if one uses samples derived from the population and one that is randomly selected–the sample should be representative of the population from which it is derived. Hence, statistical hypothesis testing is meaningless if one is using the entire population of data. This could be the case for gross domestic product, international trade, labour mobility, and financial markets. However, analysts tend to test for statistical significance even when using the entire population of data, which not appropriate. But most empirical analyses use samples of the population. These can be small or large. But what is critical is they are representative of the population from which they are drawn.

A critical exercise is building data sets that are representative or interrogating samples to determine their representativeness. One then has to determine what the sample population is representative of. This relates to the external validity of one's results. Also, if one is using samples constructed by others one has to determine if these samples are appropriately representative. This requires transparency on how data are constructed and how the sample is derived. If one uses American data, one cannot extrapolate from the American results to Canada, Japan, China, India, Lebanon or Israel, for example. What is statistically significant for the US is not statistically significant for other countries, by definition. If one runs a classroom experiment with a sample student population, one has to ask what is the sample representative of to what extent can one generalize from such a sample? The same would go for field experiments as well as for what is referred to as randomized controlled experiments. Can you really generalize from Newcastle, Australia to the rest of the world or from a village in one part of Africa to the rest of Africa or to India, for example? Tests of statistical significance and relatedly statistical hypothesis testing can only be directly pertinent to a particular representative sample. And what this is, and how the sample is constructed, is a critical prior to any legitimate to any statistical hypothesis test. Moreover, tests of statistical significance tell us nothing about the representativeness of the sample. This can only be determined by prior non-mechanical analysis. Additionally, when building one's own samples, it is critically important to construct samples so that they are as representative as possible given the target population.

It is also important to avoid the illusion that simply by massaging the data to yield statistical significance (by increasing sample size) this implicitly implies that the sample is representative. Statistical significance tests and relatedly statistical hypothesis testing is not a shortcut to determining the representativeness and overall robustness of one's sample. We are well aware that increasing sample size sufficiently makes one's results statistically significant, given how this statistic is constructed. This results in rejecting the null hypothesis. Unless one appreciates that this can simply be a matter of massaging data (increasing sample size without much thought to representativeness and robustness), this approach can inadvertently result in unscientific and frivolous conclusions, including the notion that one's results are now, finally, scientifically robust, even though all that has been done is increasing sample size.

The focus on statistical significance was, originally, to determine if one's empirical result is probably a fluke or not, given that one's sample is representative. Having a very small sample (given how statistical significance is calculated) might generate statistically insignificant results. The sample is too small to generate a reliable outcome. The result, however, would be true for the sample, even though it might be a fluke, not replicable with another sample. A statistically insignificant result might still be of importance when the result is analytically important, where the latter is related to the size effect. One danger of being too mechanical in one's use of tests of statistical significance would be dismissing analytically significant results out of hand, just because they are statistically insignificant. Increasing sample size can be of importance, however, since this has the effect of reducing the probability of one's results being fluky, which is of considerable importance. But how one increases sample size is critical. Simply having results that are statistically significant is beside the point if the sample size is grown without paying careful attention to the representativeness of the sample.

This point of context is highlighted in the most recent focus on Big Data. One reason why samples tend to be relatively small relates to the costs involved in building larger samples. One of the apparent blessings of Big Data is that it is easier and cheaper to construct larger samples, therefore, generating empirical results that can all be statistically significant. And this is vitally important, especially for those who focus on statistical significance and statistical hypothesis testing as the core of the scientific analysis. But Big Data is no magic answer to the problem of small samples. Small samples, in and of themselves, are not a problem if the sample is big enough. Just because one has a lot of data (Big Data) does not mean that one has resolved the so-called the small sample size problem. Driving up sample size with Big Data simply generates statistically significant results (so one can reject the null hypothesis). But this does not resolve the scientific challenge, where it exists, of small sample size, yielding statistically insignificant results.

The Big Data problem is exactly the same as when one blindly increases the sample size to generate (because of the nature of the calculation) statistically significant results. Any sample small or large must be constructed so that it is robust (minimize the errors, noise) and is representative. Big Data, where much of the data are not well-constructed and not representative, when accepted blindly, can generate highly biased, but statistically significant results. One can have biased samples that are statistically significant. Clearly, once again, statistical significance, even with Big Data as a companion, need not signify empirical results that are in any sense scientifically robust. Indeed, it is much better, scientifically, to have a less Big sample, that is well-structured and representative [21].

Another bit of context related to hypothesis testing and the use of correlation analysis. Firstly, the same problems apply here as they do for standard tests of statistical significance and statistical hypothesis testing. Moreover, speaking directly to Big Data, increasing sample size sufficiently will, of course yield statistically significant results. But, however strong is a correlation coefficient and even if it is also statistically significant reveals nothing about causality. This is especially the case when the correlations are drawn without a theoretical framework. One easily ends up with statistically significant spurious correlations. The same is true if the theory is wrong, not being based upon careful consideration of alternative causal variables. Big Data, in and of itself, does not resolve the sample size issue of representativeness, for example, nor can it address the important issue of causality.

4 Robust Samples, Analytical Significance and Impactfulness

Even given robust and representative samples, statistical significance, P-values, and statistical hypothesis testing cannot tells us anything with regards to analytical significance, a key point made by critics of the misuse of tests of statistical significance. To reiterate, statistical significance, P-values, and statistical hypothesis testing, also cannot tell anything with regards to the robustness and representativeness of one's sample. This can only be achieved by doing the hard work, that cannot be done mechanically through tests of statistical significance. These tests become analytically meaningless even if we get our sample size large enough. Given that the sample is representative (and large enough) the result is unlikely to be a fluke. But that is all of the information that these tests convey at their very best.

Bearing this in mind, the much neglected question in the analytical literature is, irrespective of sample size, what is the size effect; how large is the estimated coefficients or correlation. And, related to this, how small is too small to be unimportant and how big is big enough to be important [25–27]. To the extent that the sample is representative and well-constructed (robust), the critical analytical question relates to the size effect, also referred to as analytical or clinical significance. A necessary condition for the size effect to be of scientific value, generalizable to a pertinent population, is for the sample to be representative. But the point of focus should then fall on analytical, not on statistical significance. And, then, hypothesis testing moves from the realm of statistical hypothesis testing (which is all about statistical significance) to testing for analytical significance and testing for and establishing causality [4, 5, 8, 14, 20, 25–27, 33].

Hypothesis testing can take the form of testing for analytical significance. The null can be that the independent variable has no effect on the dependent variable. But this here is not related to testing the null with regards to statistical significance. The prior here is that the estimated coefficient is in some sense meaningful and real. The question is, is this real effect of analytical significance. The null can be that government macroeconomic intervention has no effect on the extent and depth of the business cycle, that managerial quality has no impact on productivity, that mental health has no effect on productivity, that individuals don't make financial decisions based on herding (follow the leader), that a medical treatment has no effect on morbidity. We might find that improved managerial quality increases productivity by 5 percent and that herding can explain (statistically) 20 percent of financial decisions (the purchase or selling of shares). Are these effects big enough to be of importance? No statistical package can address this question. Analytical significance needs to be discussed in the context of the size effect and perhaps the costs (or opportunity costs) of achieving a particular size effect. What needs to be made explicit is the size effect and what it means in terms of the impact of the independent variable or variables on the dependent variable. The narrative surrounding the size effect (analytical significance) becomes of fundamental importance. This type of hypothesis testing focuses attention on impact and raises questions and provides insights on causality (the possible causal relationship between the independent variable on the dependent variable). The size effect should be reported and discussed even if one's findings are not statistically significant as long as the sample is a representative one. A too small sample size simply suggests that one should (probably) increase the size of the sample. But the size effect still provides some insight on the relationship between the independent variable(s) and the dependent variable or the correlation between pertinent variables.

Once one pays attention to the size effect, then one must pay attention to simple but important descriptive statistics such as whether one measures the size effect in terms of the arithmetic average or in terms of the median, for example. The choice could make a big difference to one's narrative about the size effect. Reporting on the size effect using both measures could contribute to an important analytical narrative. Much hinges on the analytical hypotheses one is testing. One might also wish to report on the variance for each mean. The latter is important because the extent of variation can speak to the reliability of the impact of the size effect. Greater reliability is generated by samples with less variation, ceteris paribus. One much discussed measure of size effect has been the Cohen d statistic, which standardizes the size effect in terms of standard deviation [15]. This might be too much of a mechanical measure, but it still provides some pertinent information on size effect and, therefore, on analytical significance. Also, one can specify the size effect in terms of the percentage of one sample whose size effect is greater than some threshold to be of analytical importance. This specification is significant when the variance is relatively large. However, which way one decides to articulate the size effect or its context, this discourse is of critical importance and needs to be given superior weight to the standard focus on statistical significance and statistical hypothesis testing. Science is supposed to speak to analytical importance, the relationship between possible causal variables, and the measured relationship between independent and dependent variables. This is the purview of size effect related narratives.

5 The Role of Non-Statistical Theory and Statistical Analysis

Without carefully considered theory (non-statistical theories or models) underpinning one's empirical analysis, even if one focuses on size effects, these measured effects might be misleading and misguiding, a product of spurious correlations and the use of missing or inappropriate modelling variables. The most poorly specified models will always generate correlations and estimated coefficients for the independent variables. And if the sample size is large enough these will also be statistically significant. With Big Data, for example, we enter into a sublime world of statistical significance. But here, without well-specified theory and representative samples, the size effects, no matter how big are scientifically meaningless.

Theory is required to specify which variables are being causally linked. And, this hypothesized causal link can then be tested in terms of correlation and size effect of the independent variables. But a problem arises if the theory is not well thought out or interrogated for its reasonableness. For example, it could be possible for there to be a high and positive correlation between rainfall and clothing output in New York City in a specified period. The underlying theory is that more rain is causally linked to more clothing production and vice versa. But is this a plausible theory based on our empirical understanding of the relationship between rainfall and the level of clothing production? What one has actually measured might very well be a spurious correlation, one that might even be statistically significant. One might have a model of financial markets that predicts no major financial crisis into the future, which any individual versed in financial history might find suspicious, but financial decision makers might find scientifically valid, especially if the results are statistically significant. [32]. But such analysis removes from the data set deep financial crisis such as characterized by the Great Depression of the late 1920s. Overall, poorly modelled and ahistorical (financial crises eliminated from the data set), generates spurious and misleading results.

One might have an analysis that attempts to determine laggard economic performance, but assumes that the economy is economically efficient, following from the behavioural assumption that all rational decision makers will always be economically efficient. But this approach of conventional economics can't address causes of poor economic performance that are caused by economic inefficiency (referred to by [23] as x-inefficiency) as this aspect of the black box of the firm is assumed to be unimportant. This type of model is plagued by missing variable problems since alternative more realistic assumptions underlying the theory of the firm are ignored and are not being tested. Statistical significance and even an impressive size effect loses its scientific clout in this type of scenario. Conventional economics assumes that higher wages increases unemployment making workers worse off all other things remaining the same. But this assumes that higher wages don't incentivize individuals to increase productivity offsetting the costs of higher wages. The latter is assumed away. What one has here is another critical missing variable problem that can lead to misleading causal analysis.

In medical research, randomized controlled experiments, now used in economics [16] faces other types of methodological problems, vested in theory, which can cause significant issues irrespective of size effects and statistically significant results. When appropriately testing for the size effect for particular drugs against the control, the characteristics of the test group needs to be carefully constructed so that they match individuals who actually have a particular illness and this test group needs to be carefully mapped against the characteristics of the control group. The same holds true when testing for the impact of economic policy. This requires a theory that hypothesizes the relationship between the drug being tested and the circumstances in which (and test and control group characteristics) one would expect predicted results. As always, the theory needs to be reality-based, reasonable and realistic given likely real-world circumstances.

Theory is what one uses knowingly or unwittingly to test hypotheses (either statistically or in terms of causality non-statistical hypothesis testing). Theory provides the guideposts that frame one's empirical analysis. Therefore, it's critically important to take into consideration alternative theories and to consider the reasonableness of the theories one wishes to test in efforts to determine the causal relationship between the dependent and independent variables. On the importance of non-statistical theory as an engine of empirical analysis, Thomas Kuhn (Coase 1994: 27) writes:

> The road from scientific law to scientific measurement can rarely be travelled in the reverse direction. To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly; even then nature may not yield consistent or generalizable results without a struggle.

Models which have a high correlation statistic and, overall, impressive size effects, appear to have the capacity to predict well and be strongly suggestive of causality (independent variables cause the dependent variables). But this reasoning would be wrong. Alternative variables (independent variables) might have the same power (size effect and predictive value), but might be more causally significant. What is of critical importance is to locate those other variables and determine which are impactful and which variables are most reasonable in terms of the realism of their underlying assumptions and the realism of the model in terms of context and the reasonableness of the analytical narrative. This would allow us to distinguish between models with identical correlation statistics and models with identical measured coefficients for the independent variables. It is those models that are more realistic in assumptions, modelling and context, that allow us to identify models and variables that can address questions of causality [3, 7, 8, 31]. Then the size effect becomes analytically meaningful. But statistical analysis without theory, which amounts to correlation analysis, or statistical analysis with poorly informed theory, yields meaningless or, perhaps more specifically, misleading size effects and meaningless tests of statistical significance.

The importance of non-statistical theory to scientific empirical analysis requires highlighting given the rise of Big Data. As discussed above, Big Data is viewed by many experts and pundits as the grand solution to limitations of and issues related to empirical analysis. A theoretical empty empirical analysis invariably results in spurious correlations. Theory is required to specify possible plausible causal relationships that are being identified or tested. Moreover, Big Data (large sample size) is no guarantee of a representative sample given the hypothesis being tested. A large enough but relatively small well-structured representative sample is scientifically robust as compared to a Big Data sample that is not well-structured or representative [21]. Harford [21, pp. 15–19] makes the point that:

> a theory-free analysis of mere correlations is inevitably fragile. If you have no idea what is behind a correlation, you have no idea what might cause that correlation to break down big data do not solve the problem that has obsessed statisticians and scientists for centuries: the problem of insight, of inferring what is going on, and figuring out how we might intervene to change a system for the better.

The fascination with Big Data is in part related to the focus by experts on statistical significance and a further related focus on bigness as the key means to achieve statistical significance. But as with all statistical analysis, its not all about size and its certainly not about statistical significance if one is to achieve a better understanding of causality, possible causality and the veritable impact of the independent variables upon the dependent variable. Being bigger does not distract from the need for samples to be representative and robust. Nor does bigger distract for the importance of the empirical analysis being informed by theory. Of course, empirical evidence can suggest modification to theory. But this would be as much the case for Big Data as for well-structured, representative, smaller samples.

6 Statistical Analysis: The Implications of Going Beyond Statistical Significance

Going beyond statistical significance demands a more holistic approach to empirical analysis. It does not mean abandoning statistical significance as one input in the analytical package. However, statistical significance becomes a relatively small player in an analytical narrative that speaks to analytical significance, causality, and robustness. As discussed, even to the extent that statistical significance and its corollary, statistical hypothesis testing, remains in play, it must be clearly contextualized in terms of the variety of confidence intervals that can be used–there is no scientifically determined correct confidence interval. A P-value that generates insignificance at a 95 percent level of confidence might generate significance at a 94, 90, 89, 85, 80 percent level of confidence. The reader needs to know this context to make a judgement call on statistical significance in terms of various possible confidence intervals. This involves more work than simply looking at P through the lenses of one's preordained confidence interval. This exercise is scientifically completely meaningless if there is no prior critical work on the sample, getting the sample as representative as possible and checking for errors in data construction. And one would also have to develop a narrative surrounding the representativeness of the sample. A key question in a scientific narrative is to what extent can one actually generalize from the sample to a wider population, how wide, and within what time-frame. Are one's results time dependent? Here too, tests of statistical significance are beside the point. They tell us nothing about sample representativeness, robustness, or generalizability.

Given that one is working with an appropriate sample (or an entire population, GDP of a country, for example) then the difficult work of determining the analytical significance of one's estimates comes to fore. This has nothing to do with statistical significance. P-values provide us with no useful information here. What is critically important is that we can explain the size effect of our estimated coefficients including the correlation coefficient. We need a narrative that discusses what size effect is analytically significant. This very often is contingent on the problem one is addressing. It might also relate to costs incurred if one executes a certain policy or puts a particular medical treatment in place. What are the overall costs relative to the benefits? This would be a net size effect. Computer programmes generate size effect estimates and also related coefficients of variance. So, the estimates are there. The hard work is related to first understanding and then explaining the size effects in the context of the questions asked and the non-statistical hypotheses being tested.

But as discussed above, there is a prior to even a robust size effect analytically important narrative. This is all about model or models that inform the derivation of the estimated coefficients, more specifically the choice of coefficients to be estimated with regards to the dependent variable. If there is no model the estimated relationships including the correlation coefficients can be misleading and spurious. Correlations need not have anything to do causation. Moreover, there might be alternative theories that should be discussed or examined. Otherwise, one's estimates are plagued with omitted variable problems generating misleading results with regards to causation and even the true size effect of specified estimated coefficients.

Figure 1 illustrates some of these points. The first step is model choice, which can involve examining alternative models/theories in terms of their reasonableness and realism (assumptions) given the context of the theories being tested. This is related to non-statistical hypothesis testing. This allows for the most scientifically appropriate choice of independent variables to be chosen for empirical estimation. But one then must choose the pertinent data set and check for the robustness of the construction of the data set and the representativeness of this sample (if it is a sample). Sample size can be important. And some scholars might want to use tests of statistical significance to help determine if the current sample size is adequate [28, 30]. However, significance and estimated P-values need to be contextualized in terms of alternative confidence intervals. But note in this narrative, statistical significance can play one role, but not at all the key or core role in the scientifically robust analytical empirical narrative. Given that the model and data are appropriate, the model is run and estimates are generated. These results can then be analysed for impact (size effects and variance) and causality. This is the fundamental objective of robust empirical analysis built upon robust models, and robust and representative data. Of course, empirical analyses that challenges the models being tested, in terms of causality, sign of coefficients, correlations, or size effects, can then be used to revise models or theories when and where appropriate [19]. But it is critically important

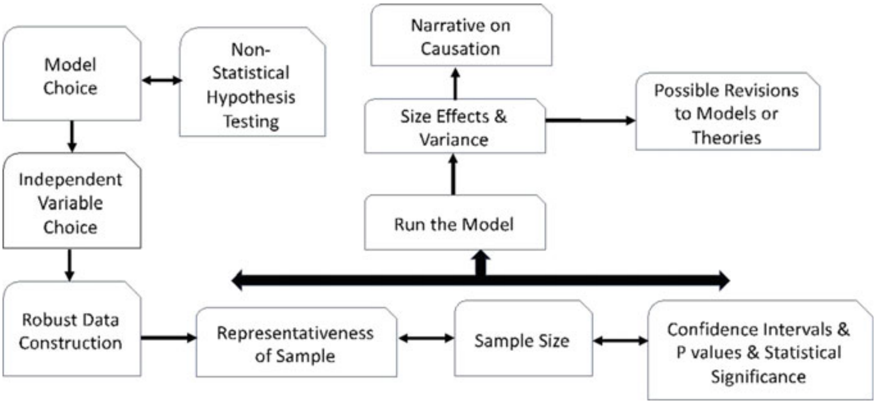

Model Choice
Non-Statistical Hypothesis Testing
Narrative on Causation
Size Effects & Variance
Possible Revisions to Models or Theories
Independent Variable Choice
Run the Model
Robust Data Construction
Representativeness of Sample
Sample Size
Confidence Intervals & P values & Statistical Significance

Fig. 1 More scientifically robust to engaging in empirical analysis

that a prior to any revisions to theory is based upon estimates derived from robust and representative data sets.

7 The Persistence of Inappropriate Approaches to Statistical Analysis

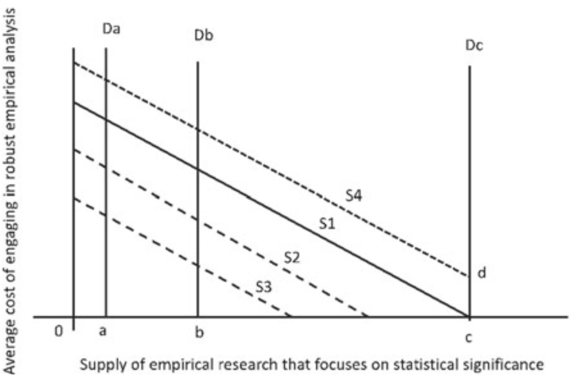
In spite of the literature and organizational outcry about the misuse of tests of statistical significance, this misuse persists. One argument put forth by economists to help explain the dominance of statistical significance and relatedly statistical hypothesis testing in empirical analysis is the relatively low costs of using these tests as a determinant of scientific importance. It's simply cheaper, in terms of time, to focus on statistical significance, which is pumped out of any basic statistical programme, to determine the truth of one's empirical analysis [4, 35, 36]. From the perspective of this modelling scenario, if most individuals are motivated by producing their empirical research by minimizing their cost of production, researchers will persist in supplying a flawed product (statistical significance and statistical hypothesis testing, with the accompanying focus on P-statistics) to the market. This is analogous to firms being able to supply bread that has a large component of sawdust at a low price.

But the price is not the only factor motivating the supply of flawed empirical output to the market of ideas. Peer pressure is important. There is respectable empirical literature on the importance of peer effects influencing behaviour [1, 13]. If your peers focus on statistical significance, so shall you. This is particularly important when individuals are uncertain about best practice approaches and rely on peers who are viewed as leaders to determine the empirical approaches to focus upon. This is a form of rational herding in a world of imperfect information (on herding: [2, 11, 22]). Related to this, researchers and teachers will abide by what their leaders decide, including journal editors, which is the best course of action[18]. This would be true even if researchers believe that the leaders are wrong. This is a form of herding that is based on power relationships. This is a form of involuntary herding. Individuals might engage in such herding because of the fear of punishment (and the related cost of punishment, hence the cost of adopting the appropriate statistical methodology). Punishment might consist of having one's research output or research grant proposals rejected and being excluded from research teams. Researchers might also focus on statistical significance and statistical hypothesis testing when they actually believe that using inappropriate or incomplete methodologies is the best practice. Here we have a case of incorrect mental models (in this case, what is the best, most scientific, approach to empirical analysis) driving the decision-making process [6]. Which mental models one holds to be true is very much a function of one's formal and informal education. One also has the costs of changing one's methodological approach. This relates to what has been referred to path dependency. If one has consistently applied sub-optimal methodologies shifting gear might involve the psychological costs of changing one's mind. But there are also the costs to analysts who have focused on statistical hypothesis testing having to re-focus and retrain as well as having to expose their past research to critical revision. This can result in research leaders imposing costs on others who would want to introduce different or broader approaches to statistical analysis [4, 8].

Focusing on the supply side of empirical research, given costs, I would argue that, overall, supply is affected by peer pressure, herding, power relationships, and mental models. Changes in these factors could change (shift the supply curve) of empirical analyses that concentrate on statistical significance as the core truth test. In a very simple model, reduced peer pressure and improved, more accurate, mental models would reduce the supply of empirical analyses that concentrate on statistical significance to determine the validity of non-statistical hypothesis (testing for size effect, for example), and vice versa.

In Fig. 2, S1, S2, and S3 are three different supply curves. Each is a function of the average cost of engaging in robust empirical research. The typical supply curve in economics is upward sloping, with supply being a positive function of price. But to better illustrate our point we focus on the cost of supplying empirical research, yielding a downward sloping supply curve. One would expect that supply would increase as cost falls, ceteris paribus. Supply can shift inward or outward depending on what happens with the various independent variables such as peer effects and mental models.

Fig. 2 Supply & demand of empiricial research



But the supply side is more complex than this. Improved mental models, by themselves, will not do the trick. Peer pressure, herding and power relationships can keep supply (the supply curve) fixed. Interestingly, this model predicts even if most analysts are dominated by incorrect mental models, they will adopt the correct approach to empirical analysis, if their peers, especially the most respected and powerful peers adopt the correct approach to empirical analysis. This would shift the supply curve inward to the left. Given average cost, less of the relatively low quality (statistical significance) research will be supplied. On the other hand, even if most researchers want to do the right thing (correct mental models), this model predicts that it is unlikely that they will do so unless their leadership, those in power, adopt the correct approach to empirical analysis. Peers with power is a necessary condition to drive improvements in the quality of empirical analysis. Given the latter, improvements in mental models (driven by education and experience) would improve the quality of empirical research supplied to the market. Key to improving the quality, the scientific robustness, of empirical output is changing the leadership and leadership structure (for herein lies the power), of empirical research.

Of course, there is always the risk-taking lone wolfs, the more entrepreneurial researchers and leaders, who will supply more scientific research. But given the material and psychological costs involved in this exercise, the entrepreneurship driver cannot be expected to yield much change on the supply side. Once again, much depends on the extent of change taking place amongst leadership, peers with power.

But the key to the dominance of statistical significance-based analysis for the lower quality low-cost option. Why would there be a market for empirical research that inappropriately uses tests of statistical significance as the cornerstone for determining the substantive or analytical importance of one's results? It is obvious, from the literature, that academic journals, government agencies, and private sector firms making use of statistical analysis affords a ready market for statistical significance-based analysis (statistical hypothesis testing). Leadership and related peer effects affect not only the supply but also the demand for empirical research, be it relatively low or high quality. If there is a demand for higher quality empirical research, given the interaction between the independent variables on the supply and demand-side, on would expect that the market for higher quality research would clear with a greater supply, and demand, of such research.

To simplify this narrative, assume a demand curve that is completely insensitive or inelastic to average cost, given by our demand curves Da, Db, and Dc. In many instances, the demand for empirical research would not be affected by the additional costs of doing empirical research more robustly, going well beyond a focus on statistical significance. These additional costs might be relatively small as compared to the total overall costs of doing such research. To the extent that there was some elasticity, this would simply enforce the use (demand) for lower quality research. On the other hand, to the extent that private sector decision makers understood that higher quality research would yield more profit, this would shift our demand curve inwards to left. This would require a mental model that recognized the importance of the size effect, for example, as critical to generating truly robust research.

In Fig. 2, given demand curve Dc, total supply at the lowest possible average cost (assumed to be zero for simplicity) would be cleared on the market (Dc = S1, at point c). But supply could be greater than demand where S4 cuts Dc at d. There is an excess supply of lower quality empirical output which could be resolved by increasing demand (more journals and organizations accepting such output). Given demand curve Dc, when supply falls below S3, there will be zero demand. In other words, all higher quality empirical output will be rejected because it will not find a market. The world is more complicated than this. There might be a market for some higher quality output, but this would not necessarily dominate the market, given our assumptions. But in this model, increasing the demand for higher quality output, given by shifting the demand curve to the left will generate an increasing market for higher quality, higher cost empirical output. The main point to be illustrated is that the driving force for the production more higher quality empirical research is the demand-side. If the demand-side does not budge, the supply side is constrained. Like the market for organic apples, you can have all the willing suppliers at a particular price (and cost), but if there is no demand, none will be sold.

It is important to reiterate that in the market for quality empirical research there is significant interaction between the demand and the supply side. Another important point to note is that the supply side need not be binary. On the demand-side, to get published, to get research grants, one need not only rely on tests of statistical significance to determine analytical significance. Topping-off a narrative of statistical significance and statistical hypothesis testing by framing this with different confidence intervals and a discourse on sample construction and representativeness as well as a narrative on size effects and variances could open some doors (increase the demand) for higher quality, more robust empirical analysis. But the demand-side door would be closed if one excluded statistical significance completely from the analytical discourse. For many in decision-making positions (those with power), a necessary condition for demand to be realized is some narrative on statistical significance, statistical hypothesis testing.

8 Conclusion

It is important to place the use of statistical significance tests and relatedly statistical hypothesis testing and P-values in their proper analytical context. When this is done, statistical significance can be seen as a minor player in empirical and statistical analysis. It is relatively insignificant. Hence, I am not arguing, as many have in the past, that statistical significance is one of the pillars of any empirical discourse. Moreover, I am arguing that current approaches to tests of statistical significance tests are inappropriate being narrowly focused and decontextualized. For example, confidence intervals are taken as God-given, when they should not be and the variety of possible and valid confidence intervals and their implications for statistical hypothesis testing and P-values should be made explicit. And, these tests are applied even when their applications are meaningless scientifically such as to populations (as opposed to samples) and to any sample (as opposed to random samples). And, this is only the tip of the iceberg. Significance tests, per se, become even more analytically problematic when one acknowledges that simply increasing sample size, even in a manner inconsistent with statistical methodology, yields statistical significance. And, this is a significant problem with the current romanticisation of Big Data as the ultimate solution to problems related to empirical analysis–no thought here about the importance of substantive significance, of size effects.

The main argument in this chapter is that there is much more to robust empirical analysis than simply determining if one's results are statistically significance or not. Even determining the size effect and focusing on the latter can yield highly misleading results unless the size effect is placed in a much broader context. This contextualization also applies to our minor analytical player: tests of statistical significance and its twin, statistical hypothesis testing.

A necessary condition to robust empirical analysis is choosing an appropriate model or models to test (non-statistical hypothesis testing, for example), which serves to direct the choice of independent variables to estimate against the dependent variable and the variables with which to estimate correlation coefficients. Modeless analysis is as scientifically questionable as poorly chosen models driving one's analyses. Appropriate modelling allows one to more robustly discuss and integrate issues of causality, avoid spurious correlation, and omitted variables issues, one of which is a spurious correlation.

Then comes the hard work of making sure that data are well-constructed and samples are representative given the non-statistical hypotheses being tested. Without the latter any tests are scientifically dubious. The data construction piece is a fundamentally important pillar of robust statistical analysis. Then comes the estimation process, running the model or models. The size effect, analytical importance, substantive importance, clinical importance, are what matters most. The narrative on size effect reveals the overall importance of one's results, but if and only if one has executed the prior steps in the analytical process appropriately. Once again, statistical significance is a minor and even an insignificant player in this process.

I also attempt to address the dominance in the use of statistical significance in empirical analytical narratives in spite of the well-know problems and limitations in its use. Key to its dominance is the power vested in those who support the use of statistical hypothesis testing as a key determinant of empirical analysis. Also important is the mental models used by practitioners and those with the power to decide what gets published, what works get commissioned and who receives research grants. But even if there is a supply of properly constructed empirical research there must be the demand for it if a more robust and holistic approach to empirical research is to have an increasing impact on the empirical literature across fields. Which approach dominates has tremendous effect on public and private decision-making with often a significant effect on the population at large. One suggested a way to generate the demand for a more robust, holistic, and scientifically meaningful approach is to retain statistical hypothesis testing whilst engaging in the more holistic approach outlined in this chapter.

One important contribution of this chapter is placing statistical hypothesis testing in a broader context thereby recognizing the fundamental importance of non-statistical theory in the empirical analytical enterprise. The same is true of the importance of robust data construction and the representativeness of the same. It’s not simply about the overall importance of the size effect and the overall insignificance of statistical significance in the larger empirical analytical enterprise. But at the end of the day, if all of the priors are done correctly, what needs to dominate the analytical exercise is a focus on analytical significance and causality even though this more holistic and scientifically robust analytical path is the more difficult and challenging one.

Acknowledgements Thanks to Hannah Altman and Louise Lamontagne for their helpful comments and suggestions.

References

1. Akerlof, G., Kranton, R.: Economics and identity. Q. J. Econ. **115**, 715–753 (2000)
2. Akerlof, G.A., Shiller, R.J.: Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism. Princeton University Press, Princeton (2009)
3. Altman, M.: The methodology of economics and the survivor principle revisited and revised: some welfare and public policy implications of modeling the economic agent. Rev. Soc. Econ. **57**, 427–449 (1999)
4. Altman, M.: Introduction to special issue on statistical significance. J. Socio-Econ. **33**, 523–525 (2004)
5. Altman, M.: Statistical significance, path dependency, and the culture of journal publication. J. Socio-Econ. **33**, 651–663 (2004)
6. Altman, M.: Mental models, bargaining power and institutional change. In: Paper present at the, World Interdisciplinary Network for Institutional Research Conference, Old Royal Naval College, Greenwich, London, UK, 11–14 September 2014
7. Altman, M.: A bounded rationality assessment of the new behavioral economics. In: Frantz, R., Chen, S.-H., Dopfer, K., Heukelom, F., Mousavi, S. (eds.) Routledge Handbook of Behavioral Economics, pp. 179–193. Routledge, London (2017)

8. Altman, M.: A more scientific approach to applied economics: reconstructing statistical, analytical significance, and correlation analysis. Available at SSRN (2018). https://ssrn.com/abstract=3126147 or https://doi.org/10.2139/ssrn.3126147
9. American Statistical Association. American statistical association releases statement on statistical significance and p-values: provides principles to improve the conduct and interpretation of quantitative science (2016). https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
10. Arrow, K.J.: Decision Theory and the Choice of a Level of Significance for the t-Test. In: Olkin, I., et al. (eds.) Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, pp. 70–78. Stanford University Press, Stanford (1960)
11. Baddeley, M.: Herding, social influence and expert opinion. J. Econ. Methodol. **20**, 35–44 (2013)
12. Bailey, D.H., Borwein, J.M., Brent, R.P., Reisi, M.: Reproducibility in computational science: a case study: randomness of the digits of Pi. Exp. Math. (2016). https://doi.org/10.1080/10586458.2016.1163755
13. Becker, G.: Accounting for Tastes. Harvard University Press, Cambridge (1996); Coase, R.: Essays on Economics and Economists. University of Chicago Press, Chicago (1994)
14. Coe, R.: Its the effect size, stupid: what effect size is and why it is important. In: Paper presented at the 2002 Annual Conference of the British Educational Research Association, University of Exeter, Exeter, Devon, England, September 12–14 (2002). http://www.leeds.ac.uk/educol/documents/00002182.htm
15. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Routledge, New York (1977)
16. Deaton, A.: Instruments, randomization, and learning about development. J. Econ. Lit. **48**, 424–455 (2010)
17. Fanelli, D.: How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLOS One **4** (2009). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738
18. Fidler, F., Thomason, N., Cumming, G., Finch, S., Lee, J.: Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol. Sci. **15**, 119–126 (2004)
19. Friedman, M.: The methodology of positive economics. In: Friedman, M. (ed.) Essays in Positive Economics, pp. 3–43. University of Chicago Press, Chicago (1953)
20. Gallo, A.: A refresher on statistical significance. Harvard Business Review (2016). https://hbr.org/2016/02/a-refresher-on-statistical-significance#comment-section
21. Harford, T.: Big data: are we making a big mistake? Financial Times (2014). https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0
22. Keynes, J.M.: The General Theory of Employment, Interest and Money. Macmillan, London (1936)
23. Leibenstein, H.: Allocative efficiency vs. X-efficiency. Am. Econ. Rev. **56**, 392–415 (1966)
24. Munafõ, M.R., Smith, G.D.: Replication is not enough. Nature: Int. J. Sci. **553**, 400–401 (2018). https://www.nature.com/magazine-assets/d41586-018-01023-3/d41586-018-01023-3.pdf
25. McCloskey, D.: The loss function has been mislaid: the rhetoric of significance tests. Am. Econ. Rev. **75**, 201–205 (1985)
26. McCloskey, D.: The insignificance of statistical significance. Sci. Am. **72**, 32–33 (1995)
27. McCloskey, D.N., Ziliak, S.: The standard error of regressions. J. Econ. Lit. **34**, 97–114 (1996)
28. McCrum-Gardner, E.: Sample size and power calculations made simple. Int. J. Ther. Rehabil. **17**, 10–14 (2010)
29. Morrison, D.E., Henkel, R.E.: The Significance Test Controversy: A Reader. Aldine, Chicago (1970)
30. Qualtrics. Calculating sample size (2018). https://www.qualtrics.com/blog/calculating-sample-size/
31. Simon, H.A.: Behavioral Economics. In: Eatwell, J., Millgate, M., Newman, P. (eds.) The New Palgrave: A Dictionary of Economics. Macmillan, London (1987)
32. Taleb, N.N.: The Black Swan: The Impact of the Highly Improbable. Random House, New York (2007)

33. Thompson, B.: Why encouraging effect size reporting is not working: the etiology of researcher resistance to changing practices. J. Psychol. **133**, 133–141 (1999)
34. Wasserstein, R.L., Lazar, N.A.: The ASA's statement on p-values: context, process, and purpose. Am. Stat. **70**, 129–133 (2016). https://doi.org/10.1080/00031305.2016.1154108
35. Ziliak, S., McCloskey, D.N.: Size matters: the standard error of regressions in the American economic. Rev. J. Socio-Econ. **33**, 527–546
36. Ziliak, S., McCloskey, D.N.: The Cult of Statistical Significance: How the Standard Error Costs Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor (2008)

Do Financial Gurus Produce Reliable Forecasts?

David H. Bailey, Jonathan M. Borwein, Amir Salehipour and Marcos López de Prado

1 Introduction and Background

Many investors rely on market experts and forecasters when making investment decisions, in a sense that the investors follow these forecasts when buying or selling securities. Needless to say, some of these forecasts turn out to be more accurate than others. Ranking and grading market forecasters provide investors with metrics on which they may choose forecasters with the best record of accuracy for their particular market exposure.

Some of these forecasts are optimistic, while others are pessimistic. One example of a relatively optimistic forecast was by Thomas Lee, who on January 3, 2015 predicted that the S&P 500 index would be at 2325 1 year hence [6]. (The S&P 500 ranged between 1867 and 2122 during this period, closing at 2012 on January 4,

J. M. Borwein passed away suddenly and unexpectedly on 2 August 2016.

D. H. Bailey (✉)
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
e-mail: david@davidhbailey.com

Department of Computer Science, University of California, Davis, CA, USA

J. M. Borwein
School of Mathematical and Physical Sciences, The Centre for Computer-assisted Research Mathematics and its Applications (CARMA), The University of Newcastle, Callaghan, NSW, Australia

A. Salehipour
School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, NSW, Australia
e-mail: amir.salehipour@uts.edu.au

M. López de Prado
Cornell University, 2 W Loop Rd, New York, NY 10044, USA
e-mail: ml863@cornell.edu

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_17

2016, well short of the goal.) One example of a relatively pessimistic forecast was made by Chapman University professor Terry Burnham, who in July 2013 forecasted that the Dow Jones Industrial Average (DJIA) would drop to 5,000 before it topped 20,000 [1]; he repeated this forecast in May 2014 [2]. (The DJIA exceeded 20,000 on January 25, 2017, having never dropped below 14,700 during the period 1 July 2013 through January 25, 2017.)

There have been several previous analyses of forecaster accuracy, both in academic literature and also in the financial press.

As a single example, recently Nir Kaissar analyzed a set of strategists' predictions from 1999 through November 2016 [3]. He found a relatively high correlation coefficient of 0.76 between the average forecast and the year-end price of the S&P 500 index for the given year. However, Kaissar also found that while the strategists' forecasts were reasonably close most of the time, they were surprisingly unreliable during major inflection points.

For example, Kaissar found that the strategists *overestimated* the S&P 500's year-end price by 26.2 percent on average during the three recession years 2000 through 2002, yet they *underestimated* the index's level by 10.6 percent for the initial recovery year 2003. A similar phenomenon was seen in 2008, when strategists *overestimated* the S&P 500's year-end level by a whopping 64.3 percent in 2008, but then *underestimated* the index by 10.9 percent for the first half of 2009. In other words, as Kaissar lamented, "the forecasts were least useful when they mattered most" [3].

There are numerous challenges to assessing the predictions of forecasters, not the least of which is collecting and assessing these predictions. One promising attempt was in a 2012 study by the CXO Advisory Group of Manassas, Virginia, who ranked 68 forecasters based on their 6,582 forecasts during 1998–2005 for the period of 2005–2012 [4]. Although that study did not provide full details of its grading, ranking, and metric methodology, it acknowledged some weaknesses: (a) the rankings were all weighted equally, or, in other words, all predictions and forecasts were considered equally significant; and (b) the analysis was not adjusted based on the number of forecasts made by a particular forecaster—some experts made only a handful of predictions, while others made many; weighting these the same may lead to distortions when their forecasting records are compared.

In this study, we propose to investigate in greater detail how market experts and forecasters can be graded and ranked, and then to develop and initially deploy an alternative and comprehensive methodology. We build on the experience of others who have collected lists of forecasters, notably the CXO Advisory Group study [4, 5]. Most of these collections are based on the frequency in which the investors or readers have referenced a particular forecaster. In particular, we will seek answers to the following questions:

- How do we recognize and prioritize predictions and forecasts? For instance, we may find different weights for short- and long-term forecasts, or for importance by a given criteria.
- What metrics and measures are most effective and meaningful?

For this study, we will focus on forecasts made for the S&P 500 index, mainly because this is the basis for the similar studies and hence it provides the same basis for comparison purposes. However, the developed methodology is a general one that is applicable to any index for which comprehensive data and forecasts are available.

2 Methodology

Our methodology has two parts. In the first part, every forecast or comment of every market forecaster is evaluated. This is performed by calculating the return of the S&P 500 index over four periods of time. Typically those four periods are 1 month, 3 months, 6 months, and 12 months. Then the correctness of the forecast, i.e., whether the forecaster has made a true or false forecast, is determined in accordance with the time frame for which the forecast is made, considering the correctness of other forecasts that are supposed to occur before or after the forecast. This part is similar to the methodology used in the study by the CXO Advisory team, and for this part, we directly use their evaluation [4, 5].

In the second part, we treat each individual forecast according to two factors: the time frame of the forecast, and its importance/specificity. This is because not all forecasts are equally important. For example, a forecast referring to the next few weeks should be treated differently than the one referring to the next few months; in particular, long-term forecasts should be treated as more significant than the short-term forecasts. After all, in the short-term anything could happen, as a matter of randomness, but in the long-term underlying trends, if any, tend to overcome short-term noise. For these reasons, we give more weight to longer term forecasts, since they imply investing skill with greater confidence. In this regard our study contrasts to the study of CXO Advisory team, which treated every forecast as equally significant.

In this study, we consider four-time frames, which are weighted as follows:

- Up to 1 month: 0.25,
- Up to 3 months: 0.50,
- Up to 9 months: 0.75,
- Beyond 9 months (up to 2 to 3 years): 1.00,
- If the forecast does not include a time frame, or unless there is an impression stating otherwise, we assign a weight of 0.25.

The parameter $w_t \in \{0.25, 0.50, 0.75, 1.00\}$ denotes the weight associated with these time frame.

Regarding the specificity of a forecast, we assign a weight of either 0.5, for a less specific forecast, or 1.0, for a more specific forecast. For example, a forecast that states “the market will be volatile in the next few days” is not a very specific forecast, because the investor may not be able to make a decision solely based on the forecast. However, the forecast “the market will experience a correction” is more specific, and hence, important. In this example, we assign a weight of 0.5 to forecasts of the first sort, and a weight of 1.0 to forecasts of the second sort. Again, in this regard our study contrasts with the earlier study by the CXO Advisory team, which did not introduce or assign specificity weightings. We use $w_s \in \{0.50, 1.00\}$ to denote specificity of a forecast.

Following definition of w_t and w_s, we may derive a weight for a forecast by multiplying those two weights:

$$w_i^+ = w_t \times w_s \quad \text{if forecast } i \text{ is correct,} \tag{1}$$

$$w_i^- = w_t \times w_s \quad \text{if forecast } i \text{ is not correct.} \tag{2}$$

Notice that w_i^+ is the combined weight for forecast i when it is true, and w_i^- is when it is false. Then, accuracy of a forecaster may be obtained by Equation (3).

$$\epsilon_j = \frac{\Sigma_{i=1}^{n_j} w_i^+}{\Sigma_{i=1}^{n_j} w_i^+ + \Sigma_{i=1}^{n_j} w_i^-}, \tag{3}$$

where j is the forecaster's index, and n_j is the total number of forecasts made by forecaster j.

Dataset

In this study, we utilize the same dataset that was previously compiled by CXO. This dataset includes 68 separate spreadsheets, each of which refers to the data of one forecaster. The information for each forecaster consists a set of forecast statements (text), the returns of the S&P 500 index and the correctness of forecast as evaluated by CXO [4, 5].

Algorithm

To apply our ranking methodology to the dataset, we have developed a program in the programming language Python 2.7. The program reads every sheet in the dataset, evaluates the texts (forecast statements) by assigning appropriate weightings, performs the calculations, i.e., Eqs. (1) to (3), and generates two outputs and saves them as two spreadsheet files. The first spreadsheet file has 68 sheets (same as the input dataset), and in addition to the original data includes the detailed outcomes of the analyses, with rankings. The second spreadsheet includes the ranking summary for all forecasters, that is, the ranking of all 68 forecasters.

To ensure an appropriate assignment of weights to every forecast, the program has two sets of keywords. The first set includes four subsets of keywords, each of which is associated with one-time frame. Each subset includes a set of words and time adverbs that represent a specific time frame. For example, the word "soon" is one keyword, which represents a very short-term time frame. The second set of keywords includes words, adjectives, and adverbs that reflect the importance and specificity of the forecasts. The algorithm analyzes every forecast by reading the associated text strings, applies both sets of keywords to find any match, and then assigns weights accordingly. A default weight of 0.25 and/or 0.5 will be assigned to a forecast if there is no matching with respect to the time frame and/or specificity.

Training the Algorithm

It is obvious that the performance of the algorithm heavily depends on those two sets of keywords. For this reason, we consider a set of 14 forecasters (about 20%) as the training dataset. More precisely, we manually analyze and evaluate every forecast in the training set. Then we apply formulas (1) through (3) to calculate the accuracy of the forecasts. Given the accuracy of the forecasters in the training set, we evaluate the performance of our algorithm. To do so, we apply the algorithm to the training dataset and compare the forecasters' accuracy obtained by the algorithm against the one obtained manually. This comparison allows for tuning the algorithm, because we can update the original sets of keywords by adding new keywords that are not already in the sets.

Testing the Algorithm

After tuning the algorithm, we applied it to the remaining 54 forecasters in the dataset, which we call the testing dataset. The results of this stage along with the outcomes of the algorithm on the training dataset (in total analyzing 68 forecasters) may be represented as the evaluation and ranking of market forecasters by our developed methodology. This is discussed in more detail in Section 3.

3 Results

After training our algorithm on the training dataset, we ran it on the entire dataset in order to derive the ranking of each market forecaster. We presented the outcomes and findings in the following sections. Notice that the accuracy of the algorithm over the training dataset has been observed to be 92.16%; in other words, the error of the algorithm on the training dataset is 7.84%.

To calculate the accuracy of the algorithm, we manually derived the accuracy of every forecaster in the training dataset. Then we ran the algorithm, which automatically calculates the accuracy of each forecaster, on the same dataset. Let ϵ_j^* denotes the manually obtained accuracy of forecaster j, and ϵ_j the one obtained by the algorithm. Then, the error of the algorithm in calculating the accuracy of forecaster j is

$$\frac{|\epsilon_j - \epsilon_j^*|}{\epsilon_j^*} \times 100.$$

The algorithm's average error over all forecasters in the training dataset can easily be calculated by averaging all errors in the training dataset.

3.1 Forecaster Accuracy

Plate 1 shows the accuracy of each of the 68 forecasters analyzed by the algorithm. Because not every forecaster has made an equal number of forecasts, the figure shows

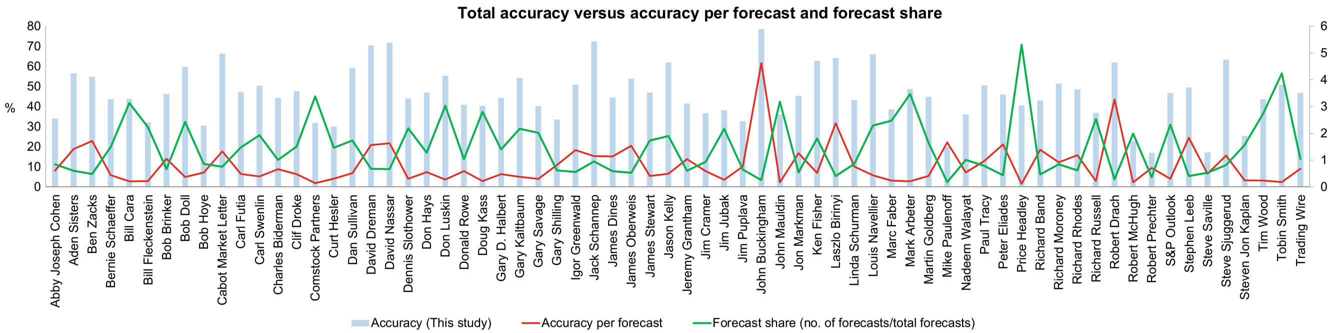


Plate 1 Accuracy of each forecaster (on the left axis) versus accuracy per forecast and forecast share (on the right axis). For forecaster j, the accuracy per forecast is obtained by dividing the accuracy by the number of forecasts (n_j), and forecast share is obtained by dividing the number of forecasts by the total number of forecasts (by all forecasters)

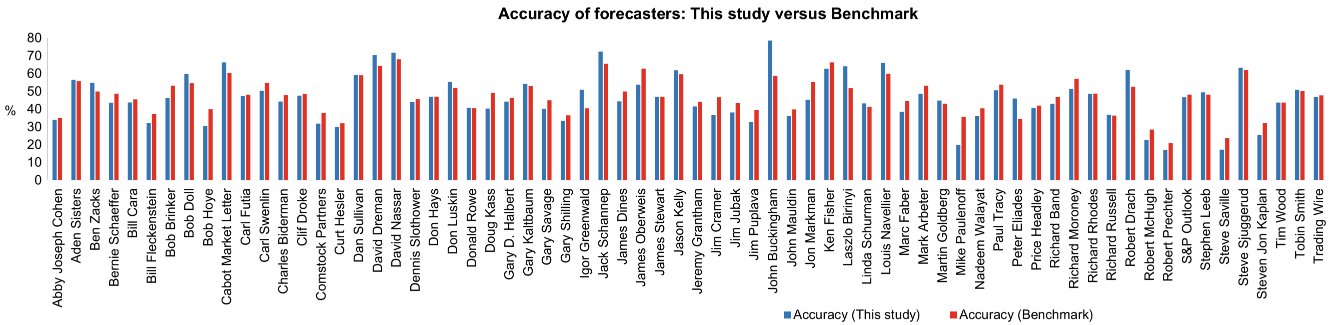


Plate 2 Comparing accuracy of forecasters obtained by our method (this study) against the accuracy obtained by the study of CXO Advisory team (Benchmark). The values of accuracy are in percent

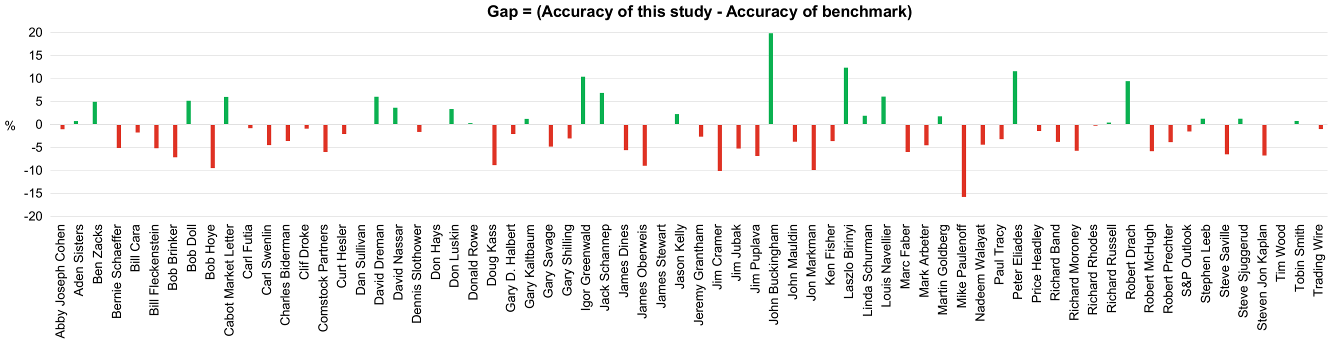


Plate 3 The accuracy gap between our method (this study) and that of the CXO Advisory team (Benchmark) for all forecasters. The accuracy gap Δ_j for forecaster j can be calculated by Equation (6). Positive values of gap reflect higher accuracy, compared with the benchmark, and negative values reflect lower accuracy. As the figure shows, the majority of forecasters have lower accuracy scores

Table 1 Comparison of rankings of the 68 forecasters obtained by our method (this study), and that of the CXO Advisory team (Benchmark). The forecasters were sorted by their values of accuracy (rankings) obtained by our method. The values for accuracy are in % (out of 100), and state the accuracy of every forecaster in predicting the market. For the comparison purposes, we also reported the ranking of the benchmark study. The last column of the table reports the values of accuracy gap (see Equation (6))

Forecaster names	No. of forecasts	Accuracy (This study)	Ranking (This study)	Ranking (Benchmark)	Gap	Forecaster names	No. of forecasts	Accuracy (This study)	Ranking (This study)	Ranking (Benchmark)	Gap
John Buckingham	17	78.69	1	11	19.87	Jon Markman	36	45.37	35	14	−9.89
Jack Schannep	63	72.51	2	3	6.89	Martin Goldberg	109	44.92	36	48	1.80
David Nassar	44	71.84	3	1	3.66	James Dines	39	44.44	37	25	−5.56
David Dreman	45	70.47	4	4	6.03	Charles Biderman	67	44.35	38	34	−3.57
Cabot Market Letter	50	66.39	5	7	6.01	Gary D. Halbert	93	44.32	39	40	−2.07
Louis Navellier	152	66.09	6	8	6.09	Dennis Slothower	145	44.03	40	41	−1.61
Laszlo Birinyi	27	64.21	7	23	12.36	Bill Cara	208	43.84	41	42	−1.74
Steve Sjuggerud	54	63.35	8	6	1.28	Tim Wood	182	43.78	42	46	0.00
Ken Fisher	120	62.80	9	2	−3.59	Bernie Schaeffer	99	43.68	43	29	−5.10
Robert Drach	19	62.07	10	21	9.44	Linda Schurman	57	43.29	44	50	1.91
Jason Kelly	126	61.96	11	9	2.27	Richard Band	31	43.10	45	38	−3.78
Bob Doll	161	59.84	12	16	5.18	Jeremy Grantham	40	41.55	46	45	−2.64
Dan Sullivan	115	59.23	13	10	0.10	Donald Rowe	69	40.89	47	51	0.31
Aden Sisters	40	56.57	14	13	0.76	Price Headley	352	40.65	48	49	−1.40
Don Luskin	201	55.35	15	22	3.39	Doug Kass	186	40.41	49	27	−8.83
Ben Zacks	32	54.95	16	26	4.95	Gary Savage	134	40.24	50	43	−4.79
Gary Kaltbaum	144	54.29	17	20	1.23	Marc Faber	164	38.60	51	44	−5.97

(continued)

Table 1 (continued)

Forecaster names	No. of forecasts	Accuracy (This study)	Ranking (This study)	Ranking (Benchmark)	Gap	Forecaster names	No. of forecasts	Accuracy (This study)	Ranking (This study)	Ranking (Benchmark)	Gap
James Oberweis	35	53.90	18	5	−8.96	Jim Jubak	144	38.22	52	47	−5.20
Richard Moroney	56	51.47	19	12	−5.67	Richard Russell	168	36.91	53	60	0.44
Tobin Smith	281	50.96	20	24	0.78	Jim Cramer	62	36.68	54	39	−10.09
Igor Greenwald	37	50.96	21	52	10.42	John Mauldin	211	36.19	55	55	−3.72
Paul Tracy	52	50.66	22	17	−3.19	Nadeem Walayat	67	36.13	56	53	−4.38
Carl Swenlin	128	50.42	23	15	−4.47	Abby Joseph Cohen	56	34.06	57	62	−1.03
Stephen Leeb	27	49.54	24	31	1.26	Gary Shilling	41	33.56	58	59	−3.03
Mark Arbeter	230	48.75	25	19	−4.50	Jim Puplava	43	32.71	59	56	−6.82
Richard Rhodes	41	48.60	26	28	−0.24	Bill Fleckenstein	148	32.17	60	58	−5.16
Clif Droke	100	47.70	27	30	−0.90	Comstock Partners	224	31.93	61	57	−5.96
Carl Futia	98	47.39	28	33	−0.79	Bob Hoye	57	30.53	62	54	−9.47
Don Hays	85	47.04	29	36	−0.02	Curt Hesler	97	30.02	63	65	−2.06
James Stewart	115	46.99	30	37	0.03	Steven Jon Kaplan	104	25.42	64	64	−6.72
Trading Wire	69	46.85	31	35	−0.98	Robert McHugh	132	22.77	65	66	−5.80
S&P Outlook	154	46.76	32	32	−1.52	Mike Paulenoff	12	20.00	66	61	−15.71
Bob Brinker	44	46.24	33	18	−7.09	Steve Saville	35	17.22	67	67	−6.46
Peter Eliades	29	46.07	34	63	11.59	Robert Prechter	24	17.02	68	68	−3.81

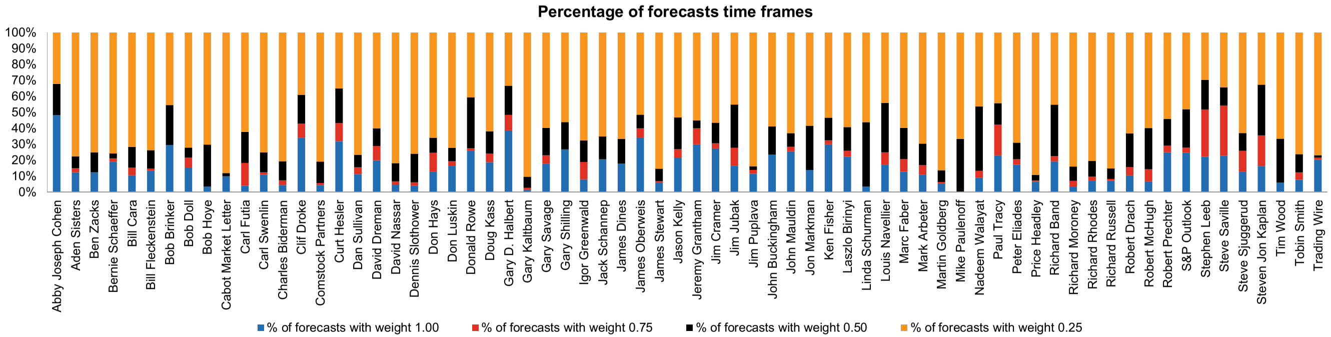


Plate 4 Time frame distribution of forecasts per forecaster. As the graph reveals the majority of forecast statements were made either over a short-term period, i.e., up to a few weeks, or without a specific time frame (associated with a weight of 0.25). Other forecasts were stated covering a long-term period, beyond 9 months (associated with a weight of 1.00). Still other forecasts predicted events over a time period from 3 and 9 months (those associated with a weight of 0.75)

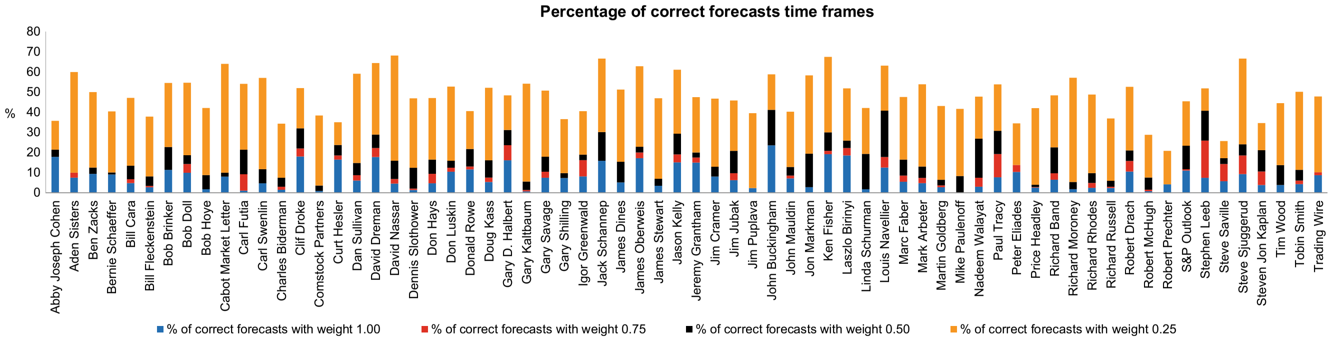


Plate 5 Time frame distribution of correct forecasts per forecaster. A similar behavior to that of Plate 4 is observed here: the majority of correct forecast statements were made over a short-term period (associated with a weight of 0.25) followed by a long-term period (associated with a weight of 1.00)

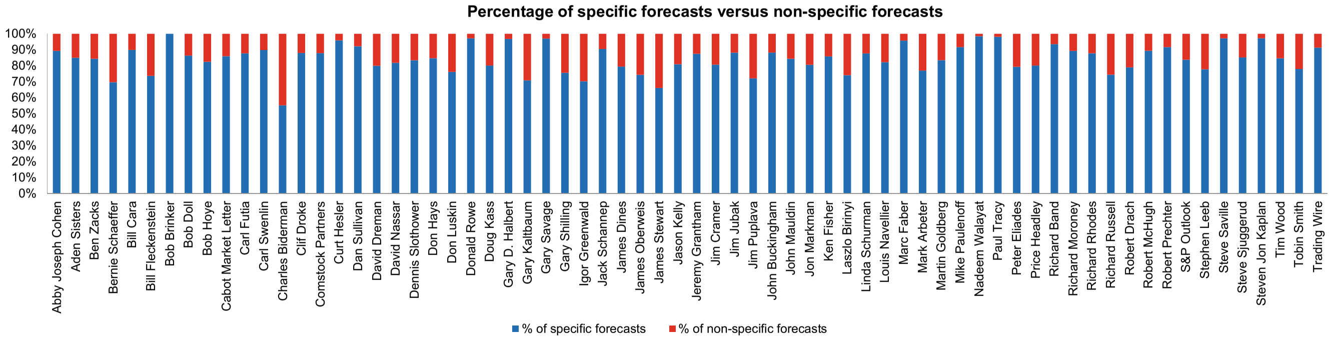


Plate 6 Distribution of specific versus non-specific forecasts per forecaster. The graph reveals that the majority of forecast statements are specific. This observation can almost be concluded for every forecaster

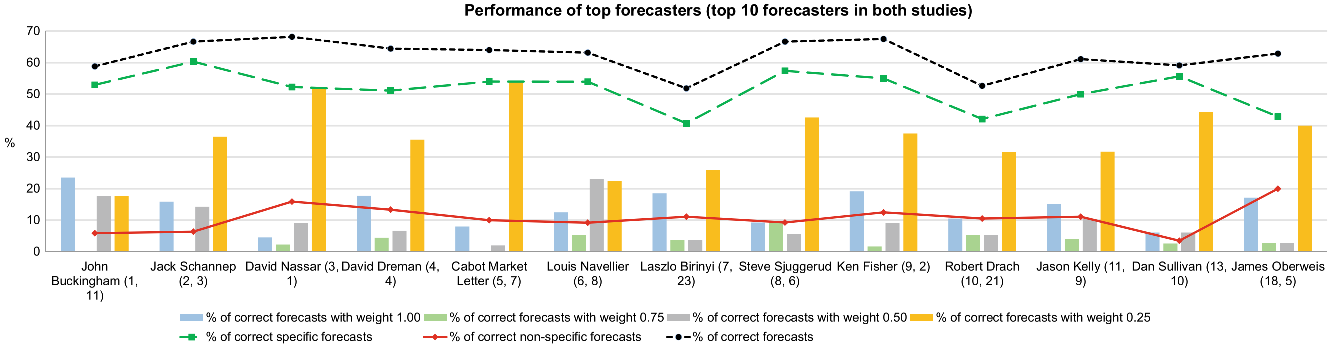


Plate 7 Analyzing performance of top 10 forecasters in each study (in total 13 forecasters were further studied). The graph analyzes the percentage of correct forecasts per time frame, as well as the percentage of correct specific and non-specific forecasts. In addition to those, the percentage of total correct forecasts is plotted. According to the plot, the number of long-term and specific forecasts that were correctly predicted impact accuracy and ranking the most. The numbers inside parenthesis next to each forecaster’s name (on the horizontal axis) state the forecaster rank obtained by this study, and by the benchmark

the accuracy per forecast, and forecast share. For forecaster j, accuracy per forecast is obtained by dividing its accuracy (which is obtained by the algorithm) by its number of forecasts, i.e., n_j. That is

$$e_j = \frac{\epsilon_j}{n_j}. \tag{4}$$

The forecast share of forecaster j, i.e., s_j can be derived by Equation (5).

$$s_j = \frac{n_j}{\Sigma_j n_j} \times 100. \tag{5}$$

Plate 1 analyzes forecasters' performance along their contribution into the forecasting process. The left axis denotes the values of accuracy, and the right axis denotes the values of accuracy per forecast and forecast share. The reader may analyze the statistic e_j (accuracy per forecast) in assessing the performance of forecaster j.

