Online at `https://digitalresearch.bsu.edu/mathexchange`

ISSN 1550-1736

Cover design by Patrick Foley.

A Word from the Editor

The editorial board of the *Mathematics Exchange* is delighted to present our latest issue, comprising nine captivating articles that explore a diverse array of mathematical topics tailored for a broad undergraduate audience. We extend our sincere appreciation to the authors for their dedicated efforts in sharing their new discoveries, inspiring, and motivating our readers to immerse themselves in the world of mathematics. We trust that you will find this collection to be a rewarding culmination of their scholarly endeavors.

The Law of Small Numbers states the convergence of the Binomial distribution to the Poisson distribution. The inaugural article vividly illustrates this law. Specifically, utilizing the programming language R, the authors delve into the total variation distance between these two distributions.

While the classification of semi-simple Lie algebras was resolved over a century ago, the challenge of categorizing solvable Lie algebras remains open, particularly in higher dimensions. The second article contributes to this ongoing discourse by delving into the classification of solvable Lie algebras in the dimension seven setting, building upon the established classifications in dimensions six and lower.

The third article offers a novel and concise proof of the well-known fact that the set of monomial matrices forms a subgroup of invertible matrices. The work not only addresses a well-known fact but also fills a gap in the literature by providing a readily available proof. In addition, the authors establish the simple yet profound result that the inverse of a nonnegative matrix is nonnegative if and only if the matrix is monomial.

Article four introduces an engaging exploration of the classic combinatorial game Cram, featuring rectangular polyominoes rather than the conventional 1×2 dominoes. This article exemplifies the discovery of innovative results by posing insightful questions, relying on elementary arguments and leveraging symmetry to articulate winning strategies. The findings may inspire further exploration of the subject with different polyominoes.

The fifth article investigates level sets of real-valued continuous functions on closed intervals, inspired by the intermediate value theorem. This inquiry delves into the behavior of functions whose endpoints converge to the same real number, providing valuable insights into the structure of these sets.

Turning to the realm of signal processing, the sixth article focuses on Independent Component Analysis (ICA) as a blind-source separation method. Specifically, it explores how ICA handles over-complete data, demonstrating its ability to consistently group sources with similar spatial maps in the presence of three sinusoidal sources and two sensors.

The seventh article delves into the Riemann Zeta function, a captivating infinite converging sum of powers of natural numbers. The authors present various irrationality proofs, with a specific focus on demonstrating the irrationality of certain values of the Zeta function.

In the eighth article, the authors generalize a result by Mortini and Rupp, offering insights into the Cauchy transform of the complex power function. Employing limits and a contour technique, they navigate around branch cuts, introducing a novel approach utilizing hyperbolic geometric functions. This paper enriches our understanding of integrating over curves with multi-functions and establishes connections to solutions of differential equations, particularly those involving the hypergeometric function.

The final article introduces a matrix iteration framework to study the Mandelbrot set and filled Julia sets. By employing a sequence of affine transformations, the authors establish an alternate form of iteration by complex polynomials. This framework enables the verification of membership in the Mandelbrot set and filled Julia sets, demonstrating that boundedness in the operator norm corresponds to belonging to these sets.

We trust that you will find this issue of the *Mathematics Exchange* to be a source of intellectual enrichment and inspiration. As always, we eagerly welcome and encourage your ideas on how we can continue to enhance our service to our valued readers.

Yayuan Xiao

11.15.2023

Managing Editor

Editorial Board

Proofreader

Call for Papers

We are always soliciting contributions for future issues of this journal. Contributions are accepted from all undergraduate students who have worked on a project beyond the classroom in any mathematical area (e.g., pure, applied, actuarial, and education). Appropriate papers from other departments and other institutions are also welcome. Often the articles are written by undergraduates individually, working in teams, or working with faculty. On occasion we also include articles written solely by faculty or graduate students as long as they are accessible to undergraduates.

To submit an article, please select ONE member from the editorial board, and forward your material in PDF form, usually prepared by LaTeX (preferred) or Microsoft Word, to the editor you selected. We use double anonymized peer review, the identities of both reviewers and authors are concealed from each other throughout the review. To facilitate this, please remove any identifying information, such as authors' names or affiliations, from your manuscript before submission. Please ensure that the title page (that include all authors' names and affiliations, a complete address of the corresponding author including an email address, acknowledgements, and conflict of interest statement) is present in your submission as a separate file. If authors are undergraduate students, please include your advisor's name and contact information in the title page. Review and selection of articles is handled by the editorial committee. Editorial changes of accepted articles are communicated through students' advisors, when appropriate.

More information, including links to all previous issues, are available online at
`https://digitalresearch.bsu.edu/mathexchange`.

Contents

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 17, No. 1 (Fall 2023)
Pages 2 – 14

An investigation into the law of small numbers using R

*Yasir Zubayr Barlas, Dudley Stark**

Yasir Zubayr Barlas pursued his undergraduate studies in the field of mathematics at Queen Mary, University of London. His academic interests encompass various areas of mathematics, with a specific focus on probability and statistics. The research presented in this paper was carried out during his undergraduate years.

Dudley Stark received his Ph.D. from University of Southern California in 1994 and is a Reader (Associate Professor) in Mathematics and Probability at Queen Mary, University of London. His research interests lie in the fields of probability and combinatorics. He enjoys teaching a variety of modules in financial mathematics, statistics, and pure mathematics.

Abstract

The Law of Small Numbers states that the Binomial distribution converges to the Poisson distribution. Using the programming language R, we investigate the total variation distance between Binomial$(n, c/n)$ and Poisson(c) when we fix c and n individually. We also look at the asymptotics for nd_{TV} for a fixed c, where nd_{TV} is the total variation distance d_{TV} multiplied by increasing values of n. Several properties of d_{TV} are looked at in this paper.

1 Introduction

'The Law of Small Numbers' is a book written by Ladislaus von Bortkiewicz [1]. Quine and Seneta [2] state that there is much misconception about the book and its contents. Assume we have a short series of N independent observations with a Poisson(λ_i) for $i \in \{1,\ldots,N\}$. Bortkiewicz found that these observations act as if they are from a sample of size N with a Poisson distribution, even with unequal λ_i's. It is known that in certain circumstances, the Binomial distribution converges to the Poisson distribution.

***Corresponding author:** d.s.stark@qmul.ac.uk

In our study, we will be using Binomial$(n, c/n)$ and Poisson(c) for our investigation of the total variation distance between the two distributions.

Definition 1. Total variation distance measures the closeness between two distributions. The distance is defined by

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \frac{1}{2}\sum_{j=0}^{\infty} |P(X = j) - P(Y = j)|$$

where X and Y are discrete random variables and $\mathscr{L}(X)$ and $\mathscr{L}(Y)$ denotes their distributions. The state space of these discrete random variables is $\{0, 1, 2, \dots\}$. It is an important statistical distance measure, which in layman's terms measures the difference between two probability distributions. It is part of a wider field that too measures the difference between two probability distributions, called 'f-divergence' [3].

We wished to find higher order expansions of the total variation distance, but this was not possible using the programming language R. Instead, we look at the first order asymptotics for nd_{TV}, where c is fixed and n is increasing. This paper reports on several plots of the total variation distance for the law of small numbers.

For interested readers, we review a number of properties of the total variation distance which include calculating d_{TV} as a finite sum and the metric axioms. We also provide the R code of our plots, if a reader would like to use the code in their own research.

In our research, we have looked at several scenarios for our calculation of the total variation distance. We manipulated n and c to observe d_{TV}, and later nd_{TV}. One scenario involves finding a convergence towards a certain value of the total variation distance multiplied by n. nd_{TV} is this scenario, where c is fixed and $n \to \infty$. The convergence can be seen when n is small and in some cases when n is large, however this depends on the value of c.

2 Properties of the Total Variation Distance

The following properties seen in this section are important when it comes to calculating the total variation distance. They support the validity of our results and are part of the research project conducted.

2.1 d_{TV} as a finite sum

The total variation distance can be found in a shorter way, by looking at only the positive parts of the original formula. The new method of calculating the total variation distance will only look at a finite number of terms, hence why it is shorter.

Definition 2. A real function $f(x)$ can be split into two parts,

$$f^{+}(x) = \begin{cases} f(x), & \text{if } f(x) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$f^{-}(x) = \begin{cases} -f(x), & \text{if } f(x) < 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $f(x) = f^{+}(x) - f^{-}(x)$.

Theorem 1. The total variation distance (d_{TV}) can be found by looking only at the positive parts

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \sum_{j=0}^{n} (P(X=j) - P(Y=j))^{+}.$$

Proof: We can separate functions into two different parts, one being the positive part and the other being the negative part. The positive part looks only at the positive values of a function and vice versa for the negative part.

By **Definition 2.**,

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \frac{1}{2} \sum_{j=0}^{\infty} |P(X=j) - P(Y=j)|$$

$$|P(X=j) - P(Y=j)| = ((P(X=j) - P(Y=j))^{+} + (P(X=j) - P(Y=j))^{-} \tag{3}$$

$$(P(X=j) - P(Y=j)) = ((P(X=j) - P(Y=j))^{+} - (P(X=j) - P(Y=j))^{-} \tag{4}$$

The distributions are discrete and we assume that we have a fixed n number of trials.

Looking at (3),

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \frac{1}{2} \sum_{j=0}^{n} (P(X=j) - P(Y=j))^{+} + \frac{1}{2} \sum_{j=0}^{n} (P(X=j) - P(Y=j))^{-}$$

$$2 \cdot d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \sum_{j=0}^{n} (P(X=j) - P(Y=j))^{+} + \sum_{j=0}^{n} (P(X=j) - P(Y=j))^{-}$$

Now looking at (4),

$$\sum_{j=0}^{n}(P(X=j)-P(Y=j))=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{+}-\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{-}$$

Left hand side,

$$\sum_{j=0}^{n}P(X=j)-\sum_{j=0}^{n}P(Y=j)=1-1=0.$$

Implying that,

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{+}=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{-}.$$

From this, we can now calculate the distance as follows

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{+}.$$

2.2 Metric Axioms

Theorem 2. The total variation distance (d_{TV}) is a metric on the space of distributions.
Proof: We review the metric axioms in order to prove this theorem.

Definition 3. A metric satisfies three axioms,

(i) Non-negativity, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))\geq 0$

(ii) The identity of indiscernibles, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=0$ if and only if $\mathscr{L}(X)=\mathscr{L}(Y)$

(iii) Symmetry, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=d_{TV}(\mathscr{L}(Y),\mathscr{L}(X))$

(iv) The triangle inequality, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Z))\leq d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))+d_{TV}(\mathscr{L}(Y),\mathscr{L}(Z))$

Axiom 1: Straightforward from the definition of total variation distance, notice that the absolute value ensures that we take the positive result.

Axiom 2: For the identity of indiscernibles, it can be easy to see that $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=0$ if and only if $\mathscr{L}(X)=\mathscr{L}(Y)$. We know that $|P(X=j)-P(Y=j)|\geq 0$ for all j. We need the left hand side to be equal to zero, so we make $P(X=j)=P(Y=j)$. The two distributions are then identical, $\mathscr{L}(X)=\mathscr{L}(Y)$, so $|P(X=j)-P(X=j)|=0$.

Axiom 3: For the axiom of symmetry, we see that $|P(X=j)-P(Y=j)|\geq 0$. If we were to swap $P(X=j)$ and $P(Y=j)$ around for $|P(Y=j)-P(X=j)|\geq 0$, the absolute value allows for the same result.

Axiom 4: For the triangle inequality, we need to show that $d_{TV}(\mathscr{L}(X),\mathscr{L}(Z))\leq d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))+d_{TV}(\mathscr{L}(Y),\mathscr{L}(Z))$. We can first display

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Z))=\frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Z=j)|.$$

We can then see the following

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Z)) \leq \frac{1}{2}\sum_{j=0}^{\infty}\left\{|P(X=j)-P(Y=j)|+|P(Y=j)-P(Z=j)|\right\}.$$

The right hand side can be separated into

$$\frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)|+\frac{1}{2}\sum_{j=0}^{\infty}|P(Y=j)-P(Z=j)|.$$

So then

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Z)) \leq d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))+d_{TV}(\mathscr{L}(Y),\mathscr{L}(Z)).$$

2.3 d_{TV} as a finite sum

Theorem 3. The total variation distance (d_{TV}) has property $0 \leq d_{TV} \leq 1$.

Proof: We have already seen that $d_{TV} \geq 0$. By **Definition 1.** the total variation distance (d_{TV}) is defined by

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y)) = \frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)|$$

The absolute value being taken ensures that there are no negative values, so we know that $d_{TV} \geq 0$. To find out why d_{TV} does not exceed 1, we have

$$\frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)| \leq \frac{1}{2}\sum_{j=0}^{\infty}(P(X=j)+P(Y=j))$$

Looking at the right hand side,

$$\frac{1}{2}\sum_{j=0}^{\infty}P((X=j)+P(Y=j)) = \frac{1}{2}\sum_{j=0}^{\infty}P(X=j)+\frac{1}{2}\sum_{j=0}^{\infty}P(Y=j)$$

$$\frac{1}{2}\sum_{j=0}^{\infty}P(X=j)+\frac{1}{2}\sum_{j=0}^{\infty}P(Y=j) = \frac{1}{2}+\frac{1}{2} = 1$$

It is seen that $0 \leq d_{TV} \leq 1$.

A value of 0 for the total variation distance tells us that both random variables are identical. This is seen from our first metric axiom, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y)) = 0$ if and only if $\mathscr{L}(X) = \mathscr{L}(Y)$.

A value of 1 can be seen as the opposite of having the total variation distance equal to 0. For a value of 1, both random variables have disjoint sample spaces.

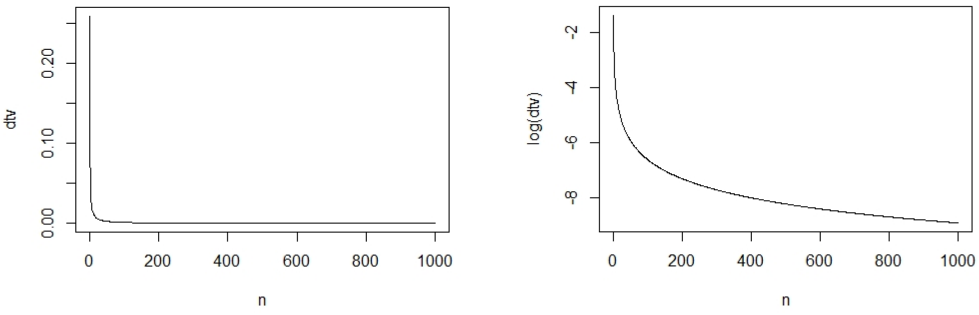
3 Main Results

We use the programming language R [4] to produce our results. The function 'TotalVarDist' from the package 'distrEx' [5] is used to calculate the total variation distance without having to write longer pieces of code. Our random variables X and Y remain the same throughout, with $X \sim \text{Binomial}(n, c/n)$ and $Y \sim \text{Poisson}(c)$.

3.1 d_{TV} for multiple values of n

We keep c as a fixed value here and we make n increase seemingly to infinity. Kennedy and Quine [6] have found an exact expression for $n \geq 1$ and $0 < np \leq 2 - \sqrt{2}$. They call this $f_1(p)$ in their paper, where $p = \frac{c}{n}$ in our calculations

$$f_1\left(\frac{c}{n}\right) = c\left(1 - \frac{c}{n}\right)^{n-1} - ce^{-c}.$$



(a) d_{TV}

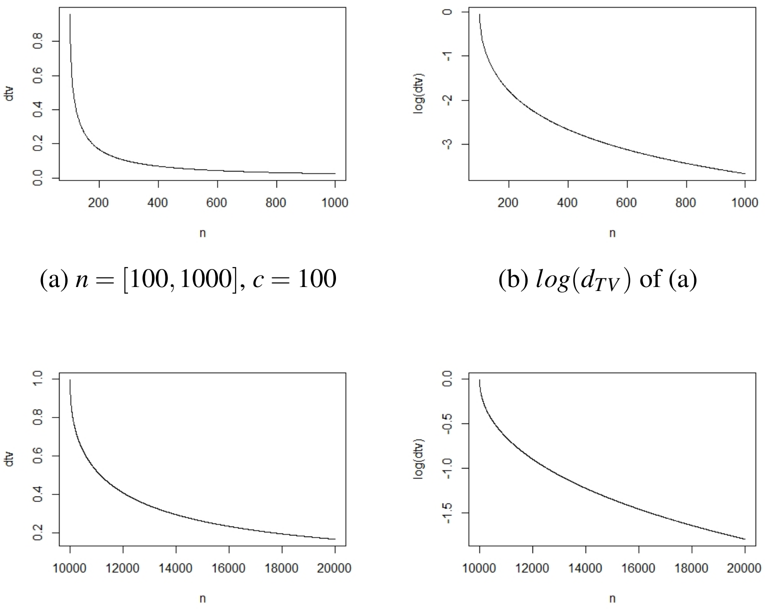
(b) $log(d_{TV})$

Figure 1: Plot for $c = 2 - \sqrt{2}$ for increasing values of n

R Code for Figure 1:

```
#dtv for multiple n
library(distrEx)
library(distr)
n = 1
results1 = list()
while (n <= 1000){
  c = 2 - sqrt(2)
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  results1 = c(results1, x)
  #once n = 1000 in the program, produce a plot of the results
  if (n == 1000){
    plot(unlist(results1), type="l", xlab = "n", ylab = "dtv")}
n = n + 1}
```