Finally, we compared the accuracy of forecasters obtained by our method against that of published previously in the study of CXO Advisory team (Benchmark). This is graphically depicted in Plate 2.

To have a better grasp of changes in the forecasters accuracy obtained by our method in this study, compared to the earlier study of CXO Advisory team (Benchmark), we define the accuracy gap, which is the difference in values of accuracy between two studies. Let Δ_j denotes the accuracy gap of forecaster j. Equation (6) shows how Δ_j may be derived.

$$\Delta_j = \epsilon_j - \epsilon'_j, \tag{6}$$

where ϵ_j is the value of accuracy for forecaster j, which is obtained by our method, and ϵ'_j is the value of accuracy for forecaster j reported in the study of CXO Advisory team. Gap scores Equation (6) have either positive or negative values. Positive values of gap reflect improvement in the accuracy over the benchmark study, and negative values reflect decreased accuracy. We analyzed the accuracy gap of all forecasters, and illustrated this in Plate 3. Later we report the values of accuracy gap for each forecaster in Table 1. According to the figure, most forecasters have lower accuracy scores with our methodology; in particular, only 36.76% of the forecasters have improved accuracy, and the remaining have lower accuracy. This may be due to the inclusion of additional information of the forecasts' time frames and specificity in our method.

In addition to this, we also analyzed the distribution of forecasters over the accuracy intervals. These were separately calculated for our method (this study) and for the study of CXO Advisory team (Benchmark), and are illustrated in Figure 1. According to the calculated values for accuracy, we considered seven intervals for the values of accuracy, and then calculated the percentage of forecasters that have their accuracy located in an interval. Those seven intervals are as follows:

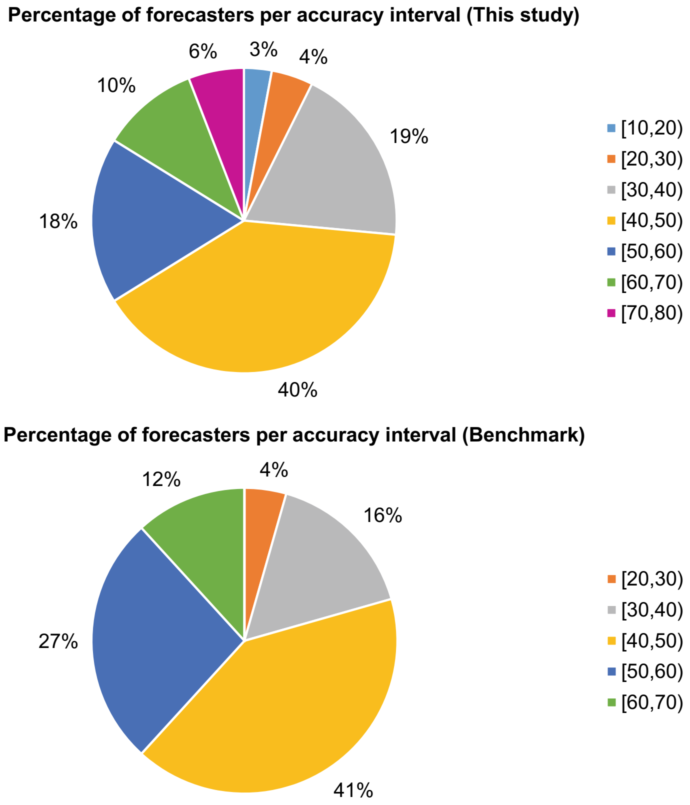


Fig. 1 Analyzing the distribution of forecasters over the accuracy intervals. Seven intervals were considered for the values of accuracy, and then percentage of forecasters in every interval was calculated. The figure on the top shows this distribution for our method (this study); the figure on the bottom shows that for the study of CXO Advisory team (Benchmark). In particular, notice that our method grouped the forecasters into seven intervals, while the benchmark study grouped them into five intervals

- [10, 20),
- [20, 30),
- [30, 40),
- [40, 50),
- [50, 60),
- [60, 70),
- [70, 80).

There are several points of interest in this data. First, in both studies about 40% of the forecasters have an accuracy score between 40% and 50%. Second, our method identifies two new intervals for accuracy values: a low accuracy interval with ranges for accuracy values between 10% and 20%, in which 3% of the forecasters are located, and a high accuracy interval with ranges for accuracy values between 70% and 80%, in which 6% of the forecasters are located. Third, while the percentage of forecasters in the accuracy interval [50%, 60%) has dropped by about 9% (from 27% in the study of CXO Advisory team to 19% in this study), the percentage of the interval [30%, 40%) has increased by 3%. This implies that our method assigns fewer forecasters in the accuracy interval of 50% to 60%, and assigns more forecasters to the interval [30%, 40%).

3.2 Time Frame and Specificity Analysis

Earlier we discussed the importance of time frame and specificity in forecast statements. It is more difficult to forecast the market's long-term behavior than its short-term behavior, and a specific forecast is more valuable than a non-specific one.

Let us start by investigating time frame distribution of a forecaster. Recall that every forecast may be categorized into one of the four-time windows. Hence, for forecaster j, we count the number of forecasts corresponding to each time window and divide this value by the total number of forecasts of forecaster j. This produces up to four percentage values per forecaster, each for one time window. If we continue this for all forecasters, we obtain the graph of Plate 4.

A similar analysis can also be performed for those accurate forecasts, that is those turned out to be "correct" forecasts. This is illustrated in Plate 5, which shows the time frames distribution of a forecaster, and only over correct forecasts. In total, only 48% of all forecasts were correct. In this evaluation, we excluded incorrect forecasts and considered the remaining (both correct or neutral) as correct forecasts.

The time frame distribution of all forecast statements is shown in Figure 2. The graph on the left is over all forecasts, and the graph on the right is over all correct forecasts. Note that the majority of the correct forecasts (around 67.56%) were stated within a short-term period; another 28% of the correct forecasts cover periods between 1 and 3 months, and for more than 9 months. Only less than 5% of the correct forecasts predicted periods from 3 to 9 months.

In addition to the time frame distribution, we analyze specificity of the forecast statements. The majority of the forecasts made by forecasters were fairly specific. This is depicted in Plate 6. Approximately 84% of the forecasts are specific, and only a small percentage (around 16%) are vague and non-specific (see Figure 3). Recall that in this study the major criterion of a forecast specificity is whether the investor can solely make a decision by that forecast.

3.3 Ranking the Forecasters

In this section, we report the ranking of the market forecasters as resulted by implementing our method. This is fully reported in Table 1. The forecasters in Table 1

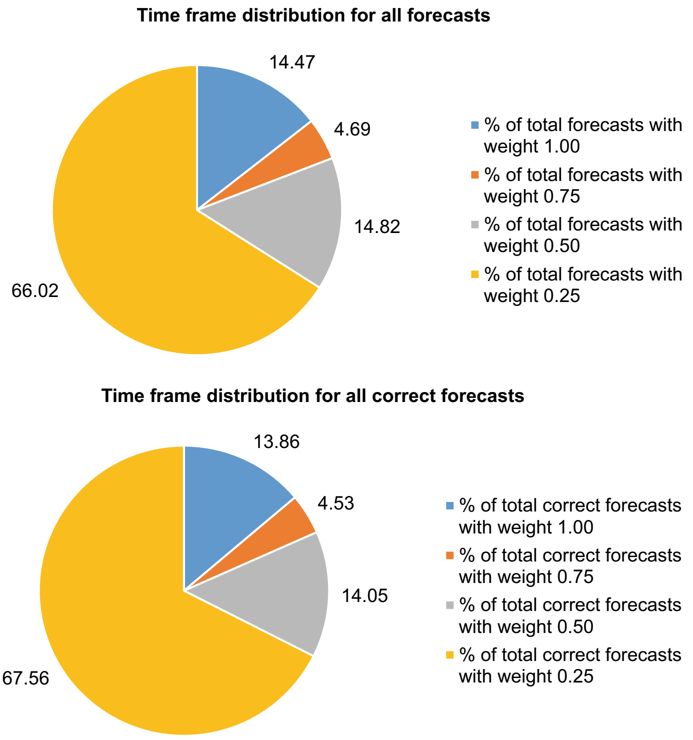


Fig. 2 Distribution of the forecasting time frame over all forecasts (figure on the left) and over all correct forecast statements (figure on the right). As the figures show the majority of forecasts are stated over a short-term time frame

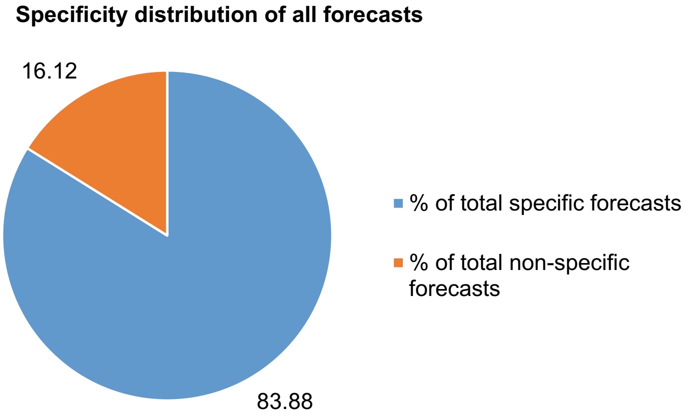


Fig. 3 Distribution of the forecasting specificity over all forecast statements. According to the figure the majority of forecasts are specific enough to assist an investor in making decisions

were ranked on the basis of their accuracy obtained by our method (this study). For comparison purposes, we reported the accuracy of each forecaster as reported in the study of CXO Advisory team (Benchmark). Also, the values of accuracy gap, which were discussed in Equation (6) are reported here. A positive value of accuracy gap means the forecaster's accuracy is improved over the benchmark, and a negative value means the accuracy has decreased.

In checking the top forecasters in each of the two studies, we observe that both share a set of 13 forecasters, so we further analyzed the performance of these 13 forecasters, and shown in Plate 7.

The figure illustrates the percentage of correct forecasts per time frame, and the percentage of correct specific and non-specific forecasts. Also, we included the percentage of total correct forecasts. According to the plot, the number of long-term and specific forecasts that were correctly predicted impact accuracy and ranking the most. For example, "John Buckingham" has a rank of 1 in our study and 11 in the benchmark study, and "David Nassar" has a rank of 3 in our study and 1 in the benchmark study. However, the majority of David's correct forecasts cover periods less than 1 month, whereas John's correct forecasts mainly cover long-term and middle-term periods. Moreover, John has more correct specific and less correct non-specific forecasts.

On the other hand, if we only consider the number of correct forecast statements in order to evaluate forecasters' performance, David's accuracy would be approximately 70%, while John's would be approximately 60%, thus ranking David before John.

4 Conclusion

Market forecasts are widely read in the investment community. Some of these forecasts turn out to be uncannily accurate, while others lead to significant losses. To better understand the extent to which various forecasters have forecasting skill, we have developed a ranking methodology to rank and grade market forecasters. This study builds upon a previous study by the CXO Advisory Group in several directions. In particular, we distinguish forecasts by their specificity, rather than considering all predictions and forecasts equally important, and we also analyze the impact of the number of forecasts made by a particular forecaster. Our results show that some forecasters have done very well, even more so than reflected in earlier studies, but the majority perform at levels not significantly different than chance.

Acknowledgements Our co-author Jonathan M. Borwein sadly passed away in August 2016, while the team was working on this research project. Jon's love of life, pursuit of quality, and extensive knowledge on a wide range of topics have inspired us to extend our reach.

Amir Salehipour is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number DE170100234) funded by the Australian Government. Also, we would like to thank the CXO Advisory team, in particular Steve LeCompte, for his support on providing us with the dataset, and for several discussions during the course of this research.

References

1. Burnham, T.: Ben Bernanke as Easter bunny: why the Fed can't prevent the coming crash. In: PBS NewsHour. http://www.pbs.org/newshour/making-sense/ben-bernanke-as-easter-bunny-why-the-fed-cant-prevent-the-coming-crash/ (2013). Accessed 11 July 2013
2. Burnham, T.: Why one economist isn't running with the bulls: dow 5,000 remains closer than you think. In: PBS NewsHour. http://www.pbs.org/newshour/making-sense/one-economist-isnt-running-bulls-dow-5000-remains-closer-think/ (2014). Accessed 21 May 2014
3. Kaissar, N.: S&P 500 forecasts: crystal ball or magic 8?. In: Bloomberg News. https://www.bloomberg.com/gadfly/articles/2016-12-23/s-p-500-forecasts-mostly-hit-mark-until-they-matter-most (2016). Accessed 23 Dec 2016
4. LeCompte, S. (ed).: Guru grades. In: CXO Advisory Group. http://www.cxoadvisory.com/gurus/ (2013)
5. LeCompte, S. (ed).: The most intriguing gurus?. In: CXO Advisory Group. http://www.cxoadvisory.com/4025/investing-expertise/the-most-intriguing-gurus/ (2009)
6. Udland, M.: Here's what 13 top wall street pros are predicting for stocks in 2015. In: Business Insider. http://www.businessinsider.com/wall-street-2015-sp-500-forecasts-2015-1 (2015). Accessed 3 Jan 2015

Entropy Maximization in Finance

Jonathan M. Borwein and Qiji J. Zhu

1 Introduction

The principle of maximum entropy appeared in statistical mechanics due to the work of Boltzmann [1] and Gibbs [12]. Statistical mechanics considers the aggregate behavior of large physical systems of microscopic elements. This aggregate behavior is the observation of the "moments" of a probability distribution of those microscopic elements. Knowing finite number of observations there could be many different probability distributions that are consistent with these observations. The principle of maximum entropy suggests to select the probability distribution that maximizes an entropy. Jaynes' work [16, 17] relates this principle to Shannon's information theory [27]. He points out that in essence the maximum entropy methods select the most uninformative distribution possible if one choose to use the Boltzmann–Shannon entropy.

The structures of such entropy maximization problems were explored in solving other application problems often with the Boltzmann–Shannon entropy replaced by other concave functions. This approach is referred to as entropy maximization method which has wide applications in diverse fields. We show in this paper that

This project started from Jonathan Borwein's comment that the "two fund" theorem in Markowitz portfolio theory can be viewed as exploring the structure of an entropy maximization problem and similar phenomena should exist in other financial problems. Sadly, he could not see its completion.

Passed away suddenly and unexpectedly on 2 August 2016. Was Laureate Professor and Director.

J. M. Borwein
School of Mathematical and Physical Sciences, The Centre for Computer-assisted Research Mathematics and its Applications (CARMA), The University of Newcastle, Callaghan, NSW, Australia

Q. J. Zhu (✉)
Department of Mathematics, Western Michigan University, Kalamazoo, MI 49009, USA
e-mail: zhu@wmich.edu

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_18

several important results in financial theory can be derived by using the entropy maximization method. They are the Markowitz portfolio theory and two fund theorem, the capital market pricing model, the fundamental theorem of asset pricing, selecting a pricing equivalent martingale measure using the entropy maximization method and determining the super/sub-hedging bounds and portfolios. The structures of the solutions to the entropy maximization problems often play a crucial role in understanding these applications.

It is not a coincidence that many financial problems can be formulated as generalized entropy maximization problems. It has been a long tradition in financial economy to model the risk aversion of a market participant using a concave utility function and assuming a rational market participant attempts to maximize his/her utility. Such a maximization problem in practice must subject to various constraints related to budget or risk control. In a simple one price economy, these constraints are often linear making the resulting problem fits the pattern of an entropy maximization problem. Another modeling principle is that agents in financial market try to minimize their risk. Since diversification reduces risk, risk measures are usually convex. Thus, minimizing risk subject to various constraints also leads to generalized entropy maximization problems where the negative of the risk measure takes the role of a generalized entropy. Entropy maximization method is a special case of the more general convex duality theory (see, e.g., [4, 5]). Indeed convex duality and general convex duality theory have wider applications in finance (see, e.g., [6]). Nevertheless, when entropy maximization method is applicable, the structure of the entropy maximization problem and its solutions provides additional information to the financial applications.

Many important results in finance can be handled using a uniform framework of entropy maximization problems is a powerful testimony to the significant impact of physical science in financial research. This is a double-edged sword. On one hand relating financial and physical models opens the door for systematically applying physical and mathematical principles and methods in financial research. This is especially beneficial in introducing effective quantitative methods into financial practice. Moreover, the relationship of entropy maximization and information theory is also highly relevant in financial problems. For example, maximizing the utility of a portfolio can be interpreted as best utilize the information contained in the market model. On the other hand, we need to recognize that in some aspects financial markets are significantly different from a physical system. That means when we use theoretical results in finance, in particular, those related to the entropy maximization, caution is warranted.

The rest of the paper is arranged as follows: we lay out preliminaries regarding the entropy maximization method and a simple one-period financial market model in the next section. Then we discuss four financial applications of the entropy maximization methods alluded to above in Sections 3–7. We conclude in Section 8.

2 Preliminaries

2.1 Entropy Optimization Problem

The mathematical formulation of an entropy maximization problem is

$$\inf_x [f(x) : Ax = b]. \tag{1}$$

Here f is a lower semicontinuous convex function representing the negative of some generalized entropy function on a Banach space X and $Ax = b$ is a linear constraint with b in a finite-dimensional space representing the finite number of observations on "moments".

Recall that, for a lower semicontinuous convex function f on X, the Fenchel conjugate of f is defined by

$$f^*(y) := \sup_x [\langle y, x\rangle - f(x)]$$

and the subdifferential of f at x is defined by

$$\partial f(x) := \{x^* \in X^* \mid f(y) - f(x) \geq \langle x^*, y - x\rangle \ \forall y \in X\}.$$

Below is a concise summary of important results on duality of entropy maximization problem emphasizing the link between dual solutions and Lagrange multipliers for the primal problem (see [3–5] for details). These results are special cases of the classical convex duality theory developed by Fenchel [11], Moreau [22] and Rockafellar [24].

If constraint qualification condition (CQ)

$$b \in \operatorname{ri} A \operatorname{dom} f \tag{2}$$

holds, where ri signifies the relative interior and dom $f := \{x : f(x) < \infty\}$ is the domain of f, then we have strong duality

$$\inf_x [f(x) : Ax = b] = \max_z [\langle z, b\rangle - f^*(A^\top z)] = (f^* \circ A^\top)^*(b). \tag{3}$$

Moreover, if $\bar{x}$ and $\bar{z}$ are solutions to the primal and dual problems, respectively, then

$$\bar{x} \in \partial f^*(A^\top \bar{z}) \tag{4}$$

$$A^\top \bar{z} \in \partial f(\bar{x}) \tag{5}$$

and

$$b \in \partial(f^* \circ A^\top)(\bar{z}). \tag{6}$$

Note that the constraint qualification condition implies the existence of a dual solution $\bar{z}$ which is the Lagrange multiplier for the primal problem. In other words, if a primal solution $\bar{x}$ exists then the Lagrangian for the primal problem

$$L(x, \bar{z}) := f(x) + \langle \bar{z}, b - Ax \rangle \tag{7}$$

as a function of x attains a minimum at $x = \bar{x}$. However, the existence of a primal solution is not always guaranteed and usually needs additional verification.

2.2 A Portfolio Model

To be concise we only deal with a simple one-period financial market model on an economy with finite status to highlight the role of the entropy maximization methods. Many of the results discussed here also extend to more general models. We refer to books [6, 26] for details of some of the generalizations and alternative approaches.

Let $S_t = (S_t^0, S_t^1, \ldots, S_t^M)$, $t = 0, 1$ be a financial market in a one period economy. Here $S_0^0 = 1$, $S_1^0 = R > 1$ represents a risk free bond and $S_t^m, m = 1, \ldots, M$ represents the price of the mth risky financial asset at time t. We assume that S_0 is a constant vector representing the prices of the assets in this financial market at $t = 0$. The risk is modeled by assuming $\hat{S}_1 = (S_1^1, \ldots, S_1^M)$ to be a random vector on a probability space $(\Omega, \mathcal{F}, P)$. A portfolio is a vector $x \in \mathbf{R}^{M+1}$ whose component x_m represents the share of the mth asset in the portfolio. Then $x \cdot S_1$ is the payoff and $x \cdot (S_1 - S_0)$ is the gain of the portfolio x both belong to $RV(\Omega, \mathcal{F}, P)$, the space of random variables on the probability space $(\Omega, \mathcal{F}, P)$. We will also use the notation $\hat{x} = (x_1, \ldots, x_M)^\top$ to denote the risky part of the portfolio.

Clearly, given a financial market S, different portfolios may correspond to the same gain. We call such portfolios equivalent. We denote $port[S]$ the space of equivalent class of portfolios, i.e., the quotient space of $\mathbf{R}^{M+1}$ with respect to the portfolio equivalent relationship. To avoid technical complications we assume in the sequel that the sample space Ω is finite. Then it is not hard to check that the minimum norm of the portfolios in each equivalent class is a norm $\|\cdot\|_p$ for $port[S]$ and $(port[S], \|\cdot\|_p)$ is a finite-dimensional Banach space. The norm $\|\cdot\|_p$ is a reasonable indication of the leverage level of a portfolio.

Having setup the model for a financial market we now turn to several important financial results that can be understood in a unified framework of entropy maximization in the next several sections.

3 Markowitz Portfolio Theory

Markowitz [21] considers a portfolio theory that involves only the risky assets. He postulates that the investors will want to minimize the risk given a fixed expected return and will attempt to maximize the expected return given a fixed risk. Markowitz uses the standard deviation to measure the risk of a portfolio. Standardizing the initial endowment to 1 and denote the expected return of a portfolio by μ, we can represent the Markowitz portfolio problem as an entropy maximization problem. Define

$$f(\hat{x}) = \frac{1}{2}Var(\hat{x} \cdot \hat{S}_1) = \frac{1}{2}\hat{x}^\top \Sigma \hat{x}, \tag{8}$$

where the covariant matrix

$$\begin{aligned} \Sigma &= \mathbf{E}[(\hat{S}_1 - \mathbf{E}(\hat{S}_1))^\top (\hat{S}_1 - \mathbf{E}(\hat{S}_1))] \\ &= (\mathbf{E}[(S_1^i - \mathbf{E}(S_1^i))(S_1^j - \mathbf{E}(S_1^j))])_{i,j=1,\ldots,M}, \end{aligned} \tag{9}$$

is assumed to be positive definite. Denote

$$\hat{A} = \begin{bmatrix} \mathbf{E}(\hat{S}_1) \\ \hat{S}_0 \end{bmatrix} \text{ and } b = \begin{bmatrix} \mu \\ 1 \end{bmatrix}. \tag{10}$$

Then we can write the Markowitz portfolio problem as

$$\min[f(\hat{x}) : \hat{A}\hat{x} = b]. \tag{11}$$

This is because minimizing f and minimizing the standard deviation σ of the payoff of the portfolio are equivalent.

Remark 1 Since $\sqrt{\hat{x}^\top \Sigma \hat{x}}$ can be viewed as an equivalent norm on the space of random vectors on probability space $(\Omega, \mathcal{F}, P)$ we can directly deal with all portfolios in $\mathbf{R}^M$ rather than the quotient space $port[S]$.

We can calculate that

$$f^*(y) = \frac{1}{2}\hat{y}\Sigma^{-1}\hat{y}. \tag{12}$$

It follows that letting $z = (z_1, z_2)$ we have

$$\begin{aligned} f^* \circ \hat{A}^\top(z) &= \frac{1}{2}z^\top \hat{A}\Sigma^{-1}\hat{A}^\top z \\ &= \frac{1}{2}z^\top \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} z \end{aligned} \tag{13}$$

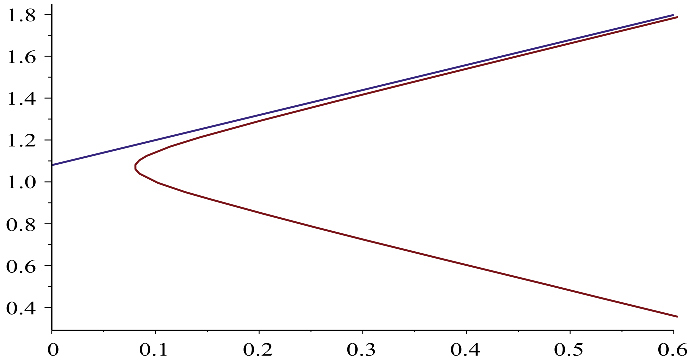


Fig. 1 Markowitz bullet

where $\alpha = \mathbf{E}(\hat{S}_1)\Sigma^{-1}\mathbf{E}(\hat{S}_1)^\top$, $\beta = \mathbf{E}(\hat{S}_1)\Sigma^{-1}\hat{S}_0^\top$ and $\gamma = \hat{S}_0\Sigma^{-1}\hat{S}_0^\top$. It is easy to calculate that

$$\begin{aligned}(f^* \circ A^\top)^*(b) &= \max_z \left\{ z^\top \begin{bmatrix} \mu \\ 1 \end{bmatrix} - \frac{1}{2} z^\top \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} z \right\} \\ &= \frac{1}{2}[\mu, 1] \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}^{-1} \begin{bmatrix} \mu \\ 1 \end{bmatrix} \\ &= \frac{1}{2} \frac{\gamma\mu^2 - 2\beta\mu + \alpha}{\alpha\gamma - \beta^2} = \frac{1}{2}\sigma^2. \end{aligned} \tag{14}$$

Markowitz represent each portfolio as a point in the (σ, μ)-plane. Thus, the optimal portfolio will be located on the curve

$$\sigma = \sqrt{\frac{\gamma\mu^2 - 2\beta\mu + \alpha}{\alpha\gamma - \beta^2}} \tag{15}$$

usually referred to as the Markowitz bullet due to its shape. A typical Markowitz bullet is shown in Fig. 1 with an asymptote

$$\mu = \frac{\beta}{\gamma} + \sigma\sqrt{\frac{\alpha\gamma - \beta^2}{\gamma}}. \tag{16}$$

In summary, we have

Theorem 1 (Markowitz Portfolio Theorem) *The effect of each portfolio $\hat{x}$ can be represented as a point in the (σ, μ)-plane. Portfolios represent optimal tradeoff between return and risk are located on the upper boundary of the Markowitz bullet given by*

$$\sigma = \sqrt{\frac{\gamma\mu^2 - 2\beta\mu + \alpha}{\alpha\gamma - \beta^2}}.$$

Remark 2 Markowitz portfolio problem (11) is defined on the portfolio space. The dimension of a portfolio space equals to the number of risky assets involved in the portfolio which can be quite large. For example considering the well-known benchmark SP500 index. This is a portfolio involving 500 stocks. That means considering Markowitz portfolio problem in a comparable universe of risky asset one has to deal with an entropy maximization problem in a 500 dimensional space. However, the dual problem is on a two dimensional space related to the two constraints on the expected return and the initial endowment. After standardizing the initial endowment we left with only one variable: the expected return μ. Thus, the performance of each portfolio can be intuitively represented by a point on the (σ, μ)-plane. In short, the key to the success of the Markowitz portfolio theory is to focus on the simpler dual problem (14) rather than the primal problem (11).

We now turn to discuss optimal portfolios on this Markowitz bullet. Let $\hat{\bar{x}}$ and $\bar{z}$ be the solutions to the primal and dual problems, respectively. Then it follows from (4) and (6) that

$$\hat{\bar{x}} = \Sigma^{-1}\hat{A}^\top \bar{z} \tag{17}$$

and

$$\begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} \bar{z} = b = \begin{bmatrix} \mu \\ 1 \end{bmatrix}. \tag{18}$$

Thus,

$$\begin{aligned} \hat{\bar{x}} &= \Sigma^{-1}\hat{A}^\top \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}^{-1} \begin{bmatrix} \mu \\ 1 \end{bmatrix} \\ &= \mu \frac{\Sigma^{-1}\hat{A}^\top}{\alpha\gamma - \beta^2} \begin{bmatrix} \gamma \\ -\beta \end{bmatrix} + \frac{\Sigma^{-1}\hat{A}^\top}{\alpha\gamma - \beta^2} \begin{bmatrix} -\beta \\ \alpha \end{bmatrix} \end{aligned} \tag{19}$$

is affine in μ. The structure of the optimal portfolio in (19) tells us that knowing two optimal portfolios one can generate any of the portfolios on the Markowitz bullet as their linear combination. This result is known as the two fund theorem.

Theorem 2 (Two Fund Theorem) *Select two distinct portfolios on the Markowitz efficient frontier. Then any portfolio on the Markowitz efficient frontier can be represented as the linear combination of these two portfolios.*

Proof Let

$$\hat{x}_i = \Sigma^{-1}\hat{A}^\top \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}^{-1} \begin{bmatrix} \mu_i \\ 1 \end{bmatrix}, i = 1, 2 \tag{20}$$

be two chosen portfolios on the Markowitz frontier. Suppose $\hat{x}$ is a portfolio on the Markowitz frontier. Then, for some μ,

$$\hat{x} = \Sigma^{-1}\hat{A}^{\top}\begin{bmatrix}\alpha & \beta \\ \beta & \gamma\end{bmatrix}^{-1}\begin{bmatrix}\mu \\ 1\end{bmatrix}. \tag{21}$$

Defining

$$\begin{bmatrix}k_1 \\ k_2\end{bmatrix} = \begin{bmatrix}\mu_1 & \mu_2 \\ 1 & 1\end{bmatrix}^{-1}\begin{bmatrix}\mu \\ 1\end{bmatrix}. \tag{22}$$

we have

$$\begin{bmatrix}\mu \\ 1\end{bmatrix} = k_1\begin{bmatrix}\mu_1 \\ 1\end{bmatrix} + k_2\begin{bmatrix}\mu_2 \\ 1\end{bmatrix}, \tag{23}$$

so that

$$\hat{x} = k_1\hat{x_1} + k_2\hat{x_2}.$$

□

Remark 3 The two fund theorem explores the fact that Markowitz optimal portfolio as a function of the return μ is affine. This is a structure of the solution of the entropy maximization problem when we have a quadratic function as the negative of the generalized entropy. In pointing out that all efficient Markowitz portfolios are generated by just two basic efficient portfolios, the two fund theorem greatly simplifies that task of determining Markowitz portfolios. In practice, one can often use two broad-based indices to approximate the two basic efficient portfolios. This can be viewed as the theoretical foundation for the passive investment strategy of buy and hold broad-based indices.

4 Capital Asset Pricing Model

Capital asset pricing model (CAPM) is a theoretical model independently proposed by Lintner [20], Mosssin [23], Sharpe [28] and Treynor [32] for pricing a risky asset according to its expected payoff and market risk, often referred to as the beta. Mathematically the core of the capital asset pricing model can be viewed as an extension of the analysis of the Markowitz portfolio theory to include a riskless bond. Thus the model is actually simpler: the function f defined below are similar to that we used in the Markowitz portfolio theory:

$$f(x) = \frac{1}{2}Var(x \cdot S_1) = \frac{1}{2}x^{\top}\begin{bmatrix}0 & 0 \\ 0 & \Sigma\end{bmatrix}x, \tag{24}$$

and

$$A = \begin{bmatrix} \mathbf{E}(S_1) \\ S_0 \end{bmatrix}, \text{ and } b = \begin{bmatrix} \mu \\ 1 \end{bmatrix}. \tag{25}$$

As discussed in Remark 1 we directly consider portfolio $x \in \mathbf{R}^{M+1}$ and the following entropy maximization problem

$$\inf[f(x) \mid Ax = b]. \tag{26}$$

Direct calculation yields

$$f^*(y) = \begin{cases} +\infty & y_0 \neq 0 \\ \frac{1}{2}\hat{y}\Sigma^{-1}\hat{y} & \text{otherwise.} \end{cases} \tag{27}$$

Using the duality relationship in (3) we can calculate the value of the entropy maximization problem (26) to be

$$\begin{aligned} f^* \circ A^\top(z) &= \begin{cases} +\infty & z_1 R + z_2 \neq 0 \\ \frac{1}{2} z^\top \hat{A}\Sigma^{-1}\hat{A}^\top z & [R, 1]z = 0 \end{cases} \\ &= \begin{cases} +\infty & z_1 R + z_2 \neq 0 \\ \frac{1}{2}(\alpha - 2\beta R + \gamma R^2) z_1^2 & z_2 = -z_1 R \end{cases}. \end{aligned} \tag{28}$$

Since Σ is positive definite, $z^\top \hat{A}\Sigma^{-1}\hat{A}^\top z > 0$ whenever $z \neq 0$. Thus $\Delta := \alpha - 2\beta R + \gamma R^2 > 0$. We can calculate that

$$\begin{aligned} (f^* \circ A^\top)^*(b) &= \max_z \left\{ z^\top \begin{bmatrix} \mu \\ 1 \end{bmatrix} - (f^* \circ A^\top)(z) \right\} \\ &= \max_{z_1} \left\{ z_1(\mu - R) - \frac{1}{2} z_1^2 \Delta \right\} = \frac{(\mu - R)^2}{2\Delta} \end{aligned} \tag{29}$$

We know it only make sense to involve risky assets when we can expect an excess return, that is, return μ should be higher than the riskless return R. Placing the optimal portfolio again in the (σ, μ)-plane we see that they are all on the straight line

$$\sigma = \frac{\mu - R}{\sqrt{\Delta}} \text{ or } \mu = R + \sigma\sqrt{\Delta}. \tag{30}$$

Again we see the affine structure of the solution. Thus, all the optimal solution, represented in the (σ, μ)-plane, should be the convex combination of two basic optimal solutions. This is rather similar to the two fund theorem in the previous section. A convenient choice for the two basic optimal solutions are taking one portfolio that contains only the riskless bond and another portfolio with only risk asset. Clearly, the portfolio that contains only risk assets has to reside on the Markowitz efficient frontier. We call this portfolio the *market portfolio*. Summarizing we get what people often referred to as the two fund separation theorem.

Theorem 3 (Two Fund Separation Theorem) *All the optimal portfolios in the CAPM model are convex combinations of the riskless bond and the market portfolio.*

Now we turn to the issue of calculating the optimal portfolio. Denoting the solutions to the primal and dual problems by $\bar{x}$ and $\bar{z}$, respectively, we have

$$A^\top \bar{z} = f'(\bar{x}) = \begin{bmatrix} 0 \\ \Sigma \hat{\bar{x}} \end{bmatrix} \tag{31}$$

or

$$\hat{\bar{x}} = \Sigma^{-1} \hat{A}^\top \bar{z} \tag{32}$$

and

$$\bar{z}_1 R + \bar{z}_2 = 0 \tag{33}$$

It follows from (29) that $\bar{z}_1 = (\mu - R)/\Delta$ so that by (33) we have

$$\bar{z} = \bar{z}_1 \begin{bmatrix} 1 \\ -R \end{bmatrix} = \frac{\mu - R}{\Delta} \begin{bmatrix} 1 \\ -R \end{bmatrix}. \tag{34}$$

Combining (32) and (34) we have a clean representation of the risky part of the optimal portfolio

$$\hat{\bar{x}} = \frac{\mu - R}{\Delta} \Sigma^{-1} \hat{A}^\top \begin{bmatrix} 1 \\ -R \end{bmatrix}. \tag{35}$$

We can calculate the capital allocated to the risky part of the portfolio to be

$$1 - \bar{x}_0 = \hat{S}_0 \cdot \hat{\bar{x}} = \frac{\beta - \gamma R}{\Delta} (\mu - R). \tag{36}$$

From (36) we see that to get an excess return $\mu > R$, we need to long risky assets when $R < \beta/\gamma$ and short risky assets when $R > \beta/\gamma$. When R is exactly β/γ, no portfolio can achieve excess return and there is no benefit involving risky assets in the portfolio.

Next we focus on the case when $R < \beta/\gamma$. We observe that when the right hand side of (36) is 1 we have a portfolio that is entirely consisting of risky assets. The corresponding optimal portfolio is the market portfolio

$$\bar{x}_M = \left(0, \frac{1}{\beta - \gamma R} \Sigma^{-1} \hat{A}^\top \begin{bmatrix} 1 \\ -R \end{bmatrix} \right). \tag{37}$$

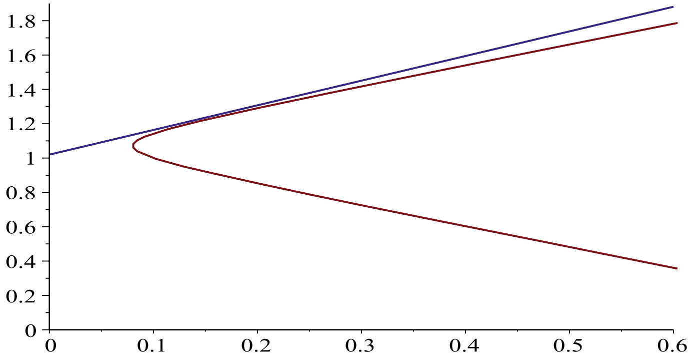


Fig. 2 Capital market line

The corresponding return μ_M and the standard deviation σ_M are given below

$$\mu_M = R + \frac{\Delta}{\beta - \gamma R} \tag{38}$$

$$\sigma_M = \frac{\sqrt{\Delta}}{\beta - \gamma R}. \tag{39}$$

We observe that the market portfolio is independent in μ. Moreover, as alluded to in the two fund separation theorem all the optimal portfolios are a combination of the market portfolio and the riskless bond. On the (σ, μ)-plane they are all located on the line (see Fig. 2)

$$\mu = R + \sqrt{\Delta}\sigma. \tag{40}$$

We call this line the *capital market line*.

We can summarize the above as:

Theorem 4 (Capital Market Line) *Optimal portfolios represented as points in the* (σ, μ)*-plane are all located on the* capital market line

$$\mu = R + \sigma\sqrt{\Delta},$$

where $\Delta = \alpha - 2\beta R + \gamma R^2$*. The capital market line is tangent to the boundary of the Markowitz bullet at*

$$(\sigma_M, \mu_M) = \left(\frac{\sqrt{\Delta}}{\beta - \gamma R}, R + \frac{\Delta}{\beta - \gamma R}\right)$$

and intercept the μ axis at $(0, R)$. *The portfolio corresponding to* (σ_M, μ_M) *is*

$$\bar{x}_M = \left(0, \frac{1}{\beta - \gamma R} \Sigma^{-1} \hat{A}^\top \begin{bmatrix} 1 \\ -R \end{bmatrix}\right).$$

and is called the capital market portfolio.

Alternatively we can write the slope of the capital market line as

$$\sqrt{\Delta} = \frac{\mu_M - R}{\sigma_M}. \tag{41}$$

This quantity is called the *price of risk* and we can rewrite the equation for the capital market line as

$$\mu = R + \frac{\mu_M - R}{\sigma_M}\sigma. \tag{42}$$

Next we discuss how to use the capital market line to price a risky asset. The capital asset pricing model assumes that adding a fair priced risky asset to the market should not change the capital market line. The price is indirectly reflected in the expected return of the asset. Thus, given a risky asset a^i, we try to determine the its expected return μ_i.

Theorem 5 (Capital Asset Pricing Model) *Suppose that we know a financial market* S *with a riskless bond returning* R. *Let* a^i *be a fair priced risky asset with expected percentage return* μ_i. *Then*

$$\mu_i = R + \beta_i(\mu_M - R). \tag{43}$$

Here $\beta_i = \sigma_{iM}/\sigma_M^2$ *is called the beta of* a^i, *where* $\sigma_{iM} = cov(a^i, \bar{x}_M \cdot S)$ *is the covariance of* a^i *and the market portfolio.*

Proof Consider a portfolio relies on the parameter α that consists the risky asset a^i and the market portfolio:

$$p(\alpha) = \alpha a^i + (1 - \alpha)\bar{x}_M \cdot S. \tag{44}$$

Denote the expected return and the standard variation of $p(\alpha)$ by μ_α and σ_α, respectively, we have

$$\mu_\alpha = \alpha\mu_i + (1 - \alpha)\mu_M, \tag{45}$$

and

$$\sigma_\alpha^2 = \alpha^2\sigma_i^2 + 2\alpha(1 - \alpha)\sigma_{iM} + (1 - \alpha)^2\mu_M^2. \tag{46}$$

The parametric curve $(\sigma_\alpha, \mu_\alpha)$ must lie below the capital market line because the latter consists of optimal portfolios. On the other hand it is clear that when $\alpha = 0$ this curve coincide with the capital market line. Thus, the capital market line is an tangent line of the parametric curve $(\sigma_\alpha, \mu_\alpha)$ at $\alpha = 0$. It follows that

$$\frac{\mu_M - R}{\sigma_M} = \left[\frac{d\mu_\alpha}{d\sigma_\alpha}\right]_{\alpha=0} = \frac{\sigma_M(\mu_i - \mu_M)}{\sigma_{iM} - \sigma_M^2}. \tag{47}$$

Solving for μ_i we derive

$$\mu_i = R + \beta_i(\mu_M - R). \tag{48}$$

□

5 Fundamental Theorem of Asset Pricing

In this section, we consider the problem of pricing a risky asset from a different perspective based on the principle of no arbitrage. This perspective leads to the fundamental theorem of asset pricing (FTAP) a fundamental result pioneered by Cox and Ross [7] and developed in progressing generality in the past several decades by many researchers (see [8, 10, 13, 14]). FTAP links the no arbitrage principle to the existence of equivalent martingale measures which can be used to price risky assets including contingent claims in a given financial market. Our discussion starts with portfolio utility (seeing as a generalized entropy) maximization problem and then view the equivalent martingale measure (also called the risk neutral measure) as the dual solution follows the idea in [6, 33]. A similar framework using directional derivatives instead of convex duality has been discussed in [25].

Gain without risk is what every investor desires. Such opportunities arguably will not last. Since when everyone tries to chase it the price will move up that will eventually eliminate the opportunity. Based on this observation, in modeling a financial market a guiding principle is that arbitrage should not exist. The following is a formal definition.

Definition 1 (Arbitrage) We say that a portfolio Θ is an arbitrage if it involves no risk, $\Theta \cdot (S_1 - S_0) \geq 0$ yet has opportunity to gain something $\Theta \cdot (S_1 - S_0) \neq 0$.

The Fundamental Theorem of Asset Pricing (FTAP) links no arbitrage with the existence of risk-neutral or martingale measures defined below:

Definition 2 (Equivalent martingale measure) We say that Q is an *Equivalent Martingale Measure (EMM)* on economy $(\Omega, \mathcal{F}, P)$ for financial market S provided that, for any atom B_i of $\mathcal{F}$, $Q(B_i) \neq 0$ if and only if $P(B_i) \neq 0$, and

$$\mathbf{E}^Q[S_1] = S_0.$$

The significance of the theorem is that knowing an equivalent martingale measure Q can be used to pricing financial assets. Suppose $\phi(S_1)$ is a function of the financial assets in the market represents the payoff of a contingent claim at time $t = 1$. Then $\phi_0 = \mathbf{E}^Q[\phi(S_1)]$ is a reasonable price for this derivative at $t = 0$ in the sense that using this price will not create any arbitrage opportunities. To understand FTAP let's denote the set of gains by

$$W := \{\Theta \cdot (S_1 - S_0) : \Theta \in port[S]\} \subset RV(\Omega, \mathcal{F}, P).$$

We can see that, in fact, W is a subspace of $RV(\Omega, \mathcal{F}, P)$. It is not hard to see that if Θ is an arbitrage portfolio then $\Theta \cdot (S_1 - S_0) \in RV(\Omega, \mathcal{F}, P)^+ \backslash \{0\}$, where $RV(\Omega, \mathcal{F}, P)^+$ is the cone of nonnegative random variables. Thus, no arbitrage can be described as

$$W \cap RV(\Omega, \mathcal{F}, P)^+ \backslash \{0\} = \emptyset.$$

Traditional proofs of the FTAP rely on applying an appropriate version of the cone separation theorem to ensure that there is a hyperplane separating W and $RV(\Omega, \mathcal{F}, P)^+$. Then, a scaling of the normal vector of such a separating hyperplane gives us an equivalent martingale measure.

The fact that such an equivalent martingale measure comes from a generic separation theorem is often interpreted as the no arbitrage price being independent of investor's preferences. However, we derive FTAP using a framework of entropy maximization where the "entropy" is a utility function that captures the risk aversion of a typical investor. This approach also shows that martingale measures are actually related to the risk aversion of investors. We consider a general extended valued upper semicontinuous utility function u that satisfies the following conditions:

(u1) (Risk aversion) u is strictly concave,
(u2) (Profit seeking) u is strictly increasing and $\lim_{t\to+\infty} u(t) = +\infty$,
(u3) (Bankruptcy forbidden) For any $t < 0$, $u(t) = -\infty$,

A rational investor with a utility function u satisfying conditions (u1)–(u3) will try to maximize the expected utility of the final wealth among all portfolios in $port[S]$. In other words, if $w_0 > 0$ is the initial wealth of the investor, he wants to solve the following portfolio utility maximization problem:

$$\sup\{\mathbf{E}[u(w_0 + \Theta \cdot (S_1 - S_0)] : \Theta \in port[S]\}. \tag{49}$$

It turns out that an arbitrage opportunity is exactly characterized by the optimal value for problem (49) to be $+\infty$.

Theorem 6 (Characterizing arbitrage with utility optimization) *The portfolio space $port[S]$ contains an arbitrage if and ony if the optimal value of the utility optimization problem is $+\infty$*

Proof The "only if" part is easy: if $\Theta \in port[S]$ is an arbitrage then so is $r\Theta$ for any $r > 0$. Then it is easy to see that $\mathbf{E}[u(w_0 + r\Theta \cdot (S_1 - S_0)] \to +\infty$ as $r \to +\infty$.

To prove the “if part” assume the optimal value for problem (49) is $+\infty$. Then there exists a sequence $\Theta^n \in port[S]$ such that $\mathbf{E}[u(w_0 + \Theta^n \cdot (S_1 - S_0)] \to +\infty$ as $n \to +\infty$. Necessarily, $t_n = \|\Theta^n \cdot (S_1 - S_0)\|_{RV} \to +\infty$ as n goes to ∞. Use the definition of the portfolio norm we can show that $\|\Theta^n/t_n\|$ is unifromly bounded. Thus, without loss of generality we may assume that Θ^n/t_n converges to some $\Theta^* \in port[S]$. Note that, for any n, $\Theta^n \cdot (S_1 - S_0) \geq -w_0$ by property (u3) of the utility function. Thus, $\Theta^* \cdot (S_1 - S_0) \geq 0$. Also,

$$\|\Theta^* \cdot (S_1 - S_0)\| \geq \liminf_{n\to\infty} \|\Theta^n \cdot (S_1 - S_0)/t_n\| = 1.$$

Therefore, Θ^* is an arbitrage. □

Given an initial wealth $w_0 > 0$, the set of all achievable wealth outcomes at the end of the one period economy $t = 1$ using all possible portfolios is

$$w_0 + \{\Theta \cdot (S_1 - S_0) : \Theta \in port[S]\} \subset RV(\Omega, \mathcal{F}, P).$$

Theorem 7 (Refined fundamental theorem of asset pricing) *Let S be a financial market, let u be a utility function that satisfies properties (u1), (u2) and (u3) and let $w_0 \geq 0$ be a given initial endowment. Then the following are equivalent:*

(i) *$port[S]$ contains no arbitrage.*
(ii) *The optimal value of the portfolio utility optimization problem (49) is finite value and attained.*
(iii) *There is an equivalent S-martingale measure proportional to the subdifferential of u at the optimal solution of (49).*

Proof We use a cyclical proof.

By Theorem 6, $port[S]$ contains no arbitrage if and only if the optimal value of problem (49) is finite and, therefore, (i) implies (ii).

Implication (ii) → (iii) is the key and we use entropy maximization. Observing that the utility optimization problem (49) can be written equivalently as

$$\begin{aligned} p := \max \quad & \mathbf{E}[u(y)] \\ \text{subject to} \quad & y \in w_0 + W. \end{aligned} \tag{50}$$

Alternatively we can write (50) as an entropy optimization problem

$$\begin{aligned} -p = \text{minimize} \quad & \mathbf{E}[(-u)(y)] \\ \text{subject to} \quad & y - \Theta \cdot (S_1 - S_0) = w_0. \end{aligned} \tag{51}$$

In this problem the variable $x = (y, \Theta)$, $f(x) = E[(-u)(y)]$ and the moment condition is $Ax = y - \Theta \cdot (S_1 - S_0) = w_0$. Thus, $\operatorname{dom} f = RV(\Omega, \mathcal{F}, P)^+ \times port[S]$ and the constrain qualification condition $w_0 \in \operatorname{ri} A \operatorname{dom} f$ holds. Thus, the dual problem to (51) has a solution λ which is the primal Lagrange multiplier. We have already known from the proof of the Theorem 7 that the primal problem has a solution (y^*, Θ^*). Then the Lagrangian

$$\begin{aligned} L((y,\Theta),\lambda) &= \mathbf{E}[(-u)(y)] + \langle \lambda, y - \Theta \cdot (S_1 - S_0) - w_0 \rangle \\ &= \mathbf{E}[(-u)(y)] + \langle \lambda, y - w_0 \rangle - \langle \lambda, \Theta \cdot (S_1 - S_0) \rangle \\ &= \mathbf{E}[(-u)(y) + \lambda(y - w_0)] - \langle \lambda, \Theta \cdot (S_1 - S_0) \rangle \end{aligned}$$

attains minimum at (y^*, Θ^*). It follows that $\langle \lambda, S_1 - S_0 \rangle = 0$ and $-\lambda(B_i) \in \partial(-u)(y^*(B_i))$, $i = 1, 2, \ldots, N$ for $P(B_i) > 0$. Since $-u$ is strictly decreasing we have $\lambda(B_i) > 0$ whenever $P(B_i) > 0$. Moreover, dividing $\langle \lambda, S_1 - S_0 \rangle = \mathbf{E}[\lambda(S_1 - S_0)] = 0$ by $\mathbf{E}[\lambda]$ and noticing that S_0 is a constant vector we get

$$\mathbf{E}[(\lambda/\mathbf{E}[\lambda])S_1] = S_0.$$

This is to say that $Q = (\lambda/\mathbf{E}[\lambda])P$ is a martingale measure equivalent to P. Thus, (ii) implies (iii). We can see that this martingale measure is indeed a scaling of the Lagrange multiplier.

Finally, if (iii) is true then there cannot be any arbitrage in $port[S]$ because adding an arbitrage to the optimal solution of (49) will improve it. Thus, (iii) implies (i) and we have completed a cyclic proof of the equivalence of (i), (ii) and (iii). □

Remark 4 Although no arbitrage is equivalent to the existence of an equivalent martingale measure is well known, as pointed out in [33] the proof of Theorem 7 using a class of utility functions says more: when the martingale measure is not unique, the dual problem actually points to one particular martingale measure. Thus, in principle, every choice of equivalent martingale measure (corresponding to a particular price of the contingent claim) can be viewed as a particular portfolio optimization problem with a corresponding concave utility function. In particular, when the market is not complete there are many possibilities in selecting the utility functions. Thus, the pricing of contingent claims does rely on the trader's preference.

6 Selecting a Pricing Martingale Measure by Entropy Maximization

We have seen in the previous section that equivalent martingale measure is related to the investor's risk aversion and, in general, not unique. Thus, for a contingent claim, its price under the no arbitrage principle with equivalent martingale measures is not unique. Question arises as to how to choose an appropriate martingale measure. Theoretically, if the investor's risk aversion, described by a utility function is specified then one can determine the martingale measure according to the refined FTAP in the previous section. The problem with this approach is that it is well known that specify or calibrate the utility function is very difficult in practice. Moreover, even if a utility function is known, deriving a corresponding martingale measure according to FTAP needs to solve the portfolio optimization problem which is also quite difficult. On the other hand, determining a equivalent martingale measure for a financial market S on a finite probability space amounts to solve a matrix equation which is easy. Therefore, in practice practitioners usually directly deal with equivalent martingale measures. When the martingale measures are not unique the question is then how to choose one that is reasonable. Using a criterion of maximizing the entropy was proposed by Stutzer [31] and Borwein, Choksi, and Lamarchel in [2]. The mathematical formulation is

$$\min\{f(Q) : Q \in \mathcal{M}\}, \tag{52}$$

where $\mathcal{M}$ is the set of all martingale measures on market S and f is the negative of an entropy. Often a choice for f is the Boltzmann–Shannon entropy

$$f(x) = \sum_{n=1}^{N} p(x_n), \tag{53}$$

where

$$p(t) := \begin{cases} t \ln t - t & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ +\infty & \text{if } t < 0, \end{cases}$$

but other entropy functions can also be used. Selecting Boltzmann–Shannon entropy means assuming no prior knowledge on investor's view on the probability distribution. It is also pointed out in [9, 30] that the minimal martingale measure is related to minimizing the relative entropy.

On the other hand, considering $f + \iota_{\mathbf{R}^N_+}$ if necessary, we can assume that $\operatorname{dom} f \subset \mathbf{R}^N_+$. Then we can rewrite problem (52) as

$$\min\{f(Q) : \langle Q, S_1 - S_0 \rangle = 0, \langle Q, 1 \rangle = 1\}. \tag{54}$$

Let (Θ, w) be the dual variable in which Θ and w are Lagrange multipliers corresponding to the constraints $\langle Q, S_1 - S_0 \rangle = 0$ and $\langle Q, 1 \rangle = 1$, respectively. Then the dual problem of (54) is

$$\max\{w - f^*(w + \Theta \cdot (S_1 - S_0))\}. \tag{55}$$

We can view (55) as a portfolio utility maximization problem where w plays the role of initial endowment. Thus, we can see that selecting a pricing martingale measure by maximizing an entropy eventually is still implicitly related to a utility maximization problem.

7 Super/Sub-hedging Bounds

As alluded to before, when the martingale measures are not unique, for a given contingent claim we can derive multiple prices of the contingent claim from those martingale measures that are consistent with the no arbitrage principle. Taking sup and inf of these prices we derive a range outside of which arbitrage opportunity emerges. Duality helps to unveil how to construct a portfolio to take such advantages. We will analyze the case when market price exceeds the sup. This will produce an opportunity for super-hedging. The discussion about the situation when price falls below the inf is similar.

Let $\phi(S_1)$ be the payoff of the contingent claim at $t = 1$. Define

$$U = \max\{\mathbf{E}^Q[\phi(S_1)] \mid Q \in \mathcal{M}\}. \tag{56}$$

Then U is the upper bound of no arbitrage pricing, called the super-hedging bound. Defining

$$f(Q) = \mathbf{E}^Q[-\phi(S_1)] + \iota_{\mathbf{R}^N_+}(Q), \tag{57}$$

we can represent U as the negative value of an entropy maximization problem

$$U = -\min_Q\{f(Q) : \langle Q, S_1 - S_0\rangle = 0, \langle Q, 1\rangle = 1\}. \tag{58}$$

Using the strong duality (54) and (55) we have

$$U = -\max_{w,\Theta}\{w - f^*(w + \Theta \cdot (S_1 - S_0))\}. \tag{59}$$

We can directly calculate that

$$f^*(y) = \iota_{[z:z\le -\phi(S_1)]}(y). \tag{60}$$

Thus,

$$\begin{aligned} U &= \min_{w,\Theta}\{-w + f^*(w + \Theta \cdot (S_1 - S_0))\} \\ &= \min_{w,\Theta}\{-w \mid w \le -\Theta \cdot (S_1 - S_0) - \phi(S_1)\} \end{aligned} \tag{61}$$

The dual representation in (61) provides us a way of finding the super-hedging portfolio to take advantage the arbitrage opportunity should the market price of the contingent claim exceeds the super-hedging bound U. In fact, the second line in (61) tells us that we can derive the super-hedging portfolio by solving the linear programming problem

$$\min_{w,\Theta}\{-w \mid w + \Theta \cdot (S_1 - S_0) \leq -\phi(S_1)\}. \tag{62}$$

It is easy to see that one can rewrite (62) as

$$U = \min_{\Theta} \sup_{\omega \in \Omega}[\Theta \cdot (S_1 - S_0)(\omega) + \phi(S_1)(\omega)]. \tag{63}$$

This is the formula derived by Kahalé in [18] using a separation theorem argument. Constructing super-hedging portfolio from martingale measure for continuous asset pricing model was discussed in [13, 15], where the main issue was representing a martingale integral.

8 Conclusion

Markowitz portfolio theory, the capital market pricing model, fundamental theorems of asset pricing, selecting equivalent pricing martingale measures using entropy maximizations and finding super/sub-hedging bounds and portfolios are several important results in financial economies. We illustrated that they can all be understood in the framework of entropy maximization. So many important results in financial economics involving this fundamental principle in physics demonstrated the heavy influence of methods in physical sciences to financial research.

This link to physical science brings about welcome rigor and quantitative precision into financial research. On the other hand, the principle of entropy maximization is proposed in statistical mechanics. Statistic mechanics deals with complex systems consist of many identical microscopic elements. The impact of each of the elements to the system as a whole is negligible. While these models resemble Aumann's idealized atom less economy, they are significantly different from the real financial market. Two of the main differences are first agents in a financial market are not uniform in their sizes and impacts to the market as a whole. Many big financial institutes can swing the market in a significant way. In particular, in a crises the failure of any of those big players can cause turmoil to the whole market as the 2008 financial crises and many of its predecessors have shown. The second main difference is that agents in a financial market are humans. Instead following a fixed physical law they interact with each other and their behaviors are also determined by human psychology. For these reasons those fundamental results in finance derived above using the entropy maximization methods should be treated as a rough sketch of a road map that needs to be used with caution in practice.

References

1. Boltzmann, L.: Über die Mechanische Bedeutung des Zweiten Hauptsatzes der Wärmetheorie. Wiener Berichte. **53**, 195–220 (1866)
2. Borwein, J.M., Choksi, R., Maréchal, P.: Probability distribution of assets inferred from option prices via the principle of maximum entropy. SIAM J. Optim. **14**, 464–478 (2003)
3. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization, 2nd edn, 2005. Springer, Berlin (2000)
4. Borwein, J.M., Zhu, Q.J.: Techniques of Variational Analysis. Springer, Berlin (2005)
5. Borwein, J.M., Zhu, Q.J.: A variational approach to Lagrange multipliers. J. Optim. Theory Appl. **171**, 727–756 (2016). https://doi.org/10.1007/s10957-015-0756-2
6. Carr, P., Zhu, Q.J.: Convex Duality and Finanacial Mathematics. Springer, Berlin (2018)
7. Cox, J., Ross, S.: The valuation of options for alternative stochastic processes. J. Financ. Econ. **3**, 144–166 (1976)
8. Delbaen, F., Schachermayer, W.: A general version of the fundamental theorem of asset pricing. Math. Ann. **300**, 463–520 (1994)
9. Delbaen, F., Grandits, P., Rheinländer, T., Samperi, D., Schweizer, M., Stricker, C.: Exponential hedging and entropic penalties. Math. Finance **12**, 99–123 (2002)
10. Dybvig, P., Ross, S.A.: Arbitrage, state prices and portfolio theory. Handbook of the Economics of Finance (2003)
11. Fenchel, W.: Conv Cones, Sets and Functions. Lecture Notes. Princeton University, Princeton, (1951)
12. Gibbs, J.W.: Elementary Principles in Statistical Mechanics. Charles Scribner's Sons, New York (1902)
13. Harrison, J.M., Pliska, S.: Martingales and stochastic integrals in the theory of continuous trading. Stoch. Process. Appl. **11**, 215–260 (1981)
14. Harrison, J.M., Kreps, D.M.: Martingales and arbitrage in multiperiod securities markets. J. Econom. Theory **20**, 381–408 (1979)
15. Jacka, S.D.: A martingales representation result and an application to incomplete financial markets. Math. Finance **2**, 239–250 (1992)
16. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. Ser. **II**(106), 620–630 (1957)
17. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. Ser. **II**(108), 171–190 (1957)
18. Kahalé, N.: Sparse calibrations of contingent claims. Math. Finance **20**, 105–115 (2010)
19. Kelly, J.L.: A new interpretation of information rate. Bell Syst. Tech. J. **35**, 917–926 (1956)
20. Lintner, J.: The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Rev. Econ. Stat. **47**, 13–37 (1965)
21. Markowitz, H.: Portfolio Selection. Cowles Monograph, vol. 16. Wiley, New York (1959)
22. Moreau, J.J.: Fonctionelles Convexes. Lecture Notes. College de France, Paris (1967)
23. Mossin, J.: Equilibrium in a capital asset market. Econometrica **34**, 768–783 (1966)
24. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
25. Rogers, L.C.G.: Equivalent martingale measures and no-abitrage. Stoch. Stoch. Rep. **51**, 41–49 (1994)
26. Roman, S.: Introduction to the Mathematics of Finance. Springer, New York (2004)
27. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
28. Sharpe, W.F.: Capital asset prices: a theory of market equilibrium under conditions of risk. J. Finance **19**, 425–442 (1964)
29. Sharpe, W.F.: Mutual fund performance. J. Bus. **1**, 119–138 (1966)
30. Schweizer, M.: A minimality property of the minimal martingale measure. Stat. Probab. Lett. **42**, 25–31 (1999)
31. Stutzer, M.: A simple nonparametric approach to derivative security valuation. J. Finance **51**, 1633–1652 (1996)

32. J. L. Treynor, Toward a theory of market value of risky assets. Unpublished manuscript 1962. A final version was published in 1999, in Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics. Korajczyk, R.A (edn) Risk Books, London, pp. 15–22

33. Zhu, Q.J.: Convex analysis in mathematical finance. Nonlinear Analysis: Theory Method and Applications **75**, 1719–1736 (2012)

Part IV
Number Theory, Special Functions, and Pi

Introduction

Richard P. Brent

At the Jonathan Borwein Commemorative Conference (JBCC), the talks were divided into several themes, one of which was “Number Theory, Special Functions and π”. Here we summarise the eight contributions to that theme included in the Proceedings. There is a considerable overlap with other themes, and with Experimental Mathematics. The eight contributions described here are as follows:

1. Michael Baake, Michael Coons and Neil Mañibo, *Binary Constant-Length Substitutions and Mahler Measures of Borwein Polynomials*;
2. Richard Brent, *The Borwein Brothers, π and the AGM*;
3. Cristian and Elena Calude, *The Road to Quantum Computational Supremacy*;
4. Karl Dilcher, *Nonlinear Identities for Bernoulli and Euler Polynomials*;
5. Mumtaz Hussain, Seyyed Mahboubi and Abolfazl Motahari, *Metrical Theory for Small Linear Forms and Applications to Interference Alignment*;
6. Dave Platt and Tim Trudgian, *Improved Bounds on Brun’s Constant*;
7. Matthew Skerritt and Paul Vrbik, *Extending the PSLQ Algorithm to Algebraic Integer Relations*;
8. Armin Straub and Wadim Zudilin, *Short Walk Adventures*.