We see here that as n increases, we move rapidly to a total variation distance of zero. Changing the value of c (and n where required) has little effect to the form of the plot shown in Figure 1. Additional plots are shown in Figure 2 for different c.



(a) $n = [100, 1000]$, $c = 100$

(b) $log(d_{TV})$ of (a)

(c) $n = [10000, 20000]$, $c = 10000$

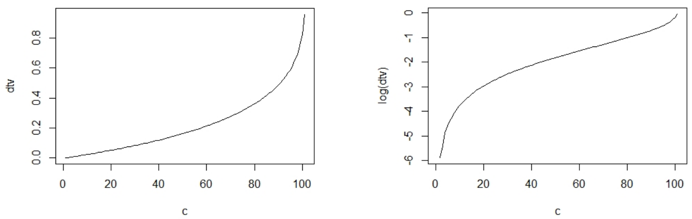
(d) $log(d_{TV})$ of (c)

Figure 2: Plots of different values of c

Comparing d_{TV} to $log(d_{TV})$, we see a similar behaviour in the plots. This behaviour is consistent at higher values of n and c, as seen by the plots. Therefore, as we increase n for a fixed c, the total variation distance decreases in value close to zero for larger values of n.

3.2 d_{TV} for multiple values of c

We keep n as the fixed value here and we increase c to a certain value of n. We cannot make c a greater value than n because that would make the probability c/n greater than one. Figure 3 shows a plot for $0 \leq c \leq 100$ and $n = 100$.



(a) d_{TV}

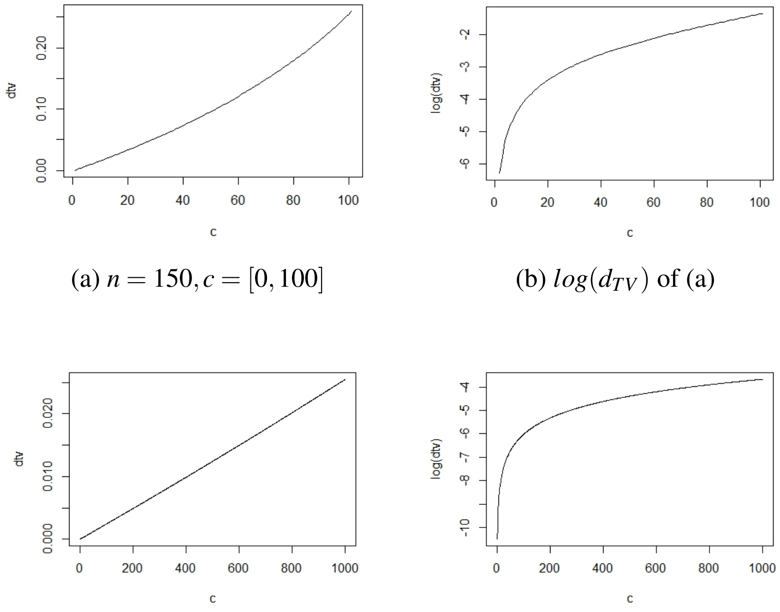
(b) $log(d_{TV})$

Figure 3: Plot for $n = 100$ for increasing values of c

R Code for Figure 3:

```
[language=R]
#dtv for multiple c
library(distrEx)
library(distr)
c = 0
results2 = list()
while (c <= 100){
  n = 100
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  results2 = c(results2, x)
  #once c = 100 in the program, produce a plot of the results
  if (c == 100){
    plot(unlist(results2), type="l", xlab = "c", ylab = "dtv")}
c = c + 1}
```

The shape of this plot in Figure 3 will not be consistent for all values of *n* (and *c* where required). When *n* is equal to the largest value of *c*, we see the behaviour of an exponential function. As *n* increases with the same boundaries for *c*, the plot eventually becomes linear. This is visualised in Figure 4. However, it remains that as *c* increases for a fixed *n* that the total variation distance increases.



(a) $n = 150, c = [0,100]$

(b) $log(d_{TV})$ of (a)

(c) $n = 10000, c = [0,1000]$

(d) $log(d_{TV})$ of (c)

Figure 4: Plots of different values of *n*

3.3 nd_{TV} for multiple values of n

This builds up on Section 3.1, where c is fixed and n is increasing. nd_{TV} has similar behaviour to what we had done in Section 3.1, as n increases we have a smaller total variation distance. Figure 5 shows a plot $7 \leq n \leq 1000$ and $c = 7$.

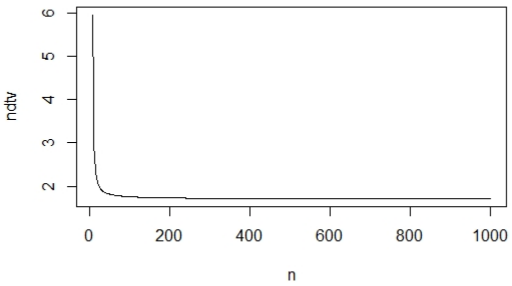


Figure 5: Plot for nd_{TV} for multiple values of n

R Code for Figure 5:

```
[language=R]
#ndtv for multiple values of n
library(distrEx)
library(distr)
n = 7
xaxis = seq(7,1000)
results3 = list()
while (n <= 1000){
  c = 7
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  nx = n*x
  results3 = c(results3, nx)
  if (n == 1000){
    plot(xaxis, unlist(results3), type="l", xlab = "n",
    ylab = "ndtv")} n = n + 1}
```

On the plot, it seems that there is a convergence towards a certain value. A result from a research paper by Prokhorov [7] tells us that

$$\lim_{n\to\infty} nd_{TV} = \frac{c}{\sqrt{2e\pi}}.$$

We investigated this and found that such a convergence exists for this particular plot at approximately 1.703 to 3 decimal places. It appears that convergence is a common feature in plots of this kind, with its visibility increasing at sufficiently large enough n. Our plot agrees with Prokhorov's result and shows that nd_{TV} is quite rapid.

3.4 nd_{TV} for multiple values of c

This builds up on Section 3.2, where we have a fixed n and c is increasing. As with Section 3.2, as we increase c we have a larger total variation distance. Figure 6 shows a plot for $0 \leq c \leq 1000$ and $n = 1000$.

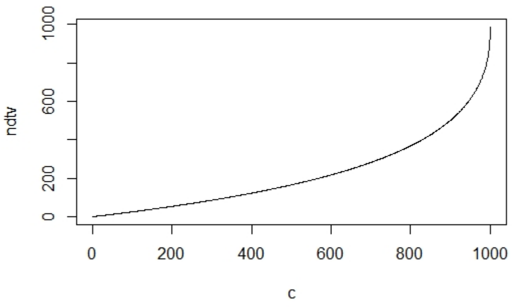


Figure 6: Plot for nd_{TV} for multiple values of c

R Code for Figure 6:

```
[language=R]
#ndtv for multiple values of c
library(distrEx)
library(distr)
c = 0
results4 = list()
while (c <= 1000){
  n = 1000
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  nx = n*x
  results4 = c(results4, nx)
  if (c == 1000){
    plot(unlist(results4), type="l", xlab = "c",
    ylab = "ndtv")} c = c + 1}
```

If we were to keep n constant, there would be no clear convergence here. The final value of nd_{TV} is 987.382 at $n = 1000$ and $c = 1000$, having $c > 1000$ would cause the probability c/n in the Binomial to be greater than one. We cannot continue calculating the total variation distance here and so we conclude that there is no convergence for a fixed n. For large enough n, we will find convergence.

3.5 The second term in the asymptotic expansion of d_{TV}

Definition 4. We have functions $f(i)$ and $g(i)$. We introduce the “little-o" notation for a function $o(g(i))$, where informally we can say “little o of g of i". $f(i) = o(g(i))$ means that

$$\forall k > 0 \, \exists m > 0 \, \forall i \geq m : 0 \leq f(i) < kg(i). \tag{5}$$

The value of m does not depend on i, but it may depend on k [8].

$f(i) = o(g(i))$ also means that

$$\lim_{i\to\infty} \frac{f(i)}{g(i)} = 0 \quad [9]. \tag{6}$$

We now define an asymptotic sequence using the “little-o" notation. A finite or infinite sequence of functions $\phi_i(z)$, $i = 1,2,\ldots$ and $z \in \mathbb{C}$ is defined to be an asymptotic sequence as $z \to z_0$ if,

$$\phi_{i+1}(z) = o(\phi_i(z)) \tag{7}$$

and also that $\lim_{z\to z_0} \frac{\phi_{i+1}(z)}{\phi_i(z)} = 0$.

From our definition of an asymptotic sequence, we say that $\sum_{i=1} a_i\phi_i(z)$, where the a_i are constants, is an asymptotic expansion or an asymptotic approximation for a function $f(z)$ if for every N

$$f(z) = \sum_{i=1}^{N} a_i\phi_i(z) + o(\phi_N(z)) \quad [10]. \tag{8}$$

The following displays a recursive method for finding the first two terms of the asymptotic expansion for d_{TV}:

$$\begin{aligned}
\phi_1(n) &= \frac{1}{n} \\
\phi_2(n) &= \frac{1}{n^2} \\
\\
a_1 &= \lim_{n\to\infty} n d_{TV} = \frac{c}{\sqrt{2e\pi}} \\
a_2 &= \lim_{n\to\infty} \left(n^2 d_{TV} - a_1 n\right) \qquad (9) \\
d_{TV} &\approx \frac{a_1}{n} + \frac{a_2}{n^2}
\end{aligned}$$

With $c = 7$, the convergence in (9) certainly occurs at $n = 100000$. The plot seen in Section 3.3 clearly indicates a convergence towards a certain value and it was found for $7 \leq n \leq 1000$. Further calculations using R prove that the error of convergence is very small, so we can use $n = 100000$. Therefore, we choose this as our n. We now substitute our values into the equations above.

$$
\begin{aligned}
\phi_1(100000) &= \frac{1}{100000} = 0.00001 \\
\phi_2(100000) &= \frac{1}{(100000)^2} = 0.0000000001 \\
a_1 &= \frac{7}{\sqrt{2e\pi}} = 1.6937950716 \\
a_2 &\doteq d_{TV}(100000)^2 - 100000a_1 = 945.68
\end{aligned}
$$

Thus, our prediction is that a_2 is about 946 for $c = 7$.

4 Conclusion

Our program allowed us to see how the total variation distance behaves for our parameters and distributions. We proved the property that the total variation distance can be written as a finite sum. Two other properties were also looked at, being that the distance is on the metric space and that $0 \leq d_{TV} \leq 1$. We have managed to find the asymptotics for a scenario with nd_{TV}.

In the future, we hope to address the following:

- The convergence is very fast for nd_{TV} where c is fixed. A program with more precision would be useful when finding say a_2.
- We would like to extend the paper by Prokhorov [7] further by finding higher order expansions.

5 Acknowledgements

Yasir Barlas is appreciative to Queen Mary, University of London, the ‘Faculty of Science and Engineering’ and the ‘School of Mathematical Sciences’ for being selected on the internship programme for this project. The ‘BAME Undergraduate Research Internship’ is sponsored by IBM and allows for undergraduate students to be involved in a research project. We are grateful to the referee for the recommended revisions during the publication process.

Bibliography

[1] Bortkiewicz, L. von. (1898) Das Gesetz der kleinen Zahlen [The law of small numbers]. Available at: `https://archive.org/details/dasgesetzderklei00bortrich/`.

[2] Quine M. P. and Seneta E. (1987) Bortkiewicz’s data and the law of small numbers. *International Statistical Review / Revue Internationale de Statistique*, **55**(2):173–181, 1987.

[3] Sason, I. (2018) On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, **20**(5):383.

[4] R Core Team. (2021) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.*

[5] Ruckdeschel P., Kohl M., Stabla T. and Camphausen F. (2006) S4 classes for distributions. *R News*, **6**(2):10-13.

[6] Kennedy, J. E., and Quine, M. P. (1989) The total variation distance between the binomial and Poisson distributions. *The Annals of Probability*, **17**(1).

[7] Prokhorov, Yu. V. (1953) Asymptotic behavior of the binomial distribution. *Uspekhi Mat. Nauk*, **8**(3(55)):135-142.

[8] Black, P. E. (2019) little-o notation. *Dictionary of Algorithms and Data Structures* [online]. Available at: `https://www.nist.gov/dads/HTML/littleOnotation.html`.

[9] de Bruijn, N. G. (1981) Asymptotic Methods in Analysis. *Bibliotheca mathematica, Dover Publications.*

[10] Murray, J. D. (1984) Asymptotic Analysis: Asymptotic expansions. *Applied Mathematical Sciences*, **48**:1-18.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 15 – 33

Classification of seven-dimensional solvable Lie algebras with five-dimensional abelian nilradicaly

*Jacksyn Bakeberg, Kate Blaine, Firas Hindeleh**

Jacksyn Bakeberg is a Ph.D candidate in mathematics at Boston University. He is interested in number theory, arithmetic geometry, and representation theory. He completed his undergraduate studies in mathematics and Arabic language at McGill University.

Kate Blaine graduated from Bard College in 2019 with a Bachelor's degree in Mathematics. She also majored in music. Since graduating, she has been taking actuarial exams to work towards the ASA credential, and she works as a Risk Adjustment Analyst at Excellus BlueCross BlueShield.

Firas Hindeleh is an Associate Professor of Mathematics at Grand Valley State University. His focus scholarly area is the classification problem for low dimensional Lie algebras. He is a strong advocate for inclusive teaching and learning.

Abstract

This paper provides a classification of seven-dimensional indecomposable solvable Lie algebras over $\mathbb{R}$ for which the nilradical is five-dimensional and abelian. We follow a technique that was first introduced by Mubarakzyanov.

1 Introduction

For the elementary theory of Lie algebras refer to [4, 6, 7]. It has to be understood that classifying solvable Lie algebras is a different exercise from studying the semisimple

*Corresponding author: hindelef@gvsu.edu

algebras. The problem of classifying all semisimple Lie algebras over the field of complex numbers was solved by Cartan in 1894 [1], and over the field of real numbers by Gantmacher in 1939 [2]. For solvable indecomposable Lie algebras the problem is much more difficult. The classification of solvable Lie algebras only exists for low dimensions and was performed by, amongst others, Mubarakzyanov for solvable Lie algebras of dimension $n \leq 5$ over the field of real and partially over the field of complex numbers in [11] and [12]. Mubarakzyanov's results are summarized in [17]. Mubarakzyanov also considered dimension six and classified solvable Lie algebras with a co-dimension one nilradical [13]. Shabanskaya and Thompson refined his results and found some missing cases in [19, 20]. Then Turkowski classified six-dimensional solvable Lie algebras with a co-dimension two nilradical in [21]. Nilpotent Lie algebras in dimension six were studied as far back as Umlauf [22], and later by Morozov [9].