The names of authors who spoke at JBCC are underlined. In cases (4), (5) and (6) they spoke on a different topic from the printed paper. In cases (3) and (7) the paper was not presented at the conference, but the editors considered it appropriate for inclusion here. Two talks do not correspond to papers in this volume, but the

R. P. Brent (✉)
Mathematical Sciences Institute, Australian National University, Canberra, Australia

School of Mathematical and Physical Sciences, The University of Newcastle, Newcastle, Australia
e-mail: JBCC@rpbrent.com

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_19

slides for these talks are available.[1] Bruce Berndt spoke on *Identities in the spirit of Jonathan Borwein*. David Harvey gave a survey on *Computing Bernoulli numbers*, but was not able to submit a paper to these Proceedings.

Jon was fascinated by the constant π, and gave many stimulating talks on this topic. The slides for some of these talks may be found on the memorial website [3]. Brent's talk and paper (2) consider the reasons for this fascination. In a nutshell, it is that theorems about π are often just the tips of 'mathematical icebergs'—much of interest lies hidden beneath the surface. The paper (2) concentrates on the analysis of superlinearly convergent algorithms for the computation of π and, more generally, the elementary functions. The key to superlinear convergence is the use of the arithmetic-geometric mean (AGM), which is the topic of the Borwein brothers' fascinating book *Pi and the AGM* [1]. Several algorithms based on the AGM are considered in Brent's paper. In a surprising new result, he shows that two algorithms that were previously considered different are actually equivalent.

Ever since Peter Shor's 1994 paper [9] on quantum algorithms for discrete logarithms and factoring, we have known that certain tasks that seem difficult with classical computers, such as integer factorisation, can be performed in polynomial time on a quantum computer. But here, 'quantum computer' refers to a mathematical model of computation, not a physical device, since physicists and engineers are still trying to overcome the problems of actually building a quantum computer of sufficient size (and reliability) to factor integers of significant size. For applications to cryptography, this means of the order of 1000 bits, i.e. numbers about as large as the tenth Fermat number $2^{1024} + 1$ (which was factored by classical means in 1995, see [4]). The 'pessimistic' view [7] is that the difficulties are insurmountable. The opposing, 'optimistic' view is that the remaining challenge is essentially of an engineering nature. The paper (3) by the Caludes presents both sides of the argument. Although Jon never published on quantum computing, we know that he was interested in it. If he could have been transported 100 years into the future, his first question might well have been 'have quantum computers capable of factoring 1000-bit numbers been built'.[2]

Jon was interested in open problems arising from the work of Kurt Mahler, such as Lehmer's problem for polynomials with small logarithmic Mahler measure. Indeed, Jon arranged for his Centre CARMA to host an online archive of Mahler's publications.[3] In the paper (1), Jon's former colleague Michael Coons and his co-authors explore the connection between Lehmer's problem and the spectral theory of binary constant-length substitutions.

[1] Slides for many of the talks presented at JBCC may be found by clicking on the author's name in the programme at https://carma.newcastle.edu.au/meetings/jbcc/programme/, and then on the 'pdf' symbol corresponding to the talk.

[2] David Hilbert is reported to have said that, under similar circumstances, he would ask if the Riemann Hypothesis had been proved. Jon would be interested in the answers to both questions.

[3] https://carma.newcastle.edu.au/mahler/collected.html.

Jon wrote several papers on Tornheim-Witten zeta functions, see for example his recent paper [2] with Karl Dilcher. Motivated by this work, Dilcher's paper (4) in this volume proves some interesting new nonlinear identities for Bernoulli and Euler polynomials.

The approximation of real numbers by rationals is a classical problem in number theory, dating from Dirichlet (1842) and even earlier. Khintchine's theorem (1924) gives an elegant approximation criterion in terms of the convergence or divergence of a certain infinite series. In paper (5), Jon's former colleague Mumtaz Hussain and his co-authors discuss Khintchine's theorem and higher-dimensional generalisations, the Khintchine-Groshev theorems. They extend some such theorems to the complex number system (where integers are replaced by Gaussian integers) and outline an interesting application to signal processing. Contrary to Hardy's opinion [6], and even excluding applications to cryptography, number theory does have practical applications![4]

The PSLQ algorithm [5] was undoubtedly one of Jon's favourite algorithms, and in his talks he often mentioned identities that were discovered using it. In (7), Jon's former student and co-author Matt Skerritt, and Jon's former postdoc Paul Vrbik, show how PSLQ can be extended to find integer relations consisting of algebraic integers from a quadratic field $\mathbb{Q}[\sqrt{D}]$, where $D \in \mathbb{Z}$ is nonzero and squarefree. The algorithm works well in cases where it is theoretically justified, and, perhaps surprisingly, also often works in cases where a theoretical justification is lacking.

Jon was interested in algorithms for the computation of constants such as π. Brun's constant B is defined by the sum of reciprocals of twin primes, but this sum converges slowly and irregularly, so standard convergence acceleration techniques work poorly and do not give rigorous upper bounds on B. Indeed, it is not even known if the sum defining B has an infinite number of terms (this is equivalent to the question of whether there are infinitely many twin primes). A lower bound on B can, of course, be obtained by summing over some finite set of twin primes, say all those smaller than 10^{16}. Finding an upper bound on B is more difficult, as any such bound implies the validity of Brun's 1919 theorem that B is finite. In paper (6), Dave Platt and Tim Trudgian improve on the previously known best lower and upper bounds on B. They also show that it will be difficult to improve the bounds much further. Roughly speaking, in time T a classical computer can find of order log log T guaranteed digits of B. Compare this with the computation of π, where the same computer can find of order T/log^2T correct digits.

Jon wrote several papers on 'short' random walks. One has the title 'A short walk can be beautiful', which no doubt expressed his opinion. In paper (8), Jon's former colleagues and co-authors[5] Armin Straub and Wadim Zudilin show that short walks can also be adventurous, in the sense that they lead us to interesting and unexpected results.

[4] For other applications of number theory, see Schroeder [8].

[5] Also, in the case of Armin Straub, former Ph.D. student.

References

1. Borwein, J.M., Borwein, P.B.: Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity. John Wiley & Sons, Toronto (1987)
2. Borwein, J.M., Dilcher, K.: Derivatives and fast evaluation of the Tornheim zeta function. Ramanujan J. **45**, 413–432 (2018)
3. Borwein, J.M.: https://www.carma.newcastle.edu.au/jon/index-talks.shtml
4. Brent, R.P.: Factorization of the tenth Fermat number. Math. Comp. **68**, 429–451 (1999)
5. Ferguson, H.R.P., Bailey, D.H., Arno, S.: Analysis of PSLQ, an integer relation finding algorithm. Math. Comp. **68**, 351–369 (1999)
6. Hardy, G.H.: A Mathematician's Apology. Cambridge University Press, Cambridge (1940) (reprinted 2004)
7. Kalai, G.: Three puzzles on mathematics, computation, and games. Not. AMS **65**, 782–784 (2018)
8. Schroeder, M.R.: Number Theory in Science and Communication: With Applications in Cryptography, Digital Information, Computing, and Self-Similarity, 5th edn. Springer, Berlin (2009)
9. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings of the 35th Annual Symposium on Foundations of Computer Science. IEEE CS Press, Washington, DC (1994)

Binary Constant-Length Substitutions and Mahler Measures of Borwein Polynomials

Michael Baake, Michael Coons and Neil Mañibo

In memory of Jonathan Michael Borwein (1951–2016)

1 Introduction

Let p be a polynomial with complex coefficients. The *logarithmic Mahler measure* of p is defined to be the logarithm of the geometric mean of $|p|$ over the unit circle; that is,

$$\mathfrak{m}(p) \;:=\; \int_0^1 \log \bigl| p\bigl(\mathrm{e}^{2\pi \mathrm{i} t}\bigr)\bigr| \, \mathrm{d}t \, . \tag{1}$$

It is well known and easily shown using Jensen's formula [42, Prop. 16.1] that the logarithmic Mahler measure satisfies

$$\mathfrak{m}(p) \;=\; \log |a_s| \, + \sum_{j=1}^{s} \log\bigl(\max\{|\alpha_j|, 1\}\bigr),$$

M. Baake · N. Mañibo
Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany
e-mail: mbaake@math.uni-bielefeld.de

N. Mañibo
e-mail: cmanibo@math.uni-bielefeld.de

M. Coons (✉)
School of Mathematical and Physical Sciences, University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
e-mail: michael.coons@newcastle.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_20

where $p(z) = a_s \prod_{j=1}^{s}(z - \alpha_j)$; see [30] for background. Here, we will only consider integer polynomials. Kronecker's lemma [32] then implies that $\mathfrak{m}(p) = 0$ if and only if p is a product of a cyclotomic polynomial (not necessarily irreducible) and a monomial. In this way, $\mathfrak{m}$ is a measure of the distance of an integer polynomial to the unit circle.

One of the most interesting and long-standing problems in this area concerns finding polynomials with small logarithmic Mahler measures. Lehmer found the polynomial

$$\ell_{\mathrm{L}}(z) \;=\; 1 + z - z^3 - z^4 - z^5 - z^6 - z^7 + z^9 + z^{10}, \tag{2}$$

which is irreducible and has precisely one root outside the unit disk. This root is real and a Salem number. The polynomial ℓ_{L} is the polynomial with the smallest known positive logarithmic Mahler measure,

$$\mathfrak{m}(\ell_{\mathrm{L}}) \;\approx\; \log(1.176281).$$

Lehmer's problem. Does there exist a constant $c > 0$ such that any irreducible non-cyclotomic polynomial p with integer coefficients satisfies $\mathfrak{m}(p) \geqslant c$?

There are some special classes of polynomials for which Lehmer's problem has long been answered in the affirmative. In particular, there is a very interesting gap result for non-reciprocal polynomials due to Smyth [43]; see also Breusch [25]. A polynomial p is *reciprocal* (in the wider sense) if $p(z) = \pm z^{\deg(p)} p(1/z)$; that is, a polynomial is reciprocal if its sequence of coefficients is palindromic, up to an overall sign. It follows from Smyth's result that, for non-reciprocal polynomials, one either has $\mathfrak{m}(q) = 0$ or $\mathfrak{m}(q) \geqslant \log(\lambda_{\mathrm{p}})$, where λ_{p} is the smallest Pisot number, which is the real root of $z^3 - z - 1$, also known as the *plastic number* [8, Ex. 2.17]. Specialising this class a bit more, Borwein, Hare and Mossinghoff [19, Cor. 1.2] showed that all non-reciprocal polynomials q with odd integer coefficients satisfy the bound

$$\mathfrak{m}(q) \;\geqslant\; \mathfrak{m}(z^2 - z - 1) \;=\; \log(\tau),$$

where $\tau = \frac{1}{2}\bigl(1 + \sqrt{5}\,\bigr)$ is the golden ratio, a well-known Pisot number. The golden ratio is characterised by the property that it is the smallest limit point of Pisot numbers. See [45] for a general survey, [22] for work on reciprocal polynomials and [24, 29, 38, 40] for more results on small Mahler measures and limit points.

One of the most studied classes of integer polynomials in relation to Lehmer's problem are the *Borwein polynomials*—polynomials of height 1 (coefficients in $\{-1, 0, 1\}$) with non-zero constant term; see [28]. Special importance is placed on this class, since for any integer polynomial p with $\mathfrak{m}(p) < \log(2)$ there is an integer polynomial q such that pq has height 1; see Pathiaux [41]. Boyd [21] notes that, in his experience, such a q can be taken to be cyclotomic and of fairly small degree relative to the degree of p; see also Mossinghoff [37]. If Boyd's observation were proved true in general, then to solve Lehmer's problem it would be enough to consider only Borwein polynomials; unfortunately, this is still unknown.

Before we continue, let us mention that there is a well-known connection between Mahler measures and algebraic dynamics. Here, logarithmic Mahler measures show up as entropies of $\mathbb{Z}^d$-shifts of algebraic origin [30, 34, 42]. The first appearance of a Mahler measure in a similar context actually dates back to a paper by Wannier [47] on the ground state entropy of the antiferromagnetic Ising model on the triangular lattice; see Remark 4 below for details. This general connection between Mahler measures and entropy has initiated many investigations and enhanced our knowledge about Mahler measures significantly; see [34, 42] and the references therein for a detailed account.

In this paper, under some quite natural assumptions, we relate the logarithmic Mahler measure of Borwein polynomials to a Lyapunov exponent from the spectral theory of substitutions; see [27] for an earlier appearance of a connection between Mahler measures and Lyapunov exponents. A *binary constant-length substitution* ϱ is defined on $\Sigma_2 := \{0, 1\}$ by

$$\varrho : \begin{cases} 0 \mapsto w_0 \\ 1 \mapsto w_1 \, , \end{cases}$$

where w_0 and w_1 are finite words over Σ_2 of equal length $|w_0| = |w_1| = L \geqslant 2$. Such substitutions are important objects of research in many areas of mathematics, ranging from dynamics and combinatorics (as substitutions) to number theory (this is the class of binary automatic sequences) and theoretical computer science (under the name of uniform morphisms).

Recall that the *substitution matrix* of ϱ is the matrix $M_\varrho = (m_{ij})_{0 \leqslant i,j \leqslant 1}$, where $m_{ij} \geqslant 0$ is the number of letters i in the word w_j. This matrix is also known as the *Abelianisation* of ϱ; compare [8, Sec. 4.1]. We say that ϱ is *primitive* if the non-negative matrix M_ϱ is primitive, and *aperiodic* if the hull (or shift) defined by ϱ does not contain any element with a non-trivial period. When ϱ is primitive, this is the case if and only if any of the two-sided fixed points of ϱ (or of ϱ^n with a suitable $n \in \mathbb{N}$) with legal core (or seed) is non-periodic. If one of these fixed points is non-periodic, they all are, due to primitivity; see [8, Sec. 4.2] for notions and further details.

Our main result is the following theorem; the relevant concepts concerning Lyapunov exponents are recalled in Section 3.

Theorem 1 *For any primitive, binary constant-length substitution ϱ, the extremal Lyapunov exponents are explicitly given by*

$$\chi^B_{\min} = 0 \quad \text{and} \quad \chi^B_{\max} = \mathfrak{m}(p_\varrho),$$

where p_ϱ is a Borwein polynomial, easily determined by ϱ. In particular, if p_ϱ is non-reciprocal, one has $\chi^B_{\max} \geqslant \log(\lambda_{\mathrm{p}})$, where λ_{p} is the plastic number.

Theorem 1 is a statement relating the logarithmic Mahler measure of Borwein polynomials to Lyapunov exponents of binary constant-length substitutions. Depending on which object one is interested in, it can be used in a couple of different ways. As it reads above, if one has a binary substitution, one can easily compute the extremal Lyapunov exponents using the associated Borwein polynomial. Alternatively, if one has a Borwein polynomial, one can determine an associated binary constant-length substitution. This relationship can be exploited to give some general results about extremal Lyapunov exponents for certain binary substitutions. For example, one now has a rather general result considering *bijective* substitutions, which are the substitutions where the letters in the words w_0 and w_1 are different at each position. At this point, [19, Cor. 1.2] in conjunction with Lemma 2 below implies the following consequence of Theorem 1; see Example 3 for more details.

Corollary 1 *Suppose that the primitive, binary constant-length substitution ϱ is bijective, and that w_0 is neither a palindrome nor an anti-palindrome. Then, one has $\chi^{B}_{\max} \geqslant \log(\tau)$, where τ is the golden ratio.*

As it turns out, primitive, binary constant-length substitutions which are periodic do not satisfy the assumptions of this corollary; in fact, for such periodic substitutions one has that $\chi^{B}_{\max} = 0$. A characterisation and further details regarding these periodic substitutions are given later; see Lemma 2 and Theorem 3.

In view of the above results, in the case of Borwein polynomials, Lehmer's problem can be restated in terms of the Lyapunov exponents for binary constant-length substitutions.

Lehmer's problem (dynamical analogue). Does there exist a constant $c > 0$ such that, for any primitive, binary constant-length substitution with $\chi^{B}_{\max} \neq 0$, we have $\chi^{B}_{\max} \geqslant c$?

Remark 1 Strong versions of both Lehmer's problem and our dynamical analogue would ask whether the constant c can be taken to be $\mathfrak{m}(\ell_{\mathrm{L}})$, the logarithmic Mahler measure of Lehmer's polynomial from Eq. (2). ◊

Viewing Lehmer's problem in a dynamical setting, as related to constant-length substitutions, has some heuristic benefits. In this area, especially at the interface with number theory, gap results are common and expected. For example, if $f(n)$ denotes the n^{th} letter of a one-sided fixed point of a constant-length substitution, then, for any positive integer $b \geqslant 2$, the number $\sum_n f(n) b^{-n}$ is either rational or transcendental [1, 3, 14]. Also, this number cannot be a Liouville number, that is, it has finite irrationality exponent [2, 14]. The partial sums $S(N) := \sum_{n \leqslant N} f(n)$ satisfy even stronger gap properties. If $S(N)$ is unbounded, there is a constant $c > 0$ such that $|S(N)| \geqslant c \, \log(N)$ for infinitely many integers N; see [15, 16]. Viewing Lehmer's problem for Borwein polynomials in this context may, at the least, take away some of the surprise of its conclusion, and provide an additional reason to believe in the conjecture for this class of polynomials.

The remainder of this paper is organised as follows. In Section 2, we give details regarding binary substitutions and their associated Fourier matrices, while we give the relevant definitions on Lyapunov exponents in Section 3. Section 4 contains the proof of Theorem 1 as a consequence of a more detailed version. In Section 5, we provide several examples including those related to Littlewood, Newman and Borwein polynomials. Finally, in Section 6, we explore extensions to higher dimensions and their relationship to logarithmic Mahler measures of multivariate polynomials.

2 Substitutions and their Fourier Matrices

As stated in the Introduction, we are concerned with binary constant-length substitutions ϱ defined on $\Sigma_2 := \{0, 1\}$ by

$$\varrho : \begin{cases} 0 \mapsto w_0 \\ 1 \mapsto w_1 , \end{cases} \tag{3}$$

where w_0 and w_1 are finite words over Σ_2 of equal length[1] $|w_0| = |w_1| = L \geqslant 2$.

We denote the m^{th} column of ϱ by

$$\mathcal{C}_m := \begin{bmatrix} (w_0)_m \\ (w_1)_m \end{bmatrix},$$

where $(w_i)_m$ is the m^{th} letter of the word w_i. We follow the convention of indexing the columns starting with 0; compare [8, Ch. 4]. A binary substitution is said to have a *coincidence* at position m, if the column at that specific position is either $\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ or $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. A binary substitution is called *bijective* if there are no coincidences.

For $0 \leqslant i, j \leqslant 1$, let T_{ij} be the set of all positions m where the letter i appears in w_j, and let $T := (T_{ij})_{0 \leqslant i,j \leqslant 1}$ be the resulting 2×2-matrix. Note that the substitution matrix M_ϱ, as defined above, satisfies

$$M_\varrho = \bigl(\operatorname{card}(T_{ij})\bigr)_{0 \leqslant i,j \leqslant 1} .$$

Using T, we build a matrix of pure point measures $\delta_T := (\delta_{T_{ij}})_{0 \leqslant i,j \leqslant 1}$, where we use $\delta_S := \sum_{x \in S} \delta_x$ with $\delta_\varnothing = 0$. This gives rise to an analytic matrix-valued function via

$$B(k) := \overline{\widehat{\delta_T}(k)},$$

which we call the *Fourier matrix* of ϱ; see [4, 5]. Note that $B(0) = M_\varrho$. The Fourier matrix provides more information than M_ϱ; it encodes the column positions of each letter in the corresponding words that contain them, whereas the entries of M_ϱ only count the letters 0 and 1 in w_0 and w_1, respectively.

[1] Those comfortable with the dynamical setting will note that, by working with two prototiles of unit length, the tiling and symbolic pictures of these systems are equivalent (topologically conjugate by a sliding block map).

The following two examples are paradigmatic for the two principal situations among aperiodic, binary constant-length substitutions.

Example 1 Consider the Thue–Morse substitution, as given by

$$\varrho_{\mathrm{TM}} : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 10. \end{cases}$$

Here, one has $T_{\mathrm{TM}} = \left(\begin{smallmatrix} \{0\} & \{1\} \\ \{1\} & \{0\} \end{smallmatrix}\right)$, which gives

$$\delta_{T_{\mathrm{TM}}} = \begin{pmatrix} \delta_0 & \delta_1 \\ \delta_1 & \delta_0 \end{pmatrix} \quad \text{and} \quad B_{\mathrm{TM}}(k) = \begin{pmatrix} 1 & \mathrm{e}^{2\pi \mathrm{i} k} \\ \mathrm{e}^{2\pi \mathrm{i} k} & 1 \end{pmatrix}. \qquad \diamondsuit$$

Example 2 On the other hand, for the period doubling substitution,

$$\varrho_{\mathrm{pd}} : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 00, \end{cases}$$

the corresponding matrices are $T_{\mathrm{pd}} = \left(\begin{smallmatrix} \{0\} & \{0,1\} \\ \{1\} & \varnothing \end{smallmatrix}\right)$ together with

$$\delta_{T_{\mathrm{pd}}} = \begin{pmatrix} \delta_0 & \delta_0 + \delta_1 \\ \delta_1 & 0 \end{pmatrix} \quad \text{and} \quad B_{\mathrm{pd}}(k) = \begin{pmatrix} 1 & 1 + \mathrm{e}^{2\pi \mathrm{i} k} \\ \mathrm{e}^{2\pi \mathrm{i} k} & 0 \end{pmatrix}. \qquad \diamondsuit$$

3 Lyapunov Exponents

Using the ergodic transformation $k \mapsto Lk$ mod 1 defined on the 1-torus $\mathbb{T}$, which is represented by [0, 1) with addition modulo 1 and equipped with Lebesgue measure, one can use the Fourier matrix $B(k)$ to build the matrix cocycle

$$B^{(n)}(k) := B(k)B(Lk) \cdots B(L^{n-1}k),$$

where the (dynamically unusual) extension to the right originates from the underlying spectral problem of binary substitutions; see Remark 2 below and [5, 6, 10]. Recall that the integer $L \geqslant 2$ is the common length of the words w_0 and w_1 from the definition of the binary substitution ϱ. We note further that the inverse cocycle $(B^{(n)}(k))^{-1}$ exists for almost every $k \in \mathbb{R}$, because $\det\bigl(B(k)\bigr) = 0$ for at most a countable subset of $\mathbb{R}$. Due to the ergodicity of the transformation $k \mapsto Lk$ on $\mathbb{T}$ relative to Lebesgue measure, Oseledec's multiplicative ergodic theorem ensures the existence of the Lyapunov exponents and the corresponding subspaces in which they represent the asymptotic exponential growth rate of the vector norms; see [12]. More precisely, if $v \in \mathbb{C}^2$ is any (fixed) row vector, the values

$$\chi^{B}(v,k) \;:=\; \lim_{n\to\infty} \frac{1}{n} \log \|v B^{(n)}(k)\| \tag{4}$$

exist for Lebesgue-almost every $k \in \mathbb{R}$ and are constant on a set of full measure. In fact, as a function of v, they take only finitely many values, at most two in this case. These values do not depend on the choice of the norm in (4). Moreover, in the non-degenerate case, there exists a filtration

$$\{0\} =: \mathcal{V}_0 \subsetneq \mathcal{V}_1 \subsetneq \mathcal{V}_2 := \mathbb{C}^2$$

such that χ_1^B is the corresponding exponent for all $0 \neq v \in \mathcal{V}_1$ and χ_2^B then for all $v \in \mathcal{V}_2 \setminus \mathcal{V}_1$. A vector v from the Oseledec subspace $\mathcal{V}_{i+1} \setminus \mathcal{V}_i$ satisfies the property that, for almost every $k \in \mathbb{R}$, the norm $\|v B^{(n)}(k)\|$ has exponential growth factor $\mathrm{e}^{n\chi_{i+1}^B}$ as $n \to \infty$. In general, these subspaces depend on k, and the filtration simplifies in the obvious way when $\chi_1^B(k) = \chi_2^B(k)$. We refer the reader to the monographs [12] for a general overview and [46] for a more elaborate discussion of linear cocycles.

It is well known that there exist at most two distinct exponents for two-dimensional cocycles, denoted by $\chi_{\max}^B(k)$ and $\chi_{\min}^B(k)$. For invertible cocycles, these exponents admit v-independent representations as

$$\begin{aligned}
\chi_{\max}^B(k) \;&:=\; \lim_{n\to\infty} \frac{1}{n} \log \bigl\| B^{(n)}(k) \bigr\| \quad \text{and} \\
\chi_{\min}^B(k) \;&:=\; -\lim_{n\to\infty} \frac{1}{n} \log \bigl\| \bigl(B^{(n)}(k)\bigr)^{-1} \bigr\|.
\end{aligned}$$

Moreover, the exponents are constant almost everywhere, hence effectively independent of k as well. This means that we are dealing with two *numbers*, $\chi_{\max}^B$ and $\chi_{\min}^B$. It turns out that one has an unexpected connection with logarithmic Mahler measures, as we discuss below.

Remark 2 These exponents come up in the spectral study of the substitutions associated to these Fourier matrices, and can be derived from a renormalisation scheme involving pair correlation functions; see [5, 6, 9]. In particular, by measure-theoretic arguments, one can conclude that, if $|\chi_{\max}^B| < \log\sqrt{L}$, the diffraction measure, for an arbitrary choice of weights, never has an absolutely continuous component. We refer to the literature for further details; in particular, see [35] for the binary constant-length case, and [4, 10] for an appropriate extension to a family of non-Pisot substitutions, via the corresponding inflation tiling. ♢

4 Proof of the Main Result

As stated in Theorem 1 in the Introduction, the maximal Lyapunov exponent can be written as a logarithmic Mahler measure. We prove this as Theorem 2 below, which is a more detailed version of Theorem 1; compare [35].

Lemma 1 *Let ϱ be a substitution as specified in Eq.* (3)*. Consider the sets*

$$P_a := \bigl\{ m \mid \mathcal{C}_m = \bigl[\begin{smallmatrix}0\\1\end{smallmatrix}\bigr] \bigr\} \quad \textit{and} \quad P_b := \bigl\{ m \mid \mathcal{C}_m = \bigl[\begin{smallmatrix}1\\0\end{smallmatrix}\bigr] \bigr\},$$

which collect bijective positions of the same type. Further, let $z = \mathrm{e}^{2\pi \mathrm{i} k}$ and set

$$Q(z) \, := \, \overline{\widehat{\delta_{P_a}}(k)} \quad \textit{and} \quad R(z) \, := \, \overline{\widehat{\delta_{P_b}}(k)}.$$

Then, $\det\bigl(B(k)\bigr) = p_{L}(z) \cdot \bigl(Q - R\bigr)(z)$*, where* $p_{L}(z) = 1 + z + \dots + z^{L-1}$*.*

Proof In analogy to the definitions of Q and R above, let us now define the sets $P_0 := \bigl\{ m \mid \mathcal{C}_m = \bigl[\begin{smallmatrix}0\\0\end{smallmatrix}\bigr] \bigr\}$ and $P_1 := \bigl\{ m \mid \mathcal{C}_m = \bigl[\begin{smallmatrix}1\\1\end{smallmatrix}\bigr] \bigr\}$, and let

$$S_0(z) \, := \, \overline{\widehat{\delta_{P_0}}(k)} \quad \text{and} \quad S_1(z) \, := \, \overline{\widehat{\delta_{P_1}}(k)}.$$

In general, the Fourier matrix of ϱ satisfies

$$B(k) \, = \, \begin{pmatrix} \bigl(S_0 + Q\bigr)(z) & \bigl(S_0 + R\bigr)(z) \\ \bigl(S_1 + R\bigr)(z) & \bigl(S_1 + Q\bigr)(z) \end{pmatrix} \quad \text{with } z = \mathrm{e}^{2\pi \mathrm{i} k}.$$

Since there are only four distinct column types, we see that

$$S_0 + S_1 + Q + R \, = \, p_{L} \, .$$

One can now verify the lemma by direct computation. □

Recall that, as a consequence of Oseledec's multiplicative ergodic theorem, our Lyapunov exponents exist, and are constant, for almost every $k \in \mathbb{R}$. We call them $\chi^{B}_{\min}$ and $\chi^{B}_{\max}$.

Theorem 2 *For any primitive, binary constant-length substitution ϱ, the extremal Lyapunov exponents are explicitly given by*

$$\chi^{B}_{\min} \, = \, 0 \quad \textit{and} \quad \chi^{B}_{\max} \, = \, \mathfrak{m}(Q - R),$$

with Q and R as in Lemma 1.

Proof Aside from the existence of the extremal Lyapunov exponents as limits, Oseledec's multiplicative ergodic theorem [12, 46] also guarantees forward Lyapunov regularity almost everywhere. That is, for almost every $k \in \mathbb{R}$, the sum of the exponents is given by

$$\chi^{B}_{\min}(k) + \chi^{B}_{\max}(k) \, = \, \lim_{n \to \infty} \frac{1}{n} \log \bigl| \det\bigl(B^{(n)}(k)\bigr) \bigr|, \tag{5}$$

where one can argue that, for the matrices above, the limit on the right-hand side converges for almost every $k \in \mathbb{R}$ to

$$\mathfrak{m}(Q-R) \;=\; \int_0^1 \log\bigl|(Q-R)\bigl(\mathrm{e}^{2\pi \mathrm{i} t}\bigr)\bigr| \,\mathrm{d}t\,.$$

This can be seen by an application of Birkhoff's ergodic theorem, because $t \mapsto Lt$ on $\mathbb{T}$ is ergodic for Lebesgue measure, and $t \mapsto (p_L \cdot (Q-R))\bigl(\mathrm{e}^{2\pi\mathrm{i}t}\bigr)$ defines a function in $L^1(\mathbb{T})$. The claim then follows from the multiplicative property of the determinant in conjunction with Lemma 1 and the fact that $\mathfrak{m}(p_L) = 0$. This value follows via Jensen's formula because all zeros of p_L are roots of unity.

Next, we note that the row vector $(1,1)$ is a left eigenvector of $B(k)$, for all $k \in \mathbb{R}$, with eigenvalue $p_L\bigl(\mathrm{e}^{2\pi\mathrm{i}k}\bigr)$. Hence, using this specific direction, we get one of the exponents to be $\chi^B_1 = \mathfrak{m}(p_L) = 0$. From the sum in Eq. (5), and from the fact that the logarithmic Mahler measure of an integer polynomial is always non-negative, we then get that the exponent corresponding to this invariant subspace is the minimal one, $\chi^B_1 = \chi^B_{\min}$, the other being $\chi^B_{\max} = \mathfrak{m}(Q-R)$. □

Note that the result of this theorem is not restricted to bijective substitutions, even though only the bijective positions matter for the exponents.

Remark 3 In the proof of Theorem 2, instead of invoking Birkhoff's ergodic theorem, one can also work with the uniform distribution of $(L^n k)_{n\in\mathbb{N}}$ modulo 1 for almost every $k \in \mathbb{R}$ and Weyl's lemma. The difficulty to overcome here is that the function defined by $k \mapsto \log\bigl|\det\bigl(B(k)\bigr)\bigr|$ generally has singularities. Fortunately, they are isolated (hence at most countable), and one can extend Weyl's result for locally Riemann integrable 1-periodic functions to this case via Sobol's theorem in conjunction with Diophantine approximation and discrepancy analysis; see [11] and references therein for a more comprehensive discussion.

It is interesting to observe that one can obtain $\chi^B_{\max}$ via the (k-independent) right eigenvector $\tilde{v} = \bigl(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\bigr)$ of $B(k)$, with corresponding eigenvalue $(Q-R)\bigl(\mathrm{e}^{2\pi\mathrm{i}k}\bigr)$ as before. One then has, for almost every $k \in \mathbb{R}$, that

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log\bigl\| B^{(n)}(k)\tilde{v}\bigr\| \;&=\; \lim_{n\to\infty} \frac{1}{n}\Bigl(\|\tilde{v}\| + \sum_{\ell=0}^{n-1} \log\bigl|(Q-R)\bigl(\mathrm{e}^{2\pi\mathrm{i}L^{\ell}k}\bigr)\bigr|\Bigr) \\
&=\; \mathfrak{m}(Q-R),
\end{aligned}$$

where the last step once again relies on Birkhoff's ergodic theorem or on the remarks of the preceding paragraph. ◇

In the general setting of Theorem 2, one gets a stronger result assuming periodicity. We require the following lemma, where we use the common shorthand $\varrho = (w_0, w_1)$ for the substitution from Eq. (3).

Lemma 2 *Let ϱ be a primitive, binary constant-length substitution that defines a periodic hull. Then, one either has $\varrho = (w,w)$ with w containing at least one copy each of the letters a and b, or ϱ is bijective, and of the form $\varrho = \bigl((ab)^m a, (ba)^m b\bigr)$ or $\varrho = \bigl((ba)^m b, (ab)^m a\bigr)$ for some $m \in \mathbb{N}$.*

Proof Clearly, any substitution $\varrho = (w, w)$ with $|w| > 1$ defines a periodic hull, and primitivity implies that w must contain both letters. Consequently, we can now focus on $\varrho = (w_0, w_1)$ with $w_0 \neq w_1$. Let us begin with the cases of equal prefix.

If $\varrho = (rus, rvs)$, where $|u| = |v|$ and arbitrary $r, s \in \Sigma_2$, we may as well consider $\varrho' = (sru, srv)$, because ϱ and ϱ' are conjugate and thus define the same hull [8, Prop. 4.6]. Since we only consider the case $w_0 \neq w_1$, we must have at least one position where they differ, and we may assume that this happens at the last position.

For $\varrho = (aua, avb)$, the words ab and ba are both legal (as u must contain the letter b by primitivity), and an iteration of the corresponding seeds under ϱ results in the sequences

$$\begin{aligned} a|b &\mapsto \ldots a|a \ldots \mapsto \ldots a|a \ldots \mapsto \cdots \\ b|a &\mapsto \ldots b|a \ldots \mapsto \ldots b|a \ldots \mapsto \cdots \end{aligned}$$

that converge to two-sided fixed points. Since they are proximal (equal to the right, but not to the left) by construction, the hull of ϱ must be aperiodic [8, Cor. 4.2]. A completely analogous argument works for $\varrho = (bua, bvb)$, which is again aperiodic.

Likewise, for $\varrho = (aub, ava)$, the word ab is legal, hence also ba. Using the latter as seed, we get the iteration

$$b|a \;\mapsto\; \ldots a|a \ldots \;\mapsto\; \ldots b|a \ldots \;\mapsto\; \cdots$$

that ultimately alternates between two elements that form a proximal pair, which implies aperiodicity. Analogously, for $\varrho = (bub, bva)$, we get a proximal pair (and hence aperiodicity) from an iteration that starts with the legal seed $b|a$, which is mapped to $a|b$ and then alternates between $b|b$ and $a|b$.

Consequently, periodic cases for $w_0 \neq w_1$ can only occur if the two words have unequal prefix *and* unequal suffix. When $\varrho = (aub, bva)$, the seed $b|a$ is legal, which under iteration alternates between $a|a$ and $b|a$; when $\varrho = (bua, avb)$, one has the matching situation with $a|b$ and $a|a$, so both cases possess proximal pairs and are thus aperiodic.

Finally, if $\varrho = (aua, bvb)$, one gets a proximal pair if and only if aa or bb is legal, and the same statement applies to $\varrho' = (bub, ava)$. The only way this can be avoided is if w_0 and w_1 both alternate between a and b, which indeed gives periodic hulls, and these substitutions are the two other cases stated. □

Theorem 3 *If the primitive, binary constant-length substitution ϱ defines a periodic hull, the extremal Lyapunov exponents satisfy $\chi^B_{\min} = \chi^B_{\max} = 0$.*

Proof In view of Lemma 2, we have to check the claim for three cases. When $\varrho = (w, w)$, where w contains both letters and $L = |w| \geqslant 2$, we consider an arbitrary starting vector $v = (\alpha, \beta) \neq 0$. For $n > 1$, one then has

$$v B^{(n)}(k) \;=\; \bigl(\alpha S_0\bigl(\mathrm{e}^{2\pi \mathrm{i} k}\bigr) + \beta S_1\bigl(\mathrm{e}^{2\pi \mathrm{i} k}\bigr)\bigr) \cdot (1, 1)\, B^{(n-1)}(Lk),$$

where $(1, 1)$ is a left eigenvector of $B^{(n-1)}(Lk)$. Since $S_0 + S_1 = p_L$ in this case, one finds

$$\|vB^{(n)}(k)\| \;=\; \bigl|\alpha S_0\bigl(\mathrm{e}^{2\pi\mathrm{i}k}\bigr) + \beta S_1\bigl(\mathrm{e}^{2\pi\mathrm{i}k}\bigr)\bigr|\, \|(1,1)\| \prod_{\ell=1}^{n-1} \bigl|p_L\bigl(\mathrm{e}^{2\pi\mathrm{i}L^{\ell}k}\bigr)\bigr|.$$

The first term on the right-hand side only vanishes at isolated (and hence countably many) values of k, which we may exclude. Then, a calculation with Birkhoff averages shows that, for almost every $k \in \mathbb{R}$, we get

$$\lim_{n\to\infty} \frac{1}{n} \log \|vB^{(n)}(k)\| \;=\; \mathfrak{m}(p_L) \;=\; 0,$$

which establishes the claim in this case.

If ϱ is bijective, we have $L = 2m + 1$ for the two remaining cases by Lemma 2. In line with our previous reasoning, the Fourier matrix is of the form

$$B(k) \;=\; \begin{pmatrix} Q(z) & R(z) \\ R(z) & Q(z) \end{pmatrix} \quad \text{with } z = \mathrm{e}^{2\pi\mathrm{i}k},$$

where the polynomials Q and R satisfy $Q(z) + R(z) = p_L(z) = 1 + z + \cdots + z^{2m}$, which is cyclotomic, so that $\mathfrak{m}(Q + R) = 0$. Also, due to the alternating structure of $\varrho(a)$ and $\varrho(b)$, one has $\bigl(Q - R\bigr)(z) = \pm\bigl(Q + R\bigr)(-z)$. This means that $Q - R$ is cyclotomic as well, and $\mathfrak{m}(Q - R) = 0$. Now, one sees that

$$(1, \pm 1)B^{(n)}(k) \;=\; (1, \pm 1) \prod_{\ell=0}^{L-1} (Q \pm R)\bigl(\mathrm{e}^{2\pi\mathrm{i}L^{\ell}k}\bigr)$$

which, for almost every $k \in \mathbb{R}$, gives the two exponents as $\mathfrak{m}(Q + R) = 0$ and $\mathfrak{m}(Q - R) = 0$ by a simple calculation as in Remark 3. This implies our claim for these two cases. □

The converse of Theorem 3 does not hold. For example, both the Thue–Morse and the period doubling substitutions, given in Examples 1 and 2, have $\chi^B_{\min} = \chi^B_{\max} = 0$; this means that the norm of the resulting vector after applying their respective cocycles to any starting vector v does not grow exponentially. However, one must be careful here, as zero Lyapunov exponents do not exclude sub-exponential growth behaviour.

5 Examples: From Polynomials to Substitutions

Theorem 1, and the more specific Theorem 2, allow one, given a binary constant-length substitution, to write down a polynomial whose logarithmic Mahler measure determines the maximal Lyapunov exponent related to the substitution. But the nature of our results allows one to go the other way as well, as we now do. We explain how, given a specific polynomial, one can build a substitution associated with the maximal Lyapunov exponent for the cocycle $B^{(n)}(k)$. We also comment on the essential uniqueness of these substitutions and the properties of their Fourier matrices. We focus on specific classes of height-1 polynomials that have been important in the study of the logarithmic Mahler measure in the context of Lehmer's problem.

Example 3 (Littlewood polynomials) Recall that a polynomial q of degree $n-1$, defined by $q(z) = \sum_{m=0}^{n-1} c_m z^m$ with coefficients $c_m \in \{-1, 1\}$, is called a *Littlewood polynomial of order* $n-1$; see [18, 20, 39]. As before, let $\mathcal{C}_m$ be the m^{th} column of ϱ. Starting with the polynomial, we choose $\mathcal{C}_m$ to be

$$\mathcal{C}_m = \begin{cases} \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right], & \text{if } c_m = 1, \\ \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right], & \text{if } c_m = -1, \end{cases}$$

and we build the substituted words for 0 and 1 by looking at the concatenation $\mathcal{C}_0 \, \mathcal{C}_1 \cdots \mathcal{C}_{L-1}$. Since there are only two possible column types, we see immediately that the sets P_a and P_b from Lemma 1 satisfy

$$P_a \cup P_b = \{0, 1, \dots, L-1\},$$

and also that the resulting substitution ϱ must be bijective. By construction, we have

$$\chi^B_{\max} = \mathfrak{m}(Q - R) = \mathfrak{m}(q) \tag{6}$$

for the cocycle defined by the Fourier matrix associated with ϱ, which explicitly reads

$$B(k) = \begin{pmatrix} Q(z) & R(z) \\ R(z) & Q(z) \end{pmatrix} \quad \text{where } z = \mathrm{e}^{2\pi \mathrm{i} k}.$$

Note that, in this case, $\chi^B_{\max}$ can also be calculated by observing that $(1, -1)$ is a k-independent left eigenvector of $B(k)$ with eigenvalue $(Q-R)\big(\mathrm{e}^{2\pi \mathrm{i} k}\big)$, thus also giving (6).

The substitution corresponding to $q = Q - R$ is essentially unique, up to the obvious freedom that emerges from the relation $\mathfrak{m}(-q) = \mathfrak{m}(q)$, that is, from changing all signs. This is the case because a given sequence of signs uniquely specifies the columns of ϱ. For example, let us consider the integer polynomial $q(z) = -1 - z + z^2 - z^3 + z^4$, hence we get the substitutions

$$\varrho_q : \begin{cases} 0 \mapsto 11010 \\ 1 \mapsto 00101 \end{cases} \quad \text{and} \quad \varrho_{-q} : \begin{cases} 0 \mapsto 00101 \\ 1 \mapsto 11010 \end{cases}$$

with associated Fourier matrices

$$B_q(k) = \begin{pmatrix} \mathrm{e}^{4\pi \mathrm{i} k} + \mathrm{e}^{8\pi \mathrm{i} k} & 1 + \mathrm{e}^{2\pi \mathrm{i} k} + \mathrm{e}^{6\pi \mathrm{i} k} \\ 1 + \mathrm{e}^{2\pi \mathrm{i} k} + \mathrm{e}^{6\pi \mathrm{i} k} & \mathrm{e}^{4\pi \mathrm{i} k} + \mathrm{e}^{8\pi \mathrm{i} k} \end{pmatrix} \quad \text{and} \quad B_{-q}(k) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} B_q(k).$$

Both induce a cocycle whose maximum Lyapunov exponent is

$$\chi^{B}_{\max} \;=\; \mathfrak{m}(q) \;\approx\; 0.656256. \qquad\qquad \Diamond$$

The Fourier matrices associated with bijective substitutions enjoy further properties such as simultaneous diagonalisibility and a k-independent expression for the Oseledec splitting. In these cases, one always has $\mathcal{V}_1 = \mathbb{C}(1, 1)$. This means that, for all vectors $v \neq 0$ in this subspace, the asymptotic exponential growth rate is 0, for almost every $k \in \mathbb{R}$. One also sees that every column sum and every row sum of $B(k)$ is the cyclotomic polynomial p_L; for rows it is due to the bijectivity of the substitution, for columns it follows from the constant-length property.

Before we continue, let us mention that adding coincident positions to a given constant-length substitution as prefix or suffix does not change the Lyapunov exponents. Conversely, any primitive, binary constant-length substitution either starts and ends with bijective positions, or can be conjugated into a substitution that either has coincident prefix or suffix positions, but not both. The period doubling case is an example of this. To avoid pathologies, we now restrict our attention to substitutions which do not end with a coincidence.

Example 4 (Newman polynomials) For the class of $\{0, 1\}$-polynomials with constant term 1, also known as *Newman polynomials* [39], one has $R = 0$, so the associated Fourier matrix is

$$B(k) \;=\; \begin{pmatrix} \bigl(S_0 + Q\bigr)(z) & S_0(z) \\ S_1(z) & \bigl(S_1 + Q\bigr)(z) \end{pmatrix} \quad \text{with } z = \mathrm{e}^{2\pi \mathrm{i} k},$$

which leads to $\chi^{B}_{\max} = \mathfrak{m}(Q)$ by Theorem 2. If either S_0 or S_1 is zero, $M_{\varrho} = B(0)$ is a triangular matrix, and cannot be primitive. This can be avoided if at least two coefficients of the polynomial are zero. If there is only one, we can still construct a primitive substitution by recalling that $\mathfrak{m}(-q) = \mathfrak{m}(q)$, so we only need to exchange the two bijective column types.

As a concrete example, consider $q(z) = 1 + z^2$. The two standard choices

$$\varrho_q : \begin{cases} 0 \mapsto 000 \\ 1 \mapsto 101 \end{cases} \qquad \text{and} \qquad \varrho_{q'} : \begin{cases} 0 \mapsto 010 \\ 1 \mapsto 111 \end{cases}$$

both give non-primitive substitutions; in fact, their substitution matrices are not even irreducible. However,

$$\varrho_{-q} : \begin{cases} 0 \mapsto 101 \\ 1 \mapsto 000 \end{cases} \qquad \text{and} \qquad \varrho_{-q'} : \begin{cases} 0 \mapsto 111 \\ 1 \mapsto 010 \end{cases}$$

are both primitive and aperiodic, and have $\chi^{B}_{\max} = \mathfrak{m}(q)$. One can see in this example that replacing q by $-q$ really means an exchange of w_0 and w_1 in the definition of ϱ.

As another example, consider the reciprocal Newman polynomial

$$q(z) \;=\; 1+z^{3}+z^{4}+z^{5}+z^{6}+z^{7}+z^{8}+z^{9}+z^{10}+z^{11}+z^{14}$$

taken from [21, p. 1375]. One choice for a substitution (with $L = 15$) is

$$\varrho_q : \begin{cases} 0 \mapsto 010000000000000 \\ 1 \mapsto 110111111111001 \end{cases}$$

which means $S_1(z) = z$ and $S_0(z) = z^2 + z^{12} + z^{13}$, together with $Q = q$. Here, $\mathfrak{m}(q) \approx \log(1.265122)$. Note that this value is strictly smaller than $\log(\lambda_{\mathrm{p}})$, where $\lambda_{\mathrm{p}} \approx 1.324718$ is the *plastic number* described earlier. Recall that $\log(\lambda_{\mathrm{p}})$ is the sharp lower bound for $\mathfrak{m}(p)$ over all non-reciprocal integer polynomials p that are not a product of a monomial with a cyclotomic polynomial. ◊

Note that, when associating a polynomial to a binary constant-length substitution ϱ, it is only the bijective columns of ϱ that are determined by the non-zero coefficients. Thus, we can extend the construction to generic height-1 polynomials, even when the constant term is zero, as in the period doubling example. However, when interested in non-trivial logarithmic Mahler measures, one can assume that the constant coefficient is non-zero.

Example 5 (Borwein polynomials) When considering Borwein polynomials, one can choose the columns of an associated substitution just as in Example 3, but with the additional freedom to vary the choice for each zero coefficient. As in Example 3, starting with the polynomial, we choose $\mathcal{C}_m$ to be

$$\mathcal{C}_m \;=\; \begin{cases} \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right], & \text{if } c_m = 1, \\ \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right], & \text{if } c_m = -1, \\ \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right] \text{ or } \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], & \text{if } c_m = 0, \end{cases}$$

and we build the substituted words for 0 and 1 by looking at the concatenation $\mathcal{C}_0\,\mathcal{C}_1 \cdots \mathcal{C}_{L-1}$. But now, there are two choices for each zero coefficient of the polynomials, so that, if p is a Borwein polynomial of degree $L-1$ with n zero coefficients, there are 2^n distinct binary constant-length substitutions of length L whose maximal Lyapunov exponents are all given by $\mathfrak{m}(p)$. On top of this freedom, we can also still work both with p or with $-p$, as we saw earlier.

As a concrete example, we consider Lehmer's polynomial ℓ_{L} from (4), where $c_2 = c_8 = 0$. Recall from the Introduction that this polynomial is irreducible, and has precisely one root outside the unit disk. This root is real and a Salem number. Recall further that ℓ_{L} is the polynomial with the smallest known positive logarithmic Mahler measure, $\mathfrak{m}(\ell_{\mathrm{L}}) \approx \log(1.176281)$. Here,

$$\varrho_{\ell_{\mathrm{L}}} : \begin{cases} 0 \mapsto 00111111000 \\ 1 \mapsto 11100000011 \end{cases}$$

is one of the eight substitutions that correspond to the polynomial ℓ_{L}. ◇

With this connection and representation, we obtain the following equivalent reformulation of the strong version of Lehmer's problem for Borwein polynomials.

Question 1 Does there exist a primitive, binary constant-length substitution ϱ with maximal Lyapunov exponent $0 < \chi^{B}_{\max} < \mathfrak{m}(\ell_{\mathrm{L}}) \approx \log(1.176280818)$?

6 Extensions and Outlook

Lyapunov exponents are neither restricted to constant-length substitutions nor to binary alphabets. In fact, there are many extensions possible; see [4, 6] and references therein for more. In general, however, the Lyapunov exponents are no longer logarithmic Mahler measures themselves, though they often still satisfy interesting estimates in such a setting.

Moreover, there is actually also a generalisation to higher dimensions, as briefly stated in [36]. Here, one considers stone inflation rules of finite local complexity [8] and selects a suitable marker (or reference point) in each prototile (such that the tiling and the point set are mutually locally derivable [8, Sec. 5.2] from each other). One particular class emerges from *block substitutions*, as those discussed in [7, 31].

Example 6 A simple bijective example is given by

$$a \;\mapsto\; \begin{matrix} b & a \\ a & b \end{matrix}\;, \quad b \;\mapsto\; \begin{matrix} a & b \\ b & a \end{matrix}$$

which is primitive and aperiodic. Here, one can represent the two letters by unit squares with a (coloured) reference point at their lower left corners. Then, with $\underline{k} = (k_1, k_2) \in \mathbb{R}^2$, one finds the Fourier matrix

$$B(\underline{k}) \;=\; \begin{pmatrix} 1+xy & x+y \\ x+y & 1+xy \end{pmatrix} \quad \text{where } (x, y) = \bigl(\mathrm{e}^{2\pi \mathrm{i} k_1}, \mathrm{e}^{2\pi \mathrm{i} k_2}\bigr),$$

which satisfies $(1, \pm 1) B(\underline{k}) = \bigl((1+xy) \pm (x+y)\bigr)(1, \pm 1)$. Since we also have $1 + x + y + xy = (1+x)(1+y)$ and $1 - x - y + xy = (1-x)(1-y)$, all factors are cyclotomic. Consequently, for almost every $\underline{k} \in \mathbb{R}^2$, one gets the Lyapunov exponents as

$$\chi^{B}_{\min} \;=\; \mathfrak{m}(1 + x + y + xy) = 0 \quad \text{and} \quad \chi^{B}_{\max} \;=\; \mathfrak{m}(1 - x - y + xy) = 0,$$

which resembles Example 1 in many ways. Here, in line with Eq. (1), the logarithmic Mahler measure of a multivariable polynomial p is defined as

$$\mathfrak{m}(p) \;= \int_{[0,1]^d} \log \bigl| p\bigl(\mathrm{e}^{2\pi \mathrm{i} t_1}, \ldots, \mathrm{e}^{2\pi \mathrm{i} t_d}\bigr)\bigr| \, \mathrm{d}t_1 \cdots \mathrm{d}t_d .$$

As another example, consider the bijective block substitution

$$a \mapsto \begin{matrix} b & a & b \\ a & a & a \\ b & a & b \end{matrix}\, , \quad b \mapsto \begin{matrix} a & b & a \\ b & b & b \\ a & b & a \end{matrix}$$

that emerges from the *squiral tiling* [7]. Here, one has

$$\chi^{B}_{\min} \; = \; \mathfrak{m}\big((1+x+x^2)(1+y+y^2)\big) = 0$$

as before, while

$$\chi^{B}_{\max} \; = \; \mathfrak{m}\big(x + y(1+x+x^2) + x\,y^2 - (1+x^2)(1+y^2)\big) \; \approx \; 0.723909$$

is strictly positive. ♢

It is clear that one can now repeat a lot of our previous analysis for the class of bijective block substitutions, in any dimension. As is implicit in [36], the blocks need not be cubes, as long as they have length at least 2 in each direction. We leave further experimentation along these lines to the interested reader. Outside the bijective class, interesting new phenomena are possible as follows.

Example 7 Consider the binary block substitution

$$a \mapsto \begin{matrix} b & a \\ b & b \end{matrix}\, , \quad b \mapsto \begin{matrix} a & a \\ a & a \end{matrix}$$

which is clearly primitive. It has a coincidence, so that the higher-dimensional analogue of Dekking's result, see [7, 13, 31], implies the pure point spectral nature of the corresponding dynamical system (under the action of the $\mathbb{Z}^2$-shift).

Here, the Fourier matrix reads

$$B(\underline{k}) \; = \; \begin{pmatrix} x\,y & (1+x)(1+y) \\ 1+x+y & 0 \end{pmatrix}, \quad \text{where } (x,y) = \big(\mathrm{e}^{2\pi \mathrm{i} k_1}, \mathrm{e}^{2\pi \mathrm{i} k_2}\big)$$

and $\det\big(B(\underline{k})\big) = -(1+x)(1+y)(1+x+y)$. As before, we get

$$\chi^{B}_{\min} \; = \; \mathfrak{m}\big((1+x)(1+y)\big) = 0,$$

while the sum then satisfies

$$\begin{aligned}
\chi^{B}_{\min} + \chi^{B}_{\max} \; &= \; \chi^{B}_{\max} \; = \; \mathfrak{m}\bigl((1+x)(1+y)(1+x+y)\bigr) \\
&= \; \mathfrak{m}(1+x+y) \; = \; \frac{3\sqrt{3}}{4\pi}\, L(2, \chi_{-3}) \; = \; L'(-1, \chi_{-3}) \\
&= \; 2 \int_{0}^{1/3} \log\bigl(2\cos(\pi t)\bigr)\, \mathrm{d}t \; \approx \; 0.323066,
\end{aligned}$$

with $L(s, \chi_{-3})$ denoting the Dirichlet L-function of the character $\chi_{-3}(n) = \bigl(\frac{-3}{n}\bigr)$, written in terms of the Legendre symbol; compare [24]. This special value of a Mahler measure also appears in [44, 45], and various other relations of this kind are known [17], such as

$$\mathfrak{m}(1+x+y+z) \; = \; \frac{7}{2\pi^{2}}\, \zeta(3).$$

The latter can be realised by a block substitution in three dimensions. ♢

Remark 4 It seems hardly known that the Mahler measure from Example 7, together with its integral representation, first appeared in Wannier's calculation of the ground state entropy of the antiferromagnetic Ising model on the triangular lattice [47]. This might be due to the fact that the numerical value originally given by him was incorrect, though corrected in an erratum 23 years later. Somewhat similar entropy calculations in terms of logarithmic integrals later appeared in various other papers on solvable models of statistical physics.

To be more precise, Wannier gets the entropy s as a double integral from which we obtain

$$\begin{aligned}
s \; &= \; \frac{1}{2} \int_{0}^{1} \int_{0}^{1} \log\bigl(1 + 4\cos(2\pi v)^{2} - 4\cos(2\pi u)\cos(2\pi v)\bigr)\, \mathrm{d}u\, \mathrm{d}v \\
&= \; \frac{1}{2} \int_{0}^{1} \int_{0}^{1} \log \bigl|(1+y^{2}-xy)(x+xy^{2}-y)\bigr|_{x=\mathrm{e}^{2\pi \mathrm{i} u},\, y=\mathrm{e}^{2\pi \mathrm{i} v}}\, \mathrm{d}u\, \mathrm{d}v \\
&= \; \frac{1}{2}\bigl(\mathfrak{m}(1+y^{2}-xy) + \mathfrak{m}(1+y^{2}-x^{-1}y)\bigr) \; = \; \mathfrak{m}(1+y^{2}-xy).
\end{aligned}$$

Now, one clearly has

$$\mathfrak{m}(1+y^{2}-xy) \; = \; \mathfrak{m}(1+y^{2}+xy) \; = \; \mathfrak{m}(1+x+y),$$

via a change of variables $u = u' + \frac{1}{2}$ for the first identity and the invariance of the Mahler measure under an invertible linear map with integer coefficients, as detailed in [30, Exc. 3.1], which is given by the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 2 \end{smallmatrix}\right)$ in this case. ♢

It is well known that the connection between the Mahler measure and special values of L-series has deep roots; in particular, see [23, 26], and [17] for a survey with several examples and references. Also, a connection between Mahler measures and Lyapunov exponents is known from [27]. On the other hand, our observation shows that these quantities occur in the spectral theory of dynamical systems as well, in a rather elementary way, and it seems an interesting problem to analyse this connection further.

Acknowledgements It is our pleasure to thank Franz Gähler, Michael J. Mossinghoff, Dan Rust and Tom Ward for discussions, and an anonymous referee for a number of helpful suggestions. This work was supported by the German Research Foundation (DFG) through CRC 1283. The second author additionally thanks both the University of Bielefeld and the Alfréd Rényi Institute of the Hungarian Academy of Sciences for hosting him during this research.