It is probably impossible to classify solvable Lie algebras in general in arbitrary dimension. The first step in classifying solvable Lie algebras in a specific dimension is to find the possible nilradicals. A general theorem asserts that if $\mathfrak{g}$ is an $n-$dimensional solvable Lie algebra, the dimension of its nilradical $\mathfrak{nil}(\mathfrak{g})$ is at least $\frac{n}{2}$ [13]. So for $n = 7$, the possible dimensions of the nilradical are seven, six, five, or four. The seven-dimensional nilradicals, called the nilpotents, were studied by Seely over $\mathbb{R}$ [18] and by Gong over $\mathbb{C}$ [3]. The four-dimensional nilradical case was studied by Hindeleh and Thompson [5]. The six-dimensional nilradical case was studied by Parry [16]. The five-dimensional nilradical case is still an open problem. A complete classification consists of all possible five-dimensional nilpotent algebras, including the decomposable ones. In this paper we study the case where the nilradical is the five-dimensional abelian algebra $\mathbb{R}^5$.

We note that Ndogmo and Winternitz outlined methods for classifying solvable Lie algebras with abelian nilradical for a general dimension in [14, 15]. Also, while this work was being finalized, Le, Vu A, et al. [8] posted in arXiv methods for the classification of seven-dimensional Lie algebras with five-dimensional nilradical. They conclude with the number of possible algebras without explicitly finding them. This paper provides a complete list of the seven-dimensional solvable Lie algebras with a five-dimensional abelian nilradical.

In section , we recall basic definitions and properties related to the classification of solvable Lie algebras. Then in section , we use Turkowski's method [21] for classifying solvable Lie algebras with abelian nilradical, that is also outlined by Ndogmo and Winternitz [14, 15]. Finally, we list the adjoint matrices corresponding to our algebras with trivial and one-dimensional centers in sections 3.1 and 3.2, respectively. The complete list of algebras can be found in tables 1, and 2.

2 A Method to Obtain the Solvable Algebras

2.1 General Concepts

A Lie algebra $\mathfrak{g}$ is *solvable* if its derived series *DS* terminates, i.e.

$$DS = \{\mathfrak{g}_0 = \mathfrak{g}, \mathfrak{g}_1 = [\mathfrak{g}, \mathfrak{g}], \ldots, \mathfrak{g}_k = [\mathfrak{g}_{k-1}, \mathfrak{g}_{k-1}] = 0\}$$

for some $k \geq 1$.

A Lie algebra $\mathfrak{g}$ is *nilpotent* if its central series CS terminates, i.e.

$$CS = \{\mathfrak{g}^{(0)} = \mathfrak{g}, \mathfrak{g}^{(1)} = [\mathfrak{g},\mathfrak{g}], \ldots, \mathfrak{g}^{(k)} = [\mathfrak{g},\mathfrak{g}^{(k-1)}] = 0\}$$

for some $k \geq 1$.

A solvable algebra $\mathfrak{g}$ has a decomposition of the form

$$\mathfrak{g} = \mathfrak{nil}(\mathfrak{g}) \oplus X,$$

satisfying

$$\begin{aligned} [\mathfrak{nil}(\mathfrak{g}), \mathfrak{nil}(\mathfrak{g})] &\subset \mathfrak{nil}(\mathfrak{g}), \\ [\mathfrak{nil}(\mathfrak{g}), X] &\subseteq \mathfrak{nil}(\mathfrak{g}), \\ [X,X] &\subset \mathfrak{nil}(\mathfrak{g}), \end{aligned} \tag{1}$$

where $\mathfrak{nil}(\mathfrak{g})$ denotes the nilradical of $\mathfrak{g}$, the vector space X is spanned by the remaining generators, and $\oplus$ denotes the direct sum of vector spaces.

An element n of $\mathfrak{g}$ is *nilpotent* if it satisfies

$$[\ldots[[x,n],n]\ldots n] = 0$$

for all $x \in \mathfrak{g}$ when the commutator is taken sufficiently many times.

A set of elements $\{x_1,\ldots,x_k\}$ of $\mathfrak{g}$ is called *nilindependent* if no non-trivial linear combination of them is nilpotent.

For $x \in \mathfrak{g}$, the *adjoint transformation* of x is a linear transformation $ad_x : \mathfrak{g} \to \mathfrak{g}$ defined by

$$ad_x(y) = [x,y],$$

for all $y \in \mathfrak{g}$. In this paper, the restriction of ad_x to the nilradical of $\mathfrak{g}$ denoted $ad_x|_{\mathfrak{nil}(\mathfrak{g})}$ is realized by matrices $A \in gl(5,\mathbb{R})$. Notice that if n is a nilpotent element of $\mathfrak{g}$, then $ad_n|_{\mathfrak{nil}(\mathfrak{g})}$ is a nilpotent matrix.

A set of matrices in $gl(n,\mathbb{R})$ will be called *linearly nilindependent* if no non-trivial linear combination of them is nilpotent.

2.2 Basic Structural Theorems

We shall choose a basis for $\mathfrak{g} = \langle e_1, e_2, \ldots, e_5, x_1, x_2\rangle$ where $e_i \in \mathfrak{nil}(\mathfrak{g}), x_\alpha \in X$, for $i = 1,\ldots,5$, and $\alpha = 1,2$.

To classify the seven-dimensional solvable Lie algebras with five-dimensional nilradical, one must start with a five-dimensional nilpotent algebra that will form $\mathfrak{nil}(\mathfrak{g})$, and add $X = \langle x_1, x_2\rangle$ satisfying the properties in (5). The following are all the nilpotent Lie algebras up to isomorphism in dimension five: $\mathbb{R}^5$, $A_{3,1} \oplus \mathbb{R}^2$, $A_{4,1} \oplus \mathbb{R}$, and $A_{5,1} - A_{5,6}$, where $\mathbb{R}^n$ denotes the $n-$dimensional abelian algebra, and $A_{n,k}$ denotes the k^{th} algebra of dimension n from Patera's list [17]. The focus of this article is on the first case, namely $\mathfrak{nil}(\mathfrak{g}) = \mathbb{R}^5$.

Since the nilradical is abelian and basis elements must satisfy the relations in (5), we have

$$[e_i, e_j] = 0 \tag{2a}$$

$$\begin{pmatrix} [x_\alpha, e_1] \\ \vdots \\ [x_\alpha, e_5] \end{pmatrix} = \begin{pmatrix} e_1 & \dots & e_5 \end{pmatrix} A^\alpha \tag{2b}$$

$$[x_1, x_2] = R^i e_i \tag{2c}$$

where $A^\alpha = ad_{x_\alpha}|_{\mathfrak{nil}(\mathfrak{g})}$, $\alpha = 1,2$ and $i, j = 1, \dots, 5$ and we use the Einstein summation notation. The classification of our Lie algebras thus amounts to classification of the matrices A^α and the constants R^i.

By the Jacobi identity involving x_1, x_2, and an e_i,

$$[[x_1,x_2],e_i] + [[x_2,e_i],x_1] + [[e_i,x_1],x_2] = 0.$$

Thus

$$\begin{aligned} ad_{[x_1,x_2]}(e_i) &= [x_1,[x_2,e_i]] - [x_2,[x_1,e_i]] \\ &= ad_{x_1}([x_2,e_i]) - ad_{x_2}([x_1,e_1]) \\ &= ad_{x_1}(ad_{x_2}(e_i)) - ad_{x_2}(ad_{x_1}(e_i)) \\ &= [ad_{x_1}, ad_{x_2}](e_i). \end{aligned}$$

Hence $[ad_{x_1}, ad_{x_2}]$ is an inner derivation of the nilradical. Since the nilradical is abelian, we have

$$[A^1, A^2] = 0. \tag{3}$$

Also, since $x_\alpha \notin \mathfrak{nil}(\mathfrak{g})$, then A^α cannot be nilpotent. In fact A^1 and A^2 are linearly nilindependent and commute pairwise.

We perform a combination of changes of basis until we reach our desired form. For $i = 1, \dots, 5$, and $\alpha = 1,2$, the following changes of basis preserve the nilradical:

(i) Absorbtion-type change of basis

$$\bar{x}_\alpha = x_\alpha + r^i_\alpha e_i \qquad r^i_\alpha \in \mathbb{R}.$$

(ii) A change of basis in X

$$\bar{x}_\alpha = G^\beta_\alpha x_\beta \qquad G \in GL(2,\mathbb{R}).$$

(iii) A change of basis in $\mathfrak{nil}(\mathfrak{g})$

$$\bar{e}_i = S^j_i e_j \qquad S \in GL(5,\mathbb{R}),$$

where $S = (S^j_i)$ is the automorphism that will change every A^α to a similar matrix $SA^\alpha S^{-1}$.

Thus our classification problem reduces to finding the derivations of the nilradical that are not nilpotent and that satisfy equation (3).

3 Classes of Solvable Algebras

We will determine all real solvable algebras $N = \mathbb{R}^5 \oplus X$ such that the $\dim X = 2$. The dimension of the center of $\mathfrak{g}$ is

$$\dim Z(\mathfrak{g}) \leq 2 \dim \mathfrak{nil}(\mathfrak{g}) - \dim \mathfrak{g} = 3$$

(see Ref. [10]). The algebras that possess a center of dimension at least two are decomposable into a direct sum of lower-dimensional algebras [10]. Therefore, in the following, the classification problem is solved for the cases $\dim Z(\mathfrak{g}) = 0, 1$.

The derivation matrices A^{α} form an abelian subalgebra of $gl(5,\mathbb{R})$ and hence a subalgebra of some maximal abelian subalgebra. This maximal abelian subalgebra cannot be a maximal abelian nilpotent subalgebra; as a matter of fact it contains no nilpotent elements at all. In his Ph.D. dissertation, Ndogmo [14] (and later in [15]) outlines a technique to find the equivalent classes of nilindependent derivations $\{A^1, A^2\}$. What we mean by "equivalent classes" is

(i) The pair $\{y_1^1 A^1 + y_2^1 A2, y_1^2 A^1 + y_2^2 A2\}$, where $y_1^1 y_2^2 - y_2^1 y_1^2 \neq 0$, is equivalent to $\{A^1, A^2\}$.

(ii) The pair $\{SA^1S^{-1}, SA^2S^{-1}\}$, where S is the automorphism of the nilradical, is equivalent to $\{A^1, A^2\}$.

Using Ndogmo and Winternitz's notation, we list our A^1, A^2 in block-diagonal form by the dimension of each block. Namely, the

$$(u_1 u_2 \cdots u_i, v_1 v_2 \cdots v_j, w_1)$$

partition consists if i real lower triangular blocks of dimension $u_i \times u_i$, j real blocks of complex conjugate type of dimension $v_j \times v_j$ (v_j is even), and w_1 stands for the dimension of a one-dimensional zero block. Note that for our cases $w_1 = 0$ for the trivial center cases, and $w_1 = 1$ for the one-dimensional center case. In our case $\sum u_i + \sum v_j + w_1 = 5$.

3.1 Algebras with trivial center

For all of the block partitions in this section, we were able to find a change of basis that produces $[x_1, x_2] = 0$.

The Jacobi identity can give a nonlinear homogeneous system of equations on the free parameters. Each solution to that system will give a subcase for that partition. We list the matrices $A^{\alpha} = ad_{x_{\alpha}}|_{\mathfrak{nil}(\mathfrak{g})}$ for each case or subcase and give the conditions on those free parameters. We summarize our list of algebras in Table 1, suppressing the conditions.

(i) The $(11111, 0, 0)$ partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 0 & a_4 & 0 \\ 0 & 0 & 0 & 0 & a_5 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & 0 & b_4 & 0 \\ 0 & 0 & 0 & 0 & b_5 \end{pmatrix}.$$

To ensure a trivial center and an indecomposable algebra, we need $a_i^2 + b_i^2 \neq 0$ for $i = 3, 4, 5$. We denote this algebra by $N_{7,1}$.

(ii) The (1112,0,0) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 0 & a_4 & 0 \\ 0 & 0 & 0 & p_1 & a_4 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & 0 & b_4 & 0 \\ 0 & 0 & 0 & p_2 & b_4 \end{pmatrix}.$$

To ensure a trivial center and an indecomposable algebra, we need $a_i^2 + b_i^2 \neq 0$ for $i = 3, 4$. We denote this algebra $N_{7,2}$.

(iii) The (111,2,0) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 0 & b_1 & c_1 \\ 0 & 0 & 0 & -c_1 & b_1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & 0 & b_2 & c_2 \\ 0 & 0 & 0 & -c_2 & b_2 \end{pmatrix}.$$

To ensure a trivial center and an indecomposable algebra as well as a complex block, we need $a_3^2 + b_3^2 \neq 0$ and $c_1^2 + c_2^2 \neq 0$. We denote this algebra $N_{7,3}$.

(iv) The (122, 0, 0) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_3 & 0 \\ 0 & 0 & 0 & p_2 & a_3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & b_3 & 0 \\ 0 & 0 & 0 & q_2 & b_3 \end{pmatrix}.$$

To ensure a trivial center and an indecomposable algebra, we need $a_3^2 + b_3^2 \neq 0$.

We denote this algebra $N_{7,4}$.

(v) The (12, 2, 0) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra $N_{7,5}$.

(vi) The (1, 22, 0) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_1 & 0 \\ 0 & 0 & 0 & 0 & b_1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix}.$$

We denote this algebra $N_{7,6}$.

(vii) The (113,0,0) partition

For this partition, the homogeneous nonlinear system imposed by the Jacobi identity has three independent solutions. Each solution will give us a subcase below. For all the subcases, we need $a_3^2 + b_3^2 \neq 0$ to ensure a trivial center and an indecomposable algebra.

(a) The first solution requires $p_1 = q_1 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 0 & a_3 & 0 \\ 0 & 0 & p_2 & p_3 & a_3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & 0 & b_3 & 0 \\ 0 & 0 & q_2 & q_3 & b_3 \end{pmatrix}.$$

We denote this algebra by $N_{7,7}$.

(b) The second solution requires $p_1 = p_3 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 0 & a_3 & 0 \\ 0 & 0 & p_2 & 0 & a_3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & q_1 & b_3 & 0 \\ 0 & 0 & q_2 & q_3 & b_3 \end{pmatrix}.$$

We denote this algebra by $N_{7,8}$.

(c) The third solution requires $q_3 = p_3 q_1$ and $p_1 = 1$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 1 & a_3 & 0 \\ 0 & 0 & p_2 & p_3 & a_3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & q_1 & b_3 & 0 \\ 0 & 0 & q_2 & q_3 & b_3 \end{pmatrix}.$$

We denote this algebra by $N_{7,9}$.

(viii) The (14,0,0) partition

For this partition, the homogeneous nonlinear system imposed by the Jacobi identity has nine independent solutions. Each solution will give us a case below.

(a) The first solution requires $q_4 q_6 \neq 0$, $p_1 = \frac{q_1 p_4}{q_4}$, $p_2 = -\frac{p_4 q_1 q_5 - p_4 q_2 q_6 - p_5 q_1 q_4}{q_4 q_6}$, and $p_6 = \frac{p_4 q_6}{q_4}$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & p_4 & 0 & 0 \\ 0 & p_3 & p_5 & p_6 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & q_4 & 1 & 0 \\ 0 & q_3 & q_5 & q_6 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,10}$.

(b) The second solution requires $p_1 = p_6 = q_1 = q_6 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_2 & p_4 & 0 & 0 \\ 0 & p_3 & p_5 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & q_2 & q_4 & 1 & 0 \\ 0 & q_3 & q_5 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,11}$.

(c) The third solution requires $q_1 q_4 \neq 0$, $p_6 = q_6 = 0$, $p_1 = \frac{q_1 p_4}{q_4}$, and $p_5 = \frac{p_4 q_5}{q_4}$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & p_4 & 0 & 0 \\ 0 & p_3 & p_5 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & q_4 & 1 & 0 \\ 0 & q_3 & q_5 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,12}$.

(d) The fourth solution requires $q_1 q_5 \neq 0$, $p_4 = q_4 = 0$, and $p_1 = -\frac{p_2 q_6 - p_5 q_1 - p_6 q_2}{q_5}$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 \\ 0 & p_3 & p_5 & p_6 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & 0 & 1 & 0 \\ 0 & q_3 & q_5 & q_6 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,13}$.

(e) The fifth solution requires $q_6 \neq 0$, $p_4 = q_4 = q_5 = 0$, and $p_2 = \frac{p_5 q_1 + p_6 q_2}{q_6}$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 \\ 0 & p_3 & p_5 & p_6 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & 0 & 1 & 0 \\ 0 & q_3 & 0 & q_6 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,14}$.