References

1. Adamczewski, B., Bugeaud, Y.: On the complexity of algebraic numbers. I. Expansions in integer bases. Ann. Math. **165**, 547–565 (2007). arXiv:math/0511674
2. Adamczewski, B., Cassaigne, J.: Diophantine properties of real numbers generated by finite automata. Compos. Math. **142**, 1351–1372 (2006). arXiv:math/0601604
3. Adamczewski, B., Faverjon, C.: Méthode de Mahler: relations linéaires, transcendance et applications aux nombres automatiques. Proc. Lond. Math. Soc. (3) **115**, 55–90 (2017). arXiv:1508.07158
4. Baake, M., Frank, N.P., Grimm, U., Robinson, E.A.: Geometric properties of a binary non-Pisot inflation and absence of absolutely continuous diffraction. Studia Math. **247**, 109–154 (2019). arXiv:1706.03976
5. Baake, M., Gähler, F.: Pair correlations of aperiodic inflation rules via renormalisation: some interesting examples. Topol. Appl. **205**, 4–27 (2016). arXiv:1511.00885
6. Baake, M., Gähler, F., Mañibo, N.: Renormalisation of pair correlation measures for primitive inflation rules and absence of absolutely continuous diffraction. Commun. Math. Phys. **370**, 591–635 (2019). arXiv:1805.09650
7. Baake, M., Grimm, U.: Squirals and beyond: substitution tilings with singular continuous spectrum. Ergodic Theory Dyn. Syst. **34**, 1077–1102 (2014). arXiv:1205.1384
8. Baake, M., Grimm, U.: Aperiodic Order. Vol. 1: A Mathematical Invitation. Cambridge University Press, Cambridge (2013)
9. Baake, M., Grimm, U.: Diffraction of a binary non-Pisot inflation tiling. J. Phys. Conf. Ser. **809**, 012026 (4pp) (2017). arXiv:1706.04448
10. Baake, M., Grimm, U., Mañibo, N.: Spectral analysis of a family of binary inflations rules. Lett. Math. Phys. **108**, 1783–1805 (2018). arXiv:1709.09083
11. Baake, M., Haynes, A., Lenz, D.: Averaging almost periodic functions along exponential sequences. In: Baake, M., Grimm, U., (eds.), Aperiodic Order. Vol. 2. Crystallography and Almost Periodicity, pp. 343–362. Cambridge University Press, Cambridge (2017). arXiv:1704.08120
12. Barreira, L., Pesin, Y.: Nonuniform Hyperbolicity. Cambridge University Press, Cambridge (2007)
13. Bartlett, A.: Spectral theory of substitutions in $\mathbb{Z}^d$. Ergodic Theory Dyn. Syst. **38**, 1289–1341 (2018). arXiv:1410.8106
14. Bell, J.P., Bugeaud, Y., Coons, M.: Diophantine approximation of Mahler numbers. Proc. Lond. Math. Soc. (3) **110**, 1157–1206 (2015). arXiv:1307.4123
15. Bell, J.P., Coons, M., Hare, K.G.: The minimal growth of a k-regular sequence. Bull. Austral. Math. Soc. **90**, 195–203 (2014). arXiv:1410.5517
16. Bell, J.P., Coons, M., Hare, K.G.: Growth degree classification for finitely generated semigroups of integer matrices. Semigroup Forum **92**, 23–44 (2016). arXiv:1410.5519
17. Bertin, M.-J., Lalín, M.: Mahler measure of multivariable polynomials. In: David, C., Lalín, M., Manes, M., (eds.), Women in Numbers 2. Research Directions in Number Theory. Contemporary Mathematics, vol. 606, pp. 125–147. American Mathematical Society, Providence

(2013); revised version available at http://www.dms.umontreal.ca/~mlalin/surveyMahlerfinal-revised.pdf

18. Borwein, P., Choi, S., Jankauskas, J.: Extremal Mahler measures and L_s norms of polynomials related to Barker sequences. Proc. Am. Math. Soc. **141**, 2653–2663 (2013)
19. Borwein, P., Hare, K.G., Mossinghoff, M.J.: The Mahler measure of polynomials with odd coefficients. Bull. Lond. Math. Soc. **36**, 332–338 (2004)
20. Borwein, P., Mossinghoff, M.J.: Barker sequences and flat polynomials. In: McKee, J., Smyth, C. (eds.) Number Theory and Polynomials, pp. 71–88. Cambridge University Press, New York (2008)
21. Boyd, D.W.: Reciprocal polynomials having small measure. Math. Comput. **35**, 1361–1377 (1980)
22. Boyd, D.W.: Reciprocal polynomials having small Mahler measures. II. Math. Comput. **53**, 355–357 (1989)
23. Boyd, D.W.: Mahler's measure and special values of L-functions. Exp. Math. **7**, 37–82 (1998)
24. Boyd, D.W., Mossinghoff, M.J.: Small limit points of Mahler's measure. Exp. Math. **15**, 403–414 (2005)
25. Breusch, R.: On the distribution of the roots of a polynomial with integral coefficients. Proc. Am. Math. Soc. **2**, 939–941 (1951)
26. Deninger, C.: Deligne periods of mixed motives, K-theory and the entropy of certain $\mathbb{Z}^d$-actions. J. Am. Math. Soc. **10**, 259–281 (1997)
27. Deninger, C.: Determinants on von Neumann algebras, Mahler measures and Ljapunov exponents. J. Reine Angew. Math. (Crelle) **651**, 165–185 (2011). arXiv:0712.0667
28. Drungilas, P., Jankauskas, J., Šiurys, J.: On Littlewood and Newman multiples of Borwein polynomials. Math. Comput. **87**, 1523–1541 (2018). arXiv:1609.07295
29. El Otmani, S., Rhin, G., Sac-Épée, J.-M.: Finding new limit points of Mahler's measure by genetic algorithms. Exp. Math. **28**, 129–131 (2019)
30. Everest, G., Ward, T.: Heights of Polynomials and Entropy in Algebraic Dynamics. Springer, London (1999)
31. Frank, N.P.: Multi-dimensional constant-length substitution sequences. Topol. Appl. **152**, 44–69 (2005)
32. Kronecker, L.: Zwei Sätze über Gleichungen mit ganzzahligen Coefficienten. J. Reine Angew. Math. (Crelle) **53**, 173–175 (1857)
33. Lehmer, D.H.: Factorization of certain cyclotomic functions. Ann. Math. **34**, 461–479 (1933)
34. Lind, D., Schmidt, K., Ward, T.: Mahler measure and entropy for commuting automorphisms of compact groups. Invent. Math. **101**, 593–629 (1990)
35. Mañibo, N.: Lyapunov exponents for binary substitutions of constant length. J. Math. Phys. **58**, 113504 (9 pp.) (2017). arXiv:1706.00451
36. Mañibo, N.: Spectral analysis of primitive inflation rules. Oberwolfach Rep. **14**, 2830–2832 (2017)
37. Mossinghoff, M.J.: Algorithms for the determination of polynomials with small Mahler measure. Ph.D. thesis, The University of Texas at Austin (1995)
38. Mossinghoff, M.J.: Polynomials with small Mahler measure. Math. Comput. **67**, 1697–1705 (1998)
39. Mossinghoff, M.J.: Polynomials with restricted coefficients and prescribed noncyclotomic factors. LMS J. Comput. Math. **6**, 314–325 (2003)
40. Mossinghoff, M.J., Rhin, G., Wu, Q.: Minimal Mahler measures. Exp. Math. **17**, 451–458 (2008)
41. Pathiaux, M.: Sur les multiples de polynômes irréductibles associés à certains nombres algébriques. Sém. Delange–Pisot–Poitou. Théor. Nombres **14**, 315–324 (1972/73)
42. Schmidt, K.: Dynamical Systems of Algebraic Origin. Birkhäuser, Basel (1995)
43. Smyth, C.J.: On the product of the conjugates outside the unit circle of an algebraic integer. Bull. Lond. Math. Soc. **3**, 169–175 (1971)
44. Smyth, C.J.: On measures of polynomials in several variables. Bull. Austral. Math. Soc. **23**, 49–63 (1981)

45. Smyth, C.: The Mahler measure of algebraic numbers: a survey. In: McKee, J., Smyth, C., (eds.) Number Theory and Polynomials, pp. 322–349. Cambridge University Press, New York (2008)
46. Viana, M.: Lectures on Lyapunov Exponents. Cambridge University Press, Cambridge (2013)
47. Wannier, G.H.: Antiferromagnetism. The triangular Ising net, Phys. Rev. **79**, 357–364 (1950); and Phys. Rev. B **7**, 5017 (1973) (Erratum)

The Borwein Brothers, Pi and the AGM

Richard P. Brent

In fond memory of Jonathan M. Borwein 1951–2016

1 Introduction

Jonathan Borwein was fascinated by the constant π, and gave many stimulating talks on this topic. The slides for most of these talks may be found on the memorial website [11]. In my talk [22] at the Jonathan Borwein Commemorative Conference I discussed the reasons for this fascination. In a nutshell, it is that theorems about π are often just the tips of "mathematical icebergs"—much of interest lies hidden beneath the surface.

This paper considers some of Jonathan and Peter Borwein's contributions to the high-precision computation of π and the elementary functions log, exp, arctan, sin, etc. The material is mainly drawn from their fascinating book *Pi and the AGM* [14]. We make no attempt to review the whole book—a reader interested in the complete contents should consult one of the reviews [2, 3, 9, 48] or, better, read the book itself. We do not try to distinguish between the contributions of Jonathan and his brother Peter—so far as we know, they contributed equally to the book, although no doubt in different ways.

We take the opportunity to present some new results that are related to the material in *Pi and the AGM*. For example, the error after a finite number of iterations

R. P. Brent (✉)
Mathematical Sciences Institute, Australian National University,
Canberra, Australia
e-mail: JBCC@rpbrent.com

School of Mathematical and Physical Sciences, The University of Newcastle,
Newcastle, Australia

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_21

of some of the quadratically and quartically convergent algorithms for π can be expressed succinctly in terms of theta functions. Inspection of these expressions suggests that some algorithms, previously considered different, are actually equivalent, in the sense that they give exactly the same sequence of approximations to π if performed using exact arithmetic. For example, one of the Borweins' quadratically convergent algorithms [14, Iteration 5.2 with $r = 4$] is equivalent to the Gauss–Legendre algorithm [18, 20, 42], and it follows that one step of the Borweins' quartically convergent algorithm [14, Iteration 5.3] is equivalent to two steps of the Gauss–Legendre algorithm. These connections between superficially different algorithms do not seem to have been noticed before.

In Sect. 2 we give some necessary definitions, discuss the *arithmetic–geometric mean* and consider its connection with elliptic integrals and Jacobi theta functions. We also mention the concept of *order of convergence* of an algorithm.

A brief history of quadratically convergent algorithms for π is given in Sect. 3.

In Sect. 4 we consider some quadratically and quartically convergent algorithms for π, including the Gauss–Legendre algorithm and several algorithms due to the Borweins. In Sect. 5 we show that some of the algorithms of Sect. 4, although superficially different, are actually equivalent when performed with exact arithmetic.

Chapter 5 of *Pi and the AGM* considers some striking *Ramanujan–Sato* formulæ for $1/\pi$ that give very fast (though linearly convergent) algorithms for computing π. The first such formulæ were given by Ramanujan [40]. Later authors include Takeshi Sato, the Borwein brothers and the Chudnovsky brothers. See [6, 7, 15] for references. In Sect. 6 we briefly consider some Ramanujan–Sato formulæ and the corresponding algorithms for computing π.

One of the “icebergs” alluded to above is the fast computation of elementary functions to arbitrary precision. The constant $\pi = 4 \arctan(1)$ is of course just a special case (the tip of the iceberg). In Sect. 7 we outline how fast algorithms for computing elementary (and some other) functions can be based on the arithmetic–geometric mean iteration.

2 Preliminaries: Means, Elliptic Integrals and Theta Functions

We define the *order of convergence* of a sequence. It will be sufficient to say that a sequence $(x_n)_{n\in\mathbb{N}}$ *converges linearly* to L (or with *order of convergence* 1) if

$$0 < \mu_0 = \liminf_{n\to\infty} \frac{|x_{n+1} - L|}{|x_n - L|} \le \limsup_{n\to\infty} \frac{|x_{n+1} - L|}{|x_n - L|} = \mu_1 < 1.$$

If $\mu_0 = \mu_1$ then μ_0 is called the *rate of convergence*.

We say that a sequence $(x_n)_{n\in\mathbb{N}}$ *converges to L with order $p > 1$* if the sequence converges to L and there exists

$$p = \lim_{n\to\infty} \frac{\log |x_{n+1} - L|}{\log |x_n - L|} > 1.$$

Quadratic, cubic and *quartic* convergence are the cases $p = 2, 3, 4$, respectively. For example, if $x_n = 2^n \exp(-3^n)$, then $(x_n)_{n\in\mathbb{N}}$ converges cubically to zero, because $\log|x_{n+1}|/\log|x_n| = (-3^{n+1} + O(n))/(-3^n + O(n)) \to 3$ as $n \to \infty$.

Roughly speaking, if a sequence converges linearly to L with rate μ, then the number of correct decimal digits in the approximation to L increases by about $\log_{10}(1/\mu)$ per term. For example, if

$$x_n = 2\sqrt{3}\sum_{j=0}^{n}\frac{(-1)^j}{(2j+1)3^j}\,, \tag{1}$$

then x_n converges linearly to π with about $\log_{10} 3 \approx 0.4771$ decimal digits per term.[1] If a sequence converges to L with order $p > 1$, then the number of correct digits is approximately multiplied by p for each additional term. For example, Newton's method for computing square roots[2]

$$x_{n+1} := \frac{1}{2}\left(x_n + \frac{S}{x_n}\right)$$

converges quadratically to $L := \sqrt{S}$, provided that x_0 and S are positive. In fact, it is easy to show that

$$x_{n+1} - L \approx \frac{1}{2L}(x_n - L)^2.$$

We now consider some well-known means. The *arithmetic mean* of $a, b \in \mathbb{R}$ is

$$\mathrm{AM}(a, b) := \frac{a+b}{2}\,,$$

and the *geometric mean* is

$$\mathrm{GM}(a, b) := \sqrt{ab}. \tag{2}$$

Assuming that a and b are positive, we have the inequality

$$\mathrm{GM}(a, b) \le \mathrm{AM}(a, b).$$

Initially we assume that a, b are positive real numbers. In Sect. 7 we permit a, b to be complex. To resolve the ambiguity in the square root in (2) we assume that $\Re(GM(a, b)) \ge 0$, and $\Im(GM(a, b)) \ge 0$ if $\Re(GM(a, b)) = 0$.

Given two positive reals a_0, b_0, we can iterate the arithmetic and geometric means by defining, for $n \ge 0$,

[1] The formula (1) is listed in Bailey's compendium [5], and is attributed to Madhava of Sangamagramma (c.1340–c.1425). It follows from the Taylor series for $\arctan(1/\sqrt{3})$.

[2] Attributed to Hero of Alexandria (c.10–70 A.D.), though also called the *Babylonian method*.

$$a_{n+1} = \mathrm{AM}(a_n, b_n)$$
$$b_{n+1} = \mathrm{GM}(a_n, b_n).$$

The sequences (a_n) and (b_n) converge quadratically to a common limit called the *arithmetic–geometric mean* (AGM) of a_0 and b_0. We denote it by $\mathrm{AGM}(a_0, b_0)$.

Gauss [27] and Legendre [36] solved the problem of expressing $\mathrm{AGM}(a, b)$ in terms of known functions. The answer may be written as

$$\frac{1}{\mathrm{AGM}(a, b)} = \frac{2}{\pi} \int_0^{\pi/2} \frac{d\theta}{\sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta}} \cdot \tag{3}$$

The right-hand side of (3) is the product of a constant (whose precise value will be significant later) and a *complete elliptic integral of the first kind*. As usual, the complete elliptic integral of the first kind is defined by

$$K(k) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k^2 \sin^2 \theta}} = \int_0^1 \frac{dt}{\sqrt{(1 - t^2)(1 - k^2 t^2)}},$$

and the *complete elliptic integral of the second kind* by

$$E(k) := \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \theta}\, d\theta = \int_0^1 \frac{\sqrt{1 - k^2 t^2}}{\sqrt{1 - t^2}}\, dt.$$

The variable k is called the *modulus*, and $k' := \sqrt{1 - k^2}$ is called the *complementary modulus*. It is customary to define

$$K'(k) := K(\sqrt{1 - k^2}) = K(k')$$

and

$$E'(k) := E(\sqrt{1 - k^2}) = E(k'),$$

so in the context of elliptic integrals a prime (′) does *not* denote differentiation. On the occasions when we need a derivative, we use operator notation

$$\mathrm{D}_k K(k) := dK(k)/dk.$$

We remark that *Pi and the AGM* uses the "dot" notation $\dot{K}(k) := dK(k)/dk$, but this is potentially ambiguous and hard to see, so we prefer to avoid it.

The moduli k and k' can in general be complex, but unless otherwise noted we assume that they are real and in the interval (0, 1).

In terms of the Gaussian hypergeometric function

$$F(a, b; c; z) := 1 + \frac{a \cdot b}{1! \cdot c} z + \frac{a(a + 1) \cdot b(b + 1)}{2! \cdot c(c + 1)} z^2 + \cdots,$$

we have

$$K(k) = \frac{\pi}{2}\, F\left(\tfrac{1}{2}, \tfrac{1}{2}; 1; k^2\right) \tag{4}$$

and

$$E(k) = \frac{\pi}{2}\, F\left(-\tfrac{1}{2}, \tfrac{1}{2}; 1; k^2\right). \tag{5}$$

From (4) and [1, 17.3.21], we also have[3]

$$K'(k) = \frac{2}{\pi}\log\left(\frac{4}{k}\right) K(k) - f(k), \tag{6}$$

where $f(k) = k^2/4 + O(k^4)$ is analytic in the disk $|k| < 1$.

Substituting $(a, b) \mapsto (1, k')$ in (3), and recalling that $k^2 + (k')^2 = 1$, we have

$$\mathrm{AGM}(1, k') = \frac{\pi}{2K(k)}\cdot \tag{7}$$

Thus, if we start from $a_0 = 1$, $b_0 = k' \in (0, 1)$ and apply the AGM iteration, $K(k)$ can be computed from

$$\lim_{n\to\infty} a_n = \frac{\pi}{2K(k)}\cdot \tag{8}$$

$E(k)$ can be computed via the AGM at the same time as $K(k)$, using the well-known result [14, (b) on pg. 15]

$$\frac{E(k)}{K(k)} = 1 - \frac{k^2}{2} - \sum_{n=0}^{\infty} 2^n\,(a_n - a_{n+1})^2.$$

It follows from (4) and (6) that, for small k,

$$K'(k) = \left(1 + O(k^2)\right)\log\left(\frac{4}{k}\right). \tag{9}$$

This will be relevant in Sect. 7. A bound on the $O(k^2)$ term is given in [14, Thm. 7.2].

The Gauss–Legendre algorithm depends on *Legendre's relation*: for $0 < k < 1$,

$$E(k)K'(k) + E'(k)K(k) - K(k)K'(k) = \frac{\pi}{2}\cdot$$

For a proof, see *Pi and the AGM*, Sec. 1.6.

A computationally important special case, obtained by taking $k = k' = 1/\sqrt{2}$, is

$$\left(2E\left(1/\sqrt{2}\right) - K\left(1/\sqrt{2}\right)\right) K\left(1/\sqrt{2}\right) = \frac{\pi}{2}\cdot \tag{10}$$

[3] Here and elsewhere, log denotes the natural logarithm.

It can be shown [14, Thm. 1.7] that the two factors in (10) are

$$K(1/\sqrt{2}) = \frac{\Gamma^2\left(\frac{1}{4}\right)}{4\pi^{1/2}} \text{ and } 2E(1/\sqrt{2}) - K(1/\sqrt{2}) = \frac{\Gamma^2\left(\frac{3}{4}\right)}{\pi^{1/2}} \cdot$$

To estimate the order of convergence and to obtain error bounds, we consider the parameterisation of the AGM in terms of *Jacobi theta functions*. We need the basic theta functions of one variable, defined for $|q| < 1$ by

$$\theta_2(q) := \sum_{n\in\mathbb{Z}} q^{(n+1/2)^2},\ \theta_3(q) := \sum_{n\in\mathbb{Z}} q^{n^2},\ \theta_4(q) := \sum_{n\in\mathbb{Z}} (-1)^n q^{n^2}.$$

The theta functions satisfy many identities [47, §21.3]. In particular, we use the following addition formulæ, due to Jacobi [32]. They are proved in [14, §2.1].

$$\theta_3^2(q) = \theta_2^2(q^2) + \theta_3^2(q^2), \tag{11}$$

$$\theta_3^4(q) = \theta_2^4(q) + \theta_4^4(q). \tag{12}$$

It is not difficult to show that

$$\frac{\theta_3^2(q) + \theta_4^2(q)}{2} = \theta_3^2(q^2) \text{ and } \sqrt{\theta_3^2(q)\theta_4^2(q)} = \theta_4^2(q^2).$$

Thus, the AGM variables (a_n, b_n) can be parameterised by $(\theta_3^2(q^{2^n}), \theta_4^2(q^{2^n}))$ if scaled suitably. More precisely, if $1 = a_0 > b_0 = \theta_4^2(q)/\theta_3^2(q) > 0$, where $q \in (0, 1)$, then the variables a_n, b_n appearing in the AGM iteration satisfy

$$a_n = \frac{\theta_3^2(q^{2^n})}{\theta_3^2(q)},\ b_n = \frac{\theta_4^2(q^{2^n})}{\theta_3^2(q)} \cdot \tag{13}$$

It is useful to define auxiliary variables $c_{n+1} := a_n - a_{n+1} = (a_n - b_n)/2$. Using the quotient for a_n and the addition formula (11), we see that

$$c_n = \frac{\theta_2^2(q^{2^n})}{\theta_3^2(q)} \tag{14}$$

holds for $n \geq 1$. We could use (14) to define c_0, but this will not be necessary.[4]

We can write q (which is called the *nome*) explicitly, in fact

$$q = \exp(-\pi K'(k)/K(k)). \tag{15}$$

[4] Salamin [42] defines c_n using the relation $c_n^2 = a_n^2 - b_n^2$. This has the advantage that c_0 is defined naturally, and for $n > 0$ it is equivalent to our definition. However, it is computationally more expensive to compute $(a_n^2 - b_n^2)^{1/2}$ than $a_n - a_{n+1}$.

This is due to Gauss/Jacobi; for a proof see [14, Thm. 2.3]. In the important special case $k = k' = 1/\sqrt{2}$, we have $K' = K$ and $q = e^{-\pi} = 0.0432139\ldots$

Because the AGM iteration converges quadratically, it offers the prospect of quadratically convergent algorithms for approximating π and, more generally, all the elementary functions. This is the topic of Sects. 4 and 7 below. First, we make some comments on the history of quadratically convergent algorithms for π.

3 Historical Remarks

An algorithm for computing $\log(4/k)$, using (7), (9) and the AGM, assuming that we know π to sufficient accuracy, was given by Salamin [8, pg. 71] in 1972. On the same page Salamin gives an algorithm for computing π, taking $k = 4/e^n$ in (9). With his choice $\pi \approx 2n\,\mathrm{AGM}(1, k)$. However, this assumes that we know e, so it is not a "standalone" algorithm for π via the AGM. Similarly, if we take $k = 4/2^n$ in (9), we obtain an algorithm for computing $\pi \log 2$ (and hence π, if we know $\log 2$).

In 1975, Salamin [42] and (independently) the present author [18, 20] discovered a quadratically convergent algorithm for computing π via the AGM *without* needing to know e or $\log 2$ to high precision. It is known as the "Gauss–Legendre" algorithm (after the discoverers of the key identities [26, 36]) or the "Brent–Salamin" algorithm (after the twentieth century discoverers [21]), and is about twice as fast as the earlier algorithms which assume a knowledge of e or $\log 2$. We abbreviate the name to *Algorithm GL*. Bailey and Borwein, in *Pi: The Next Generation* [6, Synopsis of paper 1], say "This remarkable co-discovery arguably launched the modern computer era of the computation of π".[5]

In 1984, Jon and Peter Borwein [12] (see also [14, Alg. 2.1]) discovered another quadratically convergent algorithm for computing π, with convergence about as fast as Algorithm GL. We call this the (first) *Borwein-Borwein algorithm*, or *Algorithm BB1*. Yet another quadratically convergent algorithm, which we call the (second) *Borwein-Borwein algorithm* and abbreviate as *Algorithm BB2*, dates from 1986—see [13] and [14, Iteration 5.1]. Although Algorithm BB2 appears different from Algorithm GL, we show in Sect. 5 that the two algorithms are in fact equivalent, in the sense of producing the same sequence of approximations to π. This surprising fact does not seem to have been noticed before.

4 Some Superlinearly Convergent Algorithms for π

In this section, we describe the Gauss–Legendre algorithm (GL) and two quadratically convergent algorithms (BB1 and BB2) due to Jon and Peter Borwein. We also describe a fourth order algorithm (BB4) due to the Borweins.

[5] In [10, §10], Jon Borwein says "It [Algorithm GL] is based on the arithmetic–geometric mean iteration (AGM) and some other ideas due to Gauss and Legendre around 1800, although neither Gauss, nor many after him, ever directly saw the connection to effectively computing π".

Using Legendre's relation and the formulæ that we have given for E and K in terms of the AGM iteration, it is not difficult to derive Algorithm GL. We present it in pseudocode using the same style as the algorithms in [23].

Algorithm GL
Input: The number of iterations n_{max}.
Output: A sequence of n_{max} intervals containing π.

```
$a_0 := 1;\ b_0 := 1/\sqrt{2};\ s_0 := \frac{1}{4}$.
for $n$ from 0 to $n_{max} - 1$ do
    $a_{n+1} := (a_n + b_n)/2;$
    $c_{n+1} := a_n - a_{n+1};$
    output $(a_{n+1}^2/s_n,\ a_n^2/s_n)$.
    if $n < n_{max} - 1$ then
        $b_{n+1} := \sqrt{a_n b_n};$
        $s_{n+1} := s_n - 2^n\, c_{n+1}^2$.
```

Remarks

1. Subscripts on variables such as a_n, b_n are given for expository purposes. In an efficient implementation only a constant number of real variables are needed, because a_{n+1} can overwrite a_n (after saving a_n in a temporary variable for use in the computation of b_{n+1}), and similarly for b_n, c_n and s_n.
2. The purpose of the final "if ... then" is simply to avoid unnecessary computations after the final output. Similar comments apply to the other algorithms given below.
3. Salamin [42] notes the identity $4a_{n+1}c_{n+1} = c_n^2$ which can be used to compute c_{n+1} without the numerical cancellation that occurs when using the definition $c_{n+1} = a_n - a_{n+1}$. However, this refinement costs time and is unnecessary, because the terms $2^n c_{n+1}^2$ diminish rapidly and make only a minor contribution to the overall error caused by using finite-precision real arithmetic. To obtain an accurate result it is sufficient to use $O(\log n_{max})$ guard digits.

Neglecting the effect of rounding errors, Algorithm GL gives a sequence of lower and upper bounds on π:

$$\frac{a_{n+1}^2}{s_n} < \pi < \frac{a_n^2}{s_n},$$

and both bounds converge quadratically to π. The lower bound is more accurate, so the algorithm is often stated with just the lower bound a_{n+1}^2/s_n (we call this variant *Algorithm GL1*). Table 1 shows the approximations to π given by the first few iterations. Correct digits are shown in bold. The quadratic convergence is evident.

Table 1 Convergence of algorithm GL

n	Lower bound a_{n+1}^2/s_n		Upper bound a_n^2/s_n
0	2.914213562373095048801689	$< \pi <$	4.00000000000000000000000
1	**3.14**0579250522168248311331	$< \pi <$	**3.1**87672642712108627201930
2	**3.1415926**46213542282149344	$< \pi <$	**3.141**680293297653293918070
3	**3.141592653589793238**279513	$< \pi <$	**3.141592653**895446496002915
4	**3.141592653589793238462643**	$< \pi <$	**3.14159265358979323846**6361

Recall that in Algorithm GL we have $a_0 = 1$, $b_0 = 1/\sqrt{2}$, $s_0 = \frac{1}{4}$ and, for $n \geq 0$,

$$a_{n+1} = \frac{a_n + b_n}{2}\,,\ b_{n+1} = \sqrt{a_n b_n},\ \ c_{n+1} = a_n - a_{n+1},\ \ s_{n+1} = s_n - 2^n\, c_{n+1}^2\,.$$

Take $q = e^{-\pi}$, and write

$$a_\infty := \lim_{n\to\infty} a_n = \theta_3^{-2}(q) = 2\pi^{3/2}/\,\Gamma^2(\tfrac{1}{4}) \approx 0.8472, \tag{16}$$

$$s_\infty := \lim_{n\to\infty} s_n = \theta_3^{-4}(q)/\pi \, = 4\pi^2/\,\Gamma^4(\tfrac{1}{4}) \approx 0.2285\,. \tag{17}$$

Since $c_n = \theta_2^2(q^{2^n})/\theta_3^2(q)$, we have

$$s_n - s_\infty = \theta_3^{-4}(q) \sum_{m=n}^{\infty} 2^m \theta_2^4(q^{2^{m+1}})\,. \tag{18}$$

Write $a_n/a_\infty = 1 + \delta_n$ and $s_n/s_\infty = 1 + \varepsilon_n$. Then

$$\delta_n = \theta_3^2(q^{2^n}) - 1 \sim 4q^{2^n} \ \text{ as } n \to \infty,$$

and (17), (18) give

$$\varepsilon_n = \pi \sum_{m=n}^{\infty} 2^m\, \theta_2^4(q^{2^{m+1}}) \sim 2^{n+4}\pi q^{2^{n+1}}.$$

Writing

$$\frac{a_n^2/a_\infty^2}{s_n/s_\infty} = \frac{a_n^2}{\pi s_n} = \frac{(1+\delta_n)^2}{1+\varepsilon_n}\,,$$

it is straightforward to obtain an upper bound on π:

$$0 < a_n^2/s_n - \pi \, < U(n) := 8\pi q^{2^n}. \tag{19}$$

Convergence is quadratic: if $e_n := a_n^2/s_n - \pi$, then

Table 2 Numerical values of upper and lower bounds for algorithm GL

n	$a_n^2/s_n - \pi$	$\pi - a_{n+1}^2/s_n$	$\dfrac{a_n^2/s_n - \pi}{U(n)}$	$\dfrac{\pi - a_{n+1}^2/s_n}{L(n)}$
0	8.58e-1	2.27e-1	0.790369040	0.916996189
1	4.61e-2	1.01e-3	0.981804947	0.999656206
2	8.76e-5	7.38e-9	0.999922813	0.999999998
3	3.06e-10	1.83e-19	0.999999999	1.000000000
4	3.72e-21	5.47e-41	1.000000000	1.000000000
5	5.50e-43	2.41e-84	1.000000000	1.000000000
6	1.20e-86	2.31e-171	1.000000000	1.000000000
7	5.76e-174	1.06e-345	1.000000000	1.000000000
8	1.32e-348	1.11e-694	1.000000000	1.000000000

$$\lim_{n\to\infty} e_{n+1}/e_n^2 = \tfrac{1}{8\pi}\cdot$$

Replacing a_n by a_{n+1} and δ_n by δ_{n+1}, we obtain a lower bound on π:

$$0 < \pi - \frac{a_{n+1}^2}{s_n} < L(n) := (2^{n+4}\pi^2 - 8\pi)q^{2^{n+1}}. \tag{20}$$

Pi and the AGM [(2.5.7) on page 48] gives a slightly weaker lower bound which, via (16), may be written as

$$\pi - \frac{a_{n+1}^2}{s_n} \le \frac{2^{n+4}\pi^2 q^{2^{n+1}}}{a_\infty^2}\cdot \tag{21}$$

Since $a_\infty^2 < 1$, the bound (21) is weaker than the bound (20). In (20), the factor $(2^{n+4}\pi^2 - 8\pi)$ is the best possible, since an expansion of a_{n+1}^2/s_n in powers of q gives $\pi - a_{n+1}^2/s_n = (2^{n+4}\pi^2 - 8\pi)q^{2^{n+1}} - \; O(2^n q^{2^{n+2}})$, with the minus sign before the "O" term informally indicating the sign of the remainder.

In Table 2, $U(n) := 8\pi\exp(-2^n\pi)$ and $L(n) := (2^{n+4}\pi^2 - 8\pi)\exp(-2^{n+1}\pi)$ are the bounds given in (19)–(20). It can be seen that the bounds are very accurate for $n > 1$, as expected from our analysis.

Recall that Algorithm GL gives approximations a_n^2/s_n and a_{n+1}^2/s_n to $\pi = a_\infty^2/s_\infty$. Using the expressions for a_n and s_n in terms of theta functions, we see that

$$\pi = \frac{a_n^2\,\theta_3^{-4}(q^{2^n})}{s_n - \theta_3^{-4}(q)\sum_{m=n}^{\infty} 2^m\,\theta_2^4(q^{2^{m+1}})}\,, \tag{22}$$

[or similarly with the numerator replaced by $a_{n+1}^2\theta_3^{-4}(q^{2^{n+1}})$]. The expression (22) for π is essentially of the form

$$\pi = \frac{a_n^2 - O(q^{2^n})}{s_n - O(2^n q^{2^{n+1}})}\;\left[\text{or } \frac{a_{n+1}^2 - O(q^{2^{n+1}})}{s_n - O(2^n q^{2^{n+1}})}\right].$$

This shows precisely how Algorithm GL approximates π and why it provides upper [or lower] bounds.

In *Pi and the AGM*, Jon and Peter Borwein present a quadratically convergent algorithm for π, based on the AGM, but different from Algorithm GL. It is Algorithm 2.1 in Chapter 2, and was first published in [12]. We call it *Algorithm BB1*.

Instead of using Legendre's relation, Algorithm BB1 uses the identity

$$K(k)\,\mathrm{D}_k K(k)\Big|_{k=1/\sqrt{2}} = \frac{\pi}{\sqrt{2}},$$

where D_k denotes differentiation with respect to k.

Using the connection between $K(k')$ and the AGM, the Borweins [14, (2.4.7)] prove that

$$\pi = 2^{3/2} \left. \frac{(\mathrm{AGM}(1,k'))^3}{\mathrm{D}_k\,\mathrm{AGM}(1,k')} \right|_{k=1/\sqrt{2}}.$$

An algorithm for approximating the derivative in this formula can be obtained by differentiating the AGM iteration symbolically. Details are given in [14].

We now present Algorithm BB1. Note that the algorithm given in [14] defines the upper bound $\overline{\pi}_n := \overline{\pi}_{n-1}(x_n+1)/(y_n+1)$ and omits the lower bound $\underline{\pi}_n$, but $\underline{\pi}_n$ can be obtained from [14, ex. 2.5.11]. We present a version that computes upper ($\overline{\pi}_n$) and lower ($\underline{\pi}_n$) bounds for comparison with Algorithm GL.

Algorithm BB1
Input: The number of iterations n_{max}.
Output: A sequence of n_{max} intervals containing π.

```
$x_0 := \sqrt{2};$
**output** $(\underline{\pi}_0 := x_0,\ \overline{\pi}_0 := x_0 + 2).$
$y_1 := {x_0}^{1/2};\ x_1 := \frac{1}{2}(x_0^{1/2} + x_0^{-1/2});$
**for** $n$ from 1 to $n_{max} - 1$ do
    $\underline{\pi}_n := \dfrac{2\,\overline{\pi}_{n-1}}{y_n + 1};\ \overline{\pi}_n := \underline{\pi}_n\left(\dfrac{x_n+1}{2}\right);$
    **output** $(\underline{\pi}_n,\ \overline{\pi}_n);$
    **if** $n < n_{max} - 1$ **then**
        $x_{n+1} := \frac{1}{2}(x_n^{1/2} + x_n^{-1/2});\ \ y_{n+1} := \dfrac{y_n\, x_n^{1/2} + x_n^{-1/2}}{y_n + 1}.$
```

It may be shown that $\overline{\pi}_n$ decreases monotonically to the limit π, and $\underline{\pi}_n$ increases monotonically to π. Moreover, $\overline{\pi}_n - \underline{\pi}_n$ decreases quadratically to zero. This is illustrated in Table 3.

Table 3 Convergence of algorithm BB1

n	$\underline{\pi}_n$		$\overline{\pi}_n$
0	1.414213562373095048801689	$< \pi <$	**3**.414213562373095048801689
1	**3.1**19132528827772757303373	$< \pi <$	**3.14**2606753941622600790720
2	**3.1415**48837729436193482357	$< \pi <$	**3.1415926**60966044230497752
3	**3.14159265**3436966609787790	$< \pi <$	**3.141592653589793238**645774
4	**3.14159265358979323846**0785	$< \pi <$	**3.141592653589793238462643**

It is not immediately obvious that Algorithm BB1 depends on the AGM. However, the AGM is present in *Legendre form*: if $a_0 := 1$, $b_0 := k' = 1/\sqrt{2}$, and we perform n steps of the AGM iteration to define a_n, b_n, then $x_n = a_n/b_n$ and, for $n \geq 1$, $y_n = \mathrm{D}_k b_n/\mathrm{D}_k a_n$.

Comparing Tables 1 and 3, we see that Algorithm BB1 gives better upper bounds, but worse lower bounds, than Algorithm GL, for the same value of n (i.e. same number of square roots).

As for Algorithm GL, we can express the error after n iterations of Algorithm BB1 using theta functions, and deduce the asymptotic behaviour of the error.

Consider the AGM iteration with $a_0 = 1$, $b_0 = k' = (1-k^2)^{1/2}$. Then a_n and b_n are functions of k. In *Pi and the AGM* it is shown that, for $n \geq 1$,

$$\overline{\pi}_{n-1} = \left(2^{3/2} b_n^2 a_n / \mathrm{D}_k a_n\right)|_{k=1/\sqrt{2}} . \tag{23}$$

Now a_n and b_n are given by (13) with $q = e^{-\pi}$. We differentiate a_n with respect to k, where $k = (1-b_0^2)^{1/2} = \theta_2^2(q)/\theta_3^2(q)$. This gives

$$\mathrm{D}_k a_n = \mathrm{D}_q\left(\frac{\theta_3^2(q^{2^n})}{\theta_3^2(q)}\right) \Big/ \mathrm{D}_q\left(\frac{\theta_2^2(q)}{\theta_3^2(q)}\right)\Bigg|_{q=e^{-\pi}} . \tag{24}$$

We remark that (24) gives $\mathrm{D}_k a_0 = 0$, as expected since a_0 is independent of k.

Thanks to the analyticity of the theta functions in $|q| < 1$, there is no difficulty in showing that[6]

$$\lim_{n\to\infty} \mathrm{D}_k a_n = \mathrm{D}_k \lim_{n\to\infty} a_n .$$

We denote the common value by $\mathrm{D}_k a_\infty$. Taking the limit in (23), we obtain (as also follows from [14, (2.4.7)]):

$$\mathrm{D}_k a_\infty = \frac{2^{3/2} a_\infty^3}{\pi} = 0.547486\ldots \tag{25}$$

[6] Similarly, where we exchange the order of taking derivatives and limits elsewhere in this section, it is easy to justify.

Table 4 Numerical values of upper and lower bounds for algorithm BB1

n	$\overline{\pi}_n - \pi$	$\dfrac{\overline{\pi}_n - \pi}{2^{n+4}\pi^2 q^{2^{n+1}}}$	$\pi - \underline{\pi}_n$	$\dfrac{\pi - \underline{\pi}_n}{4\pi q^{2^n}}$
1	1.01e-3	0.9896487063	2.25e-2	0.9570949132
2	7.38e-9	0.9948470082	4.38e-5	0.9998316841
3	1.83e-19	0.9974691480	1.53e-10	0.9999999988
4	5.47e-41	0.9987456847	1.86e-21	1.0000000000
5	2.41e-84	0.9993755837	2.75e-43	1.0000000000
6	2.31e-171	0.9996884727	6.01e-87	1.0000000000
7	1.06e-345	0.9998444059	2.88e-174	1.0000000000
8	1.11e-694	0.9999222453	6.59e-349	1.0000000000

Now $a_n - a_\infty = \sum_{m=n+1}^{\infty} c_m$, and differentiating both sides with respect to k gives

$$\mathrm{D}_k a_n - \mathrm{D}_k a_\infty = \sum_{m=n+1}^{\infty} \mathrm{D}_q \left(\frac{\theta_2^2(q^{2^m})}{\theta_3^2(q)} \right) \Big/ \mathrm{D}_q \left(\frac{\theta_2^2(q)}{\theta_3^2(q)} \right) \Bigg|_{q=e^{-\pi}} . \tag{26}$$

We remark that (26) is analogous to (18), which we used in the analysis of Algorithm GL. Using (23)–(26), we obtain an upper bound on π (for $n \geq 1$, $q = e^{-\pi}$)

$$0 < \overline{\pi}_n - \pi < 2^{n+4}\pi^2 q^{2^{n+1}} . \tag{27}$$

A slightly weaker bound than (27) is proved in [14, §2.5].

Similarly, we can obtain a lower bound on π:

$$0 < \pi - \underline{\pi}_n < 4\pi q^{2^n} . \tag{28}$$

We omit detailed proofs of (27) and (28); they involve straightforward but tedious expansions of power series in q. Experimental evidence is provided in Table 4.

Table 4 gives numerical values of the approximation errors $\overline{\pi}_n - \pi$ and $\pi - \underline{\pi}_n$, and the ratio of these values to the bounds (27) and (28), respectively. It can be seen that the bounds are very accurate (as expected from the expressions for the errors in terms of theta functions and the rapid convergence of the series for the theta functions). The upper bound overestimates the error by a factor of $1 + O(2^{-n})$. A computation shows that we cannot replace the bound by the function $L(n)$ defined in (20), although a similar bound appears to be valid if the constant 8π in (20) is replaced by a slightly smaller constant, e.g. 7π.

The bounds (27)–(28) can be compared with the lower bound $(2^{n+4}\pi^2 - 8\pi)q^{2^{n+1}}$ and upper bound $8\pi q^{2^n}$ for Algorithm GL. The upper bound is better for Algorithm

BB1, but the lower bound is better for Algorithm GL. This confirms the observation above regarding the comparison of Tables 1 and 3.

Since it will be needed in Sect. 5, we state another quadratic algorithm, *Algorithm BB2*, different from Algorithm BB1 but also due to Jon and Peter Borwein (iteration 5.2 on page 170 of [14] with the parameter $r = 4$).

Algorithm BB2
Input: The number of iterations n_{max}.
Output: A sequence of n_{max} approximations to π.

```
$\alpha_0 := 6 - 4\sqrt{2};\ \ k_0 := 3 - 2\sqrt{2};$
**for** $n$ from 0 to $n_{max} - 1$ do
    **output** $\widehat{\pi}_n := 1/\alpha_n$ ;
    **if** $n < n_{max} - 1$ **then**
        $k_n' := \sqrt{1 - k_n^2};\ \ k_{n+1} := \dfrac{1 - k_n'}{1 + k_n'}$ ;
        $\alpha_{n+1} := (1 + k_{n+1})^2 \alpha_n - 2^{n+2} k_{n+1}$ .
```

In Algorithm BB2, we have $\widehat{\pi}_n \to \pi$ quadratically [14, pg. 170]. We remark that it would be clearer to increase (by one) the subscripts on the variables in Algorithm BB2, so as to correspond to the usage in Algorithm GL, which implicitly has $k_0' = b_0/a_0 = 1/\sqrt{2}$ and $k_1 = (1 - k_0')/(1 + k_0') = 3 - 2\sqrt{2}$, but we have kept the notation used in [14].

The Borwein brothers did not stop at quadratic (second-order) algorithms for π. In Chapter 5 of *Pi and the AGM* they gave algorithms of orders 3, 4, 5 and 7. Of course, these algorithms are not necessarily faster than the quadratic algorithms, because we must take into account the amount of work per iteration. For a fair comparison, we can use Ostrowski's *efficiency index* [39, §3.11], defined as $\log(p)/W$, where $p > 1$ is the order of convergence and W is the work per iteration. A justification of this measure of efficiency is given in [17]. Consider a simple example—if we combine three iterations of Algorithm BB2 into one iteration of a new algorithm, then we obtain an algorithm of order 8, but with three times as much work per iteration. The efficiency index is the same in both cases, as it should be.

We refer to [14, Chapter 5] for the Borweins' cubic, quintic and higher order algorithms, and consider only their quartic algorithm, which we call *Algorithm BB4*. It is a specialisation to the case $r = 4$ of the slightly more general algorithm given in [14, iteration 5.3, pg. 170]. The same special case is given in [15, Algorithm 1] and has been used in extensive calculations of π, see for example [4, 33]. We have changed notation slightly ($a_n \mapsto z_n$) to avoid conflict with the notation used in Algorithm GL.

Algorithm BB4
Input: The number of iterations n_{max}.
Output: A sequence of n_{max} approximations to π.

```
$y_0 := \sqrt{2} - 1;\ \ z_0 := 2y_0^2;$
**for** $n$ from 0 to $n_{max} - 1$ do
    **output** $\pi_n := 1/z_n$ ;
    **if** $n < n_{max} - 1$ **then**
        $y_{n+1} := \dfrac{1 - (1 - y_n^4)^{1/4}}{1 + (1 - y_n^4)^{1/4}}$ ;
        $z_{n+1} := z_n(1 + y_{n+1})^4 - 2^{2n+3} y_{n+1}(1 + y_{n+1} + y_{n+1}^2).$
```

In Algorithm BB4, π_n converges quartically to π. A sharp error bound is

$$0 < \pi - \pi_n < \pi^2\, 4^{n+2} \exp(-2\pi\, 4^n). \tag{29}$$

This improves by a factor of two on the error bound given in [14, top of pg. 171]. We defer the proof until Sect. 5.

Table 5 shows the error $\pi - \pi_n$ after n iterations of the Borwein quartic algorithm, and the ratio of the error $\pi - \pi_n$ to the upper bound (29).

At this point the reader may well ask "which of Algorithms GL, BB1, BB2 and BB4 is the fastest?" The answer seems to depend on implementation details. All four algorithms involve the same number of square roots to obtain comparable accuracy (counting a fourth root in Algorithm BB4 as equivalent to two square roots, which is not necessarily correct[7]). Algorithm GL has the advantage that high-precision divisions are only required when generating the output (so the early divisions can be skipped if intermediate output is not required). The other three algorithms require at least one division per iteration. Borwein, Borwein and Bailey [15, pg. 202] say "[Algorithm BB4] is arguably the most efficient algorithm currently known for the extended precision calculation of π," and the times given in Bailey's paper [4, pg. 289] confirm this (28 hours for Algorithm BB4 versus 40 hours for Algorithm BB1). However, Kanada [33], who extended Bailey's computation, reached the opposite conclusion. His computation took 5 hours 57 minutes with Algorithm GL, and 7 hours 30 minutes with Algorithm BB4 (which was used for verification).

[7] For example, one might compute $x^{1/4}$ using two *inverse* square roots, i.e. $(x^{-1/2})^{-1/2}$, which is possibly faster than two square roots, i.e. $(x^{1/2})^{1/2}$, see [23, §4.2.3].

Table 5 Approximation error in algorithm BB4

n	$\pi - \pi_n$	$\frac{\pi - \pi_n}{\text{bound (29)}}$
0	2.273790912e-1	0.7710517124
1	7.376250956e-9	0.9602112619
2	5.472109145e-41	0.9900528160
3	2.308580715e-171	0.9975132040
4	1.110954934e-694	0.9993783010
5	9.244416653e-2790	0.9998445753
6	6.913088685e-11172	0.9999611438
7	3.376546688e-44702	0.9999902860
8	3.002256862e-178825	0.9999975715

5 Equivalence of Some Algorithms for π

In the following, *doubling* an algorithm A means to construct an algorithm A^2 that outputs $(x_0, x_2, x_4, \ldots)$ if algorithm A outputs $(x_0, x_1, x_2, \ldots)$. Replacing n by $2n$ in (20) and retaining only the most significant term, we see that an error bound for Algorithm GL1 doubled is

$$0 < \pi - a_{2n+1}^2/s_{2n} < \pi^2\, 4^{n+2} \exp(-2\pi\, 4^n).$$

It is suggestive that the right-hand side is the same as in the error bound (29) for the Borwein quartic algorithm after n iterations.

On closer inspection we find that the two algorithms (GL1 doubled and BB4) are *equivalent*, in the sense that they give *exactly* the same sequence of approximations to π. Symbolically,

$$\pi_n = a_{2n+1}^2/s_{2n}, \tag{30}$$

where a_n, s_n are as in Algorithm GL, and π_n is as in Algorithm BB4. This observation appears to be new—it is not stated explicitly in *Pi and the AGM* or elsewhere, so far as we know.[8]

Before proving the result, we give some empirical evidence for it, since that is how the result was discovered—in the spirit of "Experimental Mathematics," as beloved by Jon Borwein. In Table 6, $n + 1$ is the number of square roots, and the second column is the error in the approximation given by Algorithm GL1 after n iterations, or by the Algorithm BB4 after $n/2$ iterations (n even). The error is the same for both algorithms (verified to 1000 decimal digits, not all shown).

[8] For example, the equivalence is not mentioned in [4], [15], [29], [30] or [33].

Table 6 Approximation error for algorithms GL1 doubled and BB4

n	$\pi - a_{2n+1}^2/s_{2n}$ (for Algorithm GL1) or $\pi - \pi_n$ (for Algorithm BB4)
0	2.2737909121669818966095465906980480562749752399816e-1
2	7.3762509563132989512968071098827321760295030264154e-9
4	5.4721091456899418327485331789641785565936917028248e-41
6	2.3085807149343902668213207343869568303303472423996e-171
8	1.1109549335576998257002904117322306941479378545140e-694

Using the definitions of the two algorithms, equality for the first line of the table ($n = 0$) follows from

$$a_1^2/s_0 = \pi_0 = \tfrac{3}{2} + \sqrt{2} = \pi - 0.227\ldots$$

For the second line ($n = 2$) we have, with $t := 2^{-1/4}$,

$$a_3 = \frac{(t^2 + 2t + 1 + 2\sqrt{2t^3 + 2t})}{8} \quad \text{and} \quad s_2 = \frac{8t^3 - 4t^2 + 8t - 5}{16},$$

so

$$\frac{a_3^2}{s_2} = \frac{(t^2 + 2t + 1 + 2\sqrt{2t^3 + 2t})^2}{4(8t^3 - 4t^2 + 8t - 5)}. \tag{31}$$

Also, from the definition of Algorithm BB4 we find, with

$$y_1 = \frac{1 - (12\sqrt{2} - 16)^{1/4}}{1 + (12\sqrt{2} - 16)^{1/4}},$$

that

$$\pi_1 = \frac{1}{(6 - 4\sqrt{2})(1 + y_1)^4 - 8y_1 - 8y_1^2 - 8y_1^3}. \tag{32}$$

It is not obvious that the algebraic numbers given by (31) and (32) are identical, but it can be verified that they both have minimal polynomial

$$\begin{aligned}P(x) :=& 1 - 1635840576x - 343853312x^2 + 60576043008x^3 \\ &+ 1865242664960x^4 - 16779556159488x^5 + 37529045696512x^6 \\ &- 29726424956928x^7 + 6181548457984x^8.\end{aligned}$$

Using Sturm sequences [45], it may be shown that $P(x)$ has two real roots, one in the interval $[0, 1]$, and the other in $[3, 4]$. A numerical computation shows that $|a_3^2/s_2 - \pi_1| < 1$, but both a_3^2/s_2 and π_1 are real roots of $P(x)$, so they must be equal.

Clearly this "brute force" approach does not generalise. To prove the equivalence of Algorithms BB4 and GL1, we first consider the equivalence of Algorithms BB2 and GL1.

Theorem 1 *Algorithm BB2 is equivalent to Algorithm GL1, in the sense that*

$$\widehat{\pi}_n = a_{n+1}^2/s_n,$$

where $\widehat{\pi}_n = 1/\alpha_n$ is as in Algorithm BB2, and a_{n+1}, s_n are as in Algorithm GL.

Proof In the proof we take $n \geq 0$, $q = e^{-\pi}$, and assume that a_n, b_n, c_{n+1}, s_n are defined as in Algorithm GL, and $k_n, \alpha_n, \widehat{\pi}_n$ are as in Algorithm BB2.

Algorithm GL implements the recurrence

$$s_{n+1} = s_n - 2^n c_{n+1}^2, \tag{33}$$

whereas Algorithm BB2 implements the recurrence

$$\alpha_{n+1} = (1 + k_{n+1})^2 \alpha_n - 2^{n+2} k_{n+1}. \tag{34}$$

We show that the recurrences (33)–(34) are related. Noting the remark on subscripts following the statement of Algorithm BB2, we see that $k_n = c_{n+1}/a_{n+1}$, since both sides equal $\theta_2^2(q^{2^{n+1}})/\theta_3^2(q^{2^{n+1}})$. Thus

$$1 + k_{n+1} = a_{n+1}/a_{n+2}. \tag{35}$$

Define $\beta_n := a_{n+1}^2 \alpha_n$ and $\gamma_n := a_{n+2}^2 k_{n+1}$. Substituting (35) into (34) and clearing the fractions gives

$$\beta_{n+1} = \beta_n - 2^{n+2} \gamma_n. \tag{36}$$

Now

$$4\gamma_n = 4a_{n+2}^2 k_{n+1} = 4a_{n+2} c_{n+2} = \theta_2^4(q^{2^{n+2}})/\theta_3^4(q) = c_{n+1}^2,$$

so (36) is equivalent to

$$\beta_{n+1} = \beta_n - 2^n c_{n+1}^2. \tag{37}$$

This is essentially the same recurrence as (33). Also, $s_0 = 1/4$ and $\beta_0 = a_1^2 \alpha_0 = 1/4$, so $s_0 = \beta_0$. It follows that $s_n = \beta_n$ for all $n \geq 0$. Thus $s_n = a_{n+1}^2 \alpha_n$, and

$$\widehat{\pi}_n = 1/\alpha_n = a_{n+1}^2/s_n,$$

which completes the proof. □

Corollary 1 *Algorithm BB4 is equivalent to Algorithm GL1 doubled, in the sense that*

$$\pi_n = a_{2n+1}^2/s_{2n},$$

where π_n is as in Algorithm BB4, and a_n, s_n are as in Algorithm GL.

Proof The Borwein brothers noted [14, pg. 171] that Algorithm BB4 is equivalent to Algorithm BB2 doubled,[9] i.e. $\pi_n = \widehat{\pi}_{2n}$. Thus, the result follows from Theorem 1. □

Corollary 2 *For Algorithm BB4, the error bound* (29) *holds.*

Proof In view of Corollary 1, the error bound (29) follows from (30) and the error bound (20) for Algorithm GL. □

6 Some Fast (but Linear) Algorithms for π

Let $(x)_n := x(x+1)\cdots(x+n-1)$ denote the *ascending factorial*. In Chapter 5 of *Pi and the AGM*, Jon and Peter Borwein discuss *Ramanujan–Sato* series such as

$$\frac{1}{\pi} = 2^{3/2} \sum_{n=0}^{\infty} \frac{(\frac{1}{4})_n (\frac{1}{2})_n (\frac{3}{4})_n}{(n!)^3} \, \frac{(1103 + 26390n)}{99^{4n+2}} \,.$$

This is linearly convergent, with rate $1/99^4$, so adds nearly eight decimal digits per term, since $99^4 \approx 10^8$.

A more extreme example is the Chudnovsky series [24]

$$\frac{1}{\pi} = 12 \sum_{n=0}^{\infty} (-1)^n \, \frac{(6n)!\,(13591409 + 545140134n)}{(3n)!\,(n!)^3\, 640320^{3n+3/2}} \,, \tag{38}$$

which adds about 14 decimal digits per term.

Although such series converge only linearly, their convergence is so fast that they are competitive with higher order algorithms such as Algorithm GL for computing highly accurate approximations to π. Which algorithm is the fastest in practice depends on details of the implementation and on technological factors such as memory sizes and access times.

7 Fast Algorithms for the Elementary Functions

In this section, we consider the *bit-complexity* of algorithms. The bit-complexity of an algorithm is the (worst case) number of single-bit operations required to complete the algorithm. For a fuller discussion, see Chapter 6 of *Pi and the AGM*. We are interested in asymptotic results, so are usually willing to ignore constant factors.

[9] In fact, this is how Algorithm BB4 was discovered, by doubling Algorithm BB2 and then making some straightforward program optimisations.

If all operations are performed to (approximately) the same precision, then it makes sense to count *operations* such as multiplications, divisions and square roots. Algorithms based on the AGM fall into this category.

If the precision of the operations varies widely, then bit-complexity is a more sensible measure of complexity. An example is Newton's method, which is self-correcting, so can be started with low precision. Another example is summing a series with rational terms, such as $e = \sum_{k=0}^{\infty} 1/k!$.

The bit-complexity of multiplying two n-bit numbers to obtain a $2n$-bit product is denoted by $M(n)$. The classical algorithm shows that $M(n) = O(n^2)$, but various asymptotically faster algorithms exist. The best result so far, due to Harvey, van der Hoeven and Lecerf [31], is

$$M(n) = O\left(n \log n \, K^{\log^* n}\right)$$

with $K = 8$. Here the *iterated logarithm* function $\log^* n$ is defined by

$$\log^* n := \begin{cases} 0 & \text{if} \quad n \le 1; \\ 1 + \log^*(\log n) & \text{if} \quad n > 1. \end{cases}$$

It is unbounded but grows *extremely* slowly as $n \to \infty$, e.g. slower than

$$\log \log \cdots \log n \quad \text{[for any fixed number of logs]}.$$

Added in proof (December 2019): Harvey and van der Hoeven have recently announced that $M(n) = O(n \log n)$.

We follow *Pi and the AGM* and assume that $M(n)$ is non-decreasing and satisfies the weak regularity condition

$$2M(n) \le M(2n) \le 4M(n).$$

Newton's method can be used to compute reciprocals and square roots with bit-complexity

$$O\left(M(n) + M(\lceil n/2 \rceil) + M\left(\lceil n/2^2 \rceil\right) + \cdots + M(1)\right) = O(M(n)).$$

It can be shown that the bit-complexities of squaring, multiplication, reciprocation, division and root extraction are asymptotically the same, up to small constant factors [19]. All these operations have bit-complexity of order $M(n)$.

To compute π to n digits (binary or decimal) by the arctan formula (1), or to compute $1/\pi$ by the Chudnovsky series (38), we have to sum of order n terms. Using *divide and conquer*, also called *binary splitting* [19, 28],[10] this can be done with bit-complexity

[10] Somewhat more general, but based on the same idea, is E. Karatsuba's *FEE method* [34].

$$O(M(n)\log^2 n).$$

Suppose we compute π to n-digit accuracy using one of the quadratically convergent AGM algorithms. This requires $O(\log n)$ iterations, each of which has bit-complexity $O(M(n))$. Thus, the overall bit-complexity is

$$O(M(n)\log n).$$

This is (theoretically) better than series summation methods, the best of which have bit-complexity of order $M(n)\log^2 n$.

In practice, a method with bit-complexity of order $M(n)\log^2 n$ may be faster than a method with bit-complexity of order $M(n)\log n$ unless n is sufficiently large. This is one reason for the recent popularity of the Chudnovsky series (38) for high-precision computation of π, even though the AGM-based methods are theoretically (i.e. asymptotically) more efficient.

In Sect. 3, we mentioned Salamin's algorithm for computing $\log x$ for sufficiently large $x = 4/k$, i.e. sufficiently small k, using (9). We can evaluate $K'(k)/\pi$ using the AGM with $(a_0, b_0) = (1, k)$, and hence approximate $\log(4/k)$, assuming that π is precomputed. To compute $\log x$ to n-bit accuracy requires about $2\log_2(n)$ AGM iterations, or $3\log_2(n)$ iterations if we count the computation of π.

If x is not sufficiently large, we can use the identity $\log(x) = \log(2^p x) - p\log 2$, where p is a sufficiently large integer (but not too large or excessive cancellation will occur). This assumes that $\log 2$ is precomputed, and that the precision is increased to compensate for cancellation.

To obtain a small relative error when x is close to 1, say $|x-1| < 2^{-n/\log n}$, it is better to use the Taylor series for $\log(1+z)$, with $z = x-1$. The Taylor series computation can be accelerated by "splitting," see [23, §4.4.3] and [44],

The $O(k^2)$ error term in the expression (9) can be written explicitly using hypergeometric series, see [14, (1.3.10)]. This gives one way of improving the accuracy of the approximation $K'(k)$ to $\log(4/k)$. We give an alternative using theta functions, for which the series converge faster than the hypergeometric series (which converge only linearly). The result (39) follows from several identities given in Sect. 2. We collect them here for convenience:

$$\begin{aligned}
\log(1/q) &= \pi K'(k)/K(k),\\
k &= \theta_2^2(q)/\theta_3^2(q),\\
K(k) &= (\pi/2)\,\theta_3^2(q),\\
K'(k) &= (\pi/2)/\mathrm{AGM}(1,k).
\end{aligned}$$

Putting these pieces together gives the elegant result of Sasaki and Kanada [43]

$$\log(1/q) = \frac{\pi}{\mathrm{AGM}(\theta_2^2(q), \theta_3^2(q))} \cdot \tag{39}$$

In (39) we can replace q by q^4 to avoid fractional powers of q in the expansion of $\theta_2(q)$, obtaining an exact formula for all $q \in (0, 1)$:

$$\log(1/q) = \frac{\pi/4}{\mathrm{AGM}(\theta_2^2(q^4), \theta_3^2(q^4))} \cdot \tag{40}$$

As in Salamin's algorithm, we have to ensure that $x := 1/q$ is sufficiently large, but now there is a trade-off between increasing x or taking more terms in the series defining the theta functions. For example, to attain n-bit accuracy, if $x > 2^{n/36}$, we can use $\theta_2(q^4) = 2(q + q^9 + q^{25} + O(q^{49}))$ and $\theta_3(q^4) = 1 + 2(q^4 + q^{16} + O(q^{36}))$. This saves about four AGM iterations, compared to Salamin's algorithm. We remark that a result similar to (39) and (40) is given in (7.2.5) of *Pi and the AGM*, but with an unfortunate typo (a reciprocal is missing).

So far we have assumed that the initial values a_0, b_0 in the AGM iteration are real and positive. There is no difficulty in extending the results that we have used to complex a_0, b_0, provided that they are nonzero and a_0/b_0 is not both real and negative. For simplicity, we assume that $a_0, b_0 \in \mathcal{H} = \{z \mid \Re(z) > 0\}$.

In the AGM iteration (and in the definition of the geometric mean) there is an ambiguity of sign. We always choose the square root with positive real part. Thus the iterates a_n, b_n are uniquely defined and remain in the right half-plane $\mathcal{H}$.

When using (40), we may need to apply a rotation to q, say by a multiple of $\pi/3$, in order to ensure that the starting values $(\theta_2^2(q^4), \theta_3^2(q^4))$ for the AGM lie in $\mathcal{H}$.[11]

For $z \in \mathbb{C}\backslash\{0\}$, $\log(z) = \log(|z|) + i \arg(z)$, provided we use the principal values of the logarithms. Thus, if $x \in \mathbb{R}$, we can use the complex AGM to compute

$$\arctan(x) = \Im(\log(1 + ix)).$$

$\arcsin(x)$, $\arccos(x)$, etc can be computed via arctan using elementary trigonometric identities such as

$$\arccos(x) = \arctan(\sqrt{1 - x^2}/x).$$

Since we can compute log, arctan, arccos, arcsin, we can compute exp, tan, cos, sin (in suitably restricted domains) using Newton's method. The trigonometric functions can also be computed via the complex exponential. Similarly for the hyperbolic functions cosh, sinh, tanh and their inverse functions.

[11] Alternatively, we could drop the simplifying assumption that $a_0, b_0 \in \mathcal{H}$ and use the "right choice" of Cox [25, pg. 284] to implement the AGM correctly.

Although computing the elementary functions via the complex AGM is conceptually straightforward, it introduces the overhead of complex arithmetic. It is possible to avoid complex arithmetic by the use of Landen transformations (which transform incomplete elliptic integrals). See exercise 7.3.2 of *Pi and the AGM* for an outline of this approach, and [20] for more details.

Whichever approach is used, the bit-complexity of computing n-bit approximations to any of the elementary functions (log, exp, arctan, sin, cos, tan, etc) in a given compact set $A \subset \mathbb{C}$ that excludes singularities of the relevant function is $O(M(n) \log n)$. Here "n-bit approximation" means with absolute error bounded by 2^{-n}. We could require relative error bounded by 2^{-n}, but the proof would depend on a Diophantine approximation result such as Mahler's well-known result on approximation of π by rationals [38], because of the difficulty of guaranteeing a small relative error in the neighbourhood of a zero of the function.[12]

Certain non-elementary functions can be computed with bit-complexity $O(M(n) \log n)$ via the AGM. For example, we mention complete and incomplete elliptic integrals, elliptic functions, and the Jacobi theta functions $\theta_2(q)$, $\theta_3(q)$, $\theta_4(q)$. Functions that appear *not* to be in this class of "easily computable" functions include the Gamma function $\Gamma(z)$ and the Riemann zeta function $\zeta(s)$.

Algebraic functions can be computed with bit-complexity $O(M(n))$, see for example [14, Thm. 6.4]. It is plausible to conjecture that no elementary transcendental functions can be computed with bit-complexity $O(M(n))$ (or even $o(M(n) \log n)$). However, as usual in complexity theory, nontrivial lower bounds are difficult to prove and depend on the precise model of computation.

Acknowledgements I am grateful to Jon and Peter Borwein for becoming sufficiently interested in this subject to write their book *Pi and the AGM* only a few years after the publication of [18–20, 42]. Reading a copy of *Pi and the AGM* was my first introduction to the Borwein brothers, and was the start of my realisation that we shared many common interests, despite living in different hemispheres. Much later, after Jon and his family moved to Newcastle (NSW), I followed him, bringing our common interests closer together, and benefitting from frequent interaction with him.

Thanks are also due to David H. Bailey for his assistance, and to the Magma group for their excellent software [16].

The author was supported in part by an Australian Research Council grant DP140101417. Jon Borwein was the Principal Investigator on this grant, which was held by Borwein, Brent and Bailey.

[12] Mahler's result is sufficient for the usual elementary functions, whose zeros are rational multiples of π, but it is not applicable to the problem of computing combinations of these functions, e.g. $\exp(\sin x) + \cos(\log x)$, with small relative accuracy. In general, we do not know enough about the rational approximation of the zeros of such functions to guarantee a small *relative* error. However, the result that we stated for computing elementary functions with a small *absolute* error extends to finite combinations of elementary functions under the operations of addition, multiplication, composition, etc. Indeed, the set of *elementary functions* is usually considered to include such finite combinations, although precise definitions vary. See, for example, §7.3 of *Pi and the AGM*, Knopp [35, pp. 96–98], Liouville [37], Ritt [41], and Watson [46, pg. 111].

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1965). Online version at http://people.math.sfu.ca/~cbm/aands/. Accessed 7 Aug 2018
2. Andrews, G.E.: Pi and the AGM: a study in analytic number theory and computational complexity. Book review in Bull. (NS) AMS **22**, 198–201 (1990)
3. Askey, R.: Book review: Pi and the AGM. Am. Math. Mon. **95**, 895–897 (1988)
4. Bailey, D.H.: The computation of π to 29,360,000 decimal digits using Borweins' quartically convergent algorithm. Math. Comput. **50**, 283–296 (1988)
5. Bailey, D.H.: A collection of mathematical formulas involving π, Feb. 6 (2018). http://www.davidhbailey.com/dhbpapers/pi-formulas.pdf. Aaccessed 7 Aug 2018
6. Bailey, D.H., Borwein, J.M.: Pi: The Next Generation. Springer, Berlin (2016)
7. Baruah, N.D., Berndt, B.C., Chan, H.H.: Ramanujan's series for $1/\pi$: a survey. Am. Math. Mon. **116**, 567–587 (2009)
8. Beeler, M., Gosper, R.W., Schroeppel, R.: HAKMEM, AI Memo 239, MIT AI Lab (1972). (Item 143 by E. Salamin.)
9. Berndt, B.C.: Book review: Pi and the AGM. Math. Comput. **50**, 352–354 (1988)
10. Borwein, J.M.: The life of pi: from Archimedes to Eniac and beyond, prepared for Berggren Festschrift (2012). https://www.carma.newcastle.edu.au/jon/pi-2012.pdf. Accessed 7 Aug 2018
11. Borwein, J.M., Lectures and presentations. https://www.carma.newcastle.edu.au/jon/index-talks.shtml. Accessed 7 Aug 2018
12. Borwein, J.M., Borwein, P.B.: The arithmetic-geometric mean and fast computation of elementary functions. SIAM Rev. **26**, 351–365 (1984)
13. Borwein, J.M., Borwein, P.B.: More quadratically convergent algorithms for π. Math. Comput. **46**, 247–253 (1986)
14. Borwein, J.M., Borwein, P.B.: Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity. Monographies et Études de la Société Mathématique du Canada. Wiley, Toronto (1987)
15. Borwein, J.M., Borwein, P.B., Bailey, D.H.: Ramanujan, modular equations, and approximations to pi or how to compute one billion digits of pi. Am. Math. Mon. **96**, 201–219 (1989)
16. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language, J. Symb. Comput. **24**, 235–265 (1997)
17. Brent, R.P.: Some efficient algorithms for solving systems of nonlinear equations. SIAM J. Numer. Anal. **10**, 327–344 (1973)
18. Brent, R.P.: Multiple-precision zero-finding methods and the complexity of elementary function evaluation. In: Traub, J.F. (ed.) Analytic Computational Complexity, pp. 151–176. Academic, New York (1975)
19. Brent, R.P.: The complexity of multiple-precision arithmetic. In: Anderssen, R.S., Brent, R.P. (eds.) The Complexity of Computational Problem Solving, pp. 126–165. University of Queensland Press, Brisbane (1976)
20. Brent, R.P.: Fast multiple-precision evaluation of elementary functions. J. ACM **23**, 242–251 (1976)
21. Brent, R.P.: Old and new algorithms for π. Notices AMS **60**, 7 (2013)
22. Brent, R.P.: Jonathan Borwein, Pi and the AGM, Keynote Talk at the Jonathan Borwein Commemorative Conference, Newcastle, NSW (2017). http://maths-people.anu.edu.au/~brent/talks.html. Accessed 7 Aug 2018
23. Brent, R.P., Zimmermann, P.: Modern Computer Arithmetic. Cambridge University Press, Cambridge (2010)
24. Chudnovsky, D.V., Chudnovsky, G.V.: The computation of classical constants. Proc. Nat. Acad. Sci. USA **88**(21), 8178–8182 (1989)
25. Cox, D.A.: The arithmetic-geometric mean of Gauss. L'Enseignement Mathématique **30**, 275–330 (1984)

26. Gauss, C.F.: Unpublished Notebook Entry of May 1809, Reproduced in J. Arndt and C. Haenel, Pi: Algorithmen, Computer, Arithmetik, Chap. 7, p. 99. Springer, Berlin (1998)
27. Gauss, C.F.: Carl Friedrich Gauss Werke, Bd. 3, Göttingen, 1876, 362–403
28. Gourdon, X., Sebah, P.: Binary splitting method (2001). http://numbers.computation.free.fr/Constants/Algorithms/splitting.html. Accessed 7 Aug 2018
29. Guillera, J.: Easy proofs of some Borwein algorithms for π. Am. Math. Mon. **115**, 850–854 (2008)
30. Guillera, J.: New proofs of Borwein-type algorithms for Pi. Integr. Transform. Spec. Funct. **27**, 775–782 (2016)
31. Harvey, D., van der Hoeven, J., Lecerf, G.: Even faster integer multiplication. J. Complex. **36**, 1–30 (2016)
32. Jacobi, C.G.J.: Fundamenta Nova Theoriae Functionum Ellipticarum, Königsberg, 1829. Reprinted in Gesammelte Mathematische Werke **1**, 255–263 (1829)
33. Kanada, Y.: Vectorization of multiple-precision arithmetic program and 201,326,000 decimal digits of pi calculation. In Supercomputing 88. IEEE 117–128 (1988)
34. Karatsuba, E.A.: Fast evaluations of transcendental functions. Probl. Peredachi Informat. **27**, 4 (1991). Also https://en.wikipedia.org/wiki/FEE_method. Accessed 7 Aug 2018
35. Knopp, K.: The Elementary Functions, §23 in Theory of Functions Parts I and II, pp. 96–98. Dover, New York (1996) l
36. Legendre, A.M.: Exercices de Calcul Integral, vol. 1, p. 61. Paris (1811)
37. Liouville, J.: Sur la classification des Transcendantes et sur l'impossibilité d'exprimer les racines des certaines équations en fonction finie explicite des coefficients. Part 1. J. Math. Pure Appl. **2**, 56–105 (1837). Also Part 2, ibid **3**, 523–547 (1838)
38. Mahler, K.: On the approximation of π. Proc. Kon. Nederlandsche Akad. v. Wetenschappen Ser. A **56**, 30–42 (1953) = Indag. Math. **15**, 30–42 (1953). Also https://carma.newcastle.edu.au/mahler/docs/119.pdf. Accessed 7 Aug 2018
39. Ostrowski, A.M.: Solution of Equations and Systems of Equations. Academic Press, New York (1960)
40. Ramanujan, S.: Modular equations and approximations to pi. Quart. J. Math. (Oxford) **45**, 350–372 (1914)
41. Ritt, J.F.: Integration in Finite Terms. Columbia University Press, New York (1948)
42. Salamin, E.: Computation of π using arithmetic-geometric mean. Math. Comput. **30**, 565–570 (1976)
43. Sasaki, T., Kanada, Y.: Practically fast multiple-precision evaluation of $\log(x)$. J. Inf. Process. **5**, 247–250 (1982)
44. Smith, D.M.: Efficient multiple-precision evaluation of elementary functions. Math. Comput. **52**, 131–134 (1989)
45. Sturm, J.C.F.: Mémoire sur la résolution des équations numériques. Bulletin des Sciences de Férussac **11**, 419–425 (1829)
46. Watson, G.N.: A Treatise on the Theory of Bessel Functions, 2nd edn. Cambridge (1966)
47. Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis, 3rd edn. Cambridge (1920). Also http://archive.org/details/cu31924001549660. Accessed 7 Aug 2018
48. Wimp, J.: Pi and the AGM: a study in analytic number theory and computational complexity. Review in SIAM Rev. **30**, 530–533 (1988)

The Road to Quantum Computational Supremacy

Cristian S. Calude and Elena Calude

This paper is dedicated to the memory of Jon Borwein (1951–2016) whose broad mathematical interests included also quantum computing.

> A hyper-fast quantum computer is the digital equivalent of a nuclear bomb; whoever possesses one will be able to shred any encryption and break any code in existence.[1] [50]

1 Fairy Tales or More Cautionary Tales?

Following the development of Shor's quantum algorithm [81] in 1994 and Grover's quantum algorithm [44] two years later, quantum computing was seen as a bright beacon in computer science, which led to a surge of theoretical and experimental results. The field captured the interest and imagination of the large public and media, and not surprisingly, unfounded claims about the power of quantum computing and its applications proliferated.

A certain degree of pessimism began to infiltrate when experimental groups floundered while attempting to control more than a handful of qubits. Recently, a broad wave of ambitious industry-led research programmes in quantum computing—driven

[1] A typical example of incorrect, largely spread, myth quoted from a recent mystery novel.

C. S. Calude (✉)
Department of Computer Science, University of Auckland, Private Bag 92019,
Auckland, New Zealand
e-mail: c.calude@auckland.ac.nz

E. Calude
Institute of Natural and Computational Sciences, Massey University at Albany, Private Bag
102-904 North Shore MSC, Auckland, New Zealand
e-mail: e.calude@massey.ac.nz

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_22

by D-Wave Systems,[2] the tech giants Google, IBM, Microsoft, Intel and start-ups like Rigetti Computing and Quantum Circuits Incorporated—has emerged[3] and bold claims about a future revolutionised by quantum computing are resurfacing.

Governments are also involved: phase 1 (2015–2019) £330 million of the UK government programme on quantum technologies [89] is rolling and the European Commission has announced a €1 billion initiative in quantum technology [39]. The European flagship quantum programme, whose explicit goal is to stimulate a "second quantum revolution", aims to "build a universal quantum computer able to demonstrate the resolution of a problem that, with current techniques on a supercomputer, would take longer than the age of the universe" by 2035, [78]; see also Figure 1.

Undoubtedly, these programmes are extremely beneficial to the development of various quantum technologies, but are the claims about the future of quantum computing realistic? "We tend to be too optimistic about the short run, too pessimistic about the long run", said recently Preskill [73]; see also [8, 85].

2 Quantum Algorithmics

First and foremost, *quantum computing cannot compute all partial functions a universal Turing machine can calculate because only total functions can be computed by quantum circuits* [13]. Consequently, quantum computing potential advantages could come only from faster than classical computations.

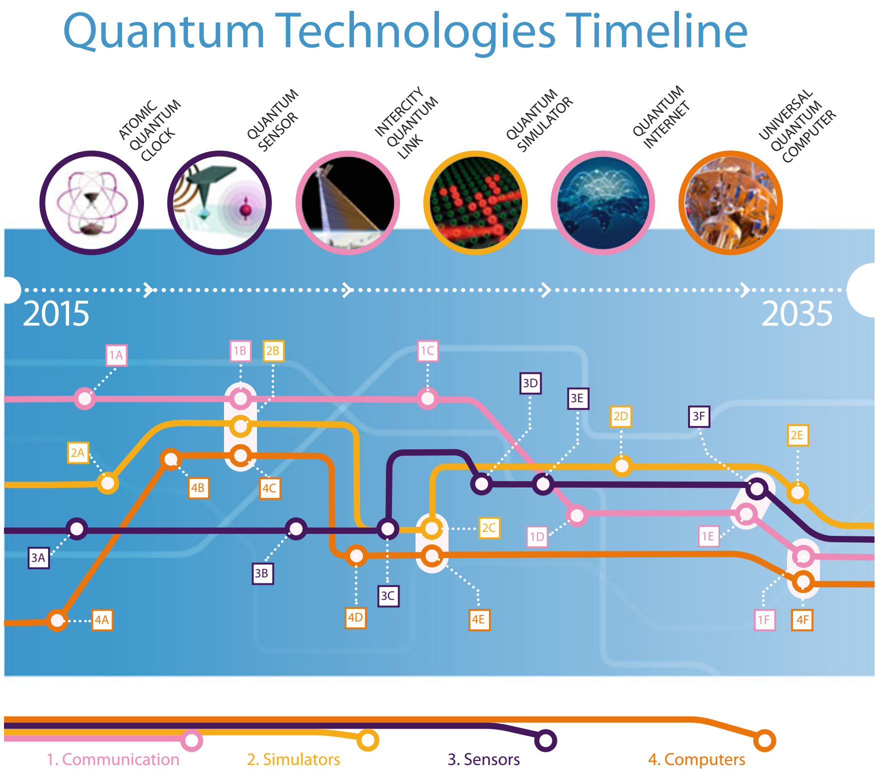
While Shor's algorithm, Deutsch–Jozsa algorithm and various others in the "black-box" paradigm[4] are believed to provide an exponential speed-up over classical computers, this is far from the case in general. We said "believed" because the superiority of Shor's quantum algorithm over classical ones is still an open problem and various techniques allowing efficient classical simulation of quantum algorithms have been successfully developed [6, 26, 43] even for some "black-box" quantum ones [5, 28, 51, 52].

In fact, since the introduction of Shor's and Grover's algorithms some twenty years ago, the development within the field of quantum algorithmics has been rather slow—see [76] for a global picture—and many of them are novel uses of a handful of core quantum algorithms. So, why are there so few quantum algorithms that offer speed-up over classical algorithms? Although written more than a decade ago, Shor's article [82] is still actual:

[2] The company's relatively steady progress in producing and selling the first series of D-Wave quantum computers has gone from 28 qubits in 2007 to more than 2,000 in their 2000Q$^{\text{TM}}$ System machine [35]. In September 2019, the 5,000-qubit D-Wave machine called "Advantage" has been delivered to the Los Alamos National Laboratory [90].

[3] Of course, the industry work is based and has continued the academic efforts, sometimes using successful experimentalists from academia, like Google does.

[4] Where access to a quantum black-box or "oracle" with certain structural properties is assumed.



	1. Communication	2. Simulators	3. Sensors	4. Computers
0 – 5 years	A Core technology of quantum repeaters B Secure point-to-point quantum links	A Simulator of motion of electrons in materials B New algorithms for quantum simulators and networks	A Quantum sensors for niche applications (incl. gravity and magnetic sensors for health care, geosurvey and security) B More precise atomic clocks for synchronisation of future smart networks, incl. energy grids	A Operation of a logical qubit protected by error correction or topologically B New algorithms for quantum computers C Small quantum processor executing technologically relevant algorithms
5 – 10 years	C Quantum networks between distant cities D Quantum credit cards	C Development and design of new complex materials D Versatile simulator of quantum magnetism and electricity	C Quantum sensors for larger volume applications including automotive, construction D Handheld quantum navigation devices	D Solving chemistry and materials science problems with special purpose quantum computer > 100 physical qubit
> 10 years	E Quantum repeaters with cryptography and eavesdropping detection F Secure Europe-wide internet merging quantum and classical communication	E Simulators of quantum dynamics and chemical reaction mechanisms to support drug design	E Gravity imaging devices based on gravity sensors F Integrate quantum sensors with consumer applications including mobile devices	E Integration of quantum circuit and cryogenic classical control hardware F General purpose quantum computers exceed computational power of classical computers

Fig. 1 Quantum timeline: 2015–2035 [78]

The first possible reason is that quantum computers operate in a manner so different from classical computers that our techniques for designing algorithms and our intuitions for understanding the process of computation no longer work. The second reason is that there really might be relatively few problems for which quantum computers can offer a substantial

> speed-up over classical computers, and we may have already discovered many or all of the important techniques for constructing quantum algorithms.

Best quantum algorithms typically provide a quadratic or low-order polynomial speed-up [42]. Furthermore, there are pointers [1, 20] suggesting that quantum computers cannot offer more than a (perhaps small) polynomial advantage for **NP**-complete problems,[5] and such a speed-up would struggle to compete with the heuristic approaches commonly used to solve them in practice. However, even a polynomial-order speed-up could be of significant benefit for problems requiring exact solutions or for problems that can classically be solved in sub-exponential time, like the graph isomorphism problem (see [31]).

Grover's quantum algorithm [44] is an interesting example: access to an unsorted quantum database that can be queried with a quantum input is given, and asked if it contains a specific entry. Grover's algorithm offers a *provable* speed-up. However, the speed-up is not exponential and, more importantly, the problem it solves is far from being realistic: the cost of constructing the quantum database could negate any advantage of the algorithm, and in many classical scenarios one could do much better by simply creating (and maintaining) an ordered database. Using Grover's algorithm as a subroutine for solving problems in image processing is more efficient because the cost of preparing the quantum "database" can be spread out over several calls [59]; this strategy motivated a new hybrid quantum-classical paradigm for embedded quantum annealing algorithms [9]. Other applications are discussed in [66].

Quantum simulation, quantum-assisted optimisation and quantum sampling are believed to offer near-term quantum solutions to hard problems that may lead even to commercialisation [65].

3 What Is Quantum Computational Supremacy?

The quantum computational advantage for simulating quantum systems was first stated by Feynman in 1982, in one of the pioneering papers in quantum computing [41] (the other one was Manin [62]). What is the justification of Feynman's insight? According to the data processing inequality [16, 34], (classical) post-processing cannot increase information. This suggests that to run an accurate classical simulation of a quantum system one must know a lot about the system before the simulation is started [12]. Manin [62] and Feynman [41] have argued that a quantum computer might not need to have so much knowledge. This line of reasoning seemingly inspired Deutsch [37] to state

> **The postulate of quantum computation**: Computational devices based on quantum mechanics will be computationally superior compared to digital computers.

[5] Perhaps the most important class of "difficult computational problems" such as the well-known travelling salesman problem, which have applications in almost every area of science and beyond, from planning and logistics to microchip manufacturing.

A spectacular support for this postulate came from Shor's 1994 polynomial factoring quantum algorithm [81] in spite of the fact that the problem whether factoring is in **P** was, and still is, open. The belief that factoring integers is computationally hard[6] is essential for much of modern cryptography and computing security. In 2002, Hemaspaandra, Hemaspaandra and Zimand [48], improving results in [21, 83], showed that there are tasks on which polynomial-time quantum machines are exponentially faster almost everywhere than any classical—even bounded-error probabilistic—machine.

In 2011, the syntagm "quantum supremacy" was coined and discussed[7] by J. Preskill in his Rapporteur talk "Quantum Entanglement and Quantum Computing" [72] at the 25th *Solvay Conference on Physics* (Brussels, Belgium, 19–22 October 2011):

> We therefore hope to hasten the onset of the era of quantum supremacy, when we will be able to perform tasks with controlled quantum systems going beyond what can be achieved with ordinary digital computers.

Recently, quantum supremacy was described in [22] as follows:

> Quantum supremacy is achieved when a formal computational task is performed with an existing quantum device which cannot be performed using any known algorithm running on an existing classical supercomputer in a reasonable amount of time.

Note the imprecision in the above formulation: the comparison is made with "any known algorithm running on an existing classical supercomputer" and the classical computation takes "a reasonable amount of time". Can this imprecision be decreased or, even better, eliminated? Just as there is no current proof that $\mathbf{P} \neq \mathbf{NP}$—one of the important open problems in classical complexity theory—there is no mathematical proof for the Postulate of quantum computation; in fact, the Postulate is not amenable to a proof. The hypothesis $\mathbf{P} \neq \mathbf{NP}$ can be used for deriving useful results; similarly, adopting assumptions in terms of both quantum physics and classical complexity theory—which can be justified heuristically or experimentally—can lead to precise statements which can be proved or disproved. The following two assumptions

> **The postulate of noise**: Quantum systems are inherently noisy.
>
> **The Extended Church–Turing Thesis**: A probabilistic Turing machine can efficiently simulate any realistic model of computation.

have been used by Kalai [53] to challenge the Postulate of quantum computation. Here "efficiently" means "with at most polynomial overhead"; the adjective "realistic" (or "reasonable" as an alternative) refers to a "physically realisable in principle". It

[6] For results pointing to the opposite assumption see [6, 19, 26, 43, 68].

[7] The use of the word "supremacy"—which denotes "the state or condition of being superior to all others in authority"—was criticised in [91] because the syntagm "white supremacy" is associated with the racial segregation and discrimination of the apartheid regime of South Africa. Proposals like "quantum advantage" or "quantum superiority" have been discussed [77], but to date, none has gained ground.

is worth mentioning that these assumptions are themselves challengeable; see, for example, [22] for the Extended Church–Turing Thesis.

A quantum computational supremacy experiment has to prove both a lower bound and an upper bound. In Google's proposed experiment—to be discussed in detail in Section 6—the upper bound is given by a quantum algorithm running on a quantum computer with 49 qubits[8]—a mathematical fact and an engineering artefact (the construction of the quantum machine); the lower bound is necessary for proving that no current classical computer can simulate the sampling in a reasonable time from the output distributions of pseudo-random quantum circuits.

Upper bounds are positive results while lower bounds are negative. Upper bounds are useful when we want to show that a problem can be solved by a "good" algorithm. But if we want to argue that no algorithm solving a problem can be better than a given one, or perhaps that some problem is so hard that we can't possibly hope to find a good solution to it, we need lower bounds.

In mathematics and theoretical computer science, it is well known that negative results are more difficult to prove than positive ones. In classical computability theory, it is more difficult to prove incomputability than computability, and in complexity theory lower, bounds are more difficult to prove than upper bounds [84]. The superiority of Shor's quantum algorithm [81] is a prime example. A methodology for proving lower bounds in quantum computing is discussed in [45, pp. 144–149]. Sometimes unproved claims about the quantum superiority of a quantum algorithm have been shown to be incorrect: an example is the superiority of Deutsch's quantum algorithm over any classical one, see [28, 37, 51].

Another issue is correctness: how do we know that the quantum computer solution is indeed correct—quantum computing is a probabilistic type of computation—if we can't check it with a reliably tested classical computer? For a promising approach see [11, 25]. Meantime we note that even classical correctness is a very difficult problem. The Ackermann A function [10] is a singular example: computing the value of $A(x, y)$ is prohibitively difficult because the function is computable but not primitive recursive, but testing the predicate $A(x, y) = z$ is very easy [27].

Finally, the discussion about quantum supremacy suggests a misleading comparison between classical and quantum computing. If a quantum computer can outdo **any** classical computer on one problem we have quantum supremacy, even if classical computers could be at least as good as quantum ones in solving many (most) other problems.

Put it bluntly, *quantum supremacy, if achieved, won't make classical computing obsolete.* In fact, the hybrid approach combining quantum and classical computing, briefly mentioned in Section 2, could be a good strategy in solving some (many) difficult problems [9].

[8] A qubit is a 2-state quantum system. There are many ways to build qubits, hence not all qubits are equal. The magic number 49 (or 50) refers to qubits in the quantum circuit model which are more difficult to control than the qubits used by the D-Wave machine [29] (to embed a complete graph of N vertices in D-Wave hardware Chimera graph we need approximately N^2 qubits, so 2,048 D-Wave qubits correspond to about fully connected 45 qubits) or the trapped atom qubits used by specialised quantum simulators [18, 93].

4 Criteria for Quantum Computational Supremacy

Harrow and Montanaro [46] have proposed a reasonable list of criteria for a quantum supremacy experiment. According to them, we need to have:

1. a well-defined computational problem,
2. a quantum algorithm solving the problem which can run on a near-term hardware capable of dealing with noise and imperfections,
3. an amount of computational resources (time/space) allowed to any classical competitor,
4. a small number of well-justified complexity-theoretic assumptions, and
5. a verification method that can efficiently distinguish between the performances of the quantum algorithm from **any** classical competitor using the allowed resources.

Large integer factoring is a typical problem for a quantum supremacy experiment. Indeed, it is well defined, it has huge practical importance, there are efficient quantum algorithms solving it (Shor's algorithm and variants [19, 81]), the complexity-theoretic assumption is that no classical algorithm can factor essentially faster than the current ones and the solution is quickly verifiable. This seems an almost ideal candidate, except for (a) the strong complexity-theoretic assumption [68] and (b) the lack of a near-term hardware running such a quantum algorithm for sufficiently large integers (say a 2,048-bit number), see [46]. A possible solution for (b) could be a hybrid (quassical) approach [9].

Harrow and Montanaro [46] state that "we do not require that the computational task[9] is of practical interest". This is a strong assumption in itself which is adequate only for a foundational study.

Table 1 in [46], p. 205, lists seven plausible approaches to quantum computational supremacy: factoring, single photons passing through a linear-optical network (boson sampling), quantum circuits on many qubits and only a few layers of quantum gates (low-depth circuits), random quantum circuits containing gates that either all commute or do not commute (instantaneous quantum polynomial time, IQP), quantum approximate optimisation algorithms (QAOA), quantum adiabatic optimisation and quantum analogue simulation. These approaches are then evaluated according to usefulness, assumption implying no classical simulation and difficulties to solve on a quantum computer and to verify. Factoring is the only useful problem, simulation is often useful, adiabatic optimisation could be useful and the remaining three problems do not seem to be useful. Factoring is the hardest to solve on a quantum computer, boson sampling, adiabatic optimisation and analogue simulation are easy and the remaining three are moderately difficult. Only factoring is easy to verify. The complexity-theoretic assumptions are generally very strong, assessing their plausibility is a very difficult task and, generally, conclusions are rather controversial. A detailed complexity-theoretic analysis of various possible quantum supremacy experiments can be found in [4]. The papers [4, 46] are exceptionally singular in offering balanced and more formal analyses.

[9] Their formulation for what we call a computational problem.

5 Is the Quest for Quantum Computational Supremacy Worthwhile?

Apart publicity and marketing, is the effort of demonstrating the quantum computational supremacy justified? What are the (possible) benefits? Can the claim of quantum computational supremacy be falsified?

We will start with the second question. The main benefit could be foundational and philosophical: a better understanding of the nature of quantum mechanics through its computational capabilities.[10] Such a gain will boost the efforts of not only building larger scale quantum computers but also, and, more importantly, developing new and powerful algorithms for these machines possibly leading to solutions to important practical problems. From this perspective, the answer to the first question is affirmative.

Let us examine closer the foundational gain. A successful quantum supremacy experiment could be a complement to Bell experiment: the latter refuted local hidden models of quantum mechanics, while the former *seems* to invalidate the Extended Church–Turing Thesis [92]. The paper [46] discusses the advantages of a successful quantum supremacy experiment, even one that barely surpasses any classical competitor, illustrated with hard-to-simulate classical systems like protein folding or fluid dynamics. Here we suggest a different perspective which motivated the tentative formulation above. The Extended Church–Turing Thesis—which incidentally has nothing to do with either Church nor Turing—is a foundational principle of classical complexity theory which ensures that the polynomial-time class **P** is well defined.[11] The Thesis places strong constraints, one of them being that *the model of computation is digital*. For example, analogue computers are excluded because they assume infinite arithmetic precision. Furthermore, it is known that an infinite precision calculator with operations $+$, $\times$, $=0?$, can factor integers in polynomial time (see [80, 87]).[12] But, are quantum computers a "reasonable" model of computation? Are quantum systems digital? At first glance quantum computers (and, more generally, quantum systems) appear to be analogue devices, since a quantum gate is described by a unitary transformation, specified by complex numbers; a more in-depth analysis is still required.

What does it take to refute the claim of quantum computational supremacy? This amounts to prove that any computation performed by any quantum computer can be simulated by a classical machine in polynomial time, a weaker form of the Extended Church–Turing Thesis. This statement cannot be proved for the same reasons the

[10] A beautiful result regarding the computational power of algorithmic random strings was proved in [33]. This was used as a test of quality for quantum randomness in [30].

[11] The Thesis equating feasible computation with polynomial-time computation has significantly less "evidence" than the Church–Turing Thesis; in fact, according to [36], it "lacks evidence".

[12] Feynman's 1982 intuition (Section 3) was substantiated in [61] by running a quantum analogue emulation. The quantum version of analogue computers, continuous-variable quantum computers, have been theoretically studied [54]; the model in [86] offers a universal gate set for both qubits and continuous variables.

Church–Turing Thesis cannot be proved: obviously, they may be disproved. The paper [69] presents efficient classical boson sampling algorithms and a theoretical analysis of the possibility of scaling boson sampling experiments; it concludes that "near-term quantum supremacy via boson sampling is unlikely".

6 Google Quantum Computational Supremacy

In the landscape of various proposals for quantum computational supremacy experiments, Google's approach is not only well documented, but had chances to be completed really very soon [67]. The proposed experiment is not about solving a problem: it is the computational task of sampling from the output distribution of pseudo-random quantum circuits built from a universal gate set.[13] This computational task is difficult because as the grid size increases, the *memory needed to store everything increases classically exponentially*.[14] The required memory for a $6 \times 4 = 24$-qubit grid is just 268 megabytes, less than the average smartphone, but for a $6 \times 7 = 42$-qubit grid it jumps to 70 terabytes, roughly 10,000 times that of a high-end PC. Google has used Edison, a supercomputer housed by the US National Energy Research Scientific Computing Center and ranked 72 in the Top500 List [40], to simulate the behaviour of the grid of 42 qubits. The classical simulation stopped at this stage because going to the next size up *was thought to be currently impossible: a* 48*-qubit grid would require 2,252 petabytes of memory, almost double that of the top supercomputer in the world.* The path to quantum computational supremacy was obvious: if Google could solve the problem with a 50-qubit quantum computer, it would have beaten every other computer in existence.

The abstract of the main paper describing the theory behind the experiment [22] reads[15]:

> A critical question for the field of quantum computing in the near future is whether quantum devices without error correction can perform a well-defined computational task beyond the capabilities of state-of-the-art classical computers, achieving so-called quantum supremacy. *We study the task of sampling from the output distributions of (pseudo-)random quantum circuits, a natural task for benchmarking quantum computers. Crucially, sampling this distribution classically requires a direct numerical simulation of the circuit, with computational cost exponential in the number of qubits.* This requirement is typical of chaotic systems. *We extend previous results in computational complexity to argue more formally that this sampling task must take exponential time in a classical computer.* We study the convergence to the chaotic regime using extensive supercomputer simulations, modelling circuits with up to 42 qubits—the largest quantum circuits simulated to date for a computational task that approaches quantum supremacy. We argue that while chaotic states are extremely sensitive to errors, quantum supremacy can be achieved in the near-term with approximately fifty superconducting qubits. We introduce cross entropy as a useful benchmark of quantum circuits which approximates the circuit fidelity. We show that the cross entropy can be efficiently

[13] For another promising quantum simulation see [32].

[14] But, do we *really need* to store everything?

[15] Our emphasis.

> measured when circuit simulations are available. *Beyond the classically tractable regime, the cross entropy can be extrapolated and compared with theoretical estimates of circuit fidelity to define a practical quantum supremacy test.*

Google was on track to deliver before the end of the year. Alan Ho, an engineer in Google's quantum AI lab, revealed the company's progress at a quantum computing conference in Munich, Germany. According to [79]:

> His team is currently working with a 20-qubit system that has a "two-qubit fidelity" of 99.5 per cent—a measure of how error-prone the processor is, with a higher rating equating to fewer errors. For quantum supremacy, Google will need to build a 49-qubit system with a two-qubit fidelity of at least 99.7 per cent. Ho is confident his team will deliver this system by the end of this year.

Let us note that many, if not most, discussions about quantum computational supremacy focus on the most exciting possibilities of quantum computers, namely, the upper bound. What about the lower bound? The article [22] refers cautiously to the lower bound in the abstract: "We extend previous results in computational complexity *to argue more formally* that this sampling task must take exponential time in a classical computer". Indeed, they do not claim to have a proof for the lower bound, just a "better formal argument". Their argument is reinforced later in the introduction:

> State-of-the-art supercomputers cannot simulate universal random circuits of sufficient depth in a 2D lattice of approximately 7×7 qubits with any known algorithm and significant fidelity.

Does Google's experiment satisfy the criteria discussed in Section 4? The problem is well defined, albeit a simulation, not a computational problem,[16] the quantum algorithm solving the problem will run on a quantum computer—promised to be built before the end of 2017[17]—capable of dealing with noise and imperfections, the classical competitor would be allowed a reasonable amount of computational resources and there is a plausible verification. The weakest part comes from the complexity-theoretic assumption [22]:

> **Memory assumption**. Sampling this distribution classically requires a direct numerical simulation of the circuit, with computational cost exponential in the number of qubits.

The assumption was corroborated by the statement:

> Storing the state of a 46-qubit system takes nearly a petabyte of memory and is at the limit of the most powerful computers. [67]

[16] One could argue that the task itself is rather uninteresting and without obvious applications. Indeed, all the time nature is doing quantum "things" that we don't know how to solve classically. For example, the structure of atoms can in general only be determined experimentally, but nature manages it with near-perfect fidelity. If Google achieved the goal—an undisputable big technical feat—the meaning of the achieved "supremacy" could still be debatable.

[17] When pressed for an update, a spokesperson [for Google] recently said that 'we hope to announce results as soon as we can, but we're going through all the detailed work to ensure we have a solid result before we announce'. Reference [15], 24 January 2018. The goal was not reached as of 30 September 2019.

7 IBM Challenge

The Memory assumption is crucial for the proposed lower bound, and, indeed, this was confirmed very soon. The paper [71] proved that a supercomputer can simulate sampling from random circuits with low depth (layers of gates) of up to 56 qubits.

> With the current rate of progress in quantum computing technologies, 50-qubit systems will soon become a reality. To assess, refine and advance the design and control of these devices, one needs a means to test and evaluate their fidelity. This in turn requires the capability of computing ideal quantum state amplitudes for devices of such sizes and larger. In this study, we present a new approach for this task that significantly extends the boundaries of what can be classically computed. We demonstrate our method by presenting results obtained from a calculation of the complete set of output amplitudes of a universal random circuit with depth 27 in a 2D lattice of 7 × 7 qubits. We further present results obtained by calculating an arbitrarily selected slice of 237 amplitudes of a universal random circuit with depth 23 in a 2D lattice of 8 × 7 qubits. Such calculations were previously thought to be impossible due to impracticable memory requirements. *Using the methods presented in this paper, the above simulations required 4.5 and 3.0 TB of memory, respectively, to store calculations, which is well within the limits of existing classical computers.*[18]

Better results have been quickly announced, see, for example, [23]. The limits of classical simulation are not only unknown but hard to predict.

In spite of this, IBM has announced a prototype of a 50-qubit quantum computer, stating that it "aims to demonstrate capabilities beyond today's classical systems" with quantum systems of this size [49].

8 Latest Developments

At 2018 Consumer Electronics Show in Las Vegas, Intel CEO Brian Krzanich reported, "the successful design, fabrication and delivery of a 49-qubit superconducting quantum test chip" [56]. The 49-qubit superconducting quantum test chip is called "Tangle Lake" after a chain of lakes in Alaska known for extreme cold temperatures. At the event, Mike Mayberry, managing director of Intel Labs said: "We expect it will be five to seven years before the industry gets to tackling engineering-scale problems, and it will likely require 1 million or more qubits to achieve commercial relevance". In [74] J. Preskill aptly said: "Quantum computers with 50-100 qubits may be able to perform tasks which surpass the capabilities of today's classical digital computers, but noise in quantum gates will limit the size of quantum circuits that can be executed reliably. …Quantum technologists should continue to strive for more accurate quantum gates and, eventually, fully fault-tolerant quantum computing". Jay Gambetta, from IBM Thomas J. Watson Research Center believes that "a universal fault-tolerant quantum computer, which has to use logical qubits, is still a long way off", [15]. E. Tang (then an 18-year-old undergraduate student at UT Austin) has recently proved [88] that classical computers can solve the "recommendation

[18] Our emphasis.

problem"—given incomplete data on user preferences for products, can one quickly and correctly predict which other products a user will prefer?—with performance comparable to that of a quantum computer. Is this significant? **Yes**, because quantum computer scientists had considered this problem to be one of the best examples of a problem that quantum computers can solve exponentially faster than their classical ones and the quantum solution in [55] was hailed as one of the first examples in *quantum machine learning and big data* that would be unlikely to be done classically… In October 2018, Bravyi, Gosset and Köning [24] have presented an argument—based on non-locality—which suggests that a certain quantum algorithm requiring only constant-depth quantum circuits can be a suitable candidate for showing quantum computational supremacy.

9 Closing Remarks

Recall that the computational power of quantum computing is less than that of a universal Turing machine [13], so quantum computing potential advantages could come only from faster than classical computations.

Does the paper [71] destroy the quest for quantum computational supremacy? Is there any incompatibility between the classical simulation reported in [71] and the IBM statement cited at the end of Section 7? Tentatively we answer with no to both questions. The following paragraph [2] is relevant:

> This paper[19] does not undercut the rationale for quantum supremacy experiments. The truth, ironically, is almost the opposite: it being possible to simulate 49-qubit circuits using a classical computer is a precondition for Google's planned quantum supremacy experiment, because it's the only way we know to check such an experiment's results! The goal, with sampling-based quantum supremacy, was always to target the "sweet spot", which we estimated at around 50 qubits, where classical simulation is still possible, but it's clearly orders of magnitude more expensive than doing the experiment itself. If you like, the goal is to get as far as you can up the mountain of exponentiality, conditioned on people still being able to see you from the base. Why? Because you can. Because it's there.[20] Because it challenges those who think quantum computing will never scale: explain this, punks! But there's no point unless you can verify the result.

Here are a few more lessons. The first is not to underestimate the importance of mathematical modelling and proving (lower bounds, in particular). As the title of the blog [2] says, "2^n is exponential, but 2^{50} is finite", the difference between exponential and polynomial running times is asymptotic and in some concrete cases it is a challenge to find finite evidence for the difference. Furthermore, proving that a problem is in **P** itself is not a guarantee that there is an algorithm in **P** that is practically useful: primality has been known to be in **P** since 2002, but all known deterministic algorithms are too slow in practice, so probabilistic tests of primality continue to be used.

[19] That is, [71].

[20] "It is not the mountain we conquer but ourselves", as Edmund Hillary aptly said.

Second, the conversation on quantum computing, quantum cryptography and their applications needs an infusion of modesty (if not humility), more technical understanding and clarity as well as less hype. Raising false expectations could be harmful for the field.

Third, a trend in quantum computing is emerging: when a problem is solved efficiently in quantum computing, it draws more attention and often produces better classical alternatives than existed before. Some of the new efficient classical solutions, see, for example, [5, 7, 28, 51, 52, 88], have been directly inspired by the quantum work.

Finally, the race quantum versus classical is running so fast—a sample is given by the references posted/published since October 2017, the month when the paper [71] was posted—that by the time this paper is printed some results discussed here could be obsolete. One fact is certain: as of 30 September 2019, the quantum computational supremacy was not (yet?) demonstrated.

10 P.S. Quantum Desperation

Two[21] important articles have been published in quantum computing on 23–24 October 2019. The first, written by a Google team and published in the prestigious journal *Nature* [14], announces the experimental realisation of quantum supremacy with a programmable machine with 53 qubits:

> *Our Sycamore processor takes about 200 seconds to sample one instance of a quantum circuit a million times our benchmarks currently indicate that the equivalent task for a state-of-the-art classical supercomputer would take approximately 10,000 years. This dramatic increase in speed compared to all known classical algorithms is an experimental realization of quantum supremacy for this specific computational task, heralding a much-anticipated computing paradigm.*

This paper has sparked a huge interest not only in the quantum community but the whole world. Announcements and comments have instantly appeared in prestigious science magazines like *New Scientist*, "Google reigns supreme" and "It's official: Google has achieved quantum supremacy", major newspapers like *The Washington Post*, "Bravo for Google's 'quantum supremacy.' Here's what needs to happen next" and worldwide broadcasters like *BBC*, "Google claims 'quantum supremacy' for computer". Not everybody was convinced even at an intuitive level of understanding: *Reuters*: "Google unveils quantum computer breakthrough; critics say wait a qubit", *The Financial Post*: "Google claims 'quantum supremacy' with quantum computer breakthrough, but skeptics don't agree", to cite only two sources.

The inventor of the concept of quantum supremacy is also cautious [75]:

> The Google team has apparently demonstrated that it's now possible to build a quantum machine that's large enough and accurate enough to solve a problem we could not solve before…

[21] Section 10 added on January 24, 2020.

The second paper, written by an IBM team, was posted in the archive [70] and is summarised as follows:

> We argue that an ideal simulation of the same task can be performed on a classical system in 2.5 days and with far greater fidelity. This is in fact a conservative, worst-case estimate, and we expect that with additional refinements the classical cost of the simulation can be further reduced.

Could both reports be correct?

Interestingly, immediately after publishing the paper [14] *Nature* published also an anonymous editorial [38] including the following significant paragraphs:

> As the world digests this achievement – including the claim that some quantum computational tasks are beyond supercomputers – it is too early to say whether supremacy represents a new dawn for information technology. … At the very least, quantum computers as a routine part of life are likely to be decades or more into the future.
>
> …
>
> Instead of proceeding with caution, a quantum gold rush is under way, with investors joining governments and companies to pour large sums of money into developing quantum technologies. Unrealistic expectations are being fuelled that powerful general-purpose quantum computers could soon be on the horizon. Such misguided optimism could be dangerous for the future of this still-fledgling field.

Undoubtedly, Google's technological achievement is remarkable, and it helps building the case for a possible quantum supremacy by achieving a high upper bound. *The real problem is that there is no formal argument for the lower bound,* see Section 4, the supplementary material to [14] and the mathematical discussion in [60]. Furthermore, *it is not for IBM*[22] *or anybody else to disprove the lower bound claimed by Google*[23]: *the onus is on Google to prove it.*

Where does "desperation" in the title of this section come from? As noted in [47]

> It has taken Google 13 years[24] to get this far. Without a profitable device, research could dry up. It happened to Apollo, programme. It has happened at times with AI.

There are very few agreements in quantum computing, but one is that the area has been showered with money in recent years but has delivered very little practical solutions. How long can the flow of money continue? There is a sense that the answer is not too encouraging, so something had/has to be done. Downgrading the mathematical notion of quantum speed-up[25] to quantum supremacy was meant to help, but not without a price. The origin, merit and pitfalls of this concept have been recently discussed by its inventor J. Preskill in a thoughtful article in *Quanta Magazine* [75]. One main objection pointed there is that the "word exacerbates the already overhyped

[22] Although they did before [71].

[23] Comments like "Tellingly, not even IBM thinks the simulation would be especially easy – nor, as of this writing, has IBM actually carried it out". Reference [3] are irrelevant.

[24] And a tone of money (our comment).

[25] Note that Grover's quantum algorithm [44] *proved a quantum speed-up* 25 years ago, yet insufficient to justify quantum computing practicality.

reporting on the status of quantum technology". This was echoed also in the IBM paper [70]:

> For the reasons stated above, and since we already have ample evidence that the term "quantum supremacy" is being broadly misinterpreted and causing ever growing amounts of confusion, we urge the community to treat claims that, for the first time, a quantum computer did something that a classical computer cannot with a large dose of skepticism due to the complicated nature of benchmarking an appropriate metric.

The current tendency seems to move the arguments from the mathematics and science to media propaganda.

> Google's demonstration should give these skeptics pause. To all appearances, a 53-qubit device really was able to harness 9 quadrillion amplitudes for computation, surpassing (albeit for a special, useless task) all the supercomputers on earth. Quantum mechanics worked: an outcome that's at once expected and mind-boggling, conservative and radical. [3]

Interestingly, this is not a new tactic in settling quantum mechanics controversies and a most prominent example is the famous Einstein–Bohr disagreement on the Copenhagen interpretation. Einstein view [17, p. 29]:

> The theory reminds me a little of the system of delusions of an exceedingly intelligent paranoiac.

opposed Bohr's "shut up and calculate!" attitude (using Mermin's expression [64]). According to Lakatos [58, pp. 59–60], [57, p. 105]:

> After 1925, Bohr and his associates introduced a new and unprecedented lowering of critical standards for scientific theories. This led to a defeat of reason within modern physics and to an anarchist cult of incomprehensible chaos.

Recently, there is an apparent change [63, p. 9]:

> while Einstein won and would continue to win all the logical battles, Bohr was decisively winning the propaganda war.

Let's hope that in quantum computing mathematics and science will prevail over propaganda.

Acknowledgements We thank N. Allen for fruitful discussions and suggestions, specifically for insight on Feynman's paper [41], and R. Brent, R. Goyal, L. Hemaspaandra, K. Pudenz, R. Hua, K. Svozil and an anonymous referee for excellent critical comments and suggestions. This work has been supported in part by the Quantum Computing Research Initiatives at Lockheed Martin.

References

1. Aaronson, S.: The limits of quantum. Sci. Am. 62–69 (2008)
2. Aaronson, S.: Shtetl-optimized – 2^n is exponential, but 2^{50} is finite (2017). https://www.scottaaronson.com/blog/?p=3512

3. Aaronson, S.: Why Google's quantum supremacy milestone matters. The New York Times (2019). https://www.nytimes.com/2019/10/30/opinion/google-quantum-computer-sycamore.html
4. Aaronson, S., Chen, L.: Complexity-theoretic foundations of quantum supremacy experiments. Technical report No. 200, Electronic Colloquium on Computational Complexity (2016)
5. Abbott, A.A.: The Deutsch-Jozsa problem: De-quantisation and entanglement. Nat. Comput. **11**(1), 3–11 (2011)
6. Abbott, A.A.: De-quantisation of the quantum Fourier transform. Appl. Math. Comput. **291**(1), 3–13 (2012)
7. Abbott, A.A., Calude, C.S.: Understanding the quantum computational speed-up via de-quantisation. EPTCS **26**, 1–12 (2010)
8. Abbott, A.A., Calude, C.S.: Limits of quantum computing: a sceptic's view (presented by Jon Borwein). Quantum for quants (2016). http://www.quantumforquants.org/quantum-computing/limits-of-quantum-computing/
9. Abbott, A.A., Calude, C.S., Dinneen, M.J., Hua, R.: A hybrid quantum-classical paradigm to mitigate embedding costs in quantum annealing. Int. J. Quantum Inf. **1950042**, 40 (2019)
10. Ackermann, W.: On Hilbert's construction of the real numbers. Math. Ann. **99**, 118 (1928)
11. Aharonov, D., Ben-Or, M., Eban, E., Mahadev, U.: Interactive proofs for quantum computations (2017). https://arxiv.org/abs/1704.04487
12. Allen, N.: Email to C. S. Calude. Accessed 19 Nov 2017
13. Allen, E.H., Calude, C.S.: Quassical computing. Int. J. Unconv. Comput. **14**, 43–57 (2018)
14. Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J.C., Barends, R., Biswas, R., Boixo, S., Brandao, F.G.S.L., Buell, D.A., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Collins, R., Courtney, W., Dunsworth, A., Farhi, E., Foxen, B., Fowler, A., Gidney, C., Giustina, M., Graff, R., Guerin, K., Habegger, S., Harrigan, M.P., Hartmann, M.J., Ho, A., Hoffmann, M., Huang, T., Humble, T.S., Isakov, S.V., Jeffrey, E., Jiang, Z., Kafri, D., Kechedzhi, K., Kelly, J., Klimov, P.V., Knysh, S., Korotkov, A., Kostritsa, F., Landhuis, D., Lindmark, M., Lucero, E., Lyakh, D., Mandrá, S., McClean, J.R., McEwen, M., Megrant, A., Mi, X., Michielsen, K., Mohseni, M., Mutus, J., Naaman, O., Neeley, M., Neill, C., Niu, M.Y., Ostby, E., Petukhov, A., Platt, J.C., Quintana, C., Rieffel, E.G., Roushan, P., Rubin, N.C., Sank, D., Satzinger, K.J., Smelyanskiy, V., Sung, K.J., Trevithick, M.D., Vainsencher, A., Villalonga, B., White, T., Yao, Z.J., Yeh, P., Zalcman, A., Neven, H., Martinis, J.M.: Quantum supremacy using a programmable superconducting processor. Nature **574**, 505–510 (2019)
15. Ball, P.: The era of quantum computing is here. Outlook: Cloudy, Quanta Magazine (2018). https://www.quantamagazine.org/the-era-of-quantum-computing-is-here-outlook-cloudy-20180124
16. Beaudry, N.J., Renner, R.: An intuitive proof of the data processing inequality. Quantum Inf. Comput. **12**(5–6), 432–441 (2012)
17. Becker, A.: What Is Real? The Unfinished Quest for the Meaning of Quantum Physics. Basic Books, New York (2018)
18. Bernien, H., Schwartz, S., Keesling, A., Levine, H., Omran, A., Pichler, H., Choi, S., Zibrov, A.S., Endres, M., Greiner, M., Vuletić, V., Lukin, M.D.: Probing many-body dynamics on a 51-atom quantum simulator. Nature **551**, 579, EP –, 11 (2017)
19. Bernstein, D.J., Heninger, N., Lou, P., Valenta, L.: Post-quantum RSA (2017). https://cr.yp.to/papers/pqrsa-20170419.pdf
20. Bernstein, E., Vazirani, U.: Quantum complexity theory. In: Proceedings of the 25th Annual ACM Symposium on Theory of Computing, San Diego, California, 16–18 May 1993, pp. 11–20. ACM Press (1993)
21. Berthiaume, A., Brassard, G.: Oracle quantum computing. J. Mod. Opt. **41**, 195–199 (1992)
22. Boixo, S., Isakov, S.V., Smelyanskiy, V.N., Babbush, R., Ding, N., Jiang, Z., Bremner, M.J., Martinis, J.M., Neven, H.: Characterizing quantum supremacy in near-term devices (2017). arXiv:1608.00263 [quant-ph]
23. Boixo, S., Isakov, S.V., Smelyanskiy, V.N., Neven, H.: Simulation of low-depth quantum circuits as complex undirected graphical models (2018). https://arxiv.org/pdf/1712.05384.pdf

24. Bravyi, S., Gosset, D., Köning, R.: Quantum advantage with shallow circuits. Science **362**, 308–311 (2018)
25. Broadbent, A.: How to verify a quantum computation. Theory Comput. **14**, 1–37 (2018)
26. Browne, D.E.: Efficient classical simulation of the quantum Fourier transform. New J. Phys. **9**(5), 146 (2007)
27. Calude, C.: Super-exponentials nonprimitive recursive, but rudimentary. Inf. Process. Lett. **25**(5), 311–316 (1987)
28. Calude, C.S.: De-quantizing the solution of Deutsch's problem. Int. J. Quantum Inf. **5**(3), 409–415 (2007)
29. Calude, C.S., Calude, E., Dinneen, M.J.: Adiabatic quantum computing challenges. ACM SIGACT News **46**(1), 40–61 (2015)
30. Calude, C.S., Dinneen, M.J., Dumitrescu, M., Svozil, K.: Experimental evidence of quantum randomness incomputability. Phys. Rev. A **82**(2), 022102 (2010)
31. Calude, C.S., Dinneen, M.J., Hua, R.: QUBO formulations for the graph isomorphism problem and related problems. Theor. Comput. Sci. 1950042–40 (2019). https://doi.org/10.1142/S0219749919500424
32. Campbell, E.: Random compiler for fast Hamiltonian simulation. Phys. Rev. Lett. **123**, 070503 (2019)
33. Chaitin, G.J., Schwartz, J.T.: A note on Monte Carlo primality tests and algorithmic information theory. Commun. Pure Appl. Math. **31**(4), 521–527 (1978)
34. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
35. D-Wave Systems (2017). https://www.dwavesys.com/d-wave-two-syste
36. Davis, M.: Interview with Martin Davis. Not. Am. Math. Soc. **55**(560–571) (2008)
37. Deutsch, D.: Quantum theory, the Church-Turing principle and the universal quantum computer. Proc. R. Soc. Lond. Ser. A Math. Phys. Sci (1934–1990) **400**(1818), 97–117 (1985)
38. Editorial: A precarious milestone for quantum computing quantum computing will suffer if supremacy is overhyped. Everyday quantum computers are still decades away. Nature **574**, 453–454 (2019)
39. European flagship quantum programme (2017)
40. Edison supercomputer in TOP 500 ranking (2017). https://www.top500.org/list/2017/06/?page=1
41. Feynman, R.P.: Simulating physics with computers. Int. J. Theor. Phys. **21**, 467–488 (1982)
42. Fürer, M.: Solving NP-Complete problems with quantum search. In: Laber, E.S., Bornstein, C., Nogueira, L.T., Faria, L. (eds.) LATIN 2008: Theoretical Informatics. LNCS, vol. 4957, pp. 784–792. Springer, Berlin (2008)
43. Griffiths, R., Niu, C.: Semiclassical Fourier transform for quantum computation. Phys. Rev. Lett. **76**(17), 3228–3231 (1996)
44. Grover, L.K.: A fast quantum mechanical algorithm for database search. In: Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, pp. 212–219. ACM Press (1996)
45. Gruska, J.: Quantum Computing. McGraw-Hill, London (1999)
46. Harrow, A.W., Montanaro, A.: Quantum computational supremacy. Nature **549**(7671), 203–209 (2017)
47. Havean, D.: Beware quantum winter. Google's quantum breakthrough is the first step on a long road. Let's make sure we don't stumble. New Scientist, p. 21. Accessed 2 Nov 2019
48. Hemaspaandra, E., Hemaspaandra, L.A., Zimand, M.: Almost-everywhere superiority for quantum polynomial time. Inf. Comput. **175**(2), 171–181 (2002)
49. IBM builds 50-qubit quantum computer (2017). http://techvibesnow.com/ibm-builds-50-qubit-quantum-computer/
50. Ignatius, D.: The Quantum Spy. W. W. Norton, New York (2018)
51. Johansson, N., Larsson, J.-Å.: Efficient classical simulation of the Deutsch-Jozsa and Simon's algorithms. Quantum Inf. Process. **16**(9), 233 (2017)
52. Johansson, N., Larsson, J.-Å.: Quantum simulation logic, oracles, and the quantum advantage. Entropy **21**(8), 800 (2019)

53. Kalai, G.: How quantum computers fail: quantum codes, correlations in physical systems, and noise accumulation (2011). arXiv:1106.0485 [quant-ph]
54. Kendon, V.M., Nemoto, K., Munro, W.J.: Quantum analogue computing. Philos. Trans. R. Soc. Lond. A: Math. Phys. Eng. Sci. **368**(1924), 3609–3620 (2010)
55. Kerenidis, I., Prakash, A.: Quantum recommendation system. In: Papadimitrou, C.H. (ed.) 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), pp. 49:1–49:21. Dagstuhl Publishing, Germany, Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2017)
56. Krzanich, B.: 2018 CES: Intel advances quantum and neuromorphic computing research (2018). https://newsroom.intel.com/news/intel-advances-quantum-neuromorphic-computing-research/
57. Lakatos, I.: Falsification and the methodology of scientific research programmes. In: Can Theories be Refuted?, pp. 59–60. Springer, Dordrecht (1976)
58. Lakatos, I.: Falsification and the methodology of scientific research programmes. In: The Methodology of Scientific Research Programmes. Philosophical Papers Volume 1. Cambridge University Press, Cambridge (1978). Online publication (2012)
59. Lanzagorta, M., Uhlmann, J.K.: Hybrid quantum-classical computing with applications to computer graphics. In: ACM SIGGRAPH 2005 Courses, SIGGRAPH '05. ACM, New York (2005)
60. Lipton, R.J., Regan, K.W.: Quantum supremacy at last? https://rjlipton.wordpress.com/2019/10/27/quantum-supremacy-at-last/. Accessed 27 Oct 2019
61. Lloyd, S.: Universal quantum simulators. Science **273**(5278), 1073–1078 (1996)
62. Manin, Y.I.: Vychislimoe i nevychislimoe [Computable and Noncomputable] (in Russian). Sov. Radio. 13–15 (1980). (Last assecced 30 Nov 2017). http://www.worldcat.org/title/vychislimoe-i-nevychislimoe/oclc/11674220
63. Maudlin, T.: The defeat of reason. Boston Rev. (2018). http://bostonreview.net/science-nature-philosophy-religion/tim-maudlin-defeat-reason
64. Mermin, N.D.: What is wrong with this pillow? Phys. Today **42**(4), 9 (1989)
65. Mohseni, M., Read, P., Neven, H., Boixo, S., Denchevand, V., Babbushand, R., Fowler, A., Smelyanskiy, V., Martinis, J.: Commercialize early quantum technologies. Nature **543**, 171–174 (2017)
66. Montanaro, A.: Quantum algorithms: an overview. Npj Quantum Inf. **2**, 15023, EP –, 01 (2016)
67. Neill, C., Roushan, P., Kechedzhi, K., Boixo, S., Isakov, S.V., Smelyanskiy, V., Barends, R., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Dunsworth, A., Fowler, A., Foxen, B., Graff, R., Jeffrey, E., Kelly, J., Lucero, E., Megrant, A., Mutus, J., Neeley, M., Quintana, C., Sank, D., Vainsencher, A., Wenner, J., White, T.C., Neven, H., Martinis, J.M.: A blueprint for demonstrating quantum supremacy with superconducting qubits. arXiv:1709.06678 [quant-ph]
68. Nene, M.J., Upadhyay, G.: Shor's algorithm for quantum factoring. In: Choudhary, R.K., Mandal, J.K., Auluck, N., Nagarajaram, H.A. (eds.) Advanced Computing and Communication Technologies: Proceedings of the 9th ICACCT, 2015, pp. 325–331. Springer Singapore, Singapore (2016)
69. Neville, A., Sparrow, C., Clifford, R., Johnston, E., Birchall, P.M., Montanaro, A., Neville, A.L.A., Sparrow, C., Clifford, R., Johnston, E., Birchall1, P.M., Montanaro4, A., Laing, A.: No imminent quantum supremacy by boson sampling (2017). https://arxiv.org/pdf/1705.00686.pdf
70. Pednault, E., Gambetta, J.: On "Quantum Supremacy". https://www.ibm.com/blogs/research/2019/10/on-quantum-supremacy/. Accessed 24 Oct 2019
71. Pednault, E., Gunnels, J.A., Nannicini, G., Horesh, L., Magerlein, T., Solomonik, E., Wisnieff, R.: Breaking the 49-qubit barrier in the simulation of quantum circuits (2017). https://arxiv.org/abs/1710.05867
72. Preskill, J.: Quantum computing and the entanglement frontier. In: Gross, D., Henneaux, M., Sevrin, A. (eds.) The Theory of the Quantum World, pp. 63–80. World Scientific Publishing, Singapore (2012). arXiv:1203.5813 [quant-ph]
73. Prreskill, J.: BES Roundtable on quantum computing opportunities in chemical and materials sciences. http://www.theory.caltech.edu/~preskill/talks/DOE_BES_2017_Preskill.pdf. Accessed 31 Oct 2017

74. Preskill, J.: Quantum computing in the NISQ era and beyond (2018). https://arxiv.org/abs/1801.00862
75. Preskill, J.: Why I Called It ‘Quantum Supremacy’ (2019). https://www.quantamagazine.org/john-preskill-explains-quantum-supremacy-20191002/
76. Quantum algorithm zoo (2017). http://math.nist.gov/quantum/zoo/
77. Quantum advantage. The quantum pontiff (2017). http://dabacon.org/pontiff/?p=11863
78. Quantum Manifesto (2016). http://qurope.eu/system/files/u7/93056_Quantum%20Manifesto_WEB.pdf
79. Reynolds, M.: Google on track for quantum computer breakthrough by end of 2017 (2017). https://www.newscientist.com/article/2138373-google-on-track-for-quantum-computer-breakthrough-by-end-of-2017/
80. Shamir, A.: Factoring numbers in O(log n) arithmetic steps. Inf. Process. Lett. **8**(1), 28–31 (1979)
81. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings of the 35th Annual Symposium of on Foundations of Computer Science, Santa Fe, NM, 20–22 Nov 1994. IEEE Computer Society Press (1994). arXiv:quant-ph/9508027
82. Shor, P.W.: Why haven’t more quantum algorithms been found? J. ACM **50**(1), 87–90 (2003)
83. Simon, D.: On the power of quantum computation. SIAM J. Comput. **26**(5), 1474–1483 (1997)
84. Sipser, M.: Introduction to the Theory of Computation, 1st edn. International Thomson Publishing (1996); 3rd edn. (2013)
85. Svozil, K.: Quantum hocus-pocus. Ethics Sci. Environ. Polit. (ESEP) **16**(1), 25–30 (2016)
86. Takeda, S., Furusawa, A.: Universal quantum computing with measurement-induced continuous-variable gate sequence in a loop-based architecture. Phys. Rev. Lett. **119**, 120504 (2017)
87. Tamma, V.: Analogue algorithm for parallel factorization of an exponential number of large integers: Ii–optical implementation. Quantum Inf. Process. **15**(12), 5243–5257 (2016)
88. Tang, E.: A quantum-inspired classical algorithm for recommendation systems. In: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, pp. 217–228. ACM, New York (2019)
89. UK programme on quantum technologies (2017). http://uknqt.epsrc.ac.uk
90. Wheatley, M.: D-Wave debuts new 5,000-qubit quantum computer. https://siliconangle.com/2019/09/24/d-wave-debuts-new-5000-qubit-quantum-computer/. Accessed 24 Sept 2019
91. Wiesner, K.: The careless use of language in quantum information (2017). https://arxiv.org/abs/1705.06768
92. Yao, A.C.-C.: Classical physics and the Church-Turing thesis. J. ACM (JACM) **50**(1), 100–105 (2003)
93. Zhang, J., Pagano, G., Hess, P.W., Kyprianidis, A., Becker, P., Kaplan, H., Gorshkov, A.V., Gong, Z.X., Monroe, C.: Observation of a many-body dynamical phase transition with a 53-qubit quantum simulator. Nature **551**, 601–604, 11 (2017)

Nonlinear Identities for Bernoulli and Euler Polynomials

Karl Dilcher

Dedicated to the memory of my friend and mentor
Jonathan M. Borwein

1 Introduction

Various types of multiple zeta functions and Euler sums played an important role in Jonathan Borwein's work in experimental mathematics. A particularly interesting class of such series is the Mordell-Tornheim-Witten zeta function

$$\mathcal{W}(r, s, t) := \sum_{m,n\geq 1} \frac{1}{m^r n^s (m+n)^t}, \tag{1}$$

which converges for all complex r, s, t with $\mathrm{Re}(r+t) > 1$, $\mathrm{Re}(s+t) > 1$, and $\mathrm{Re}(r+s+t) > 2$, and can be meromorphically continued to all of $\mathbb{C}$. While Jonathan Borwein and his co-authors studied the series (1) (see, e.g., [2, 6, 7]), he also considered multi-dimensional analogues, especially

$$\mathcal{W}(r_1, \ldots, r_n, t) := \sum_{m_1,\ldots,m_n\geq 1} \frac{1}{m_1^{r_1} \ldots m_n^{r_n} (m_1 + \cdots + m_n)^t}; \tag{2}$$

see [2–5].