(f) The sixth solution requires $q_1 \neq 0$, $p_4 = q_4 = q_5 = q_6 = 0$, and $p_5 = -\frac{q_2 p_6}{q_1}$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 \\ 0 & p_3 & p_5 & p_6 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & 0 & 1 & 0 \\ 0 & q_3 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,15}$.

(g) The seventh solution requires $p_6 q_5 \neq 0$, $q_1 = q_4 = q_6 = 0$, and $p_1 = \frac{q_2 p_6}{q_5}$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & p_4 & 0 & 0 \\ 0 & p_3 & p_5 & p_6 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & q_2 & 0 & 1 & 0 \\ 0 & q_3 & q_5 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,16}$.

(h) The eighth solution requires $p_6 \neq 0$ and $q_1 = q_2 = q_4 = q_5 = q_6 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & p_4 & 0 & 0 \\ 0 & p_3 & p_5 & p_6 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & q_3 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,17}$.

(i) The ninth solution requires $p_1 \neq 0$ and $p_6 = q_1 = q_4 = q_5 = q_6 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 & 0 \\ 0 & p_2 & p_4 & 0 & 0 \\ 0 & p_3 & p_5 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & q_2 & 0 & 1 & 0 \\ 0 & q_3 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,18}$.

(ix) The (23, 0, 0) partition

For this partition, the homogeneous nonlinear system imposed by the Jacobi identity has three independent solutions. Each solution will give us a case below.

(a) The first solution requires $p_2 = q_2 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_3 & p_4 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & q_3 & q_4 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,19}$.

(b) The second solution requires $p_2 = p_4 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_3 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & q_2 & 1 & 0 \\ 0 & 0 & q_3 & q_4 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,20}$.

(c) The third solution requires $q_4 = q_2 p_4$ and $p_2 = 1$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & p_3 & p_4 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & q_2 & 1 & 0 \\ 0 & 0 & q_3 & q_4 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,21}$.

(x) The (3, 2, 0) partition

For this partition, the homogeneous nonlinear system imposed by the Jacobi identity has three independent solutions. Each solution will give us a case below.

(a) The first solution requires $p_1 = q_1 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ p_2 & p_3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ q_2 & q_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,22}$.

(b) The second solution requires $p_1 = p_3 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ p_2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ q_2 & q_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,23}$.

(c) The third solution requires $q_3 = p_3q_1$ and $p_1 = 1$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ p_2 & p_3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ q_2 & q_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We denote this algebra by $N_{7,24}$.

3.2 Algebras with one-dimensional center

In this section, the center of the Lie algebra $Z(\mathfrak{g}) = \langle e_5 \rangle$. For all of the block partitions in this section, we were able to find a change of basis that produces $[x_1, x_2] = e_5$.

Similarly, the Jacobi identity can give a nonlinear homogeneous system of equations on the free parameters. Each solution to that system will give a subcase for that partition. We list the matrices $A^{\alpha} = ad_{x_{\alpha}}|_{\mathfrak{nil}(\mathfrak{g})}$ for each case or subcase and give the conditions on those free parameters. We summarize our list of algebras in Table 2, suppressing the conditions.

(i) The (1111, 0, 1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & 0 & a_4 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & 0 & b_4 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

To ensure a one-dimensional center and an indecomposable algebra, we need $a_i^2 + b_i^2 \neq 0$ for $i = 3,4$. We denote this algebra by $N_{7,25}$.

(ii) The (2, 2, 1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,26}$.

(iii) The (13, 0, 1) partition

For this partition, the homogeneous nonlinear system imposed by the Jacobi identity has three independent solutions. Each solution will give us a case below.

(a) The first solution requires $p_1 = q_1 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_2 & p_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & q_2 & q_3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,27}$.

(b) The second solution requires $p_1 = p_3 = 0$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & q_3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,28}$.

(c) The third solution requires $q_3 = p_3 q_1$ and $p_1 = 1$. For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & p_2 & p_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & q_1 & 1 & 0 & 0 \\ 0 & q_2 & q_3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,29}$.

(iv) The (22,0,1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & q_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,30}$.

(v) The (112,0,1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0 & 0 \\ 0 & 0 & p_1 & a_3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & q_1 & b_3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

To ensure a one-dimensional center and an indecomposable algebra, we need $a_3^2 + b_3^2 \neq 0$. We denote this algebra by $N_{7,31}$.

(vi) The (11,2,1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,32}$.

(vii) The (0, 4, 1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ p_1 & p_3 & 0 & 1 & 0 \\ p_2 & p_4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ q_1 & -q_2 & 1 & 0 & 0 \\ q_2 & q_1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,33}$.

(viii) The (0,22,1) partition

For this case our adjoint matrices are given by

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & b_2 & 0 & 0 \\ 0 & 0 & 0 & b_2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We denote this algebra by $N_{7,34}$.

4 Conclusion

This completes the classification for the seven-dimensional solvable Lie algebras with five-dimensional abelian nilradical. Significant progress has been made on the remaining five-dimensional nilradicals and will be submitted separately.

5 Acknowledgements

This work was partially supported by National Science Foundation grant DMS-1659113, which funds a Research Experiences for Undergraduates program at Grand Valley State University.

1 Bracket relations for the real seven-dimensional solvable Lie algebras with five-dimensional Abelian nilradical and a trivial center

Table 1: Non-zero bracket relations for the real seven-dimensional solvable Lie algebras with five-dimensional Abelian nilradical and a trivial center. The elements $\{e_1,\ldots,e_5\}$ form a basis for the nilradical and $\{x_1,x_2\}$ are the remaining basis elements.

		$[x_\alpha,e_1]$	$[x_\alpha,e_2]$	$[x_\alpha,e_3]$	$[x_\alpha,e_4]$	$[x_\alpha,e_5]$
2*$N_{7,1}$	$[x_1,e_i]$	e_1		a_3e_3	a_4e_4	a_5e_5
2-7	$[x_2,e_i]$		e_2	b_3e_3	b_4e_4	b_5e_5
2*$N_{7,2}$	$[x_1,e_i]$	e_1		a_3e_3	$a_4e_4+p_1e_5$	a_4e_5
2-7	$[x_2,e_i]$		e_2	b_3e_3	$b_4e_4+p_2e_5$	b_4e_5
2*$N_{7,3}$	$[x_1,e_i]$	e_1		a_3e_3	$b_1e_4-c_1e_5$	$c_1e_4+b_1e_5$
2-7	$[x_2,e_i]$		e_2	b_3e_3	$b_2e_4-c_2e_5$	$c_2e_4+b_2e_5$
2*$N_{7,4}$	$[x_1,e_i]$	e_1	p_1e_3		$a_3e_4+p_2e_5$	a_3e_5
2-7	$[x_2,e_i]$		$e_2+q_1e_3$	e_3	$b_3e_4+q_2e_5$	b_3e_5
2*$N_{7,5}$	$[x_1,e_i]$	e_1	p_1e_3		$-e_5$	e_4
2-7	$[x_2,e_i]$		$e_2+q_1e_3$	e_3	e_4	e_5
2*$N_{7,6}$	$[x_1,e_i]$	e_1	$-e_3$	e_2	b_1e_4	b_1e_5
2-7	$[x_2,e_i]$				$-e_5$	e_4
2*$N_{7,7}$	$[x_1,e_i]$	e_1		$a_3e_3+p_2e_5$	$a_3e_4+p_3e_5$	a_3e_5
2-7	$[x_2,e_i]$		e_2	$b_3e_3+q_2e_5$	$b_3e_4+q_3e_5$	b_3e_5
2*$N_{7,8}$	$[x_1,e_i]$	e_1		$a_3e_3+p_2e_5$	a_3e_4	a_3e_5
2-7	$[x_2,e_i]$		e_2	$b_3e_3+q_1e_4+q_2e_5$	$b_3e_4+q_3e_5$	b_3e_5
2*$N_{7,9}$	$[x_1,e_i]$	e_1		$a_3e_3+e_4+p_2e_5$	$a_3e_4+p_3e_5$	a_3e_5
2-7	$[x_2,e_i]$		e_2	$b_3e_3+q_1e_4+q_2e_5$	$b_3e_4+q_3e_5$	b_3e_5
2*$N_{7,10}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	$p_4e_4+p_5e_5$	p_6e_5	
2-7	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4+q_3e_5$	$e_3+q_4e_4+q_5e_5$	$e_4+q_6e_5$	e_5
2*$N_{7,11}$	$[x_1,e_i]$	e_1	$p_2e_4+p_3e_5$	$p_4e_4+p_5e_5$		
2-7	$[x_2,e_i]$		$e_2+q_2e_4+q_3e_5$	$e_3+q_4e_4+q_5e_5$	e_4	e_5
2*$N_{7,12}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	$p_4e_4+p_5e_5$		
2-7	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4+q_3e_5$	$e_3+q_4e_4+q_5e_5$	e_4	e_5
2*$N_{7,13}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	p_5e_5	p_6e_5	
2-7	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4+q_3e_5$	$e_3+q_5e_5$	$e_4+q_6e_5$	e_5
2*$N_{7,14}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	p_5e_5	p_6e_5	
2-7	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4+q_3e_5$	e_3	$e_4+q_6e_5$	e_5
2*$N_{7,15}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	p_4e_5	p_6e_5	
2-7	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4+q_3e_5$	e_3	e_4	e_5
2*$N_{7,16}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	$p_4e_4+p_5e_5$	p_6e_5	
2-7	$[x_2,e_i]$		$e_2+q_2e_4+q_3e_5$	$e_3+q_5e_5$	e_4	e_5