Research supported in part by the Natural Sciences and Engineering Research Council of Canada, Grant # 145628481.

K. Dilcher (✉)
Department of Mathematics and Statistics, Dalhousie University, Halifax,
NS B3H 4R2, Canada
e-mail: dilcher@mathstat.dal.ca

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_23

An interesting method repeatedly used in the papers cited above, both for theoretical results and high-precision computations, is due to Crandall and is based on a free parameter; see, e.g., [6, 7] for some details. As a particular application of this method, the results on (1) obtained in [7] were first generalized by H. Tomkins [18] to (2) in the case $n = 3$, and then very recently to arbitrary n in [9].

For the main results in this last paper, the following identity is required: For all integers $n \geq 1$ we have

$$\sum_{m=1}^{n} \binom{n+1}{m} \sum_{\substack{j_1,\dots,j_m \geq 1 \\ j_1+\cdots+j_m=n}} \prod_{i=1}^{m} \frac{B_{j_i}(1)}{j_i!} = 1. \tag{3}$$

Here $B_k(x)$ is the kth *Bernoulli polynomial*, which can be defined by the generating function

$$\frac{te^{xt}}{e^t - 1} = \sum_{k=0}^{\infty} B_k(x) \frac{t^k}{k!}, \qquad |t| < 2\pi. \tag{4}$$

Equivalently it can be defined by

$$B_k(x) = \sum_{j=0}^{k} \binom{k}{j} B_{k-j} x^j, \tag{5}$$

where B_k is the kth *Bernoulli number*, defined by the generating function

$$\frac{t}{e^t - 1} = \sum_{k=0}^{\infty} B_k \frac{t^k}{k!}, \qquad |t| < 2\pi. \tag{6}$$

For the first few Bernoulli numbers and polynomials, see Table 1.

Table 1 B_n, $B_n(x)$ and $E_n^{(n+1)}(x)$ for $0 \leq n \leq 6$

n	B_n	$B_n(x)$	$E_n^{(n+1)}(x)$
0	1	1	1
1	$-1/2$	$x - \frac{1}{2}$	$x - 1$
2	$1/6$	$x^2 - x + \frac{1}{6}$	$x^2 - 3x + \frac{3}{2}$
3	0	$x^3 - \frac{3}{2}x^2 + \frac{1}{2}x$	$x^3 - 6x^2 + 9x - 2$
4	$-1/30$	$x^4 - 2x^3 + x^2 - \frac{1}{30}$	$x^4 - 10x^3 + 30x^2 - 25x - \frac{5}{2}$
5	0	$x^5 - \frac{5}{2}x^4 + \frac{5}{3}x^3 - \frac{1}{6}x$	$x^5 - 15x^4 + 75x^3 - 135x^2 + \frac{75}{2}x + \frac{99}{2}$
6	$1/42$	$x^6 - 3x^5 + \frac{5}{2}x^4 - \frac{1}{2}x^2 + \frac{1}{42}$	$x^6 - 21x^5 + \frac{315}{2}x^4 - 490x^3 + \frac{945}{2}x^2 + 294x - 357$

It is the main purpose of this paper to prove a polynomial analogue of (3), namely the following result.

Theorem 1 *For any integer $n \geq 1$ we have*

$$\sum_{m=1}^{n}\binom{n+1}{m}\sum_{\substack{j_1,\dots,j_m\geq 1\\ j_1+\cdots+j_m=n}}\prod_{i=1}^{m}\frac{B_{j_i}(x)}{j_i!}=\frac{1}{n!}\prod_{j=1}^{n}\big((n+1)x-j\big). \tag{7}$$

Setting $x=1$, we immediately obtain (3). Similarly, with $x=0$ and using the fact that $B_k(0)=B_k$, we have the following identity for Bernoulli numbers.

Corollary 1 *For any integer $n \geq 1$ we have*

$$\sum_{m=1}^{n}\binom{n+1}{m}\sum_{\substack{j_1,\dots,j_m\geq 1\\ j_1+\cdots+j_m=n}}\prod_{i=1}^{m}\frac{B_{j_i}}{j_i!}=(-1)^n. \tag{8}$$

We illustrate Theorem 1 with the first few cases.

Example. For $n=1,2,3$ we have, respectively,

$$\begin{aligned}2\,B_1(x)&=2x-1,\\ \tfrac{3}{2}\,B_2(x)+3\,B_1(x)^2&=\tfrac{1}{2}(3x-1)(3x-2),\\ \tfrac{2}{3}\,B_3(x)+6\,B_1(x)B_2(x)+4\,B_1(x)^3&=\tfrac{1}{6}(4x-1)(4x-2)(4x-3).\end{aligned}$$

In connection with extending two interesting identities of Matiyasevich [13] and Miki [14], expressions similar in nature to the left-hand side of (7) have been studied before (see [1, 10]), but the right-hand side has never been as easy as that of (7). We, therefore, believe that this identity is new.

We conclude this introduction by rewriting (7) in terms of the multinomial coefficient defined by

$$\binom{n}{j_1,\dots,j_m}=\frac{n!}{j_1!\cdots j_m!}.$$

Upon multiplying both sides of (7) by $n!$, we then get

$$\sum_{m=1}^{n}\binom{n+1}{m}\sum_{\substack{j_1,\dots,j_m\geq 1\\ j_1+\cdots+j_m=n}}\binom{n}{j_1,\dots,j_m}B_{j_1}(x)\cdots B_{j_m}(x)=\prod_{j=1}^{n}\big((n+1)x-j\big). \tag{9}$$

It is this identity which we will prove below. We begin with some auxiliary results in Section 2 and complete the proof in Section 3. We conclude this paper with some further remarks in Section 4, including an analogue of Theorem 1 for Euler polynomials.

2 Some Auxiliary Results

The multiple sum on the left of (9), namely

$$T_m(n;x) := \sum_{\substack{j_1,\dots,j_m\ge 1\\ j_1+\dots+j_m=n}} \binom{n}{j_1,\dots,j_m} B_{j_1}(x)\cdots B_{j_m}(x), \tag{10}$$

is very similar to the higher-order convolution

$$S_m(n;x) := \sum_{\substack{j_1,\dots,j_m\ge 0\\ j_1+\dots+j_m=n}} \binom{n}{j_1,\dots,j_m} B_{j_1}(x)\cdots B_{j_m}(x). \tag{11}$$

A slightly more general form of this last expression was evaluated by the present author [8], and then by several other authors, including Huang and Huang [12] who used a different method, and Petojević [16] who evaluated the sum in terms of Stirling numbers of the first kind. Both papers, and numerous others, contain evaluations of other related expressions of the type of (11).

In what follows, we will use the higher-order Bernoulli polynomials, defined as follows: Given an integer m (not necessarily positive), the kth *Bernoulli polynomial of order* m, denoted $B_k^{(m)}(x)$, is defined by the generating function

$$\left(\frac{t}{e^t-1}\right)^m e^{xt} = \sum_{k=0}^{\infty} B_k^{(m)}(x)\frac{t^k}{k!}, \qquad |t| < 2\pi. \tag{12}$$

By comparing this with (4), we see that $B_k^{(1)}(x) = B_k(x)$. Raising both sides of (4) to the power m and using the identities (11) and (12), we get

$$S_m(n;x) = B_n^{(m)}(mx). \tag{13}$$

This fact was earlier used in [8, 12].

Next we need to connect the sums $S_m(n;x)$ and $T_m(n;x)$ with each other.

Lemma 1 *For any integers $m, n \ge 1$ we have*

$$S_m(n;x) = \sum_{j=1}^{m} \binom{m}{j} T_j(n;x), \tag{14}$$

$$T_m(n;x) = \sum_{j=1}^{m} (-1)^{m-j} \binom{m}{j} S_j(n;x). \tag{15}$$

Proof To obtain (14), we subdivide the sum $S_m(n;x)$ according to the number of indices j_i that are 0. If none of them is 0, we simply have $T_m(n;x)$. If exactly one of them is 0, then we have m copies of $T_{m-1}(n;x)$. If exactly two of them are 0, we get $\binom{m}{2}$ copies of $T_{m-2}(n;x)$, and so on, until we reach the case where exactly $m-1$ of the indices are 0; this happens $\binom{m}{m-1}$ times, giving m copies of $T_1(n;x)$. Adding everything, we get (14).

The identity (15) can be obtained in different ways: Either directly by an inclusion/exclusion argument, or by solving a linear system that is inherent in (14), or, most easily by appealing to a general result on inverting finite sums; see, e.g., [17, p. 43]. □

Towards the eventual proof of (9), we now evaluate the following sum.

Lemma 2 *For any integer $n \geq 1$ we have*

$$\sum_{m=1}^{n}\binom{n+1}{m}T_m(n;x) = \sum_{k=1}^{n}(-1)^{n-k}\binom{n+1}{k}B_n^{(k)}(kx). \tag{16}$$

Proof We use (15) and change the order of summation:

$$\begin{aligned}\sum_{m=1}^{n}\binom{n+1}{m}&\sum_{k=1}^{m}(-1)^{m-k}\binom{m}{k}S_k(n;x)\\ &= \sum_{k=1}^{m}(-1)^{k}S_k(n;x)\sum_{m=k}^{n}(-1)^{m}\binom{n+1}{m}\binom{m}{k}.\end{aligned}$$

The inner sum of this last expression is an alternating analogue of the Vandermonde convolution, and can be evaluated as $(-1)^n\binom{n+1}{k}$; see, e.g., [11, (3.119)]. With this and (13), we immediately get (16). □

3 The Proof of Theorem 1

By Lemma 2, in order to finish the proof of (9), and thus of Theorem 1, we need to evaluate the right-hand side of (16). Using the generating function (12), we rewrite

$$\begin{aligned}\sum_{k=1}^{n}(-1)^{n-k}\binom{n+1}{k}B_n^{(k)}(kx) &= \sum_{k=1}^{n}(-1)^{n-k}\binom{n+1}{k}\frac{d^n}{dt^n}\left(\frac{te^{tx}}{e^t-1}\right)^k\Bigg|_{t=0}\\ &= \frac{d^n}{dt^n}\sum_{k=1}^{n}(-1)^{n-k}\binom{n+1}{k}\left(\frac{te^{tx}}{e^t-1}\right)^k\Bigg|_{t=0}.\end{aligned} \tag{17}$$

To simplify notation, we set $A(t) := te^{tx}/(e^t-1)$. Using a binomial expansion, we then have

$$\begin{aligned}\sum_{k=1}^{n}&(-1)^{n-k}\binom{n+1}{k}A(t)^k\\ &=-\sum_{k=0}^{n+1}(-1)^{n+1-k}\binom{n+1}{k}A(t)^k+(-1)^{n+1}+A(t)^{n+1}\\ &=-(A(t)-1)^{n+1}+(-1)^{n+1}+A(t)^{n+1}.\end{aligned}\tag{18}$$

We note that the constant coefficient in the Maclaurin expansion of $A(t)$ as a function of t is 1. Therefore, we can write

$$\big(A(t)-1\big)^{n+1}=\big(tB(t)\big)^{n+1}=t^{n+1}B(t)^{n+1},$$

where $B(t)$ is analytic at $t=0$. Hence

$$\frac{d^n}{dt^n}\,\big(t^{n+1}B(t)^{n+1}\big)\big|_{t=0}=0,$$

while

$$\frac{d^n}{dt^n}\,A(t)^{n+1}\big|_{t=0}=\frac{d^n}{dt^n}\left(\left(\frac{t}{e^t-1}\right)^{n+1}e^{(n+1)xt}\right)\bigg|_{t=0}=B_n^{(n+1)}((n+1)x),$$

where we have again used (12). This, together with (18), (17) and (16) gives the intermediate result

$$\sum_{m=1}^{n}\binom{n+1}{m}T_m(n;x)=B_n^{(n+1)}((n+1)x).\tag{19}$$

Finally we use a well-known explicit formula for $B_n^{(n+1)}(x)$ (see, e.g., [15, p. 130]), which immediately gives

$$B_n^{(n+1)}((n+1)x)=\prod_{j=1}^{n}\big((n+1)x-j\big).\tag{20}$$

With (19), this completes the proof of (9) and of Theorem 1.

4 Further Remarks

1. If we set $m=n+1$ in (14), we get

$$S_{n+1}(n;x)=\sum_{j=1}^{n}\binom{n+1}{j}T_j(n;x)+T_{n+1}(n;x).$$

From (10) it is clear that $T_{n+1}(n; x) = 0$ since it is an empty sum. Therefore, (19) and (20) lead to the following consequence concerning the convolution sum defined in (11).

Corollary 2 *For any $n \geq 1$ we have*

$$S_{n+1}(n; x) = \prod_{j=1}^{n} \big((n+1)x - j\big).$$

2. Whenever a result on Bernoulli polynomials is obtained, it is a natural question to ask whether there are analogues for Euler polynomials. The *Euler polynomial of order m and degree k*, $E_k^{(m)}(x)$, is defined by the generating function

$$\left(\frac{2}{e^t + 1}\right)^m e^{xt} = \sum_{k=0}^{\infty} E_k^{(m)}(x) \frac{t^k}{k!}, \qquad |t| < \pi,$$

and the (ordinary) *Euler polynomial* of degree k by $E_k(x) := E_k^{(1)}(x)$. Various properties, including recurrence relations, of these polynomials can be found, e.g., in [15, p. 143ff].

If we replace each 'B' by 'E' in (7) and (9), then all details of the proof carry through, up to the equivalent of (19). We, therefore, get the following result.

Theorem 2 *For any integer $n \geq 1$ we have*

$$\sum_{m=1}^{n} \binom{n+1}{m} \sum_{\substack{j_1,\dots,j_m \geq 1 \\ j_1+\dots+j_m=n}} \prod_{i=1}^{m} \frac{E_{j_i}(x)}{j_i!} = \frac{1}{n!} E_n^{(n+1)}((n+1)x). \tag{21}$$

In contrast to Theorem 1, however, the right-hand side of (21) does not have an easy evaluation. The first few polynomials $E_n^{(n+1)}(x)$ are listed in Table 1.

We finish by deriving an analogue of Corollary 1 for Euler numbers. The kth *Euler number of order n* is defined by

$$E_k^{(n)} := 2^k E_k^{(n)}(\tfrac{n}{2});$$

see, e.g., [15, p. 143]. In particular, this implies

$$E_k(\tfrac{1}{2}) = 2^{-k} E_k, \qquad E_n^{(n+1)}(\tfrac{n+1}{2}) = 2^{-n} E_n^{(n+1)},$$

where E_k is the kth (ordinary) *Euler number*. Setting $x = \frac{1}{2}$ in (21) and multiplying both sides by $2^n n!$, we get the following identity, written in a form analogous to (9).

Corollary 3 *For any integer $n \geq 1$ we have*

$$\sum_{m=1}^{n} \binom{n+1}{m} \sum_{\substack{j_1,\ldots,j_m \geq 1 \\ j_1+\cdots+j_m=n}} \binom{n}{j_1,\ldots,j_m} E_{j_1} \cdots E_{j_m} = E_n^{(n+1)}.$$

References

1. Agoh, T., Dilcher, K.: Higher-order convolutions for Bernoulli and Euler polynomials. J. Math. Anal. Appl. **419**(2), 1235–1247 (2014)
2. Bailey, D.H., Borwein, D., Borwein, J.M.: On Eulerian log-gamma integrals and Tornheim-Witten zeta functions. Ramanujan J. **36**(1–2), 43–68 (2015)
3. Bailey, D.H., Borwein, J.M., Crandall, R.: Computation and theory of extended Mordell-Tornheim-Witten sums. Math. Comput. **83**(288), 1795–1821 (2014)
4. Bailey, D.H., Borwein, J.M.: Computation and theory of Mordell-Tornheim-Witten sums II. J. Approx. Theory **197**, 115–140 (2015)
5. Bailey, D.H., Borwein, J.M.: Computation and experimental evaluation of Mordell-Tornheim-Witten sum derivatives. Exp. Math. **27**(3), 370–376 (2018)
6. Borwein, J.M.: Hilbert's inequality and Witten's zeta-function. Am. Math. Mon. **115**(2), 125–137 (2008)
7. Borwein, J.M., Dilcher, K.: Derivatives and fast evaluation of the Tornheim zeta function. Ramanujan J. **45**(2), 413–432 (2018)
8. Dilcher, K.: Sums of products of Bernoulli numbers. J. Number Theory **60**(1), 23–41 (1996)
9. Dilcher, K., Tomkins, H.: Derivatives and special values of higher-order Tornheim zeta functions. In preparation
10. Dilcher, K., Vignat, C.: General convolution identities for Bernoulli and Euler polynomials. J. Math. Anal. Appl. **435**(2), 1478–1498 (2016)
11. Gould, H.W.: Combinatorial Identities, revised edition, Gould Publications, West Virginia, Morgantown (1972)
12. Huang, I.-C., Huang, S.-Y.: Bernoulli numbers and polynomials via residues. J. Number Theory **76**(2) 178–193 (1999)
13. Matiyasevich, Y.: Identities with Bernoulli numbers. http://logic.pdmi.ras.ru/~yumat/personaljournal/identitybernoulli/bernulli.htm
14. Miki, H.: A relation between Bernoulli numbers. J. Number Theory **10**(3), 297–302 (1978)
15. Milne-Thomson, L.M.: The Calculus of Finite Differences. Macmillan, London (1951)
16. Petojević, A.: New sums of products of Bernoulli numbers. Integr. Transform. Spec. Funct. **19**(1–2), 105–114 (2008)
17. Riordan, J.: Combinatorial Identities. Wiley, New York (1968)
18. Tomkins, H.: An exploration of multiple zeta functions. Honours Thesis, Dalhousie University (2016)

Metrical Theory for Small Linear Forms and Applications to Interference Alignment

Mumtaz Hussain, Seyyed Hassan Mahboubi and Abolfazl Seyed Motahari

1 Metric Diophantine Approximation

At its most fundamental level, the theory of Diophantine approximation is concerned with the question of how well a real number can be approximated by rationals. Qualitatively the answer is somewhat trivial as the set of rationals $\mathbb{Q}$ is dense in the reals. In other words, for any real number r we can construct a sequence of rational numbers r_n such that $r_n \to r$ as $n \to \infty$. Quantifying the density of rationals in the reals, however, is non-trivial. A well-known theorem of Dirichlet is fundamental in the theory of Diophantine approximation and gives a rate of approximation that works for all real numbers.

Theorem 1 **(Dirichlet 1842)** *Given $x \in \mathbb{R}$ and $t > 1$, there exist integers p, q such that*

$$|x - p/q| \leq 1/qt \quad \text{and} \quad 1 \leq q < t. \tag{1}$$

An important consequence of this theorem is the following statement.

Corollary 1 *For any $x \in \mathbb{R}$, there exist infinitely many (i.m.) integers p and $q > 0$ such that*

$$|x - p/q| < 1/q^2. \tag{2}$$

Replacing the right-hand side of (2) with a faster decreasing function $\psi(q) \to 0$ as $q \to \infty$ raises the question of 'size' of the corresponding set

M. Hussain (✉) · S. H. Mahboubi
Department of Mathematics and Statistics, La Trobe University, PO Box 199,
3552 Bendigo, Australia
e-mail: m.hussain@latrobe.edu.au

A. S. Motahari
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
e-mail: motahari@sharif.edu

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_24

$$W(\psi) := \{x \in \mathbb{R} : |x - p/q| < \psi(q) \text{ for infinitely many } (p, q) \in \mathbb{Z} \times \mathbb{N}\}$$

of ψ-approximable numbers. Here and throughout ψ will be referred to as an *approximating function* and will always be monotonic unless stated otherwise. Khintchine's theorem (1924) asserts that the Lebesgue measure of this set is either zero or full if the sum $\sum q\psi(q)$ converges or diverges, respectively. Here, by 'full' we mean that the complement of the set has zero measure. Khintchine's theorem is a very delicate statement which, for example, implies that $W(\psi)$ has full measure for $\psi(q) = 1/q^2$ but $W(\psi)$ has zero Lebesgue measure for $\psi(q) = 1/q^{2+\epsilon}$, for any $\epsilon > 0$.

This paper falls within the metric theory of Diophantine approximation and is about estimating the Lebesgue measure of the set of real (or complex) points approximable infinitely often by rational numbers (or ratios of Gaussian integers) with a given error of approximation.

In higher dimensional settings, two classical categories of metric Diophantine approximation are *simultaneous* (approximation by rational points) and *dual* (approximation by rational hyperplanes). These are unified in the theory of systems of linear forms,

$$|\mathbf{q} \cdot \mathbf{x}_i - p_i| < \psi(|\mathbf{q}|), \quad (1 \le i \le n). \tag{3}$$

Here $X = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{I}^{mn} := [-1/2, 1/2]^{mn}$ is a real matrix, $\mathbf{q} \in \mathbb{Z}^m$ and $\mathbf{p} \in \mathbb{Z}^n$ are integer vectors and $|\mathbf{q}| = \max\{|q_1|, \dots, |q_m|\}$ is the supremum norm.

Historically, interest has been concentrated upon the set $W(m, n; \psi)$ of matrices X for which the system of inequalities (3) has infinitely many solutions in integer vectors $\mathbf{q}$ and $\mathbf{p}$. The main result in the linear form settings is the Khintchine–Groshev theorem, which gives an elegant answer to the question of the size of the set $W(m, n; \psi)$. The following statement is the modernised version of the Khintchine–Groshev theorem for Lebesgue measure [1]. Throughout, for any set $A \subset \mathbb{R}^l$, let $|A|_l$ denote the l-dimensional Lebesgue measure of the set A.

Theorem 2 (Khintchine–Groshev) *Let ψ be an approximating function. Then*

$$|W(m, n; \psi)|_{mn} = \begin{cases} 0, \text{ if} & \sum_{r=1}^{\infty} r^{m-1}\psi^n(r) < \infty; \\ 1, \text{ if} & \sum_{r=1}^{\infty} r^{m-1}\psi^n(r) = \infty. \end{cases}$$

The proof of the convergence case of the Khintchine–Groshev theorem is easily established by a straightforward application of the Borel–Cantelli lemma [2, Lemma 2.1] and is free from any assumption on ψ, i.e. valid for non-monotonic approximating functions. The divergence part constitutes the main substance of the theorem and requires the monotonicity assumption on the function ψ for $m = n = 1$. For all other values of m and n it can be removed, see [1, 20] and references therein.

2 Statements of Results

In this decade, a new branch of Diophantine approximation has emerged; namely, the *absolute value theory* obtained by fixing $p_i = 0$ in (3). Whilst remaining relatively undeveloped, this branch has significant potential for impact due to its connections with problems in signal processing and electronic communications [16, 30].

2.1 Mixed Type Linear Forms

We consider a variant of the absolute value theory by fixing the vector $(p_1, \ldots, p_n)$ in (3) as $(p, \ldots, p)$. To this end, let ψ be an approximating function and let $\widetilde{W}(m, n; \psi)$ be the set of $\mathbf{X} \in \mathbb{I}^{mn}$ such that the system of equations

$$|q_1 x_{1,i} + q_2 x_{2,i} + \ldots + q_m x_{m,i} - p| < \psi(|\mathbf{q}|) \quad 1 \le i \le n \tag{4}$$

is satisfied for infinitely many $(p, \cdots, p, q_1, \cdots, q_m) \in \mathbb{Z}^n \times \mathbb{Z}^m \setminus \{\mathbf{0}\}$.

Sets of similar nature have been studied by several authors, see for example [7, 10, 14, 18, 19]. A variant of this setup has been seen to be connected with the solutions to the inhomogeneous partial differential equations [14].

In the last few years, there has been significant progress in developing new strategies in Multiple Input Multiple Output (MIMO) communication schemes. Number theoretic results such as Khintchine–Groshev theorems have been used in various situations in calculating the Degrees of Freedom (DoF) in alignment schemes within MIMO X-channels, see [26] and [21, Sect. 4.7.1] for more explicit discussions. The main aim of this paper is to establish Khintchine–Groshev type theorems for $\widetilde{W}(m, n; \psi)$ for real and complex numbers. In the final section of the paper, these theorems are used in a brief sketch of an application in interference alignment. In a forthcoming article, we elaborate this application in detail for the signal processing community.

Our first result is the Lebesgue measure criterion for $\widetilde{W}(m, n; \psi)$.

Theorem 3 *Let $m \ge n$ and ψ be an approximating function; then*

$$|\widetilde{W}(m, n; \psi)|_{mn} = \begin{cases} 0, \text{ if } \sum\limits_{r=1}^{\infty} \psi^n(r) r^{m-n} < \infty; \\ 1, \text{ if } \sum\limits_{r=1}^{\infty} \psi^n(r) r^{m-n} = \infty. \end{cases}$$

By setting $p = 0$ in the above setup, a similar application of the convergence half already exists in achieving MIMO capacity within a constant gap [30]. In fact, the result in [30] can be extended to cover the complex number system by using the Khintchine–Groshev type result produced in [16, Theorem 1].

It is worth demonstrating that for $\psi(r) = r^{-\frac{m+1}{n}+1-\epsilon}$,

$$\sum_{r=1}^{\infty} \psi(r)^n r^{m-n} = \sum_{r=1}^{\infty} r^{-1-n\epsilon} \begin{cases} < \infty, & \text{if } \epsilon > 0; \\ = \infty, & \text{if } \epsilon \leq 0. \end{cases}$$

2.2 Diophantine Approximation Over Complex Numbers

Most of the complex Diophantine approximation theory is analogous to what we have discussed in the previous subsection. Surprisingly, analogues of Khintchine–Groshev theorems for systems of linear forms over complex numbers are not yet proved. We prove them here along with analogous results for mixed type linear forms. To keep the exposition compact we state only the important changes.

In the nineteenth century, Hermite and Hurwitz studied the approximation of complex numbers by the ratios of Gaussian integers, a natural analogue of approximation of real numbers by rationals,

$$\mathbf{Z}[i] = \{p_1 + ip_2 \in \mathbb{C} : p_1, p_2 \in \mathbb{Z}\}.$$

However, complex Diophantine approximation appears to be more difficult than the real case. For example, continued fractions, so simple and effective for real numbers, are not so straightforward for complex numbers. In other words, the best possible analogue of Dirichlet's theorem cannot be derived by a straightforward extension of the continued fraction expansion approach that works in the real case.

We will discuss the problem for the linear form setup and will list the recent developments so far for the particular cases. Let $\Psi : \mathbb{N} \to \mathbb{R}^+$ be a monotonically decreasing function such that $\Psi(r) \to 0$ as $r \to \infty$. An $m \times n$ matrix $\mathbf{Z} = (z_{i,j}) \in \mathbb{C}^{mn}$ is said to be Ψ-approximable if the system of inequalities

$$|q_1 z_{1,j} + q_2 z_{2,j} + \cdots + q_m z_{m,j} - p_j| < \Psi(|\mathbf{q}|) \quad (1 \leq j \leq n) \tag{5}$$

is satisfied for infinitely many vectors $\mathbf{p} \times \mathbf{q} \in \mathbb{Z}^n[i] \times \mathbb{Z}^m[i] \backslash \{\mathbf{0}\}$. Throughout, the system (5) will be written more concisely as

$$|\mathbf{qZ} - \mathbf{p}| < \Psi(|\mathbf{q}|).$$

Here

$$|\mathbf{q}| = \max\{\lfloor |q_1|_2 \rfloor, \cdots, \lfloor |q_m|_2 \rfloor\},$$

where

$$|q_k|_2 = \sqrt{|q_{k_1}|^2 + |q_{k_2}|^2}, \text{for } q_k = q_{k_1} + iq_{k_2} \in \mathbb{Z}[i],$$

and $\lfloor x \rfloor$ denotes the integer part of the real number x.

As in the real case, the starting point of such approximation properties is Dirichlet's theorem. A short proof using geometry of numbers of the complex version of Dirichlet's theorem is given below. Although the constant obtained by this approach is not best possible, the result is all that is needed to prove the complex analogue of the Khintchine–Groshev theorem.

Theorem 4 *Given any* $\mathbf{Z} \in \mathbb{C}^{mn}$ *and* $N \in \mathbb{N}$*, there exist Gaussian integers* $\mathbf{p} \in \mathbb{Z}^n[i]$ *and* $\mathbf{q} \in \mathbb{Z}^m[i]$ *with* $0 < |\mathbf{q}| \leq N$ *such that*

$$|\mathbf{q}\mathbf{Z} - \mathbf{p}| < \frac{c}{N^{m/n}}, \tag{6}$$

where $c > 0$ *is a constant (independent of* $\mathbf{Z}$ *and* N*). Moreover, there are infinitely many* $(\mathbf{p}, \mathbf{q}) \in \mathbb{Z}^n[i] \times \mathbb{Z}^m[i] \backslash \{\mathbf{0}\}$ *such that*

$$|\mathbf{q}\mathbf{Z} - \mathbf{p}| < \frac{c}{|\mathbf{q}|^{m/n}}.$$

From now onwards we restrict ourselves to the mn-dimensional unit disc $D := (\mathbb{C} \cap \Omega)^{mn}$, where $\Omega = \{a + ib : 0 \leq a, b < 1\}$, instead of considering the full space $\mathbb{C}^{mn}$. The reason behind this restriction is that it is convenient to work in the unit disc, and the approximability properties are invariant under translation by Gaussian integers. Let $W_{\mathbb{C}}(m, n; \Psi)$ denote the set of Ψ-approximable points in D, i.e.

$$W_{\mathbb{C}}(m, n; \Psi) := \left\{\mathbf{Z} \in D : |\mathbf{q}\mathbf{Z} - \mathbf{p}| < \Psi(|\mathbf{q}|) \text{for i.m.} (\mathbf{p}, \mathbf{q}) \in \mathbb{Z}^n[i] \times \mathbb{Z}^m[i] \backslash \{\mathbf{0}\}\right\}.$$

Our next result is a complex analogue of Theorem 2.

Theorem 5 *Let* Ψ *be an approximating function. Then*

$$|W_{\mathbb{C}}\left(m, n; \Psi\right)|_{mn} = \begin{cases} 0, \text{ if } & \sum\limits_{r=1}^{\infty} r^{2m-1} \Psi^{2n}(r) < \infty; \\ \text{Full, if } & \sum_{r=1}^{\infty} r^{2m-1} \Psi^{2n}(r) = \infty. \end{cases}$$

Here, $|W_{\mathbb{C}}\left(m, n; \Psi\right)|_{mn}$ denotes the complex mn-dimensional Lebesgue measure of the set $W_{\mathbb{C}}\left(m, n; \Psi\right)$. For $m = n = 1$, Theorem 5 was proved in 1952 by LeVeque [25], who combined Khintchine's continued fraction approach with ideas from hyperbolic geometry. In 1982, Sullivan [31] used Bianchi groups and some powerful hyperbolic geometry arguments to prove more general Khintchine type theorems for real and for complex numbers. In the latter case, the result includes approximation of complex numbers by ratios p/q of integers p, q from the imaginary quadratic fields $\mathbb{R}(i\sqrt{d})$, where d is a square-free natural number. The case $d = 1$ corresponds to the Picard group and approximation by Gaussian rationals. The result was also derived by Beresnevich et al. as a consequence of the ubiquity framework in [1, Theorem 7].

Next we discuss the analogue of $\widetilde{W}(m, n; \psi)$ for complex numbers. Let Ψ be an approximating function. An $m \times n$ matrix $\mathbf{Z} \in \mathbb{C}^{mn}$ is said to be Ψ-approximable if the system of inequalities

$$|q_1 z_{1,j} + q_2 z_{2,j} + \cdots + q_m z_{m,j} - p| < \Psi(|\mathbf{q}|) \quad (1 \le j \le n) \tag{7}$$

is satisfied for infinitely many vectors $(p, \dots, p, q_1, \dots, q_m) \in \mathbb{Z}^n[i] \times \mathbb{Z}^m[i] \backslash \{\mathbf{0}\}$. That is, the system (7) is obtained by keeping the nearest Gaussian integer vector $(p, \dots, p)$ the same for all the linear forms. Since the results are very similar to $\widetilde{W}(m, n; \psi)$ and can be proved analogously, they are only stated here. The first result is a Dirichlet type theorem which also serves the purpose of finding the minimum distance between $\mathbf{qZ}$ and $\mathbf{p}$.

Theorem 6 *Given any* $\mathbf{Z} \in \mathbb{C}^{mn}$ *and* $N \in \mathbb{N}$, *there exist Gaussian integers* $\mathbf{p} = (p_1 + ip_2, \dots, p_1 + ip_2) \in \mathbb{Z}^n[i]$ *and* $\mathbf{q} = (q_{11} + iq_{12}, \dots, q_{m1} + iq_{m2}) \in \mathbb{Z}^m[i]$ *with* $0 < |\mathbf{q}| \le N$ *such that*

$$|\mathbf{qZ} - \mathbf{p}| < cN^{-\frac{m+1}{n}+1},$$

where $c > 0$ *is a constant (independent of* $\mathbf{Z}$ *and* N*). Moreover, there are infinitely many* $(\mathbf{p}, \mathbf{q}) \in \mathbb{Z}^n[i] \times \mathbb{Z}^m[i] \backslash \{\mathbf{0}\}$ *such that*

$$|\mathbf{qZ} - \mathbf{p}| < c|\mathbf{q}|^{-\frac{m+1}{n}+1}.$$

Let $\widetilde{W}_{\mathbb{C}}(m, n; \Psi)$ denote the set of Ψ-approximable points in D, i.e. the set of points that satisfy the system (7). Then, one has the analogue of the Khintchine–Groshev theorem for this setup.

Theorem 7 *Let* Ψ *be an approximating function and let* $m \ge n$. *Then*

$$|\widetilde{W}_{\mathbb{C}}(m, n; \Psi)|_{mn} = \begin{cases} 0, \text{ if} & \sum\limits_{r=1}^{\infty} \left(r^{m-n}\Psi^n(r)\right)^2 < \infty; \\ \text{Full, if} & \sum_{r=1}^{\infty} \left(r^{m-n}\Psi^n(r)\right)^2 = \infty. \end{cases}$$

3 Some Proofs

3.1 *Proof of Theorem 3*

The proof of Theorem 3 splits into two parts, the convergence case and the divergence case.

3.1.1 The Convergence Case

The convergence half follows from the well-known Borel–Cantelli lemma by constructing a suitable cover for the set $\widetilde{W}(m, n; \psi)$. It does not rely on whether $m \ge n$ or $m < n$, and it is free from the monotonicity assumption on the approximating

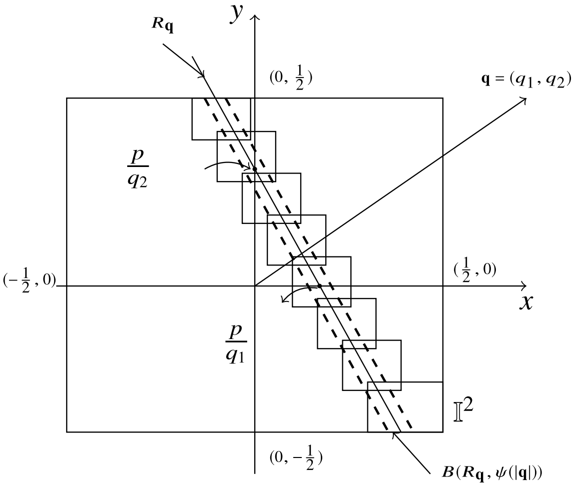


Fig. 1 The resonant set $R_{\mathbf{q}}$ is a line for $m = 2$ and $n = 1$. The resonant set $R_{\mathbf{q}}$ is a line $q_1x + q_2y - p = 0$, intercepting the x and y axes at $\frac{p}{q_1}$ and $\frac{p}{q_2}$, respectively. The set $B\left(R_{\mathbf{q}}, \psi(|\mathbf{q}|)\right)$ is the $\frac{\psi(|\mathbf{q}|)}{|\mathbf{q}|}$ neighbourhood of $R_{\mathbf{q}}$

function. It is worth pointing out that in applications the convergence case is all that matters.

Define a family of *resonant sets* as

$$\mathcal{R} := \{R_q : \mathbf{q} \in \mathbb{Z}^m \backslash \{\mathbf{0}\}\},$$

where

$$R_q = \left\{\mathbf{X} \in \mathbb{I}^{mn} : \mathbf{qX} - \mathbf{p} = 0\right\}.$$

Thus, the resonant sets are $(m-1)n$-dimensional hyperplanes passing through the point $\mathbf{p}$. The set $\widetilde{W}(m, n; \psi)$ can be written as a lim sup set using the resonant sets in the following way Fig. 1.

$$\widetilde{W}(m, n; \psi) = \bigcap_{N=1}^{\infty} \bigcup_{r>N} \bigcup_{R_q : |\mathbf{q}|=r} B\left(R_q, \psi(|\mathbf{q}|)\right),$$

where

$$B\left(R_q, \psi(|\mathbf{q}|)\right) = \left\{\mathbf{X} \in \mathbb{I}^{mn} : \text{dist}\left(X, R_q\right) \leq \frac{\psi(|\mathbf{q}|)}{|\mathbf{q}|}\right\}.$$

Thus, for each $N \in \mathbb{N}$ the family $\left\{ \bigcup_{R_q : |\mathbf{q}|=r} B\left(R_q, \psi(|\mathbf{q}|)\right) : r = N, N+1, \ldots \right\}$ is a cover for the set $\widetilde{W}(m, n; \psi)$. Now, for each resonant set $R_{\mathbf{q}}$, let $\Delta(\mathbf{q})$ be a collection of mn-dimensional closed hypercubes C with disjoint interiors and side length comparable with $\psi(|\mathbf{q}|)/|\mathbf{q}|$ and diameter at most $\psi(|\mathbf{q}|)/|\mathbf{q}|$, such that

$$C \cap \bigcup_{R_{\mathbf{q}}:|\mathbf{q}|=r} B\left(R_{\mathbf{q}}, \psi(|\mathbf{q}|)\right) \neq \emptyset$$

and

$$B\left(R_{\mathbf{q}}, \psi(|\mathbf{q}|)\right) \subset \bigcup_{C \in \Delta(\mathbf{q})} C.$$

Then

$$\#\Delta(\mathbf{q}) \ll (\psi(|\mathbf{q}|)/|\mathbf{q}|)^{-(m-1)n},$$

where # denotes cardinality. Note that

$$\widetilde{W}(m,n;\psi) \subset \bigcup_{r>N} \bigcup_{R_q:|\mathbf{q}|=r} B\left(R_{\mathbf{q}}, \Psi(|\mathbf{q}|)\right) \subset \bigcup_{r>N} \bigcup_{\Delta(\mathbf{q}):|\mathbf{q}|=r} \bigcup_{C \in \Delta(\mathbf{q})} C.$$

Hence,

$$\begin{aligned}\left|\widetilde{W}(m,n;\psi)\right|_{mn} &\leq \sum_{r>N} \sum_{\Delta(\mathbf{q}):|\mathbf{q}|=r} \sum_{C \in \Delta(\mathbf{q})} |C|_{mn} \\ &\ll \sum_{r>N} r^m \left(\frac{\psi(r)}{r}\right)^{mn} \left(\frac{\psi(r)}{r}\right)^{-(m-1)n} \\ &= \sum_{r>N} r^{m-n} \psi(r)^n.\end{aligned}$$

The sum $\sum_{r\geq 1} r^{m-n}\psi(r)^n$ is convergent, which gives zero Lebesgue measure by the Borel–Cantelli lemma.

3.1.2 The Divergence Case

To prove the divergence part of the theorem the idea of a locally ubiquitous system is used. We present a simplified version of a more abstract framework developed in [1, 3]. For the current setup, the required measure and intersection conditions in [1] are trivially satisfied. Let $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ be a function such that $\rho(r) \to 0$ as $r \to \infty$ and let

$$\Delta(\rho, t) := \bigcup_{|\mathbf{q}| \leq k^t} B(R_q, \rho(k^t))$$

where $k > 1$ is a fixed real number.

Definition 1 Let $B(X, r)$ be an arbitrary ball with centre $X \in \mathbb{I}^{mn}$ and radius $r \leq r_0(m,n)$. Suppose there exists a function ρ and an absolute constant $\kappa > 0$ such that

$$|B \cap \Delta(\rho, t)|_{mn} \geq \kappa |B|_{mn} \text{ for } t \geq t_0(B).$$

Then $\mathcal{R}$ is said to be a *locally ubiquitous* system relative to ρ.

Loosely speaking, the definition of local ubiquity says that the set $\Delta(\rho, t)$ locally approximates the underlying space $\mathbb{I}^{mn}$ in terms of the Lebesgue measure. The function ρ will be referred to as the *ubiquity function*. The actual value of κ in the above definition is irrelevant, only its existence is important. In practice, local ubiquity is usually established using standard results such as Dirichlet's Theorem or Minkowski's Convex Body Theorem, regarding the distribution of the resonant sets, from which the function ρ arises naturally. Clearly if $|\Delta(\rho, t)|_{mn} \to 1$ as $t \to \infty$ then $\mathcal{R}$ is locally ubiquitous.

The following theorem is a simplified version of Theorem 1 from [3].

Theorem 8 *Assume that there exists* $0 < \lambda < 1$ *such that the function* ρ *satisfies* $\rho(2^{t+1}) < \lambda\rho(2^t)$ *for all* $t \in \mathbb{N}$. *Suppose that* $\mathcal{R}$ *is locally ubiquitous relative to* ρ *and* ψ *is an approximating function. Then*

$$\widetilde{W}(m, n; \psi)\,|_{mn} = 1 \quad \text{if} \quad \sum_{t=1}^{\infty} \frac{\Psi(2^t)^n}{\rho(2^t)^n} = \infty.$$

To establish ubiquity two technical lemmas (Lemma 1 and Lemma 2) are needed. The work is similar to [7]; therefore, we only prove one of them and refer the interested reader to the aforementioned article [7]. Most of the metric results (Khintchine–Groshev, Jarnik, Jarnik–Besicovitch and Schmidt theorems) stem from the Dirichlet type result which is stated and proved below for the current settings. In the lemma below, we set $N = 2^t : t \in \mathbb{N}$.

Lemma 1 *For sufficiently large* N_0 *and* $N > N_0$, *for each* $\mathbf{X} \in \mathbb{I}^{mn}$ *there exists a non-zero integer vector* $\mathbf{q}$ *in* $\mathbb{Z}^m$ *and* $\mathbf{p} \in \mathbb{Z}^n$ *with* $|\mathbf{q}|, |\mathbf{p}| \leq N$ *such that*

$$|\mathbf{qX} - \mathbf{p}| < 2(m+2)N^{-\frac{m+1}{n}+1}.$$

Proof of Lemma 1. For $|\mathbf{p}| < N$ and those $\mathbf{q}$ with non-negative components, there are $(N+1)^m\, N$ possible vectors of the form $\mathbf{qX} - p$ for which

$$-\frac{m+2}{2}N \leq \mathbf{qX} - \mathbf{p} \leq \frac{m+2}{2}N.$$

Divide the cube with centre $\mathbf{0}$ and side length $(m+2)N$ in $\mathbb{R}^n$ into N^{m+1} smaller cubes of volume $(m+2)^n N^{n-m-1}$ and side length $(m+2)N^{1-\frac{m+1}{n}}$. Since $N^m < (N+1)^m$, there are at least two vectors $\mathbf{q}_1\mathbf{X} - \mathbf{p}_1$, $\mathbf{q}_2\mathbf{X} - \mathbf{p}_2$, say, in one small cube. Therefore

$$\left|(\mathbf{q}_1 - \mathbf{q}_2)\mathbf{X} - (\mathbf{p}_1 - \mathbf{p}_2)\right| < 2(m+2)N^{-\frac{m+1}{n}+1}.$$

Evidently $\mathbf{q}_1 - \mathbf{q}_2 \in \mathbb{Z}^m$ and $|\mathbf{q}_1 - \mathbf{q}_2| \leq N$. Also, $\mathbf{p}_1 - \mathbf{p}_2 \in \mathbb{Z}$ and $|\mathbf{p}_1 - \mathbf{p}_2| \leq N$ by choices of $\mathbf{p}_1$ and $\mathbf{p}_2$.

Lemma 2 *Let* $\omega(t)$ *be a positive real increasing function such that* $\omega(t) \to \infty$ *as* $t \to \infty$. *Then the family* $\mathcal{R}$ *is locally ubiquitous with respect to the function*

$$\rho(t) = 2(m+2)N^{-\frac{m+1}{n}}\omega(t).$$

Using Theorem 8, it follows that

$$\widetilde{W}(m,n;\psi)\,|_{mn} = 1 \text{ if } \sum_{t=1}^{\infty}\frac{\Psi(2^t)^n}{\rho(2^t)^n} \asymp \sum_{t=1}^{\infty} t^{m-n}\psi(t)^n = \infty.$$

Remark 1 In view of Lemma 1, it is natural to consider the following badly approximable set. Let Bad(m,n) denote the set of $\mathbf{X} \in \mathbb{I}^{mn}$ for which there exists a constant $C(\mathbf{X}) > 0$ such that

$$|\mathbf{qX} - \mathbf{p}| > C(\mathbf{X})|\mathbf{q}|^{-\frac{m+1}{n}+1} \quad \text{for all } (\mathbf{p},\mathbf{q}) \in \mathbb{Z}^{m+n},\ \mathbf{q} \neq \mathbf{0}.$$

More generally, from the convergence part of Theorem 3, it is then clear that for almost every $X \in \mathbb{I}^{mn}$ there exists a constant $C(\mathbf{X}) > 0$ such that

$$|\mathbf{qX} - \mathbf{p}| \geq C(\mathbf{X})\psi(|\mathbf{q}|) \quad \text{for all } (\mathbf{p},\mathbf{q}) \in \mathbb{Z}^{m+n},\ \mathbf{q} \neq \mathbf{0}. \tag{8}$$

Denote the set of all such numbers as Bad(c,m,n), where $C(X) = c$, and

$$\cup_{c>0}\mathrm{Bad}(c,m,n) = \mathbb{I}^{mn} \setminus \widetilde{W}(m,n;\psi)\,.$$

Now since $|\widetilde{W}(m,n;\psi)\,|_{mn} = 0$, we have $|\cup_{c>0}\mathrm{Bad}(c,m,n)|_{mn} = 1$. The question of finding the Hausdorff dimension and measure of each Bad(c,m,n) is not dealt with here. However, for the set Bad(m,n) it is straightforward to establish the following result.

Theorem 9 *If $m \geq n$, then*

$$\dim \boldsymbol{Bad}(m,n) = mn,$$

and if $m < n$, *then*

$$|\boldsymbol{Bad}(m,n)|_{mn} = 1.$$

The proof of Theorem 9 follows by setting $u = 1$ in [15, 17]. Now, for $m \geq n$, since $\mathbf{Bad}(m,n) \subseteq \mathbb{I}^{mn} \setminus \widetilde{W}(m,n;\psi)$, we have $|\mathbf{Bad}(m,n)|_{mn} = 0$.

Remark 2 Loosely speaking, Bad(m,n) consists of all those points that stay clear of $(m-1)n$-dimensional hyperplanes having diameters proportional to $|\mathbf{q}|^{-\frac{m+1}{n}+1}$ centred at the hyperplanes R_q. Note that if the exponent $-\frac{m+1}{n}+1$ is replaced by $-\frac{m+1}{n}+1-\epsilon$ for $\epsilon > 0$, then the set Bad(m,n) is of full Lebesgue measure.

Remark 3 In the case $m < n$, the set $\widetilde{W}(m,n;\psi)$ is overdetermined and lies in a subset of strictly lower dimension than mn. To see this, consider the case $m = n$ and $\det\mathbf{X} \neq 0$. This would imply that the defining inequalities (4) take the form

$$|\mathbf{q} - \mathbf{p}\mathbf{X}^{-1}| \leq C(\mathbf{X})\psi(|\mathbf{q}|),$$

which is obviously not true for sufficiently large $\mathbf{q}$.

The same logic extends to all other cases. For each $m \times n$ matrix $\mathbf{X} \in \mathbb{R}^{mn}$ with column vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ define $\tilde{\mathbf{X}}$ to be the $m \times (n-1)$ matrix with column vectors $\mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$. The set $\Gamma \subset \mathbb{R}^{mn}$ is the set of $\mathbf{X} \in \mathbb{R}^{mn}$ such that the determinant of each $m \times m$ minor of $\tilde{\mathbf{X}}$ is zero, i.e. $\operatorname{rank}(\tilde{\mathbf{X}}) < m$.

It can easily be proved that $\widetilde{W}(m, n; \psi) \subset \Gamma$ when $m < n$, which leads to further investigations of metric theory for the cases $m < n$. However, this is not within the scope of the present paper. We refer the interested reader to [7, 19], which comprehensively discuss such cases.

3.2 Proof of Theorem 4

The proof of this theorem is a straightforward application of Minkowski's linear forms theorem, which we state below for completeness.

Theorem 10 (Minkowski's linear forms theorem [13]) *Let $\mathfrak{C}$ be an n-dimensional lattice of determinant* $\det(\mathfrak{C})$ *and let a_{ij} $(1 \leq i, j \leq n)$ be real numbers. Suppose that $c_j > 0$ for $1 \leq j \leq n$, are numbers such that*

$$c_1 \cdot sc_n \geq \det(a_{ij}) \det(\mathfrak{C}).$$

Then there is a non-zero integer point $u = (u_1, u_2 \cdots, u_n) \in \mathfrak{C}$ satisfying

$$\left|\sum_{j=1}^{n} a_{1j}u_j\right| \leq c_1 \text{ and } \left|\sum_{j=1}^{n} a_{ij}u_j\right| < c_i \quad (1 < i \leq n).$$

Having Minkowski's theorem at our disposal, we are now in a position to prove Theorem 4. To avoid complicated expressions, we prove it for $m = 2, n = 1$ as the higher dimensional cases follow on similar lines. The proof of the case $m = n = 1$ can be found in [8].

Let $\mathbf{Z} = (x_1 + iy_1, x_2 + iy_2)$, $\mathbf{q} = (q_{1,1} + iq_{1,2}, q_{2,1} + iq_{2,2})$, and $p = (p_1 + ip_2)$. Then

$$\begin{aligned}|\mathbf{qZ} - \mathbf{p}| =& |q_{1,1}x_1 + q_{2,1}x_2 - q_{1,2}y_1 - q_{2,2}y_2 - p_1 + \\ & i(q_{1,2}x_1 + q_{2,2}x_2 + q_{1,1}y_1 + q_{2,1}y_2 - p_2)| < c/N^2\end{aligned}$$

holds if

$$\max\{|q_{1,1}x_1 + q_{2,1}x_2 - q_{1,2}y_1 - q_{2,2}y_2 - p_1|,\\ |q_{1,2}x_1 + q_{2,2}x_2 + q_{1,1}y_1 + q_{2,1}y_2 - p_2|\} < \frac{c}{\sqrt{2}N^2}.$$

By Minkowski's linear forms theorem, the system of inequalities

$$\begin{gathered}|q_{1,1}x_1 + q_{2,1}x_2 - q_{1,2}y_1 - q_{2,2}y_2 - p_1| < \frac{c}{\sqrt{2}N^2},\\ |q_{1,2}x_1 + q_{2,2}x_2 + q_{1,1}y_1 + q_{2,1}y_2 - p_2| < \frac{c}{\sqrt{2}N^2},\\ |q_{1,1}| \le \frac{N}{\sqrt{2}},\quad |q_{1,2}| \le \frac{N}{\sqrt{2}},\quad |q_{2,1}| \le \frac{N}{\sqrt{2}},\quad |q_{2,2}| \le \frac{N}{\sqrt{2}},\end{gathered}$$

has a non-zero solution in integers for $c \ge 8$. Hence the equation (6) has a non-zero integer solution with $0 < |\mathbf{q}| \le N$.

Remark 4 The complex points for which Theorem 4 cannot be improved by an arbitrary constant are called badly approximable. That is, a point $\mathbf{Z} \in \mathbb{C}^{mn}$ is said to be *badly approximable* if there exists a constant $C(\mathbf{Z}) > 0$ such that

$$|\mathbf{qZ} - \mathbf{p}| > C(\mathbf{Z})|\mathbf{q}|^{-\frac{m}{n}}$$

for all $(\mathbf{p}, \mathbf{q}) \in \mathbb{Z}^n[i] \times \mathbb{Z}^m[i]\backslash\{\mathbf{0}\}$. Let $\mathbf{Bad}_{\mathbb{C}}(m, n)$ denote the set of badly approximable points in $\mathbb{C}^{mn}$. The best possible value for the constant $C(\mathbf{Z})$ for $m = n = 1$ is $1/\sqrt{3}$, see [11]. For higher dimensions, just as in the real case, best possible values for the constants are not known.

The Hausdorff dimension of the set $\mathbf{Bad}_{\mathbb{C}}(1, 1)$ has been studied by various authors in different frameworks; see, for instance, [24, Sect. 5.3] in which authors determined the Hausdorff dimension for $\mathbf{Bad}_{\mathbb{C}}(1, n)$, i.e.

$$\dim \mathbf{Bad}_{\mathbb{C}}(1, n) = n.$$

In fact, as a consequence of the general framework in their paper, they proved the Hausdorff dimension to be maximal in the weighted analogue of $\mathbf{Bad}_{\mathbb{C}}$ intersected with any compact subset of $\mathbb{C}^n$. Their framework cannot be applied for the dual setup at hand. However, it is reasonable to suspect that the Hausdorff dimension for $\mathbf{Bad}_{\mathbb{C}}(m, n)$ is maximal. More generally, for any compact subset $K \subset \mathbb{C}^{mn}$, we have the following conjecture.

Conjecture 1 $\dim \mathbf{Bad}_{\mathbb{C}}(m, n) \cap K = \dim K$.

The treatment required to deal with this problem involves delicate number theoretic tools which are beyond the scope of this paper.

3.3 Proof of Theorem 5

As for Theorem 4, Theorem 5 is proved for the case $m = 2, n = 1$, leaving to the reader the obvious modifications to deal with higher dimensions. First, the convergence case is dealt with. The resonant set is defined as

$$\begin{aligned} C_{\mathbf{q}} &:= \{\mathbf{Z} \in D : |\mathbf{q}\mathbf{Z} - \mathbf{p}| = 0\} \\ &= \Big\{(x_1 + iy_1, x_2 + iy_2) \in D : \\ &\qquad |(q_{1,1} + iq_{1,2}, q_{2,1} + iq_{2,2}) \cdot (x_1 + iy_1, x_2 + iy_2) - (p_1 + ip_2)| = 0\Big\} \\ &= \left\{(x_1 + iy_1, x_2 + iy_2) \in D : \begin{array}{l} q_{1,1}x_1 + q_{2,1}x_2 - q_{1,2}y_1 - q_{2,2}y_2 = p_1 \text{ and} \\ q_{1,2}x_1 + q_{2,2}x_2 + q_{1,1}y_1 + q_{2,1}y_2 = p_2 \end{array}\right\}. \end{aligned}$$

The set $W_{\mathbb{C}}(2, 1; \Psi)$ can be written using the resonant sets

$$W_{\mathbb{C}}(2, 1; \Psi) = \bigcap_{N=1}^{\infty} \bigcup_{r>N} \bigcup_{C_{\mathbf{q}}:|\mathbf{p}|<|\mathbf{q}|=r} B\left(C_{\mathbf{q}}, \Psi(|\mathbf{q}|)\right)$$

where

$$B\left(C_{\mathbf{q}}, \Psi(|\mathbf{q}|)\right) = \left\{\mathbf{Z} \in D : \operatorname{dist}\left(\mathbf{Z}, C_{\mathbf{q}}\right) \leq \frac{\Psi(|\mathbf{q}|)}{|\mathbf{q}|}\right\}.$$

It follows that

$$W_{\mathbb{C}}(2, 1; \Psi) \subseteq \bigcup_{r>N} \bigcup_{C_{\mathbf{q}}:|\mathbf{p}|<|\mathbf{q}|=r} B\left(C_{\mathbf{q}}, \Psi(|\mathbf{q}|)\right).$$

In other words, $W_{\mathbb{C}}(2, 1; \Psi)$ has a natural cover $C = \left\{B\left(C_{\mathbf{q}}, \Psi(|\mathbf{q}|)\right) : |\mathbf{q}| > N\right\}$ for each $N = 1, 2, \cdots$. It can further be covered by a collection of four-dimensional hypercubes with disjoint interior and side length comparable with $\Psi(|\mathbf{q}|)/|\mathbf{q}|$. The number of such hypercubes is clearly $\ll (\Psi(|\mathbf{q}|)/|\mathbf{q}|)^{-2}$. Thus,

$$\begin{aligned} |W_{\mathbb{C}}(2, 1; \Psi)|_2 &\leq \sum_{r=N}^{\infty} \sum_{C_{\mathbf{q}}:|\mathbf{p}|<|\mathbf{q}|=r}^{\infty} \left|B\left(C_{\mathbf{q}}, \Psi(|\mathbf{q}|)\right)\right| \\ &\ll \sum_{r=N}^{\infty} \sum_{r<|\mathbf{q}|\leq r+1} |\mathbf{q}|^2 \left(\Psi(|\mathbf{q}|)/|\mathbf{q}|\right)^{-2} \left(\Psi(|\mathbf{q}|)/|\mathbf{q}|\right)^4 \\ &= \sum_{r=N}^{\infty} \Psi(r)^2 \sum_{r<|\mathbf{q}|\leq r+1} 1 \ll \sum_{r=N}^{\infty} r^3 \Psi(r)^2. \end{aligned}$$

Here, we follow an argument from [13, Th. 386] to conclude that

$$\sum_{r<|\mathbf{q}|\leq r+1} 1 \ll r^3.$$

Now, since the sum $\sum_{r=N}^{\infty} r^3\Psi(r)^2 < \infty$, the tail of the series can be made arbitrarily small. Hence, by the Borel–Cantelli lemma, $|W_{\mathbb{C}}(2, 1; \Psi)|_2 = 0$.

The divergence case of the above theorem can similarly be proved by following arguments as in the real case. More precisely, one would need to utilise the ubiquity framework to extend [1, Theorem 7] for the linear forms setup. Theorem 4 would again be used to prove the ubiquity lemma.

3.4 Proofs of Theorems 6 and 7

The proofs of Theorems 6 and 7 are very similar to the proofs of Theorems 4 and 5, respectively, with obvious modifications, and therefore are not given here.

4 Applications to Interference Alignment

Theorems 3 and 7 can be used to obtain several fascinating results in the field of Information Theory. These result are comprehensively discussed in another manuscript by the authors in [28]. Here, we provide basic setups to show how the new theorems (in particular Theorem 3) can be applied in communication systems to achieve the best performance in terms of achievable rates.

Let $(-Q, Q)_{\mathbb{Z}} = (-Q, Q) \cap \mathbb{Z}$ be the set of integers between $-Q$ and Q. In our model of a point to point communication system, a user selects and transmits a number $x \in (-\sqrt{P}, \sqrt{P})$ to a receiver where a number y is received as $y = x + z$. Here P is a power constraint which limits the maximum allowable transmittable absolute values and z is a normal random variable. For simplicity, we assume it to have zero mean and unit variance. The receiver attempts to recover x from the observed number y. In [29], it is shown that there exists a communication strategy such that the receiver can recover the original number with high probability if $x \in 2\mathbb{Z} \cap (-\sqrt{P}, \sqrt{P})$. The ability of the recovery comes from the fact that the selected points for x have minimum distance two, which is comparable to the noise variance which is one. The receiver in fact is able to distinguish with high probability between the points if an appropriate coding strategy is used. In this way, one can transfer roughly $0.5 \log_2 P$ bits to the receiver. This is due to fact that there are $\sqrt{P}$ possible points in $2\mathbb{Z} \cap (-\sqrt{P}, \sqrt{P})$ which can be indexed by $0.5 \log_2 P$ bits.

In a Multiple Access Channel (MAC), a set of users sends information to a single receiver. Assuming availability of three users the received signal can be denoted by $y = x_1 + ax_2 + bx_3 + z$, where z is a normal random variable with zero mean and unit variance. x_i for $i \in \{1, 2, 3\}$ is the transmitted number from the ith user. Similar to the point to point system, the receiver can recover the combination $x_1 + ax_2 + bx_3$ if the minimum distance between all possible combinations is greater than two. Let us assume x_i selects its numbers according to $2P^{1/3}\mathbb{Z} \cap (-P^{1/6}, P^{1/6})$. If the receiver can recover the transmitted signals from all the users, then the ith user can transmit $\frac{1}{6}\log_2 P$ bits to the receiver. It is sufficient to compute the minimum distance called $d_{\min}$ and shows that it is greater than or equal to a constant. From the selected points,

$$d_{\min} = 2P^{1/3} \times \min |u_1 + au_2 + bu_3|,$$

where u_i are integers in $(-2P^{1/6}, 2P^{1/6})$. Here, the Khintchine–Groshev theorem comes to play as it provides a lower bound on the minimum distance. The convergence part of the theorem implies that there exists a constant $\kappa > 0$ such that for almost all a and b

$$|u_1 + au_2 + bu_3| > \frac{\kappa}{P^{1/3}}.$$

In this way, the minimum distance is at least κ and being a constant number shows that the receiver can decode the messages of all users.

Next, we consider a communication scenario where the direct application of the Khintchine–Groshev theorem cannot provide the best achievable rate and, therefore, Theorem 3 is needed to attain the capacity of the system. In particular, we consider a MAC with three single-antenna users and a two-antenna receiver. The channel can be modelled by

$$\begin{cases} y_1 = x_1 + ax_2 + bx_3 + z_1, \\ y_2 = x_1 + \hat{a}x_2 + \hat{b}x_3 + z_2, \end{cases} \tag{9}$$

where z_1 and z_2 are normal random variables with zero means and unit variances. Since the capacity region of this channel is fully characterised, it can be shown without difficulty that each user can transmit $\frac{1}{3}\log_2 P$ bits to the receiver. The naive application of the Khintchine–Groshev theorem results in a shortcoming in achieving this rate. To see this, let us assume that all three users communicate with the receiver using a single data stream. The data streams are modulated by the constellation $\mathcal{U} = A\mathbb{Z} \cap (-Q, Q)_{\mathbb{Z}}$, where A is a factor controlling the minimum distance of the received combinations.

The received combination, which is a set of points in a two-dimensional space, consists of points $(v, \hat{v})$ such that $v = A(u_1 + au_2 + bu_3)$ and $\hat{v} = A(u_1 + \hat{a}u_2 + \hat{b}u_3)$, where u_i's are members of $\mathcal{U}$. Let us choose two sets of distinct points $(v_1, \hat{v_1})$ and $(v_2, \hat{v_2})$ in the received signal. The Khintchine–Groshev theorem provides a lower bound on any linear combination of integers. It also provides some bound on the distance between any integer vector and the linear combination of rationally independent vectors. Using the Khintchine–Groshev theorem (Theorem 2 for $m = 2, n = 1$), one can obtain $d_{\min} \approx \frac{A}{Q^2}$, where $d_{\min}$ is the minimum distance in the received combination, for precise calculation of min distance we refer to [29, Sect. A].

Borrowing from the results in [29], the noise can be removed if $d_{\min}$ is constant. Hence, it is sufficient to have $A \approx Q^2$. In a noise-free environment, each receiver antenna can decode the three messages if there is a one-to-one map from the received constellation to the transmit constellations. Mathematically, one can satisfy the separability condition by enforcing the following: each received antenna is able to decode all three messages, if the channel coefficients associated with that antenna are rationally independent. In the above multiple access channel, for instance, the receiver can decode all messages by using the signal from the first antenna if $u_1 + au_2 + bu_3 = 0$ has no non-trivial solution in integers for u_1, u_2 and u_3.

User i's rate is equal to $\log_2(2Q - 1)$. Because of the power constraint, $P = A^2Q^2$. It was shown earlier that $A \approx Q^2$. Therefore, $P \approx Q^6$. Hence, the achievable rate is $\frac{1}{6}\log_2 P$ bits.

The reason for this loss of rate comes from the poor estimate of the minimum distance obtained from the classical Khintchine–Groshev theorem. In Theorem 3, we have obtained a better bound on the minimum distance in (9) as the equations have constraints similar to (4). Therefore, the minimum distance can be obtained as $d_{\min} \approx \frac{A}{\sqrt{Q}}$ and similar calculations reveal that $\frac{1}{3}\log_2 P$ bits can be transmitted from each user and decoded by the receiver. In this way, the capacity of the channel is achieved. For a detailed description of this application, we refer the reader to our paper [28].

Acknowledgements The authors are listed in alphabetical order. We would like to thank Professor Amir Khandani for many useful discussions and guidance. We thank the anonymous referee and Richard Brent for useful comments which have improved the quality and the presentation of the paper. Mumtaz Hussain would like to thank his late mentor Laureate Professor Jonathan Borwein for many fruitful discussions surrounding topics of this paper in particular and number theory in general during his career at the University of Newcastle (Australia). Jon left a lifelong impression on many of us with his breadth of knowledge, sharpness and innovative ideas on nearly every topic.

References

1. Beresnevich, V., Dickinson, D., Velani, S.: Measure theoretic laws for lim sup sets. Mem. Am. Math. Soc. **179**(846), x+91 (2006)
2. Beresnevich, V., Ramírez, F., Velani, S.: Metric diophantine approximation: aspects of recent work. Dynamics and Analytic Number Theory, London Mathematical. Society Lecture Notes Series, vol. 437, pp. 1–95. Cambridge University Press, Cambridge (2016)
3. Beresnevich, V., Velani, S.: Ubiquity and a general logarithm law for geodesics. In: Dynamical Systems and Diophantine Approximation. Séminars Congress, vol. 19, pp. 21–36. Society Mathematical, France, Paris (2009)
4. Bresler, G., Parekh, A., Tse, D.N.: The approximate capacity of the many-to-one and one-to-many Gaussian interference channels. IEEE Trans. Inf. Theory **56**(9), 4566–4592 (2010)
5. Cadambe, V.R., Jafar, S.A.: Interference alignment and the degrees of freedom for the k user interference channel. IEEE Trans. Inf. Theory **58**(8), 5130–5150 (2012)
6. Cadambe, V.R., Jafar, S.A.: Degrees of freedom of wireless *X* networks. IEEE Trans. Inf. Theory **55**(9), 3893–3908 (2009)
7. Dickinson, D., Hussain, M.: The metric theory of mixed type linear forms. Int. J. Number Theory **9**(2), 77–90 (2013)

8. Dodson, M.M. Kristensen, S.: Hausdorff dimension and Diophantine approximation. In: Fractal Geometry and Applications: A Jubilee of Benoît Mandelbrot. Part 1. Proceedings of Symposia in Pure Mathematics, vol. 72, pp. 305–347. American Mathematical Society, Providence, RI (2004)
9. Etkin, R., Ordentlich, E.: On the degrees-of-freedom of the K-user Gaussian interference channel. IEEE Trans. Inf. Theory **55**(11), 4932–4946 (2009)
10. Fischler, S., Hussain, M., Kristensen, S., Levesley, J.: A converse to linear independence criteria, valid almost everywhere. Ramanujan J. **38**(3), 513–528 (2015)
11. Ford, L.R.: On the closness of approach of complex rational fractions to a complex irrational number. Trans. Am. Math. Soc. **27**, 146–154 (1925)
12. Groshev, A.V.: A theorem on a system of linear forms. Doklady Akad. Nauk SSSR **19**, 151–152 (1938). (in Russian)
13. Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers, 5th edn. The Clarendon Press, Oxford University Press, New York (1979)
14. Harrap, S., Hussain, M., Kristensen, S.: A problem in non-linear Diophantine approximation. Nonlinearity **31**, 1734–1756 (2018)
15. Hussain, M.: A note on badly approximable linear forms. Bull. Aust. Math. Soc. **83**(2), 262–266 (2011)
16. Hussain, M.: A Khintchine-Groshev type theorem in absolute value over complex numbers. New Zealand J. Math. **47**, 57–67 (2017)
17. Hussain, M., Kristensen, S.: Badly approximable systems of linear forms in absolute value. Unif. Dist. Theory **8**(1), 7–15 (2013)
18. Hussain, M., Kristensen, S.: Metrical results on systems of small linear forms. Int. J. Number Theory **9**(3), 769–782 (2013)
19. Hussain, M., Levesley, J.: The metrical theory of simultaneously small linear forms. Funct. Approx. Comment. Math. **48**(2), 167–187 (2013)
20. Hussain, M., Yusupova, T.: A note on weighted Khintchine-Groshev theorem. J. Théor. Nombres Bordeaux **26**(2), 385–397 (2014)
21. Jafar, S.: Interference alignment – a new look at signal dimensions in a communication network. Found. Trends Commun. Inf. Theory **7**(1) (2010)
22. Jafar, S., Shamai, S.: Degrees of freedom region of the MIMO X-channels. IEEE Trans. Inf. Theory **54**(1), 151–170 (2008)
23. Khintchine, A.: Zur metrischen Theorie der Diophantischen Approximationen. Math. Z. **24**(1), 706–714 (1926)
24. Kristensen, S., Thorn, R., Velani, S.: Diophantine approximation and badly approximable sets. Adv. Math. **203**(1), 132–169 (2006)
25. LeVeque, W.J.: Continued fractions and approximations in $k(i)$. I, II. Nederl. Akad. Wetensch. Proc. Ser. A. **55** = Indag. Math. **14**, 526–535, 536–545 (1952)
26. Maddah-Ali, M., Motahari, A., Khandani, A.: Communication over MIMO X-channels: Interference alignment, decomposition, and performance analysis. IEEE Trans. Inf. Theory **54**(8), 3457–3470 (2008)
27. Maddah-Ali, M.A.: On the degrees of freedom of the compound MISO broadcast channels with finite states. In: Proceedings 2010 IEEE International Symposium on Information Theory, pp. 2273–2277. IEEE (2010)
28. Mahboubi, S.H., Hussain, M., Motahari, A.S., Khandani, A.K.: Layered interference alignment: achieving the total DOF of MIMO X-channels. Preprint: arXiv:1412.7188
29. Motahari, A.S., Oveis-Gharan, S., Maddah-Ali, M.-A., Khandani, A.K.: Real interference alignment: exploiting the potential of single antenna systems. IEEE Trans. Inf. Theory **60**(8), 4799–4810 (2014)
30. Ordentlich, O., Erez, U.: Precoded integer-forcing universally achieves the MIMO capacity to within a constant gap. IEEE Trans. Inf. Theory **61**(1), 323–340 (2015)
31. Sullivan, D.: Disjoint spheres, approximation by imaginary quadratic numbers, and the logarithm law for geodesics. Acta Math. **149**(3–4), 215–237 (1982)

Improved Bounds on Brun's Constant

Dave Platt and Tim Trudgian

Dedicated to the memory of Jon Borwein

1 Introduction

Brun [4] showed that the sum of the reciprocals of the twin primes converges. That is, if P_2 denotes the set of primes p such that $p+2$ is also prime, the sum $B := \sum_{p \in P_2} 1/p + 1/(p+2)$ is finite.

Various estimates for Brun's constant have been given based on calculations of $\pi_2(x)$, where $\pi_2(x)$ denote the number of twin primes not exceeding x—see Brent [3, pp. 50–53] and Klyve [7, Table 1.2.3] for some historical references. Brent [3] computed $\pi_2(8 \cdot 10^{10}) = 182\,855\,913$, and, conditional on some assumptions about the random distribution of twin primes, conjectured that

$$B = 1.9021604 \pm 5 \cdot 10^{-7}. \tag{1}$$

Platt—Supported by Australian Research Council Discovery Project DP160100932 and EPSRC Grant EP/K034383/1.
Trudgian—Supported by Australian Research Council Discovery Project DP160100932 and Future Fellowship FT160100094.

D. Platt
School of Mathematics, University of Bristol, University Walk,
Bristol BS8 1TW, UK
e-mail: dave.platt@bris.ac.uk

T. Trudgian (✉)
School of Sciences, UNSW Canberra,
Canberra, BC 2610, Australia
e-mail: t.trudgian@adfa.edu.au

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_25

Additional computations were performed by Gourdon and Sebah [16] and Nicely[1] [11], who showed

$$\pi_2(2 \cdot 10^{16}) = 19\,831\,847\,025\,792. \tag{2}$$

Additionally, Nicely conjectured that

$$B = 1.902160583209 \pm 0.000000000781. \tag{3}$$

As far as we are aware the most comprehensive results on the enumeration of $\pi_2(x)$ are by Oliveira e Silva [12], who computed $\pi_2(k \cdot 10^n)$ for $k = 1, \ldots, 10\,000$ and $n = 1, \ldots, 14$ and $\pi_2(k \cdot 10^{15})$ for $k = 1, \ldots, 4\,000$.

Some explanation is required for these conjectured bounds in (1) and (3). These results are not strict error bounds, but rather, confidence intervals (in the probabilistic sense). One can obtain a lower bound on B by merely summing $B(N) := \sum_{p \in P_2, p \le N} 1/p + 1/(p+2)$ for large values of N. One can then plot this as a function of N, make assumptions about the random distribution of twin primes, and try to ascertain the rate of convergence. This is what has been done by Brent, Nicely, and others.

It is another matter to ask for a rigorous upper bound for Brun's constant; clearly computing the sum $B(N)$ for any N gives a lower bound. The first upper bound appears to be $B < 2.347$ as stated by Crandall and Pomerance [5]. A proof of this is given in a thesis by Klyve [7] who also shows that under the assumption of the Generalised Riemann Hypothesis we have $B < 2.1754$.

It is perhaps curious that the method of Crandall and Pomerance produces an upper bound for B that depends on the lower bound. When one increases N, the corresponding increase in $B(N)$ yields a better upper bound for B.

In this paper we do two things: we compute $B(N)$ for a larger N than was done previously, and using some optimisation improve the upper bound for B. The result is

Theorem 1 $1.840503 < B < 2.288490$.

The previous best lower bound was computed by Nicely [11], who, using his calculations of (2) showed that $B(2 \cdot 10^{16}) > 1.831808$. We remark that the lower bound of $B(10^{16}) > 1.83049$ by Gourdon and Sebah [16] was used by Klyve.

In Section 4.1 we give details of using the tables by Oliveira e Silva in [12] to compute $B(4 \cdot 10^{18})$. This proves the lower bound in Theorem 1. We remark here that this computation on its own would give an upper bound of 2.292 in Theorem 1.

In §2 we list two results in the literature, one an explicit bound on a sum of divisors, and another an improvement on a sieving inequality used by Montgomery

[1] We cannot resist referencing an anecdote from Jon Borwein (and his co-authors). Nicely's calculations on Brun's constant are mentioned in [2, p. 40]. Nicely discovered a bug in an Intel Pentium chip, which, according to [2] 'cost Intel about a billion dollars' although the actual amount written off was a mere US$475 million. We believe Jon would have seen this as an excellent application of pure mathematics in the modern world.

and Vaughan [9]. In §3 we introduce Riesel and Vaughan's bounds for $\pi_2(x)$. Finally, in §4 we perform our calculations that prove the upper bound in Theorem 1, and outline some of the difficulties facing future investigations into this problem.

2 Preparatory Results

We require two results from the literature. The first is an explicit estimate on $\sum_{n\le x} d(n)/n$, where $d(n)$ is the number of divisors function; the second is a large-sieve inequality.

2.1 *Bounds on the Number of Divisors*

The classical bound on $\sum_{n\le x} d(n)$ and partial summation show that

$$\sum_{n\le x}\frac{d(n)}{n}\sim\frac{1}{2}\log^2 x. \tag{4}$$

It is also possible to give an asymptotic expansion of the above relation. First, for k a non-negative integer, define the Stieltjes constants γ_k as

$$\gamma_k=\lim_{N\to\infty}\left\{-\frac{(\log N)^{k+1}}{k+1}+\sum_{n\le N}\frac{(\log n)^k}{n}\right\}.$$

Here $\gamma_0=\gamma$, which is Euler's constant. In what follows we only need the following bounds: more precision is possible, but the estimates in (5) are more than sufficient.

$$0.5772156<\gamma_0<0.5772157,\quad -0.0728159<\gamma_1<-0.0728158. \tag{5}$$

Riesel and Vaughan give a more refined estimation of (4), namely, if

$$E(x)=\sum_{n\le x}\frac{d(n)}{n}-\frac{1}{2}\log^2 x-2\gamma_0\log x-\gamma_0^2+2\gamma_1, \tag{6}$$

then by Lemma 1 [14]

$$|E(x)|<1.641x^{-1/3},\quad (x>0). \tag{7}$$

We note that an improvement is claimed in Corollary 2.2 in [1], which gives

$$|E(x)| < 1.16x^{-1/3}, \quad (x > 0).$$

This, however, appears to be in error, since, as shown in [14, p. 50] the error $|E(x)|x^{1/3}$ has a maximum of $-1.6408\ldots$ around $7.345 \cdot 10^{-4}$. We also note that one only need prove a result like (7) for $x \geq 1$ to follow the proof of Lemma 2 in [14]. (We thank Richard Brent and the anonymous referee for pointing this out.) Finally, it is possible to improve (7) by choosing an exponent smaller than $-1/3$. We will use $-2/5$ so we require a lemma.

Lemma 1 *Let $E(x)$ be as in (6). Then, for all $x \geq 1$ we have $|E(x)| \leq 0.6877x^{-2/5}$.*

Proof We proceed as in the proof of Lemma 1 in [14]. There, the authors consider three ranges, $x \geq 2$, $1 \leq x < 2$ and $0 < x < 1$. The idea with such a proof is by considering sufficiently many ranges, one can show that the global maximum of $|E(x)|x^{\alpha}$ occurs in $0 < x < 1$. By reducing α we reduce this maximum value. We find that writing $(1, \infty)$ as the union of $[n, n+1)$ for $1 \leq n \leq 7$ and $[8, \infty]$ keeps the other contributions sufficiently small. The maximum value is at $x = 6^{-}$, which establishes the lemma. □

We remark that the proof is easily adaptable to finding, for a given α, the optimal constant $c = c(\alpha)$ such that $|E(x)|x^{\alpha} \leq c$ for all $x \geq 1$. However, as we show in §4.3, the effects of further improvements are minimal.

2.2 *A Large Sieve Inequality*

Riesel and Vaughan make use of the following, which is Corollary 1 in [9].

Theorem 2 (Montgomery and Vaughan) *Let $\mathcal{N}$ be a set of Z integers contained in $[M+1, M+N]$. Let $\omega(p)$ denote the number of residue classes mod p that contain no element of $\mathcal{N}$. Then $Z \leq L^{-1}$, where*

$$L = \sum_{q \leq z} \left(N + \frac{3}{2}qz\right)^{-1} \mu^2(q) \prod_{p|q} \frac{\omega(p)}{p - \omega(p)}, \tag{8}$$

where z is any positive number.

Actually, Theorem 2 is derived from the investigations of Montgomery and Vaughan into Hilbert's inequality[2]. Specifically, Theorem 2 follows from Theorem 1 in [10].

[2] We were reminded by the referee that Jon Borwein had worked on Hilbert's inequality, although we do not believe his results to be applicable here.

That result was improved by Preissmann [13]. The upshot of all this is that Preissmann's work allows one to take $\rho = \sqrt{1 + 2/3\sqrt{6/5}} \approx 1.315\ldots$ in place[3] of 3/2 in (8).

Riesel and Vaughan choose $z = (2x/3)^{1/2}$ in (8). With Preissman's improvement we set $z = (x/\rho)^{1/2}$; it is trivial to trace the concomitant improvements.

3 Riesel and Vaughan's Bounds on $\pi_2(x)$

Riesel and Vaughan give a method to bound $\pi_2(x)$. Actually, their method is much more general and can bound the number of primes $p \le x$ such that $ap + b$ is also prime. We present below their method for the case of interest to us, namely, that of $a = 1, b = 2$. One may also consult [17]—we thank Olivier Ramaré for making us aware of this.

We first let C denote the twin prime constant

$$C = 2 \prod_{p>2} \frac{p(p-2)}{(p-1)^2}. \tag{9}$$

Note that in some sources the leading factor of 2 may be absent. Wrench [18] computed C to 45 decimal places. For our purposes the bound given by Riesel and Vaughan below is sufficient

$$1.320323 < C < 1.320324.$$

Lemma 2 *For any $s > -1/2$ we define $H(s)$ by*

$$H(s) = \sum_{n=1}^{\infty} \frac{|g(n)|}{n^s},$$

where $g(n)$ is a multiplicative function defined by

$$g(p^k) = 0 \textit{ for } k > 3, \quad g(2) = 0, \quad g(4) = -3/4, \quad g(8) = 1/4,$$
$$g(p) = \frac{4}{p(p-2)}, \quad g(p^2) = \frac{-3p-2}{p^2(p-2)}, \quad g(p^3) = \frac{2}{p^2(p-2)}, \quad (\textit{when } p > 2).$$

[3] We remark that Selberg conjectured that (8) holds with 1 in place of 3/2. It seems difficult to improve further on Preissmann's work.

Now define the constants A_i by

$$\begin{aligned}A_6 &= 9.27436 - 2\log\rho\\ A_7 &= -5.6646 + \log^2\rho - 9.2744\log\rho\\ A_8 &= 16Cc(\alpha)H(-\alpha)\rho^{\alpha/2}\\ A_9 &= 24.09391\rho^{1/2},\end{aligned}$$

where $c(\alpha)$ is such that $|E(x)|x^{\alpha} \le c(\alpha)$ for all $x > 0$. Now let

$$F(x) = \max\left\{0,\ A_6 + \frac{A_7}{\log x} - \frac{A_8}{x^{\alpha/2}\log x} - \frac{A_9}{x^{1/2}\log x}\right\}. \tag{10}$$

Then

$$\pi_2(x) < \frac{8Cx}{(\log x)(\log x + F(x))} + 2x^{1/2}. \tag{11}$$

Proof See [14], equation (3.20). □

This leads directly to the following lemma.

Lemma 3 *Let $F(x)$ be defined in (10). Choose x_0 large enough so that $F(x_0) > 0$ and set*

$$B(x_0) = \sum_{\substack{p\in P_2\\ p\le x_0}} \frac{1}{p} + \frac{1}{p+2}.$$

Then

$$B \le B(x_0) - 2\frac{\pi_2(x_0)}{x_0} + \int_{x_0}^{\infty} \frac{16C}{t\log(t)(\log(t) + F(t))} + 4t^{-\frac{3}{2}}\,\mathrm{d}t.$$

Proof We start from

$$B \le B(x_0) + \sum_{\substack{p\in P_2\\ p> x_0}} \frac{2}{p} = B(x_0) + 2\int_{x_0}^{\infty} \frac{\mathrm{d}\pi_2(t)}{t},$$

integrate by parts and apply Lemma 2. □

Riesel and Vaughan calculate $H(-1/3)$ so that they may use (7); we proceed to give an upper bound for $H(-2/5)$ in order to use Lemma 1.

Lemma 4 *Let H be as defined above, then*

$$H\left(-\frac{2}{5}\right) < 950.05.$$

Proof Write

$$g(2,s) = \log\left(1 + \frac{3}{4}2^{-2s} + \frac{1}{4}2^{-3s}\right)$$

and for $t > 2$

$$g(t,s) = \log\left(1 + \frac{4}{t(t-2)}t^{-s} + \frac{3t+2}{t^2(t-2)}t^{-2s} + \frac{2}{t^2(t-2)}t^{-3s}\right)$$

so that for $s > -1/2$ we have the Euler product

$$H(s) = \exp\left[\sum_p g(p,s)\right].$$

Now fix $P > 2$ and split the sum into

$$S_1(P,s) = \sum_{p \le P} g(p,s)$$

and

$$S_2(P,s) = \sum_{p > P} g(p,s).$$

Then by direct computation using interval arithmetic we find

$$S_1\left(10^{10}, -\frac{2}{5}\right) = 6.8509190277\ldots.$$

To estimate S_2 we write

$$\sum_{p>P} g(p,s) = \int_P^\infty g(t,s)\mathrm{d}\pi(t) \le \int_P^\infty \log\left(1 + k_1 t^{-\frac{6}{5}}\right)\mathrm{d}\pi(t),$$

where k_1 is chosen so that $\log\left(1 + k_1 t^{-\frac{6}{5}}\right) \ge g\left(P, -\frac{2}{5}\right)$. For $P = 10^{10}$ we find that $k_1 = 3.000403$ will suffice. We then integrate by parts to get

$$S_2\left(P, -\frac{2}{5}\right) \le -\log\left(1 + k_1 t^{-\frac{6}{5}}\right)\pi(P) + \frac{6}{5}\int_P^\infty \frac{k_1}{t^{11/5} + k_1 t}\pi(t)\mathrm{d}t.$$

We compute the first term using $\pi\left(10^{10}\right) = 455\,052\,511$ and for the second term we note that for $x \ge P$ we have

$$\pi(x) \leq \frac{x}{\log x}\left(1+\frac{1.2762}{\log P}\right) = k_2\frac{x}{\log x}.$$

The integral is now

$$\frac{6}{5}k_1k_2\int_P^\infty \frac{\mathrm{d}t}{\log t\left(t^{6/5}+k_1\right)} \leq \frac{6}{5}k_1k_2\int_P^\infty \frac{\mathrm{d}t}{t^{6/5}\log t} = -\frac{6}{5}k_1k_2\mathrm{Ei}\left(-\frac{\log P}{5}\right),$$

where Ei is the exponential integral

$$\mathrm{Ei}(x) = -\int_{-x}^\infty \frac{\exp(-t)}{t}\mathrm{d}t.$$

Putting this all together we have

$$\begin{aligned} S_1\left(10^{10},-\frac{2}{5}\right)+S_2\left(10^{10},-\frac{2}{5}\right) &< 6.8509191-0.0013653+0.0069531 \\ &= 6.8565069 \end{aligned}$$

and thus $H\left(-\frac{2}{5}\right) < 950.05$. □

4 Calculations

We now have everything we require to prove Theorem 1. We first proceed to the lower bound.

4.1 Computing $B(4\cdot 10^{18})$: The Lower Bound in Theorem 1

We first note the following.

Lemma 5 *We have*

$$\pi_2\left(4\cdot 10^{18}\right) = 3\,023\,463\,123\,235\,320.$$

Proof See [12], table '2d15.txt'. □

Furthermore, typical entries in the tables in [12] ('2d12.txt' for this example) look like

$$1000d12 \quad 1177209242304 \quad 1177208491858.251\ldots$$

$$1001d12 \quad 1178316017996 \quad 1178315253072.811\ldots,$$

where the second column gives the count of prime pairs below the value given in the first column, interpreting, for example, '1001d12' as $1001 \cdot 10^{12}$. From this we conclude that there are $1\,178\,316\,017\,996 - 1\,177\,209\,242\,304 = 1\,106\,775\,692$ prime pairs between $1000 \cdot 10^{12}$ and $1001 \cdot 10^{12}$. The contribution these will make to the constant B is at least

$$1\,106\,775\,692 \times \frac{2}{1001 \cdot 10^{12}} > 1.0567 \cdot 10^{-6}$$

and at most

$$1\,106\,775\,692 \times \frac{2}{1000 \cdot 10^{12}} < 1.0678 \cdot 10^{-6}.$$

We take the value of $B(10^{12}) \in [1.8065924, 1.8065925]$ from [11] and add on the contributions from the entries in the tables from [12] to conclude the following.

Lemma 6

$$B\left(4 \cdot 10^{18}\right) \in [1.840503, 1.840518].$$

We note that the lower bound in Theorem 1 follows from Lemma 6. We note further that we are 'off' by at most $1.5 \cdot 10^{-5}$, which shows that there is limited applicability for a finer search of values of $\pi_2(x)$ for $x \leq 4 \cdot 10^{18}$.

4.2 The Upper Bound in Theorem 1

We shall use Lemma 3 to bound B. Using $s = -2/5$ to get $H(-2/5) < 950.05$ (Lemma 4) and $c(2/5) < 1.0503$ (Lemma 1) we get

$$A_6 > 8.72606, \quad A_7 > -8.13199, \quad A_8 < 14580.01753, \quad A_9 < 27.63359.$$

We chose $x_0 = 4 \cdot 10^{18}$ so that $\pi_2(x_0) = 3\,023\,463\,123\,235\,320$ (Lemma 5) and $B(x_0) < 1.840518$ (Lemma 6). This leaves the evaluation of

$$\int_{x_0}^{\infty} \frac{\mathrm{d}t}{t \log t (F(t) + \log t)}.$$

We proceed using rigorous quadrature via the techniques of Molin [8] implemented using the Arb package [6] to compute

$$\int_{x_0}^{\exp(20\,000)} \frac{\mathrm{d}t}{t\log t(F(t)+\log t)}$$

and then we bound the remainder by

$$\int_{\exp(20\,000)}^{\infty} \frac{\mathrm{d}t}{t\log t(F(t)+\log t)} \le \int_{\exp(20\,000)}^{\infty} \frac{\mathrm{d}t}{t\log^2 t} = \frac{1}{20\,000}.$$

This establishes Theorem 1.

4.3 Potential Improvements

We close this section by considering potential improvements whilst still relying on Riesel and Vaughan's method. One approach is to attempt to improve the constants A_i. A second would be to compute $B(x_0)$ for larger values of x_0 than the $4 \cdot 10^{18}$ used above.

4.3.1 Improving the Constants A_i

In the following, all calculations were done with $x_0 = 4 \cdot 10^{18}$, cutting off at $\exp(20\,000)$, and using Preissmann's value for ρ in §2.2.

1. The '2' that appears in (11) is a result of the term $2\pi(z)+1$ appearing on [14, p. 54]. With the choice of $z = (x/\rho)^{1/2}$, and using the bound $\pi(x) < 1.25506x/\log x$ from Rosser and Schoenfeld [15, (3.6)], we could replace the 2 by
$$x_0^{-1/2} + \frac{5.03}{\rho^{1/2}\log\frac{x_0}{\rho}} = 0.10305\ldots.$$
2. We can replace the constant A_9 by $19.638\rho^{1/2} < 22.523$ by a careful examination of the final part of the proof of Lemma 3 in [14].
3. We could investigate other versions of Lemma 1. This would have the effect of reducing A_8. It should be noted that for larger values of α one can obtain smaller constants $c(\alpha)$ at the expense of a larger, and more slowly converging, $H(-\alpha)$. We did not pursue the optimal value of α.

However, we observe that setting $A_6 = 9.27436$ (that is, assuming Selberg's conjecture, in the footnote on page 399, that $\rho = 1$), setting $A_7 = A_8 = A_9 = 0$ and deleting the $x^{1/2}$ term from (11) altogether only reduces the upper bound for B to $2.28545\ldots$.

Table 1 Projected upper bounds on B

k	$B(10^k)$	$\pi_2(10^k)$	Upper bound for B
19	1.84181	$7.2376 \cdot 10^{15}$	2.2813
20	1.84482	$6.5155 \cdot 10^{16}$	2.2641
80	1.8878	$3.9341 \cdot 10^{75}$	1.9998

4.3.2 Increasing x_0

Knowledge of $B(x_0)$ and $\pi_2(x_0)$ for larger x_0 would allow us to further improve on our bounds for B. To quantify such improvements, recall that results such as (1) and (3) are obtained by assuming the Hardy–Littlewood conjecture, namely,

$$\pi_2(x) \sim C \int_2^x \frac{dx}{\log^2 x}, \tag{12}$$

(where C is the twin prime constant in (9)), and assuming properties on the distribution of twin primes. This leads to the hypothesis that

$$B(n) \approx B - \frac{2C}{\log n}. \tag{13}$$

Using (12) and (13), one can 'predict' the value of $\pi_2(10^k)$ and $B(10^k)$ for higher values of k. Of course one can object at this point: we are assuming a value of B in order to obtain an upper bound on B! A valid point, to be sure. The purpose of this commentary is instead to show that without new ideas, this current method is unlikely to yield 'decent' bounds on B even using infeasible computational resources.

We ran the analysis from §4 (not optimised for each k) to obtain the following Table 1.

Therefore, proving even that $B < 2$ is a good candidate for the 13th Labour of Hercules, a man referenced frequently in puzzles by the late Jon Borwein.

Acknowledgements We wish to thank Richard Brent, Tomás Oliveira e Silva, Carl Pomerance, Olivier Ramaré and the anonymous referee for their comments and contributions.

References

1. Berkane, D., Bordellès, O., Ramaré, O.: Explicit upper bounds for the remainder term in the divisor problem. Math. Comput. **81**(278), 1025–1051 (2012)
2. Borwein, J., et al.: Organic Mathematics: Proceedings of the Organic Mathematics Workshop, December 12–14, 1995, Simon Fraser University, Burnaby. British Columbia. AMS, Providence (1997)

3. Brent, R.P.: Tables concerning irregularities in the distribution of primes and twin primes up to 10^{11}. Math. Comput. **30**(134), 379 (1976)
4. Brun, V.: La série $1/5 + 1/7 + 1/11 + 1/13 + 1/17$ [etc.] où les dénominateurs sont nombres premiers jumeaux est convergente ou finie. Bull. Sci. Math. **43**, 124–128 (1919)
5. Crandall., Pomerance.: Prime Numbers: A Computational Perspective, 2nd edn. Springer, New York (2005)
6. Johansson, F.: Arb: efficient arbitrary-precision midpoint-radius interval arithmetic. IEEE Trans. Comput. **66**, 1281–1292 (2017)
7. Klyve, D.: Explicit bounds on twin primes and Brun's constant. Ph.D thesis, Dartmouth College (2007)
8. Molin, P.: Intégration numérique et calculs de fonctions L. Ph.D thesis, Institut de Mathématiques de Bordeaux (2010)
9. Montgomery, H.L., Vaughan, R.C.: The large sieve. Mathematika **20**, 119–134 (1973)
10. Montgomery, H.L., Vaughan, R.C.: Hilbert's inequality. J. Lond. Math. Soc. **2**(8), 73–82 (1974)
11. Nicely, T.R.: Prime constellations research project (2010). http://www.trnicely.net/counts.html
12. Oliveira e Silva, T.: Tables of values of pi(x) and of pi2(x) (2015). http://sweet.ua.pt/tos/primes.html
13. Preissmann, E.: Sur une inégalité de Montgomery et Vaughan. Enseign. Math. **30**, 95–113 (1984)
14. Riesel, H., Vaughan, R.C.: On sums of primes. Ark. Mat. **21**(1–2), 45–74 (1983)
15. Rosser, J.B., Schoenfeld, L.: Approximate formulas for some functions of prime numbers. Illinos J. Math. **6**, 64–94 (1962)
16. Sebah, P., Gourdon, X.: Introduction to twin primes and Brun's constant computation (2002). http://numbers.computation.free.fr/constants/Constants.html
17. Siebert, H.: Montgomery's weighted sieve for dimension two. Monatsh. Math. **82**(4), 327–336 (1976)
18. Wrench Jr., J.W.: Evaluation of Artin's constant and the twin-prime constant. Math. Comput. **15**(76), 396–398 (1961)

Extending the PSLQ Algorithm to Algebraic Integer Relations

Matthew P. Skerritt and Paul Vrbik

1 Introduction

The Euclidean algorithm for real numbers [7, Book X, Prop 3] is perhaps the simplest example of an *integer relation algorithm*. Given $a, b, \in \mathbb{R}$ the algorithm computes $g \in \mathbb{R}$ such that $a = mg$ and $b = ng$ for some $m, n \in \mathbb{Z}$. If we let $s = n$ and $t = -m$ then we have found the relation $as + bt = 0$. It was Ferguson and Forcade's efforts to generalise this to the case where $a_1, \ldots, a_n \in \mathbb{R}$ in 1979 [5] that eventually led to the PSLQ algorithm by Ferguson and Bailey in 1991 [4].

This general case is attractive. One may determine if a number α is algebraic by finding an integer relation for $\left(\alpha^0, \alpha^1, \ldots, \alpha^n\right)$ for some $n \in \mathbb{N}$. Furthermore, searching for such relations involving π led to the discovery of the Bailey–Borwein–Plouffe (BBP) formula [2].

A further extension of the integer relation problem is from real numbers and integers to complex numbers and Gaussian integers respectively. This extension was shown to be handled by the PSLQ algorithm in the 1999 paper by Ferguson, Bailey and Arno [6] in which they analysed the algorithm and proved bounds on the number of iterations required to find a relation. The complex case is rarely mentioned in the literature, although we note that it is handled by *Maple*'s implementation of the algorithm.

M. P. Skerritt (✉)
School of Mathematical and Physical Sciences, Centre for Computer-Assisted
Research Mathematics and its Applications (CARMA), The University of Newcastle,
Newcastle, NSW, Australia
e-mail: matthew.skerritt@uon.edu.au

P. Vrbik
Department of Mathematical and Computational Sciences,
University of Toronto, Mississauga, ON, Canada
e-mail: paul.vrbik@utoronto.ca

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_26

The integer relation cases handled by the PSLQ algorithm are covered by the following definition.

Definition 1 (Integer Relation) Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, and let

$$\mathcal{O} = \begin{cases} \mathbb{Z} & \text{if } \mathbb{F} = \mathbb{R} \\ \mathbb{Z}[\sqrt{-1}] & \text{if } \mathbb{F} = \mathbb{C}. \end{cases}$$

For $x \in \mathbb{F}^n$, an *integer relation* of x is a vector $a \in \mathcal{O}^n$, $a \neq 0$, such that $a_1 x_1 + \cdots + a_n x_n = 0$.

We will further generalise the integer relation problem in this paper. In order to talk about the algorithm in more generality we will use the following notation.

Notation 1 $(\mathbb{F}, \mathcal{O})$ When discussing PSLQ and our generalisations we will denote by $\mathbb{F}$ the field from which the input to the algorithm is taken, and by $\mathcal{O}$ the ring of integers from which the elements of the integer relation belong.

Observe that for the linear combination property of an integer relation to be well defined, it must be the case that $\mathcal{O} \subset \mathbb{F}$. As such, we may consider the notion a nearest integer to a given element of the field. This is important for the PSLQ algorithm.

Definition 2 (Nearest Integer) Let $x \in \mathbb{F}$. An integer $a \in \mathcal{O}$ is a *nearest integer* to x if $|x - a|$ is minimal. We consider a function $\lceil \cdot \rfloor : \mathbb{F} \to \mathcal{O}$ to be a *nearest integer function* if it maps each $x \in \mathbb{F}$ to one of its nearest integers. When the ring of integers needs to be specified, we will denote a nearest integer function by $\lceil \cdot \rfloor_{\mathcal{O}}$.

1.1 Algorithm Overview

We provide a high-level description of the unmodified PSLQ algorithm which is sufficient to understand the modifications we have made. For an alternative and slightly more detailed introduction the reader is referred to Straub [10].

We show the mathematical details of the algorithm, but omit many technical considerations needed for a practical and effective implementation. Details and analysis suitable for a practical implementation can be found in the literature, in particular: Borwein [3], and Bailey and Broadhurst [1].

The PSLQ algorithm has parameters τ, γ, and ρ that must satisfy

$$\frac{1}{\rho} \geq |x - \lceil x \rfloor| \quad \forall x \in \mathbb{F} \tag{1}$$

$$1 < \tau \leq \rho \tag{2}$$

$$\frac{1}{\tau^2} = \frac{1}{\gamma^2} + \frac{1}{\rho^2} \tag{3}$$

in order to establish runtime bounds on the algorithm [6].

For a given $\mathbb{F}$, so long as $\mathcal{O}$ is a lattice, there exists ρ such that the inequality (1) is sharp. Using this value for ρ gives the most flexibility with the other parameters. From (3) we see that $\tau \to \rho$ as $\gamma \to \infty$ and that for fixed ρ there will be a greatest lower bound for γ such that $\tau > 1$.

Definition 3 (γ_1) Let ρ be such that (1) is sharp. Then γ_1 is the value of γ that satisfies $1 = 1/\gamma^2 + 1/\rho^2$.

We use the value of ρ such that (1) is sharp, and choose any $\gamma > \gamma_1$. So long as $\rho >$ 1 (i.e., $1/\rho < 1$) then all three conditions will be satisfied.

Note that when $\mathbb{F} = \mathbb{R}$ and $\mathcal{O} = \mathbb{Z}$ then the above strategy gives $\rho = 2$ and $\gamma_1 = \sqrt{4/3}$. This value of γ_1 is precisely the lower bound of γ given in the literature.

Similarly when $\mathbb{F} = \mathbb{C}$ and $\mathcal{O} = \mathbb{Z}[\sqrt{-1}]$ (i.e., Gaussian integers) then $\rho = \sqrt{2}$ and $\gamma_1 = \sqrt{2}$. This is precisely the bound on γ given in the literature for the complex case.

The PSLQ algorithm is presented in Algorithm 1, below. In order to make sense of it, we need the following definitions.

Definition 4 (Lower Trapezoidal) Let $H = \left(h_{i,j}\right)$ be an $m \times n$ matrix. If $h_{i,j} = 0$ whenever $j > i$ then H is *lower trapezoidal.*

Note that a lower trapezoidal square matrix is exactly a lower triangular matrix.

Definition 5 (H_x) Let $x \in \mathbb{F}^n$. Then the $n \times (n-1)$ matrix $H_x = \left(h_{i,j}\right)$ is defined by

$$h_{i,j} = \begin{cases} 0 & \text{if } i < j \\ s_{i+1}/s_i & \text{if } i = j \\ -\overline{x_i} x_j/(s_j s_{j+1}) & \text{if } i > j \end{cases} \qquad \text{where } s_i = \sqrt{\sum_{k=i}^{n} x_k \overline{x_k}}.$$

Note that the complex conjugates are needed for full generality to cope with the complex case. Often the literature will present only the real case of PSLQ in which case $x_k \overline{x_k} = x_k^2$ and is reported as such. Similarly for the conjugates in Definition 7, below.

Definition 6 (Hermite Reduction D_H) Let $A = (a_{i,j})$ be a lower trapezoidal $m \times n$ matrix with $a_{j,j} \neq 0$ for all j. Then the $m \times m$ matrix $D_A = (d_{i,j})$ where

$$d_{i,j} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } i = j \\ \left\lceil \dfrac{-1}{a_{j,j}} \displaystyle\sum_{k=j}^{i} d_{i,k} a_{k,j} \right\rfloor & \text{if } i > j \end{cases}$$

is the *reducing matrix* of A. The matrix $D_A\ A$ is the *Hermite reduction* of A.

Observe that D_H is a lower triangular matrix containing invertible integers on its diagonal. It is therefore an invertible matrix whose inverse is also integer valued.

Definition 7 ($Q_{[A,k]}$) Let $A = (a_{i,j})$ be an $m \times n$ matrix with $m > n$, and let $1 \leq k \leq n$. Let $\beta = a_{k,k}$, $\lambda = a_{k,k+1}$, and $\delta = \sqrt{\beta\overline{\beta} + \lambda\overline{\lambda}}$ Then the $n \times n$ block diagonal matrix

$$Q_{[A,k]} = \begin{cases} I_n & \text{if } k = n \\ \left(q_{i,j}\right) & \text{otherwise} \end{cases}$$

where $\left(q_{i,j}\right)$ is the block diagonal matrix with submatrix

$$\begin{pmatrix} q_{k,k} & q_{k,k+1} \\ q_{k+1,k} & q_{k+1,k+1} \end{pmatrix} = \frac{1}{\delta} \begin{pmatrix} \overline{\beta} & -\lambda \\ \overline{\lambda} & \beta \end{pmatrix}$$

and 1's for all other diagonal entries.

Observe that multiplication on the right by $Q_{[A,k]}$ changes only columns k and $k+1$ in a way that is effectively multiplying those columns as a submatrix by the submatrix explicitly stated in the definition.

When used in Algorithm 1 (line 8) $Q_{[H',r]}$ is an orthogonal matrix. The swapping of rows that occurs in the prior steps will usually cause H' to cease to be lower trapezoidal. The postmultiplication with $Q_{[H',r]}$ ensures that H' is once again lower trapezoidal [3, 6]. The only case where the row swap does not remove the lower trapezoidal property of H' is when $r = n - 1$ in which case $Q_{[H',r]}$ is the identity matrix and so H' is unaffected.

Finally, we use the following notation to refer to rows and columns of matrices, when needed.

Notation 2 (col_k, row_k) For a matrix M we denote by $\mathrm{col}_k(M)$ the kth column of M and by $\mathrm{row}_k(M)$ the kth row of M.

After each iteration the value $1/\max |H'_{r,r}|$ is a lower bound for the norm of *any* integer relation of x. Furthermore if a is the integer relation found by the algorithm, then $\|a\| \leq \gamma^{n-2}M$ where M is the norm of the smallest possible integer relation [6, Theorem 3].

Note that the algorithm as presented above does not terminate if there is no integer relation for the input x. This can be remedied either by specifying termination after a maximum number of iterations are performed, or after the lower bound for the norm of an integer relation exceeds some value.

The algorithm is exact if the individual steps can be performed exactly. That is to say, if we could compute with all real numbers exactly then the algorithm would always calculate an integer relation if there is one to be found. Furthermore, it will find an integer relation in a polynomially bounded number of iterations [3, 6]. In practice, however, an implementation of the PSLQ algorithm must use floating-point arithmetic and so numerical error may prevent the detection of a valid integer relation. Nonetheless, PSLQ has shown remarkable numerically stability.

Algorithm 1. PSLQ

```
input  : $x \in \mathbb{F}^n$, $\gamma > \gamma_1$
output: $a \in \mathcal{O}^n$
algocf
/* ———————— Initialisation ———————— */
$H' \leftarrow H_{x/\|x\|}$   $A \leftarrow I_n$
/* ———————— Main Calculation ———————— */
repeat
    $H' \leftarrow D_{H'}\, H'$                          /* Hermite reduce $H'$ */
    $A \leftarrow D_{H'}\, A$                            /* Update $A$ */
    $r \leftarrow \mathrm{argmax}_{1 \le r \le n-1}(\gamma^r |H'_{r,r}|)$   /* Find $r$ such that $\gamma^r |H'_{r,r}|$ is maximal */
    $\mathrm{row}_r(H') \leftrightarrow \mathrm{row}_{r+1}(H')$   /* Exchange rows $r$ and $r+1$ in $H'$ */
    $\mathrm{row}_r(A) \leftrightarrow \mathrm{row}_{r+1}(A)$     /* Exchange rows $r$ and $r+1$ in $A$ */
    $H' \leftarrow H'\, Q_{[H',r]}$                      /* Make sure $H'$ is lower trapezoidal */
until $r = n - 1$ and $H'_{n-1,n-1} = 0$
return $\mathrm{col}_n(A^{-1})$
```

Finally, we reiterate that the algorithm as presented here lacks the details needed for practical numeric application. There are many optimisations that can, and should, be implemented in order for an implementation to be effective. The interested reader should consult the literature [1, 3].

1.2 Algebraic Number Theory

We introduce only enough algebraic number theory as is needed. The reader is referred to the literature for a more thorough study [9, e.g.].

Definition 8 **(Algebraic Number)** A number $\alpha \in \mathbb{C}$ is an *algebraic number* (or simply *algebraic*) if it is a zero of a polynomial with rational coefficients.

Definition 9 **(Algebraic Integer)** A number $\alpha \in \mathbb{C}$ is an *algebraic integer* if it is a zero of a monic polynomial with integer coefficients. The ring of all algebraic integers is denoted by $\mathcal{A}$.

Definition 10 **(Algebraic Extension)** A field, $\mathbb{K} \supset \mathbb{Q}$, is an *algebraic extension field* (or simply an *algebraic extension*) if k is algebraic for all $k \in \mathbb{K}$.

We may now talk of the algebraic integers of a particular algebraic extension field.

Definition 11 Let $\mathbb{K}$ be an algebraic extension field. The *ring of integers of* $\mathbb{K}$, denoted $\mathcal{O}_{\mathbb{K}}$, is the intersection $\mathbb{K} \cap \mathcal{A}$ of the extension field with the ring of all algebraic integers.

For the purposes of this paper we consider only simple quadratic extension fields. That is, fields of the form $\mathbb{Q}[\sqrt{D}] := \{q_1 + q_2\sqrt{D} \mid q_1, q_2 \in \mathbb{Q}\}$. Without loss of generality we may assume $D \in \mathbb{Z}$ is square free. The ring of integers of such fields are known [9] to be $\mathcal{O}_{\mathbb{Q}[\sqrt{D}]} = \mathbb{Z}[\omega] = \{\alpha + \beta\,\omega \mid \alpha, \beta \in \mathbb{Z}\}$ where

$$\omega = \begin{cases} \sqrt{D} & \text{if } D \equiv 2, 3 \pmod 4 \\ (1+\sqrt{D})/2 & \text{if } D \equiv 1 \pmod 4 \end{cases}. \tag{4}$$

2 Extension to Algebraic Integers

In order to extend PSLQ to allow for algebraic integers, we first establish the relationship between algebraic integers, algebraic extension fields, and integer relations. We want to generalise, and thus wish to encapsulate the cases already handled by the existing theory.

A naïve strategy would be to replace $\mathbb{F}$ in Definition 1 with an arbitrary extension field, and to replace $\mathcal{O}$ with the ring of integers of that extension field. However, observe that the integers ($\mathbb{Z}$) are not the ring of integers of the field of real numbers. Similarly, the Gaussian integers ($\mathbb{Z}[\sqrt{-1}]$) are not the ring of integers of the field of complex numbers. So this strategy will not capture the pre-existing cases.

We instead generalise by introducing an intermediate extension field, according to the following definition.

Definition 12 (Algebraic Integer Relation) Let $x \in \mathbb{F}^n$ and $\mathbb{K} \subseteq \mathbb{F}$ be an algebraic extension field. An *algebraic integer relation* of x is a vector $a \in (\mathcal{O}_{\mathbb{K}})^n$, $a \neq 0$, such that $a_1x_1 + \cdots + a_nx_n = 0$.

Observe that algebraic integer relations are indeed a generalisation of integer relations. When $\mathbb{F} = \mathbb{R}$ and $\mathbb{K} = \mathbb{Q}$ (thinking of $\mathbb{Q}$ as a trivial extension field) then an algebraic integer relation is also an integer relation satisfying Definition 1. The same is true for the complex case when $\mathbb{F} = \mathbb{C}$ and $\mathbb{K} = \mathbb{Q}[\sqrt{-1}]$).

Since we have stated above that we are only concerning ourselves with simple quadratic extension fields, we correspondingly restrict our attention to algebraic integer relations where $\mathbb{K} = \mathbb{Q}[\sqrt{D}]$ is a quadratic extension field, and $\mathbb{F}$ is the Archimedean norm closure of $\mathbb{K}$ (i.e., $\mathbb{R}$ if $D \geq 0$ and $\mathbb{C}$ if $D < 0$).

2.1 Reduction

One approach to computing algebraic integer relations is to reduce the problem to an integer relation problem. We may then solve the problem with an existing integer relation finding algorithm, such as PSLQ.

Observe that for $\alpha + \beta\,\omega \in \mathcal{O}_{\mathbb{Q}[\sqrt{D}]}$ we have $(\alpha + \beta\,\omega)\ x = \alpha\,x + \beta\,(x\,\omega)$. This suggests a method of reduction.

Given an algebraic extension field $\mathbb{Q}[\sqrt{D}] \subset \mathbb{F}$, and input $(x_1, \ldots, x_n) \in \mathbb{F}^n$ we compute $(x_1, x_1\omega, \ldots, x_n, x_n\omega)$ which we give as input to PSLQ producing an integer relation $\left(a'_1, \ldots, a'_{2n}\right)$ from which we attempt to reconstruct an algebraic integer relation $(a_1, \ldots, a_n)$ where $a_k = a'_{2k-1} + a'_{2k}\omega$.

When $\mathbb{F} = \mathbb{R}$ it is straightforward to see that each $a_k \in \mathcal{O}_{\mathbb{Q}[\sqrt{D}]}$, and so the reconstructed relation is, indeed, an algebraic integer relation.

However, when $\mathbb{F} = \mathbb{C}$ the a'_k are Gaussian integers $\alpha_k + \beta_k\,i$ where $\alpha_k, \beta_k \in \mathbb{Z}$. Then

$$a_k = (\alpha_{2k-1} + \beta_{2k-1}\,i) + (\alpha_{2k} + \beta_{2k}\,i)\ \omega = (\alpha_{2k-1} + \alpha_{2k}\,\omega) + (\beta_{2k-1} + \beta_{2k}\,\omega)\,i$$

which will not always be an algebraic integer in $\mathcal{O}_{\mathbb{Q}[\sqrt{D}]}$.

Ideally, we want the a'_k to only ever be integer valued. In some cases, it may be possible to transform $\left(a'_1, \ldots, a'_n\right)$ into an equivalent (for the purposes of algebraic integer relation detection) integer-valued vector, such as dividing by a common Gaussian integer divisor. We have not yet found a reliable way to detect such cases in general.

2.2 *Algebraic PSLQ*

An alternative approach to computing algebraic integer relations is to modify the PSLQ algorithm to compute them directly. We call this modified algorithm *Algebraic* PSLQ, or APSLQ.

We observe that the reducing matrix is the source of integers in the algorithm. The reducing matrix, in turn, relies on the nearest integer function. The theorems bounding the number of iterations needed to find an integer relation rely only on the τ, ρ, and γ parameters, the latter of which is arbitrarily chosen and the others of which are determined by the properties of the integer lattice.

In order to utilise as much of the existing theory as possible, we replace the nearest integer function in the computation of the reducing matrix with a nearest algebraic integer function. Additionally, we require the specification of the intermediate quadratic extension field as input to the algorithm. The algorithm remains otherwise unmodified.

This immediately causes a problem. In the case of a real quadratic extension field (when $D > 0$) the algebraic integers are dense in $\mathbb{R}$. This leaves us without a well defined nearest integer, and hence no integer lattice. We put this case away pending further algorithmic modifications and restrict our attention to complex quadratic extension fields $D < 0$.

In order to calculate the nearest integer for an arbitrary $z \in \mathbb{C}$, we first rewrite $z = \alpha + \beta\,\omega$ and use α and β to compute $\lceil z \rfloor$. There are two cases.

When $D \equiv 2, 3 \pmod 4$, then $\alpha = \Re(z)$ and $\beta = \Im(z)/\sqrt{|D|}$. We have

$$\lceil z \rfloor = \lceil \alpha \rfloor_{\mathbb{Z}} + \lceil \beta \rfloor_{\mathbb{Z}}\, \omega$$

When $D \equiv 1 \pmod 4$, then $\beta = 2\Im(z)/\sqrt{|D|}$ and $\alpha = \Re(z) - \beta/2$. We have two candidates for the nearest integer and choose the one which is closest to z.

$$\lceil z \rfloor = \lceil \alpha \rfloor_{\mathbb{Z}} + \lfloor \beta \rfloor\, \omega \quad \text{or} \quad \lceil z \rfloor = \lceil \alpha + 1/2 \rfloor_{\mathbb{Z}} + \lceil \beta \rceil\, \omega$$

We bound $|z - \lceil z \rfloor| \leq \epsilon$ for all $z \in \mathbb{C}$ using the geometric properties of the lattices.

$$\epsilon = \begin{cases} \frac{1}{2}\sqrt{|D| + 1} & \text{if } D \equiv 2, 3 \pmod 4 \\ \frac{1}{4}\frac{|D|+1}{\sqrt{|D|}} & \text{if } D \equiv 1 \pmod 4 \end{cases}$$

And so we can compute the corresponding value of ρ

$$\rho = \begin{cases} \frac{2}{\sqrt{|D|+1}} & \text{if } D \equiv 2, 3 \pmod 4 \\ \frac{4\sqrt{|D|}}{|D|+1} & \text{if } D \equiv 1 \pmod 4. \end{cases}$$

However, as $|D|$ increases, the value of ρ decreases, and eventually $\rho < 1$ making it impossible to satisfy condition (2), and causing γ_1 to become complex. This leaves us with $D = -2$, $D = -3$, $D = -7$, and $D = -11$ as the only values of D for which the existing theory holds.

We will see that even when the conditions do not hold the algorithm can still be effective (see Section 3.4, Table 5).

In this paper, we examine the efficacy APSLQ and the reduction method. We leave, for now, the question of additional modifications which may handle the problems described above.

3 Experimental Results

We tested the efficacy of the above two methods experimentally. To do so we used *Maple*'s native PSLQ implementation for reduction, and our own implementation of APSLQ (written in *Maple*). The code and results are available in a GitHub repository [8].

Our implementation of APSLQ is described in Algorithm 2. Recall that for APSLQ the matrices $D_{H'}$ are constructed using an algebraic nearest integer function.

The particulars are a little different from the algorithm presented in Section 1.1 (Algorithm 1). It is effectively the algorithm as described by Borwein [3, fig. B.5], although we note that our implementation correctly handles the complex case as described above, whereas the algorithm given by Borwein is specialised to the real case.

Algorithm 2. APSLQ

```
input  : $x \in \mathbb{F}^n, D \in \mathbb{Z}, \gamma \geq 0, \epsilon > 0, max_i > 0$
output: A vector in $\mathcal{O}_{\mathbb{K}}^n$ (where $\mathbb{K} = \mathbb{Q}[\sqrt{D}]$) or FAIL
algocf
/* ———————— Initialisation ———————— */
$y \leftarrow x/\|X\|$                                          /* Normalise input vector */
$H' \leftarrow D_{H_y}\, H_y \quad B \leftarrow D_{H_y}^{-1} \quad y \leftarrow y\, D_{H_y}^{-1}$      /* Initial Hermite reduction */
$i \leftarrow 0$                                                /* Loop counter */
/* ———————— Main Calculation ———————— */
repeat
    $r \leftarrow \text{argmax}_{1 \leq r \leq n-1} \left(\gamma^r |H'_{r,r}|\right)$        /* Find $r$ s.t. $\gamma^r |H'_{r,r}|$ is maximal */
    $\text{row}_r(H') \leftrightarrow \text{row}_{r+1}(H')$                /* Swap rows $r$ and $r+1$ in $H'$ */
    $\text{col}_r(B) \leftrightarrow \text{col}_{r+1}(B)$                  /* Swap columns $r$ and $r+1$ in $B$ */
    $y_r \leftrightarrow y_{r+1}$                                   /* Swap elements $r$ and $r+1$ in $y$ */
    $H' \leftarrow H'\, Q_{[H',r]}$                              /* Make sure $H'$ is lower trapezoidal */
    $H' \leftarrow D_{H'}\, H'$                                  /* Hermite reduce $H'$ */
    $B \leftarrow B\, D_{H'}^{-1} \quad y \leftarrow y\, D_{H'}^{-1}$                /* Update $B$ and $y$ */
    $k \leftarrow \text{argmin}_{1 \leq k \leq n}(|y_k|)$                   /* Find $k$ s.t. $|y_k|$ is minimal */
    $i \leftarrow (i+1)$                                        /* Increment loop counter */
until $y_k/\|\text{col}_k(B)\| < \epsilon$ or $i > max_i$
if $y_k/\|\text{col}_k(B)\| < \epsilon$ then return $\text{col}_k(B)$ else return FAIL
```

To understand the differences, first note that the matrix B is simply the matrix A^{-1} from Algorithm 1. Each column of B is considered a possible integer relation of x, and the vector y is kept updated so that $y = (x/\|x\|)\, B$. As such, if $y_k = 0$ for some k, then $\text{col}_k(B)$ must be an integer relation for $x/\|x\|$ and thus also for x. We terminate if we find such a relation, or if we exceed a specified number of iterations. This relation, a say, will not necessarily have the properly $\|a\| \leq \gamma^{n-2} M$ that is guaranteed for a relation given by Algorithm 1, however.

Note that because we are performing numeric (floating point) computations we are unlikely to exactly compute a 0 element in y. To detect termination, therefore, we consider only the smallest $|y_k|$ as the best candidate for a linear combination, and look to see if it is sufficiently close to 0 (i.e., less than some threshold ϵ). We scale the value of $|y_k|$ by $\|\text{col}_k(B)\|$ in order to avoid missing a possible relation if the norm the column of B is particularly large. For more details, the reader should refer to Borwein [3, appendix 1].

3.1 Methodology

We created collections of instances of algebraic integer problems. Each collection, referred to as a *test set*, consisted of 1000 algebraic integer relation problems.

For each test set we chose a quadratic extension field $\mathbb{K}$, a set of constants from which we created each of the individual problems within the set, and a size for the coefficients of any algebraic integers used as part of the individual problem creation.

We will speak of the choice of extension field in more detail when we describe the results, below.

Two sets of constants were used: one containing real constants, the other complex. The real set was

$$\left\{\pi^k : k \in \mathbb{N}, k \leq 9\right\} \cup \left\{e^k : k \in \mathbb{N}, k \leq 9\right\} \cup \left\{\gamma^k : k \in \mathbb{N}, k \leq 9\right\} \\ \cup \{\sin k : k \in \mathbb{N}, k \leq 9\} \cup \{\log 2, \log 3, \log 5, \log 7\}\,.$$

The complex set was generated by randomly choosing an integer modulus between 1 and 9 for each integer argument from -9 to 9.

$$\left\{5\,e^{-9\,i}, 4\,e^{-8\,i}, 9\,e^{-7\,i}, 5\,e^{-6i}, 2\,e^{-5\,i}, 9\,e^{-4\,i}, 8\,e^{-3\,i}, 3\,e^{-2\,i}, 2\,e^{-i}, \\ 4, 4\,e^{i}, 5\,e^{2\,i}, 2\,e^{3\,i}, 7\,e^{4\,i}, 6\,e^{5\,i}, 3\,e^{6\,i}, 3\,e^{7\,i}, 5\,e^{8\,i}, 5\,e^{9\,i}\right\}.$$

Each constant set was used in multiple test sets.

The size of the coefficients of the algebraic integers fell into two cases: *small* (coefficients in the range $[-9, 9]$ thus having exactly 1 decimal digit) and *large* (coefficients in the range $[-999999, 999999]$ thus having up to 6 decimal digits).

Once the above choices were made for a particular test set, the problems within that set were randomly generated as follows:

1. Randomly choose an integer $2 \leq k \leq 10$.
2. Randomly choose k constants, $C_1, \ldots, C_k$, from the set of constants for the test set.
3. For each C_i, randomly choose integers α_i and β_i within the specified size. Let $z_i = \alpha_i + \beta_i\,\omega$.
4. Let $C_0 = \sum_{i=1}^{k} z_i C_i$.

The problem instance was the input vector $x = (C_0, C_1, \ldots, C_k)$ which, by construction, had algebraic integer relation $\mathfrak{a} = (-1, z_1, \ldots, z_k)$.

For each test set, we attempted to solve the problems within it using PSLQ, reduction, and/or APSLQ as appropriate. Our aim was to see if the algorithm could recover the known algebraic integer relation from the input vector. Any algebraic integer multiple of the known relation was considered to be an equivalent relation for this purpose.

Test sets that used small coefficients for algebraic integers were tested using 75 decimal digits of floating-point precision. Test sets that used large coefficients were tested using 175 decimal digits of floating-point precision.

The result of a computation on an individual test instance was classified as outlined in Table 1. We simply counted the number of occurrences of each result.

No UNEXPECTED results were found during our testing. This classification was originally introduced in the testing of an early implementation as a result of an over-

Table 1 Result Classifications

GOOD	The generated algebraic integer relation was recovered.
UNEXPECTED	A different, correct algebraic integer relation was found.
BAD	An incorrect algebraic integer relation was found.
FAIL	The algorithm produced no result.

sight in which log 2, log 3 and log 6 were together in some problems. This oversight has since been corrected, yet it remains possible (although unlikely) that other unexpected relations may still be computed, so we keep the classification as a possibility.

To assess each result classification, we first note that a FAIL condition is immediate if no result is produced (usually because the maximum number of iterations was exceeded). Assuming this is not the case, let $a = (a_1, \dots, a_n)$ be the computed algebraic integer relation. Let $\mathfrak{a} = (-1, z_1, \dots, z_k)$ be the known relation from above. Recall that we are considering any algebraic multiple of $\mathfrak{a}$ to be correct and observe that if $a = \lambda\mathfrak{a}$ then it must be the case that $\lambda = -a_1$. We therefore look to see if $(-a_1)\mathfrak{a} = a$, and if so we diagnose a GOOD result. If that is not the case, we then test the computed algebraic integer relation to 1000 decimal digits of precision, and if the result is within 10^{-998} of 0 we diagnose an UNEXPECTED result. If none of the above apply, then we diagnose a BAD result.

Observe that the problem with the reconstructed relation for the reduction method in the complex case as described in Section 2.1 is not addressed at all by this diagnosis method. It is entirely possible that $(-a_1)\mathfrak{a} = a$ even if a_1 is not a valid algebraic integer for the extension field in question. We describe how we accounted for this below (see Section 3.4).

When testing sets appropriate for APSLQ we performed each computation multiple times with different values of γ and different thresholds for detecting integer relations in A^{-1} (as described above). Specifically, we used $\gamma = \gamma_1$, $\gamma = 2.0$, and $\gamma = 3.0$. Note that although the strict conditions from Section 1.1 require $\gamma > \gamma_1$ the choice of $\gamma = \gamma_1$ seems to be common in practice, and the results below do not seem to suffer.

The thresholds used were $10^{-(d-1)}$, $10^{-(d-4)}$, and $10^{-(d-\log_{10} n)}$ where d is the floating-point precision in decimal digits, and n is the number of elements in the input vector. Note that the latter of these, copied from *Maple*'s implementation, varies slightly with the number of elements of the input vector. These different thresholds made almost no difference whatsoever. For the cases where there is no γ_1 (see Table 5) the latter threshold sometimes had one fewer GOOD and one more FAIL result when compared to the other thresholds. We do not consider this significant and report the results for the first threshold ($10^{-(d-1)}$) only.

The test sets fell into three broad categories, described separately in the subsections that follow.

Table 2 Direct comparison of PSLQ and APSLQ

	Small Coefficients				Large Coefficients			
Field	PSLQ	Algebraic PSLQ			PSLQ	Algebraic PSLQ		
		$\gamma = \gamma_1$	$\gamma = 2.0$	$\gamma = 3.0$		$\gamma = \gamma_1$	$\gamma = 2.0$	$\gamma = 3.0$
Real C_i								
$\mathbb{Q}$	1000G	1000G	1000G	1000G	1000G	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-1}]$	1000G	1000G	1000G	1000G	1000G	1000G	1000G	1000G
Complex C_i								
$\mathbb{Q}[\sqrt{-1}]$	1000G	1000G	1000G	1000G	1000G	1000G	1000G	1000G

3.2 Real and Complex PSLQ

We tested our implementation of APSLQ against *Maple*'s PSLQ implementation for the cases that PSLQ was already known to work for. That is for the trivial case $\mathbb{K} = \mathbb{Q}$ and the case $\mathbb{K} = \mathbb{Q}[\sqrt{-1}]$. This testing acted as a sanity check that our implementation was correct in the known cases.

The results are tabulated in Table 2. Note that it is impossible to create test sets that use complex C_i and $\mathbb{K} = \mathbb{Q}$, so we were only able to test a single field with complex constants.

3.3 Real Quadratic Extension Fields

For the real quadratic algebraic integer relations we tested the following real quadratic extension fields:

$$\mathbb{K} = \mathbb{Q}[\sqrt{D}] \quad \text{for} \quad D \in \{2, 3, 5, 6, 7, 10, 11\}.$$

Recall that APSLQ is not appropriate for these extension fields, so only the reduction method was tested.

The results are tabulated in Table 3. We note that since we are testing real quadratic extension fields we are in the case where $\mathbb{F} = \mathbb{R}$ and so, as stated in Section 2.1, we definitely have found algebraic integer relations. Contrast this to the complex quadratic extension field testing, below.

3.4 Complex Quadratic Extension Fields

For the real quadratic algebraic integer relations we were able to test both the reduction method, and APSLQ. We tested the following complex quadratic extension fields:

Table 3 Real quadratic fields, Real C_i

Field	Small Coefficients		Large Coefficients	
	Reduction	APSLQ	Reduction	APSLQ
$\mathbb{Q}[\sqrt{2}]$	1000G	N/A	1000G	N/A
$\mathbb{Q}[\sqrt{3}]$	1000G	N/A	1000G	N/A
$\mathbb{Q}[\sqrt{5}]$	1000G	N/A	1000G	N/A
$\mathbb{Q}[\sqrt{6}]$	1000G	N/A	1000G	N/A
$\mathbb{Q}[\sqrt{7}]$	1000G	N/A	1000G	N/A
$\mathbb{Q}[\sqrt{10}]$	1000G	N/A	1000G	N/A
$\mathbb{Q}[\sqrt{11}]$	1000G	N/A	1000G	N/A

Table 4 Complex quadratic fields with γ_1

	Small Coefficients				Large Coefficients			
Field	Reduction	Algebraic PSLQ			Reduction	Algebraic PSLQ		
		$\gamma = \gamma_1$	$\gamma = 2.0$	$\gamma = 3.0$		$\gamma = \gamma_1$	$\gamma = 2.0$	$\gamma = 3.0$
Real C_i								
$\mathbb{Q}[\sqrt{-2}]$	912G88F	1000G	1000G	1000G	952G48F	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-3}]$	919G81F	1000G	1000G	1000G	923G77F	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-7}]$	956G44F	1000G	1000G	1000G	949G51F	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-11}]$	975G25F	1000G	1000G	1000G	981G19F	1000G	1000G	1000G
Complex C_i								
$\mathbb{Q}[\sqrt{-2}]$	911G1B88F	1000G	1000G	1000G	957G43F	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-3}]$	904G96F	1000G	1000G	1000G	924G76F	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-7}]$	939G61F	1000G	1000G	1000G	961G39F	1000G	1000G	1000G
$\mathbb{Q}[\sqrt{-11}]$	979G21F	1000G	999G1F	1000G	975G2B23F	1000G	995G5F	1000G

$$\mathbb{K} = \mathbb{Q}[\sqrt{D}] \quad \text{for} \quad D \in \{-2, -3, -5, -6, -7, -10, -11\}.$$

As we have tested both reduction and APSLQ for these fields, we may compare the relative efficacy of the two methods.

We accounted for the reduction problem described in Section 2.1 by checking to see if the entries in the recovered relation consisted only of valid algebraic integers from the appropriate field. This check was performed after the usual diagnosis, so that we could compare these FAIL results with the originally diagnosed result. If any entries were not appropriate algebraic integers then we changed the diagnosed result to a FAIL and also recorded the old result. We note that all such FAIL results reported for our reduction tests were initially GOOD results.

The cases where $D \in \{-2, -3, -7, -11\}$ are cases where γ_1 exists and so the three conditions (1), (2), and (3) from Section 1.1 are satisfied. These results are summarised in Table 4. Both reduction and APSLQ perform superbly for these cases.

Table 5 Complex quadratic fields without γ_1

	Small Coefficients			Large Coefficients		
Field	Reduction	Algebraic PSLQ		Reduction	Algebraic PSLQ	
		$\gamma = 2.0$	$\gamma = 3.0$		$\gamma = 2.0$	$\gamma = 3.0$
Real C_i						
$\mathbb{Q}[\sqrt{-5}]$	994G6F	997G3F	1000G	1000G	983G2B15F	992G2B6F
$\mathbb{Q}[\sqrt{-6}]$	996G4F	997G3F	999G1F	998G2F	986G1B13F	996G1B3F
$\mathbb{Q}[\sqrt{-10}]$	1000G	999G1F	999G1F	1000G	993G7F	994G6F
Complex C_i						
$\mathbb{Q}[\sqrt{-5}]$	995G5F	158G842F	187G813F	997G3F	164G836F	182G818F
$\mathbb{Q}[\sqrt{-6}]$	999G1F	136G864F	143G857F	1000G	59G941F	60G940F
$\mathbb{Q}[\sqrt{-10}]$	1000G	40G960F	42G958F	1000G	1000F	1000F

Observe that when testing the field $\mathbb{Q}[-11]$ with complex C_i and $\gamma = 2.0$ the results were slightly worse than when $\gamma = \gamma_1$. This is likely because for this field $\gamma_1 = \sqrt{22}/2 > 2$, so $\gamma = 2.0$ is too small to satisfy the required constraints in Section 1.1. This supposition is strengthened by the observation that when $\gamma = 3.0 > \sqrt{22}/2$ the results are good again.

We note a couple of BAD results for the reduction method with complex C_i. In none of these cases did APSLQ produce anything but a GOOD result (if we ignore the case described in the previous paragraph). Nonetheless, one or two bad results out of a pool of one thousand is hardly a poor result.

The cases where $D \in \{-5, -6, -10\}$ are cases where γ_1 does not exist and so the three conditions (1), (2), and (3) from Section 1.1 are not satisfied. These results are summarised in Table 5.

Observe that for real C_i the results are mostly good, despite the algorithm conditions not being satisfied. This is similar to the results for $\mathbb{Q}[\sqrt{-11}]$, highlighted above, that also failed those conditions. Contrast these to the cases with complex C_i.

The cases with complex C_i perform exceptionally poorly for APSLQ. This ought not be especially surprising since these fields do not satisfy the required conditions. It is perhaps more remarkable that the results for the real C_i case are so good. However, the reduction method gives consistently good results. If we can find a way to reliably find correct algebraic integer relations from the incorrect ones often given by this method, it should prove to be remarkably robust.

4 Further Work

Further tests are being run which look more closely at the relationship between integer coefficient size, input vector size, and the precision necessary to find an integer relation. These tests also examine how the algorithm performs with problems consisting of extra constants than those that are known to be in the integer relation (i.e., relations with constants whose coefficient will be 0 in the integer relation).

We suspect, based on some early proof-of-concept tests performed while implementing APSLQ, that the reduction method will require more precision than APSLQ for the same problem instance. The above further tests should quantify that if it is correct.

Work is ongoing to find a theoretical framework with which to further modify the APSLQ algorithm so that we may handle the real quadratic integer case, and the complex quadratic integer cases that do not satisfy the requirements from Section 1.1.

Work is also ongoing to ascertain a method of reliably extracting algebraic integers in the complex quadratic reduction case.

References

1. Bailey, D.H., Broadhurst, D.J.: Parallel integer relation detection: techniques and applications. Math. Comput. **70**(236), 1719–1736 (2001). https://doi.org/10.1090/S0025-5718-00-01278-3
2. Bailey, D.H., Borwein, P., Plouffe, S.: On the rapid computation of various polylogarithmic constants. Math. Comput. **66**(218), 903–913 (1997). https://doi.org/10.1090/S0025-5718-97-00856-9
3. Borwein, P.: Computational Excursions in Analysis and Number Theory. CMS Books in Mathematics. Springer, New York (2002). https://doi.org/10.1007/978-0-387-21652-2
4. Ferguson, H.R.P., Bailey, D.H.: A polynomial time, numerically stable integer relation algorithm. Technical report RNR-91-032, NAS Applied Research Branch, NASA Ames Research Center (1991)
5. Ferguson, H.R.P., Forcade, R.W.: Generalization of the Euclidean algorithm for real numbers to all dimensions higher than two. Bull. Am. Math. Soc. (N.S.) **1**(6), 912–914 (1979)
6. Ferguson, H.R.P., Bailey, D.H., Arno, S.: Analysis of PSLQ, an integer relation finding algorithm. Math. Comput. **68**(225), 351–369 (1999). https://doi.org/10.1090/S0025-5718-99-00995-3
7. Heath, T.L.: The Thirteen Books of Euclid's Elements, vol. 3. Cambridge University Press, Cambridge (1908)
8. Skerritt, M.P., Vrbik, P.: Algebraic-PSLQ: testing and results (Version JBCC) (2019). https://doi.org/10.5281/zenodo.3346895
9. Stewart, I., Tall, D.: Algebraic Number Theory and Fermat's Last Theorem. AK Peters Series, 3rd edn. Taylor & Francis, New York (2001)
10. Straub, A.: A gentle introduction to PSLQ (2010). http://arminstraub.com/math/pslq-intro

Short Walk Adventures

Armin Straub and Wadim Zudilin

To the memory of Jon Borwein, who convinced us that a short walk can be adventurous

1 Introduction

At some stages of our careers, we were approached by Jon Borwein to collaborate on a theme that sounded rather off-topic to us, who had interests in number theory, combinatorics and related special functions. Somewhat unexpectedly, the theme has become a remarkable research project with several outcomes (including [9–11], to list a few), a project which we continue to enjoy after the sudden loss of Jon… This note serves as a summary to our recent discoveries that certain 'probabilistic'

The original version of this chapter was revised: The cross-citations in square brackets has now been updated. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-36568-4_28

A. Straub
Department of Mathematics and Statistics, University of South Alabama,
411 University Blvd N, MSPB 325, Mobile, AL 36688, USA
e-mail: straub@southalabama.edu

W. Zudilin (✉)
Department of Mathematics, IMAPP, Radboud University,
PO Box 9010, 6500 GL Nijmegen, Netherlands
e-mail: w.zudilin@math.ru.nl; wadim.zudilin@newcastle.edu.au

School of Mathematical and Physical Sciences, The University of Newcastle,
Callaghan, NSW 2308, Australia

Laboratory of Mirror Symmetry and Automorphic Forms,
National Research University Higher School of Economics,
6 Usacheva str., 119048 Moscow, Russia

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_27

techniques apply usefully to tackling difficult problems on the border of analysis, number theory and differential equations; in particular, in evaluating multivariable Mahler measures. Our principal novelties are given in Theorems 1–3; these include hypergeometric reduction of the Mahler measures of the three-variable polynomials

$$1 + x_1 + x_2 + x_3 + x_2x_3 \quad \text{and} \quad (1 + x_1)^2 + x_2 + x_3,$$

as well as the (hypergeometric) factorisation of a related differential operator for the Apéry-like sequence

$$\sum_{k=0}^{n} \binom{n}{k}^2 \binom{2k}{k}^2, \quad \text{where } n = 0, 1, 2, \dots .$$

Echoing Jon's 'a short walk can be beautiful' [8], we add that 'a short walk can be adventurous.'

2 Uniform Random Walks

An N-step uniform random walk is a planar walk that starts at the origin and consists of N steps of length 1 each taken into a uniformly random direction. Let X_N be the distance to the origin after these N steps. The s-th moments $W_N(s)$ of X_N can be computed [10] via the formula

$$\begin{aligned} W_N(s) &= \int \cdots \int_{[0,1]^N} |e^{2\pi i\theta_1} + \cdots + e^{2\pi i\theta_N}|^s \, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_N \\ &= \int \cdots \int_{[0,1]^{N-1}} |1 + e^{2\pi i\theta_1} + \cdots + e^{2\pi i\theta_{N-1}}|^s \, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{N-1} \end{aligned}$$

and are related to the (probability) density function $p_N(x)$ of X_N via

$$W_N(s) = \int_0^\infty x^s p_N(x) \, \mathrm{d}x = \int_0^N x^s p_N(x) \, \mathrm{d}x.$$

That is, $p_N(x)$ can then be obtained as the inverse Mellin transform of $W_N(s-1)$. Finally, note that the even moments $W_3(2n)$ and $W_4(2n)$ (which are, clearly, positive integers) can be identified with the odd moments of $I_0(t)K_0(t)^2$ and $I_0(t)K_0(t)^3$, respectively, where $I_0(t)$ and $K_0(t)$ denote the modified Bessel functions of the first and second kind. Namely, for $n = 1, 2, \dots$ we have [6]

$$W_3(2n) = \frac{3^{2n+3/2}}{\pi \, 2^{2n} \, n!^2} \int_0^\infty t^{2n+1} I_0(t) K_0(t)^2 \, \mathrm{d}t$$

and

$$W_4(2n) = \frac{4^{2n+2}}{\pi^2\, n!^2} \int_0^\infty t^{2n+1} I_0(t) K_0(t)^3 \,\mathrm{d}t.$$

3 Zeta Mahler Measures

For a non-zero Laurent polynomial $P(x_1, \ldots, x_N) \in \mathbb{C}[x_1^{\pm1}, \ldots, x_N^{\pm1}]$, its zeta Mahler measure [3] is defined by

$$Z(P; s) = \int \cdots \int_{[0,1]^N} |P(e^{2\pi i\theta_1}, \ldots, e^{2\pi i\theta_N})|^s \,\mathrm{d}\theta_1 \cdots \mathrm{d}\theta_N,$$

and its logarithmic Mahler measure is

$$\mathrm{m}(P) = \left.\frac{\mathrm{d}Z(P; s)}{\mathrm{d}s}\right|_{s=0} = \int \cdots \int_{[0,1]^N} \log|P(e^{2\pi i\theta_1}, \ldots, e^{2\pi i\theta_N})| \,\mathrm{d}\theta_1 \cdots \mathrm{d}\theta_N.$$

A straightforward comparison of the two definitions reveals that

$$W_N(s) = Z(x_1 + \cdots + x_N; s) = Z(1 + x_1 + \cdots + x_{N-1}; s)$$

and

$$W_N'(0) = \mathrm{m}(x_1 + \cdots + x_N) = \mathrm{m}(1 + x_1 + \cdots + x_{N-1}) = \int_0^N p_N(x) \log x \,\mathrm{d}x, \tag{1}$$

where the derivative is with respect to s. The latter Mahler measures are known as linear Mahler measures. The evaluations $W_2'(0) = 0$,

$$W_3'(0) = L'(\chi_{-3}; -1) = \frac{3\sqrt{3}}{4\pi} L(\chi_{-3}; 2), \quad W_4'(0) = -14\zeta'(-2) = \frac{7\zeta(3)}{2\pi^2}$$

are known [24], while the following conjectural evaluations, due to Rodriguez-Villegas [13] and verified to several hundred digits [5], remain open:

$$W_5'(0) \overset{?}{=} -L'(f_3; -1) = 6\left(\frac{\sqrt{15}}{2\pi}\right)^5 L(f_3; 4),$$
$$W_6'(0) \overset{?}{=} -8L'(f_4; -1) = 3\left(\frac{\sqrt{6}}{\pi}\right)^6 L(f_4; 5),$$

where

$$f_3(\tau) = \eta(\tau)^3\eta(15\tau)^3 + \eta(3\tau)^3\eta(5\tau)^3 \quad \text{and} \quad f_4(\tau) = \eta(\tau)^2\eta(2\tau)^2\eta(3\tau)^2\eta(6\tau)^2$$

are cusp eigenforms of weight 3 and 4, respectively. Here and in what follows, Dedekind's eta function

$$\eta(\tau) = q^{1/24}\prod_{m=1}^{\infty}(1-q^m) = \sum_{n=-\infty}^{\infty}(-1)^n q^{(6n+1)^2/24}, \quad \text{where } q = e^{2\pi i\tau},$$

serves as a principal constructor of modular forms and functions. No similar formulae are known for $W_N'(0)$ when $N \geq 7$, though the story continues at a different level—see [14, 30, 31] for details.

4 Generic Two-Step Random Walks

Let X_1 and X_2 be two (sufficiently nice, independent) random variables on $[0,\infty)$ with probability density $p_1(x)$ and $p_2(x)$, respectively, and let θ_1 and θ_2 be uniformly distributed on $[0, 1]$. Then $X = e^{2\pi i\theta_1}X_1 + e^{2\pi i\theta_2}X_2$ describes a two-step random walk in the plane with a first step of length X_1 and a second step of length X_2. As in [11, eq. (3-3)], an application of the cosine rule shows that the s-th moment of $|X|$ is

$$W(s) = \mathsf{E}(|X|^s) = \int_0^\infty\int_0^\infty g_s(x,y)p_1(x)p_2(y)\,\mathrm{d}x\,\mathrm{d}y,$$

where

$$g_s(x,y) = \frac{1}{\pi}\int_0^\pi (x^2+y^2+2xy\cos\theta)^{s/2}\,\mathrm{d}\theta.$$

Observe that

$$\left.\frac{\mathrm{d}g_s(x,y)}{\mathrm{d}s}\right|_{s=0} = \frac{1}{\pi}\int_0^\pi \log\sqrt{x^2+y^2+2xy\cos\theta}\,\mathrm{d}\theta = \max\{\log|x|, \log|y|\},$$

so that, in particular,

Lemma 1 *We have*

$$W'(0) = \mathsf{E}(\log|X|) = \int_0^\infty\int_0^\infty p_1(x)p_2(y)\max\{\log x, \log y\}\,\mathrm{d}y\,\mathrm{d}x.$$

Alternative equivalent expressions, that will be useful in what follows, include

$$\begin{aligned}\mathsf{E}(\log|X|) &= \int_0^\infty\int_0^x p_1(x)p_2(y)\log x\,\mathrm{d}y\,\mathrm{d}x + \int_0^\infty\int_x^\infty p_1(x)p_2(y)\log y\,\mathrm{d}y\,\mathrm{d}x\\ &= \mathsf{E}(\log X_1) + \int_0^\infty\int_x^\infty p_1(x)p_2(y)(\log y - \log x)\,\mathrm{d}y\,\mathrm{d}x\\ &= \mathsf{E}(\log X_2) + \int_0^\infty\int_0^x p_1(x)p_2(y)(\log x - \log y)\,\mathrm{d}y\,\mathrm{d}x. \qquad (2)\end{aligned}$$

5 Linear Mahler Measures

Let N, M be integers such that $N > M > 0$. By decomposing an N-step random walk into two walks with $N - M$ and M steps, and applying Lemma 1 in the form (2), we find that

$$W'_N(0) = W'_M(0) + \int_0^{N-M} p_{N-M}(x)\left(\int_0^x p_M(y)(\log x - \log y)\,\mathrm{d}y\right)\mathrm{d}x.$$

This formula, together with known formulae for the densities [10], like $p_1(x) = \delta(x-1)$ (the Dirac delta function) and $p_2(x) = 2/(\pi\sqrt{4-x^2})$ for $0 < x < 2$, allows one to produce new expressions for linear Mahler measures. Indeed, taking $M = 1$ we get

$$W'_N(0) = \int_1^{N-1} p_{N-1}(x)\log x\,\mathrm{d}x \tag{3}$$

(which can be also derived using Jensen's formula), while $M = 2$ results in

$$W'_N(0) = \int_2^{N-2} p_{N-2}(x)\log x\,\mathrm{d}x + \frac{1}{\pi}\int_0^2 p_{N-2}(x)x \cdot {}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12 \\ \frac32, \frac32\end{matrix}\,\middle|\,\frac{x^2}{4}\right)\mathrm{d}x \tag{4}$$

(see also [20, eq. (2.1)]). Here, and in what follows, the hypergeometric notation

$${}_mF_{m-1}\left(\begin{matrix}a_1, a_2, \ldots, a_m \\ b_2, \ldots, b_m\end{matrix}\,\middle|\,z\right) = \sum_{n=0}^{\infty} \frac{(a_1)_n(a_2)_n\cdots(a_m)_n}{(b_2)_n\cdots(b_m)_n}\frac{z^n}{n!}$$

is used, where

$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)} = \begin{cases} a(a+1)\cdots(a+n-1), & \text{for } n \geq 1, \\ 1, & \text{for } n = 0, \end{cases}$$

denotes the Pochhammer symbol (the rising factorial). Note that we deduce (4) from

$$\int_0^x p_2(y)(\log x - \log y)\,\mathrm{d}y = \frac{x}{\pi} \cdot {}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12 \\ \frac32, \frac32\end{matrix}\,\middle|\,\frac{x^2}{4}\right),$$

which is valid if $0 \leq x \leq 2$.

Equations (3) and (4) and the formula

$$p_4(x) = \frac{2\sqrt{16-x^2}}{\pi^2 x}\,\mathrm{Re}\,{}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12 \\ \frac56, \frac76\end{matrix}\,\middle|\,\frac{(16-x^2)^3}{108x^4}\right)$$

obtained in [10, Theorem 4.9], provide the formulae

$$W_5'(0) = \frac{7\zeta(3)}{2\pi^2} - \frac{1}{\pi^2}\int_0^1 \sqrt{16-x^2}\,\mathrm{Re}\,{}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12\\ \frac56, \frac76\end{matrix}\,\middle|\, \frac{(16-x^2)^3}{108x^4}\right)\mathrm{d}(\log^2 x)$$

and

$$\begin{aligned}W_6'(0) &= \frac{7\zeta(3)}{2\pi^2} - \frac{1}{\pi^2}\int_0^2 \sqrt{16-x^2}\,\mathrm{Re}\,{}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12\\ \frac56, \frac76\end{matrix}\,\middle|\, \frac{(16-x^2)^3}{108x^4}\right)\mathrm{d}(\log^2 x)\\ &+ \frac{2}{\pi^3}\int_0^2 \sqrt{16-x^2}\,\mathrm{Re}\,{}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12\\ \frac56, \frac76\end{matrix}\,\middle|\, \frac{(16-x^2)^3}{108x^4}\right)\cdot {}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12\\ \frac32, \frac32\end{matrix}\,\middle|\, \frac{x^2}{4}\right)\mathrm{d}x.\end{aligned}$$

These *single* integrals can be used to numerically confirm the conjectural evaluations of $W_5'(0)$ and $W_6'(0)$.

A similar application of Lemma 1, upon decomposing a 6-step walk into two walks with 3 steps, yields the alternative reduction

$$W_6'(0) = 2\int_0^3 p_3(x)\log x\left(\int_0^x p_3(y)\,\mathrm{d}y\right)\mathrm{d}x, \tag{5}$$

where [10]

$$p_3(x) = \frac{2\sqrt{3}x}{\pi(3+x^2)}\cdot {}_2F_1\left(\begin{matrix}\frac13, \frac23\\ 1\end{matrix}\,\middle|\, \frac{x^2(9-x^2)^2}{(3+x^2)^3}\right).$$

We discuss this formula further in Section 6.

Finally, we mention that equation (3) and a modular parametrisation of $p_4(x)$ (which we indicate in Section 7) were independently cast in [23] to produce a double L-value expression for $W_5'(0)$.

6 Modular Parametrisation of $p_3(x)$ and Related Formulae

Note that formula (5) can be written as

$$W_6'(0) = \int_0^3 \log x\,\mathrm{d}(P_3(x)^2) = \log 3 - \int_0^3 P_3(x)^2\,\frac{\mathrm{d}x}{x},$$

featuring the cumulative density function

$$P_3(x) = \int_0^x p_3(y)\,\mathrm{d}y.$$

The related modular parametrisation of $p_3(x)$ is given by

$$x = x(\tau) = 3\frac{\eta(\tau)^2\eta(6\tau)^4}{\eta(2\tau)^4\eta(3\tau)^2} : (i\infty, 0) \to (0, 3),$$

so that

$$p_3(x) = \frac{2\sqrt{3}}{\pi}\,\frac{\eta(2\tau)^2\eta(6\tau)^2}{\eta(\tau)\eta(3\tau)}, \quad \mathrm{d}x = 3\pi i\,\frac{\eta(\tau)^6\eta(3\tau)^2\eta(6\tau)^2}{\eta(2\tau)^6}\,\mathrm{d}\tau$$

and

$$P_3(x) = 6i\sqrt{3}\int_{i\infty}^{\tau}\frac{\eta(\tau)^5\eta(3\tau)\eta(6\tau)^4}{\eta(2\tau)^4}\,\mathrm{d}\tau$$

is the anti-derivative of a weight 3 holomorphic Eisenstein series

$$\frac{\eta(\tau)^5\eta(3\tau)\eta(6\tau)^4}{\eta(2\tau)^4} = E_{3,\chi_{-3}}(\tau) - 8E_{3,\chi_{-3}}(2\tau),$$

where

$$\begin{aligned} E_{3,\chi_{-3}}(\tau) &= \frac{\eta(3\tau)^9}{\eta(\tau)^3} = \sum_{m,n=1}^{\infty}\left(\frac{-3}{m}\right)n^2q^{mn}, \\ \chi_{-3}(m) &= \left(\frac{-3}{m}\right) = \frac{e^{2\pi im/3} - e^{-2\pi im/3}}{i\sqrt{3}}. \end{aligned}$$

Though the anti-derivative $P_3(x)$,

$$\begin{aligned} P_3(x) &= \frac{3\sqrt{3}}{\pi}\left(\sum_{m,n=1}^{\infty}\left(\frac{-3}{m}\right)\frac{n}{m}\,q^{mn} - 4\sum_{m,n=1}^{\infty}\left(\frac{-3}{m}\right)\frac{n}{m}\,q^{2mn}\right) \\ &= \frac{9i}{\pi}\,\log\prod_{n=1}^{\infty}\left(\frac{(1-e^{2\pi i/3}q^{2n})^4(1-e^{-2\pi i/3}q^n)}{(1-e^{-2\pi i/3}q^{2n})^4(1-e^{2\pi i/3}q^n)}\right)^n, \end{aligned}$$

is not considered to be sufficiently ‘natural’, it shows up as the elliptic dilogarithm thanks to Bloch’s formula; see [17, 19] for the details. Note that

$$\begin{aligned} E_{3,\chi_{-3}}\left(-\frac{1}{3\tau}\right) &= \frac{i\tau^3}{3\sqrt{3}}\,\tilde{E}_{3,\chi_{-3}}(\tau), \\ \tilde{E}_{3,\chi_{-3}}(\tau) = \frac{\eta(\tau)^9}{\eta(3\tau)^3} &= 1 - 9\sum_{m,n=1}^{\infty}\left(\frac{-3}{n}\right)n^2q^{mn}; \end{aligned}$$

and, in addition, we have

$$\begin{aligned}\frac{1}{2\pi i}\,\frac{\mathrm{d}x/\mathrm{d}\tau}{x} = \frac{1}{2}\left(\frac{\eta(\tau)^2\eta(3\tau)^2}{\eta(2\tau)\eta(6\tau)}\right)^2 &= \frac{1}{18}\big(E_{1,\chi_{-3}}(\tau) - 4E_{1,\chi_{-3}}(4\tau)\big)^2 \\ &= \frac{1}{54\tau^2}\left(E_{1,\chi_{-3}}\left(-\frac{1}{12\tau}\right) - E_{1,\chi_{-3}}\left(-\frac{1}{3\tau}\right)\right)^2,\end{aligned}$$

where

$$E_{1,\chi_{-3}}(\tau) = 1 + 6\sum_{m,n=1}^{\infty}\left(\frac{-3}{m}\right)q^{mn}.$$

7 Modular Computation for $W_5'(0)$ and $W_6'(0)$

As (partly) shown in [10] the density $p_4(x)$ can be parameterised as follows (we make a shift of τ by half):

$$p_4(x(\tau)) = -\operatorname{Re}\left(\frac{2i(1+6\tau+12\tau^2)}{\pi}\,p(\tau)\right),$$

where

$$p(\tau) = \frac{\eta(2\tau)^4\eta(6\tau)^4}{\eta(\tau)\eta(3\tau)\eta(4\tau)\eta(12\tau)} \quad\text{and}\quad x(\tau) = \left(\frac{2\eta(\tau)\eta(3\tau)\eta(4\tau)\eta(12\tau)}{\eta(2\tau)^2\eta(6\tau)^2}\right)^3.$$

The path for τ along the imaginary axis from 0 to $i/(2\sqrt{3})$ (or from $i\infty$ to $i/(2\sqrt{3})$) corresponds to x ranging from 0 to 2, while the path from $i/(2\sqrt{3})$ to $-1/4 + i/(4\sqrt{3})$ along the arc centred at 0 corresponds to the real range (2, 4) for x. (The arc admits the parametrisation $\tau = e^{\pi i\theta}/(2\sqrt{3})$, $1/2 < \theta < 5/6$.) Note that $x(i/(2\sqrt{15})) = 1$ and

$$p_4(x(\tau)) = \begin{cases} -\dfrac{2i\cdot 6\tau}{\pi}\,p(\tau), & \text{for } \tau \text{ on the imaginary axis,} \\ -\dfrac{2i(1+6\tau+12\tau^2)}{\pi}\,p(\tau), & \text{for } \tau \text{ on the arc,}\end{cases}$$

and

$$-\frac{2i(1+6\tau+12\tau^2)}{\pi}\,p(\tau) = \frac{2\sqrt{16-x^2}}{\pi^2 x}\cdot {}_3F_2\left(\begin{matrix}\frac12,\frac12,\frac12\\ \frac56,\frac76\end{matrix}\,\middle|\,\frac{(16-x^2)^3}{108x^4}\right)$$

(this is a general form of [10, Theorem 4.9]). Formulas (1), (3) and (4) reduce the conjectural evaluations of $W_5'(0)$ and $W_6'(0)$ to the following ones:

$$\frac{7\zeta(3)}{2\pi^2} + L'(f_3;-1) \overset{?}{=} \frac{12}{\pi}\int_0^{1/(2\sqrt{15})} yp(iy)\log x(iy)\,\mathrm{d}x(iy)$$

and

$$\begin{aligned}\frac{7\zeta(3)}{2\pi^2} + 8L'(f_4;-1) \overset{?}{=} {} & \frac{12}{\pi}\int_0^{1/(2\sqrt{3})} yp(iy)\log x(iy)\,\mathrm{d}x(iy) \\ & - \frac{12}{\pi^2}\int_0^{1/(2\sqrt{3})} yp(iy)x(iy)\cdot {}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12 \\ \frac32, \frac32\end{matrix}\,\middle|\,\frac{x(iy)^2}{4}\right)\mathrm{d}x(iy).\end{aligned}$$

Furthermore, note that the Atkin–Lehner involutions $w_{12}\colon \tau \mapsto -1/(12\tau)$ and $w_6\colon \tau \mapsto (6\tau-5)/(12\tau-6)$ act on the modular function $x(\tau)$ as follows: $x(w_{12}\tau) = x(\tau)$ and $x(w_6\tau) = -8/x(\tau)$, and we also have $p(w_{12}\tau) = -\tau^2 p(\tau)$. The point $i/(2\sqrt{3})$ is fixed by w_{12}. Thus, the change of variable $y \mapsto 1/(12y)$ leads to

$$\int_0^{1/(2\sqrt{3})} yp(iy)\log x(iy)\,\mathrm{d}x(iy) = -\int_{1/(2\sqrt{3})}^{\infty} yp(iy)\log x(iy)\,\mathrm{d}x(iy).$$

8 Mahler Measures Related to a Variation of Random Walk

In [23] the Mahler measures $\mathrm{m}(1+x_1+x_2)$ and $\mathrm{m}(1+x_1+x_2+x_3)$ are computed using the modular parametrisations of

$$\sum_{n=0}^{\infty} W_3(2n)z^n = \sum_{n=0}^{\infty} \mathrm{CT}\big((1+x_1+x_2)(1+x_1^{-1}+x_2^{-1})\big)^n z^n$$

and

$$\sum_{n=0}^{\infty} W_4(2n)z^n = \sum_{n=0}^{\infty} \mathrm{CT}\big((1+x_1+x_2+x_3)(1+x_1^{-1}+x_2^{-1}+x_3^{-1})\big)^n z^n,$$

where $\mathrm{CT}(L)$ denotes the constant term of a Laurent polynomial $L \in \mathbb{Z}[x_1^{\pm}, x_2^{\pm}, \ldots]$. Note that the Picard–Fuchs linear differential equations for the two generating functions give rise to the ones for the densities $p_3(x)$ and $p_4(x)$ together with their explicit hypergeometric and modular expressions (see [10, eq. (3.2) and Remark 4.10]), though it remains unclear whether the latter information can be used to compute $W_N'(0)$ in (1) for $N = 3, 4$. This is itself an interesting question to not only assist in computing of $W_N'(0)$ for $N > 4$ but also in relation with another famous conjecture of Boyd:

$$\mathrm{m}(1+x_1+x_2+x_3+x_2x_3) \overset{?}{=} -2L'(f_2;-1) = \frac{15^2}{4\pi^4}L(f_2;3) \qquad (6)$$
$$= 0.4839979734\ldots,$$

where $f_2(\tau)=\eta(\tau)\eta(3\tau)\eta(5\tau)\eta(15\tau)$.

In analogy with the case of linear Mahler measures, we define

$$\begin{aligned}\widetilde{W}(s) &= \iiint_{[0,1]^3} |1+e^{2\pi i\theta_1}+e^{2\pi i\theta_2}+e^{2\pi i\theta_3}+e^{2\pi i(\theta_2+\theta_3)}|^s \,\mathrm{d}\theta_1\,\mathrm{d}\theta_2\,\mathrm{d}\theta_3 \\ &= Z(1+x_1+x_2+x_3+x_2x_3;s)\end{aligned}$$

as the s-th moment of a random 5-step walk for which the direction of the final step is completely determined by the two previous steps. Then the even moments

$$\begin{aligned}\widetilde{W}(2n) &= \mathrm{CT}\big((1+x_1+x_2+x_3+x_2x_3)(1+x_1^{-1}+x_2^{-1}+x_3^{-1}+(x_2x_3)^{-1})\big)^n \\ &= \sum_{k=0}^{n}\binom{n}{k}^2\binom{2k}{k}^2\end{aligned}$$

satisfy a rather lengthy recurrence equation, which is equivalent to a Picard–Fuchs differential equation of order 4. The latter splits into the tensor product of two differential equations of order 2 and, with some effort, we obtain the following result.

Theorem 1 *We have*

$$\begin{aligned}&\sum_{n=0}^{\infty}\widetilde{W}(2n)\left(\frac{t}{(4+t)(1+4t)}\right)^n \\ &= \frac{(4+t)(1+4t)}{4(1+4t+t^2)}\,{}_2F_1\left(\begin{matrix}\frac12,\frac12\\1\end{matrix}\;\middle|\;\frac{t(4+t)}{1+4t+t^2}\right)\cdot{}_2F_1\left(\begin{matrix}\frac12,\frac12\\1\end{matrix}\;\middle|\;\frac{t^2}{1+4t+t^2}\right)\end{aligned}$$

and, more generally,

$$\begin{aligned}&\frac{b}{(b+t)(1+bt)}\sum_{n=0}^{\infty}\left(\frac{t}{(b+t)(1+bt)}\right)^n\sum_{k=0}^{n}\binom{n}{k}^2\binom{2k}{k}^2\left(\frac{b}{4}\right)^{2k} \\ &= {}_2F_1\left(\begin{matrix}\frac12,\frac12\\1\end{matrix}\;\middle|\;-t(b+t)\right)\cdot\frac{1}{(1+bt)^{1/2}}\,{}_2F_1\left(\begin{matrix}\frac12,\frac12\\1\end{matrix}\;\middle|\;-\frac{t^2}{1+bt}\right) \\ &= \frac{1}{1+bt+t^2}\,{}_2F_1\left(\begin{matrix}\frac12,\frac12\\1\end{matrix}\;\middle|\;\frac{t(b+t)}{1+bt+t^2}\right)\cdot{}_2F_1\left(\begin{matrix}\frac12,\frac12\\1\end{matrix}\;\middle|\;\frac{t^2}{1+bt+t^2}\right).\end{aligned}$$

Proof Once a factorisation of this type is written down, it is a computational routine to prove it. In other words, a principal issue is discovering such a formula rather than proving it. Our original discovery of Theorem 1 involved a lot of experimental mathematics; however, we later realised that it is deducible from known formulae as follows:

$$\begin{aligned}\sum_{n=0}^{\infty} z^n \sum_{k=0}^{n} \binom{n}{k}^2 \binom{2k}{k}^2 x^k &= \sum_{k=0}^{\infty} \binom{2k}{k}^2 x^k \sum_{m=0}^{\infty} \binom{k+m}{k}^2 z^{k+m} \\ &= \sum_{k=0}^{\infty} \binom{2k}{k}^2 (xz)^k \, {}_2F_1\left(\begin{matrix} k+1,\ k+1 \\ 1 \end{matrix} \,\middle|\, z\right) \\ &= \sum_{k=0}^{\infty} \binom{2k}{k}^2 \frac{(xz)^k}{(1-z)^{k+1}} \, {}_2F_1\left(\begin{matrix} -k,\ k+1 \\ 1 \end{matrix} \,\middle|\, -\frac{z}{1-z}\right) \\ &= \frac{1}{1-z} \sum_{k=0}^{\infty} \binom{2k}{k}^2 \left(\frac{xz}{1-z}\right)^k \cdot P_k\left(\frac{1+z}{1-z}\right),\end{aligned}$$

where P_k denotes the k-th Legendre polynomial, and the latter generating function is a particular instance of the Bailey–Brafman formula [15, 34]. □

We remark that, using the general Bailey–Brafman formula and its generalisation from [29], the proof above extends to the factorisation of the two-variable generating functions

$$\sum_{n=0}^{\infty} z^n \sum_{k=0}^{n} \binom{n}{k}^2 \frac{(s)_k (1-s)_k}{k!^2} x^k$$

as well as of

$$\sum_{n=0}^{\infty} z^n \sum_{k} \binom{n}{2k}^2 \binom{2k}{k}^2 x^k \quad \text{and} \quad \sum_{n=0}^{\infty} z^n \sum_{k} \binom{n}{3k}^2 \frac{(3k)!}{k!^3} x^k,$$

and even of

$$\sum_{n=0}^{\infty} z^n \sum_{k=0}^{n} \binom{n}{k}^2 u_k \, x^k$$

for an Apéry-like sequence $u_0, u_1, u_2, \dots$.

Furthermore, we expect that Theorem 1 can lead to a hypergeometric expression for the density function $\widetilde{p}(x)$ (piecewise analytic, with finite support on the interval $0 < x < 5$), which is the inverse Mellin transform of $\widetilde{W}(s-1)$, hence to the Mahler measure evaluation

$$\mathrm{m}(1 + x_1 + x_2 + x_3 + x_2 x_3) = \widetilde{W}'(0) = \int_0^{\infty} \widetilde{p}(x) \log x \, \mathrm{d}x = \int_0^{5} \widetilde{p}(x) \log x \, \mathrm{d}x.$$

On the other hand, the reduction technique of Sections 4 and 5 suggests a different approach to computing $\widetilde{W}'(0)$, resulting in the following hypergeometric evaluation of the Mahler measure.

Theorem 2 *We have*

$$\mathrm{m}(1+x_1+x_2+x_3+x_2x_3) = -\frac{1}{2\pi}\int_0^1 {}_2F_1\left(\begin{matrix}\frac12, \frac12\\ 1\end{matrix}\,\middle|\, 1-\frac{x^2}{16}\right)\log x\,\mathrm{d}x.$$

Proof Define a related density $\widehat{p}(x)$ by

$$\begin{aligned}\int_0^4 x^s\,\widehat{p}(x)\,\mathrm{d}x = \widehat{W}(s) &= \iint_{[0,1]^2} |1+e^{2\pi i\theta_2}+e^{2\pi i\theta_3}+e^{2\pi i(\theta_2+\theta_3)}|^s\,\mathrm{d}\theta_2\,\mathrm{d}\theta_3\\ &= W_2(s)^2 = \frac{\Gamma(1+s)^2}{\Gamma(1+s/2)^4}.\end{aligned}$$

By an application of the Mellin transform calculus, we find that, for $0<x<4$,

$$\widehat{p}(x) = \frac{1}{2\pi}\cdot {}_2F_1\left(\begin{matrix}\frac12, \frac12\\ 1\end{matrix}\,\middle|\, 1-\frac{x^2}{16}\right).$$

Then it follows from Lemma 1 that

$$\widetilde{W}'(0) = \int_1^4 \widehat{p}(x)\log x\,\mathrm{d}x = -\int_0^1 \widehat{p}(x)\log x\,\mathrm{d}x,$$

where we use the evaluation

$$\int_0^4 \widehat{p}(x)\log x\,\mathrm{d}x = \mathrm{m}(1+x_2+x_3+x_2x_3) = \mathrm{m}(1+x_2)+\mathrm{m}(1+x_3) = 0.$$

The above proof extends to the general formula

$$\begin{aligned}\mathrm{m}(1+bx_1+x_2+x_3+x_2x_3) &= \log b\int_0^b \widehat{p}(x)\,\mathrm{d}x + \int_b^4 \widehat{p}(x)\log x\,\mathrm{d}x\\ &= \frac{1}{2\pi}\int_0^b {}_2F_1\left(\begin{matrix}\frac12, \frac12\\ 1\end{matrix}\,\middle|\, 1-\frac{x^2}{16}\right)\log\frac{b}{x}\,\mathrm{d}x\end{aligned}$$

for $0<b\le 4$. A related computation

$$\mathrm{m}(1+bx_1+x_2+x_3+x_2x_3) = \log b + \frac{8}{\pi^2}\int_b^4 \frac{\arccos(b/x)\,\log(x/(2\sqrt{b}))}{\sqrt{16-x^2}}\,\mathrm{d}x$$

valid for $0<b\le 4$ was given by Wan [27]; he also pointed out that $\mathrm{m}(1+bx_1+x_2+x_3+x_2x_3) = \log b$ for $b>4$ follows from Jensen's formula.

The left-hand side of another Mahler measure conjecture [13]

$$\mathrm{m}((1+x_1)^2+x_2+x_3) \overset{?}{=} -L'(\tilde{f}_2;-1) = \frac{72}{\pi^4}\,L(\tilde{f}_2;3) = 0.7025655062\ldots,$$

where $\tilde{f}_2(\tau)=\eta(2\tau)\eta(4\tau)\eta(6\tau)\eta(12\tau)$ is a cusp form of level 24, can be treated by a similar reduction, using that the densities for $(1+x_1)^2$ and x_2+x_3 are $p_2(t^{1/2})/(2t^{1/2})$ on [0, 4] and $p_2(t)$ on [0, 2], respectively. The final result is the elegant formula

$$\mathrm{m}((1+x_1)^2+x_2+x_3) = \frac{2G}{\pi} + \frac{2}{\pi^2}\int_0^1 \arcsin(1-x)\,\arcsin x\,\frac{\mathrm{d}x}{x}, \tag{7}$$

where G is Catalan's constant, and, with some further work, we can express the right-hand side hypergeometrically.

Theorem 3 *We have*

$$\mathrm{m}((1+x_1)^2+x_2+x_3) = \frac{8\Gamma(\frac{3}{4})^2}{\pi^{5/2}}\,{}_5F_4\left(\begin{matrix}\frac14,\frac14,\frac14,\frac34,\frac34\\ \frac12,\frac54,\frac54,\frac54\end{matrix}\,\middle|\,\frac14\right) + \frac{\Gamma(\frac14)^2}{54\pi^{5/2}}\,{}_5F_4\left(\begin{matrix}\frac34,\frac34,\frac34,\frac54,\frac54\\ \frac32,\frac74,\frac74,\frac74\end{matrix}\,\middle|\,\frac14\right).$$

Proof Notice that, for $0<x<1$,

$$\arcsin(1-x) = \frac{\pi}{2} - \arccos(1-x) = \frac{\pi}{2} - \sqrt{2x}\,{}_2F_1\left(\begin{matrix}\frac12,\frac12\\ \frac32\end{matrix}\,\middle|\,\frac{x}{2}\right),$$

and that, for $n>-1/2$,

$$\int_0^1 x^{n-1/2}\arcsin x\,\mathrm{d}x = \frac{\sqrt{\pi}}{2n+1}\left(\sqrt{\pi} - \frac{\Gamma(\frac{n}{2}+\frac34)}{\Gamma(\frac{n}{2}+\frac54)}\right).$$

Therefore,

$$\int_0^1 \arcsin(1-x)\,\arcsin x\,\frac{\mathrm{d}x}{x} = \frac{\pi}{2}\int_0^1 \arcsin x\,\frac{\mathrm{d}x}{x} - \pi\sqrt{2}\sum_{n=0}^{\infty}\frac{(\frac12)_n^2}{n!\,(\frac32)_n(2n+1)}\,\frac{1}{2^n} + \sqrt{2\pi}\sum_{n=0}^{\infty}\frac{(\frac12)_n^2\Gamma(\frac{n}{2}+\frac34)}{n!\,(\frac32)_n(2n+1)\,\Gamma(\frac{n}{2}+\frac54)}\,\frac{1}{2^n}.$$

From this and (7) we deduce

$$\begin{aligned}\mathrm{m}((1+x_1)^2+x_2+x_3) &= \frac{2G}{\pi}+\frac{\log 2}{2}-\frac{2\sqrt{2}}{\pi}\,{}_3F_2\left(\begin{matrix}\frac12,\frac12,\frac12\\ \frac32,\frac32\end{matrix}\,\middle|\,\frac12\right)\\ &\quad+\frac{8\sqrt{2}\,\Gamma(\frac34)}{\pi^{3/2}\Gamma(\frac14)}\,{}_5F_4\left(\begin{matrix}\frac14,\frac14,\frac14,\frac34,\frac34\\ \frac12,\frac54,\frac54,\frac54\end{matrix}\,\middle|\,\frac14\right)+\frac{\sqrt{2}\,\Gamma(\frac14)}{54\pi^{3/2}\Gamma(\frac34)}\,{}_5F_4\left(\begin{matrix}\frac34,\frac34,\frac34,\frac54,\frac54\\ \frac32,\frac74,\frac74,\frac74\end{matrix}\,\middle|\,\frac14\right).\end{aligned}$$

It remains to use

$$G+\frac14\pi\log 2=\sqrt{2}\,{}_3F_2\left(\begin{matrix}\frac12,\frac12,\frac12\\ \frac32,\frac32\end{matrix}\,\middle|\,\frac12\right)$$

(see [1, Entry 30]) and $\Gamma(\frac14)\Gamma(\frac34)=\pi\sqrt{2}$.

9 Conclusion

A goal of this final section is to highlight relevance for and links with other research and open problems.

The (hypergeometric) factorisation in Theorem 1 and similar results outlined after its proof are part of a general phenomenon of arithmetic differential equations of order 4. These are the first instances 'beyond modularity' in the sense that arithmetic differential equations of order 2 and 3 are always supplied by modular parametrisation. In order 4, we have to distinguish two particular novel situations (though our knowledge about either is imperfect and incomplete): (the Zariski closure of) the monodromy group is the orthogonal group $O_4 \simeq O_{2,2}$ of dimension 6 or the symplectic group Sp_4 of dimension 10. The example given in Theorem 1 corresponds to the first (orthogonal) situation: on the level of Lie groups, $O_{2,2}$ can be realised as the tensor product of two copies of SL_2 (or GL_2). There is a limited amount of further examples of this type [21, 29, 33] though we expect that all underlying Picard–Fuchs differential equations with such monodromy can be represented as tensor products of two arithmetic differential equations of order 2. There is a natural hypergeometric production of such orthogonal cases using Orr-type formulae (see [18, 28]) but there are plenty of other cases coming from classical work of W. N. Bailey and its recent generalisations [29, 34]. Many such cases, mostly forecast by Sun [25], are still awaiting their explicit factorisation. Though these situations do not cover symplectic monodromy instances, they can still be viewed as an intermediate step between classical modularity and Sp_4: the antisymmetric square of the latter happens to be $O_5 \simeq O_{3,2}$ (see [4]).

More in the direction of three-variable Mahler measure, the conjectural evaluation in (6) and Theorem 2 brings us to the expectation

$$\frac{1}{2\pi}\int_0^1 {}_2F_1\left(\begin{matrix}\frac12,\frac12\\ 1\end{matrix}\,\middle|\,1-\frac{x^2}{16}\right)\log x\,\mathrm{d}x \overset{?}{=} 2L'(f_2;-1). \tag{8}$$

This one highly resembles the evaluation

$$\frac{1}{2}\int_0^1 {}_2F_1\left(\begin{matrix}\frac12, \frac12\\ 1\end{matrix}\;\middle|\; \frac{x^2}{16}\right)\mathrm{d}x = \frac{1}{2}\cdot {}_3F_2\left(\begin{matrix}\frac12, \frac12, \frac12\\ 1, \frac32\end{matrix}\;\middle|\; \frac{1}{16}\right) = 2L'(f_2;0) \tag{9}$$

conjectured in [12] and established in [22]. The related modular parametrisation

$$x = x(\tau) = 16\left(\frac{\eta(\tau)\eta(4\tau)^2}{\eta(2\tau)^3}\right)^4$$

corresponds to

$$1-\frac{x^2}{16} = \left(\frac{\eta(\tau)^2\eta(4\tau)}{\eta(2\tau)^3}\right)^8,$$

$$F\left(\frac{x^2}{16}\right) = \frac{\eta(2\tau)^{10}}{\eta(\tau)^4\eta(4\tau)^4} \quad\text{and}\quad F\left(1-\frac{x^2}{16}\right) = -2i\tau F\left(\frac{x^2}{16}\right),$$

where F denotes the corresponding ${}_2F_1$ hypergeometric series. Note that x ranges from 0 to 4 when τ runs from $i\infty$ to 0 along the imaginary axis; however, the point $\tau_0 = i\,0.8774376613482\ldots$, at which $x(\tau_0)=1$, is not a quadratic irrationality. Furthermore, Cohen [16] observes another step in the ladder (9), (8):

$$\frac{6}{\pi^2}\int_0^1 {}_2F_1\left(\begin{matrix}\frac12, \frac12\\ 1\end{matrix}\;\middle|\; \frac{x^2}{16}\right)\log^2 x\,\mathrm{d}x \overset{?}{=} 2L'(f_2;-2) = \frac{3\cdot 15^3}{8\pi^6}L(f_2;4) \tag{10}$$
$$= 1.2165632526\ldots,$$

though not linked to a particular Mahler measure.

The expression in Theorem 3 is somewhat different from the one in Theorem 2, and resembles the hypergeometric evaluation of the L-value

$$\begin{aligned}-L'(\hat{f}_2;-1) &= \frac{128}{\pi^4}L(\hat{f}_2;3)\\ &= \frac{\Gamma(\frac14)^2}{6\sqrt{2}\pi^{5/2}}\,{}_4F_3\left(\begin{matrix}1, 1, 1, \frac12\\ \frac74, \frac32, \frac32\end{matrix}\;\middle|\; 1\right) + \frac{4\Gamma(\frac34)^2}{\sqrt{2}\pi^{5/2}}\,{}_4F_3\left(\begin{matrix}1, 1, 1, \frac12\\ \frac54, \frac32, \frac32\end{matrix}\;\middle|\; 1\right)\\ &\quad + \frac{\Gamma(\frac14)^2}{2\sqrt{2}\pi^{5/2}}\,{}_4F_3\left(\begin{matrix}1, 1, 1, \frac12\\ \frac34, \frac32, \frac32\end{matrix}\;\middle|\; 1\right),\end{aligned}$$

where $\hat{f}_2(\tau) = \eta(4\tau)^2\eta(8\tau)^2$ is a cusp form of level 32, obtained in [32, Theorem 3].

Finally, we remark that the integral

$$W_3'(0) = \int_0^3 \log x\,\mathrm{d}P_3(x) = \log 3 - \int_0^3 P_3(x)\,\frac{\mathrm{d}x}{x}$$

in the notation of Section 6, with $P_3(x)$ related to eta quotients, is visually linked to the following result in [7] (also discussed in greater generality in [2, 26])

$$\int_0^1 \frac{1}{9}\left(1 - \frac{\eta(\tau)^9}{\eta(3\tau)^3}\right)\frac{\mathrm{d}q}{q} = \lim_{q\to 1^-} \sum_{m,n=1}^{\infty} \left(\frac{-3}{n}\right)\frac{n}{m}\, q^{mn} = L'(\chi_{-3}; -1).$$

However, apart from the fact that the two quantities coincide we could not find a direct link between the two integrals.

Acknowledgements We thank H. Cohen for supplying us with the numerical observation (10) whose origin remains completely mysterious to us. We also thank the referee for their valuable feedback.

Both authors acknowledge support of the Max Planck Society during their stays at the Max Planck Institute for Mathematics (Bonn, Germany) in the years 2015–2017. The second author is partially supported by Laboratory of Mirror Symmetry NRU HSE, RF government grant, ag. no. 14.641.31.0001.

References

1. Adamchik, V.S.: Integral and series representations for Catalan's constant. Unpublished note. http://www.cs.cmu.edu/~adamchik/articles/catalan.htm
2. Ahlgren, S., Berndt, B.C., Yee, A.Y., Zaharescu, A.: Integrals of Eisenstein series and derivatives of L-functions. Int. Math. Res. Not. **2002**(32), 1723–1738 (2002)
3. Akatsuka, H.: Zeta Mahler measures. J. Number Theory **129**(11), 2713–2734 (2009)
4. Almkvist, G., van Straten, D., Zudilin, W.: Generalizations of Clausen's formula and algebraic transformations of Calabi–Yau differential equations. Proc. Edinb. Math. Soc. **54**(2), 273–295 (2011)
5. Bailey, D.H., Borwein, J.M.: Hand-to-hand combat with multi-thousand-digit integrals. J. Comput. Sci. **3**, 77–86 (2012)
6. Bailey, D.H., Borwein, J.M., Broadhurst, D.J., Glasser, M.L.: Elliptic integral evaluations of Bessel moments and applications. J. Phys. A **41**(20), 5203–5231 (2008)
7. Berndt, B.C., Zaharescu, A.: An integral of Dedekind eta-functions in Ramanujan's lost notebook. J. Reine Angew. Math. **551**, 33–39 (2002)
8. Borwein, J.M.: A short walk can be beautiful. J. Humanist. Math. **6**(1), 86–109 (2016)
9. Borwein, J.M., Nuyens, D., Straub, A., Wan, J.: Some arithmetic properties of short random walk integrals. Ramanujan J. **26**(1), 109–132 (2011)
10. Borwein, J.M., Straub, A., Wan, J., Zudilin, W.: Densities of short uniform random walks, with an appendix by D. Zagier. Can. J. Math. **64**(5), 961–990 (2012)
11. Borwein, J.M., Straub, A., Wan, J.: Three-step and four-step random walk integrals. Exp. Math. **22**(1), 1–14 (2013)
12. Boyd, D.: Mahler's measure and special values of L-functions. Exp. Math. **7**(1), 37–82 (1998)
13. Boyd, D., Lind, D., Rodriguez-Villegas, F., Deninger, C.: The many aspects of Mahler's measure. Final report of the Banff workshop 03w5035 (26 April–1 May 2003). http://www.birs.ca/workshops/2003/03w5035/report03w5035.pdf
14. Broadhurst, D.: Feynman integrals, L-series and Kloosterman moments. Commun. Number Theory Phys. **10**(3), 527–569 (2016)
15. Chan, H.H., Wan, J., Zudilin, W.: Legendre polynomials and Ramanujan-type series for $1/\pi$. Isr. J. Math. **194**(1), 183–207 (2013)
16. Cohen, H.: Personal communication (23 March 2018)

17. Duke, W., Imamoḡlu, Ö.: On a formula of Bloch. Funct. Approx. **37**(1), 109–117 (2007)
18. Guillera, J.: A family of Ramanujan-Orr formulas for $1/\pi$. Integral Transforms Spec. Funct. **26**(7), 531–538 (2015)
19. Paşol, V., Zudilin, W.: A study of elliptic gamma function and allies. Res. Math. Sci. **5**(4), Art. 39, p 11 (2018)
20. Rogers, M.D.: A study of inverse trigonometric integrals associated with three-variable Mahler measures, and some related identities. J. Number Theory **121**, 265–304 (2006)
21. Rogers, M.D., Straub, A.: A solution of Sun's $520 challenge concerning $520/\pi$. Int. J. Number Theory **9**, 1273–1288 (2013)
22. Rogers, M.D., Zudilin, W.: On the Mahler measure of $1 + X + 1/X + Y + 1/Y$. Int. Math. Res. Not. **2014**(9), 2305–2326 (2014)
23. Shinder, E., Vlasenko, M.: Linear Mahler measures and double L-values of modular forms. J. Number Theory **142**, 149–182 (2014)
24. Smyth, C.J.: On measures of polynomials in several variables. Bull. Aust. Math. Soc. **23**, 49–63 (1981)
25. Sun, Z.-W.: List of conjectural series for powers of π and other constants (2014). arXiv:1102.5649v47 [math.CA]
26. Takloo-Bighash, R.: A remark on a paper of S. Ahlgren, B.C. Berndt, A.J. Yee, and A. Zaharescu: "Integrals of Eisenstein series and derivatives of L-functions" [2]. Int. J. Number Theory **2**(1), 111–114 (2006)
27. Wan, J.G.: Personal communication (26 July 2011)
28. Wan, J.G.: Series for $1/\pi$ using Legendre's relation. Integral Transforms Spec. Funct. **25**(1), 1–14 (2014)
29. Wan, J.G., Zudilin, W.: Generating functions of Legendre polynomials: a tribute to Fred Brafman. J. Approx. Theory **164**, 488–503 (2012)
30. Zhou, Y.: On Borwein's conjectures for planar uniform random walks. J Austral. Math. Soc. **107**(3), 392–411 (2019)
31. Zhou, Y.: Wick rotations, Eichler integrals, and multi-loop Feynman diagrams. Commun. Number Theory Phys. **12**(1), 127–192 (2018)
32. Zudilin, W.: Period(d)ness of L-values. In: Borwein, J.M., et al. (eds.) Number Theory and Related Fields, In Memory of Alf van der Poorten. Springer Proceedings in Mathematics and Statistics, vol. 43, pp. 381–395. Springer, New York (2013)
33. Zudilin, W.: A generating function of the squares of Legendre polynomials. Bull. Aust. Math. Soc. **89**(1), 125–131 (2014)
34. Zudilin, W.: Hypergeometric heritage of W.N. Bailey. With an appendix: Bailey's letters to F. Dyson. Not. Intern. Congr. Chin. Mathematicians **7**(2), 32–46 (2019)

Correction to: Short Walk Adventures

Armin Straub and Wadim Zudilin

Correction to:
Chapter "Short Walk Adventures"
in: D. H. Bailey et al. (eds.), *From Analysis to Visualization*, Springer Proceedings in Mathematics & Statistics 313, https://doi.org/10.1007/978-3-030-36568-4_27

The original version of Chapter "Short Walk Adventures" was inadvertently published with incorrect cross-citations in square brackets "[4] and $p_2(t)$ on [2]". This has now been amended correctly as on [0,4] and $p_2(t)$ on [0,2]

$p_2(t^{1/2})/(2t^{1/2})$ on [4] and $p_2(t)$ on [2],

to

$p_2(t^{1/2})/(2t^{1/2})$ on [0,4] and $p_2(t)$ on [0,2],
The erratum chapter and the book have been updated with the change.

The updated version of the book can be found at
https://doi.org/10.1007/978-3-030-36568-4_27

D. H. Bailey et al. (eds.), *From Analysis to Visualization*,
Springer Proceedings in Mathematics & Statistics 313,
https://doi.org/10.1007/978-3-030-36568-4_28