Continued on next page

Table 1 – *Continued from previous page*

		$[x_\alpha,e_1]$	$[x_\alpha,e_2]$	$[x_\alpha,e_3]$	$[x_\alpha,e_4]$	$[x_\alpha,e_5]$
2*$N_{7,17}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	$p_4e_4+p_5e_5$	p_6e_5	
2-7	$[x_2,e_i]$		$e_2+q_3e_5$	e_3	e_4	e_5
2*$N_{7,18}$	$[x_1,e_i]$	e_1	$p_1e_3+p_2e_4+p_3e_5$	$p_4e_4+p_5e_5$		
2-7	$[x_2,e_i]$		$e_2+q_2e_4+q_3e_5$	e_3	e_4	e_5
2*$N_{7,19}$	$[x_1,e_i]$	$e_1+p_1e_2$	e_2	p_3e_5	p_4e_5	
2-7	$[x_2,e_i]$	q_1e_2		$e_3+q_3e_5$	$e_4+q_4e_5$	e_5
2*$N_{7,20}$	$[x_1,e_i]$	$e_1+p_1e_2$	e_2	p_3e_5		
2-7	$[x_2,e_i]$	q_1e_2		$e_3+q_2e_4+q_3e_5$	$e_4+q_4e_5$	e_5
2*$N_{7,21}$	$[x_1,e_i]$	$e_1+p_1e_2$	e_2	$e_4+p_3e_5$	p_4e_5	
2-7	$[x_2,e_i]$	q_1e_2		$e_3+q_2e_4+q_3e_5$	$e_4+q_4e_5$	e_5
2*$N_{7,22}$	$[x_1,e_i]$	$e_1+p_2e_3$	$e_2+p_3e_3$	e_3	$-e_5$	e_4
2-7	$[x_2,e_i]$	q_2e_3	q_3e_3		e_4	e_5
2*$N_{7,23}$	$[x_1,e_i]$	$e_1+p_2e_3$	e_2	e_3	$-e_5$	e_4
2-7	$[x_2,e_i]$	$q_1e_2+q_2e_3$	q_3e_3		e_4	e_5
2*$N_{7,24}$	$[x_1,e_i]$	$e_1+e_2+p_2e_3$	$e_2+p_3e_3$	e_3	$-e_5$	e_4
2-7	$[x_2,e_i]$	$q_1e_2+q_2e_3$	q_3e_3		e_4	e_5

2 Bracket relations for the real seven-dimensional solvable Lie algebras with five-dimensional Abelian nilradical and a one-dimensional center

Table 2: Non-zero bracket relations for the real seven-dimensional solvable Lie algebras with five-dimensional Abelian nilradical and a one-dimensional center $Z(\mathfrak{g}) = \langle e_5\rangle$. For all of the following algebras, $[x_1,x_2] = e_5$.

		$[x_\alpha,e_1]$	$[x_\alpha,e_2]$	$[x_\alpha,e_3]$	$[x_\alpha,e_4]$
2*$N_{7,25}$	$[x_1,e_i]$	e_1		a_3e_3	a_4e_4
2-6	$[x_2,e_i]$		e_2	b_3e_3	b_4e_4
2*$N_{7,26}$	$[x_1,e_i]$	$e_1+p_1e_2$	e_2	$-e_4$	e_3
2-6	$[x_2,e_i]$	q_1e_2		e_3	e_4
2*$N_{7,27}$	$[x_1,e_i]$	e_1	p_2e_4	p_3e_4	
2-6	$[x_2,e_i]$		$e_2+q_2e_4$	$e_3+q_3e_4$	e_4
2*$N_{7,28}$	$[x_1,e_i]$	e_1	p_2e_4		
2-6	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4$	$e_3+q_3e_4$	e_4
2*$N_{7,29}$	$[x_1,e_i]$	e_1	$e_3+p_2e_4$	p_3e_4	
2-6	$[x_2,e_i]$		$e_2+q_1e_3+q_2e_4$	$e_3+q_3e_4$	e_4
2*$N_{7,30}$	$[x_1,e_i]$	$e_1+p_1e_2$	e_2	p_2e_4	
2-6	$[x_2,e_i]$	q_1e_2		$e_3+q_2e_4$	e_4
2*$N_{7,31}$	$[x_1,e_i]$	e_1		$a_3e_3+p_1e_4$	a_3e_4
2-6	$[x_2,e_i]$		e_2	$b_3e_3+q_1e_4$	b_3e_4
2*$N_{7,32}$	$[x_1,e_i]$	e_1		$-e_4$	e_3
2-6	$[x_2,e_i]$		e_2	e_3	e_4

Continued on next page

Table 2 – *Continued from previous page*

		$[x_\alpha, e_1]$	$[x_\alpha, e_2]$	$[x_\alpha, e_3]$	$[x_\alpha, e_4]$
2*$N_{7,33}$	$[x_1, e_i]$	$-e_2 + p_1 e_3 + p_2 e_4$	$e_1 + p_3 e_3 + p_4 e_4$	$-e_4$	e_3
2-6	$[x_2, e_i]$	$e_1 + q_1 e_3 + q_2 e_4$	$e_2 - q_2 e_3 + q_1 e_4$	e_3	e_4
2*$N_{7,34}$	$[x_1, e_i]$	$-e_2$	e_1	$b_2 e_3$	$b_2 e_4$
2-6	$[x_2, e_i]$			$-e_4$	e_3

Bibliography

[1] E.Cartan. Sur la Reduction a sa Forme Canonique de la Structure d'un Groupe de Transformations Fini et Continu. *Amer. J. Math.*, 18(1):1–61, 1896.

[2] Felix Gantmacher. On the classification of real simple Lie groups. *Rec. Math. [Mat. Sbornik] N.S.,* 5 (47):217-250, 1939.

[3] M. Gong. *Classification of Nilpotent Lie Algebras of Dimension 7 (Over Algebraically Closed Fields and R)*. PhD thesis, University of Waterloo, 1998.

[4] Sigurdur Helgason. *Differential geometry, Lie groups, and symmetric spaces*, volume 80 of *Pure and Applied Mathematics*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978.

[5] F. Hindeleh and G. Thompson. Seven dimensional Lie algebras with a four-dimensional nilradical. *Algebras Groups Geom.*, 25(3):243-265, 2008.

[6] James E. Humphreys. *Introduction to Lie algebras and representation theory*, volume 9 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin, 1978. Second printing, revised.

[7] Nathan Jacobson. *Lie algebras*. Interscience Tracts in Pure and Applied Mathematics, No. 10. Interscience Publishers (a division of John Wiley Sons), New York-London, 1962.

[8] Vu A. Le, Tuan A. Nguyen, Tu T. C. Nguyen, Tuyen T. M. Nguyen, and Thieu N. Vo. Classification of 7-dimensional solvable lie algebras having 5-dimensional nilradicals, 2021.

[9] V. V. Morozov. Classification of nilpotent lie algebras of sixth order. *Izv. Vysš. Učebn. Zaved. Matematika*, 1958(4(5)):161-171, 1958.

[10] G. M. Mubarakzjanov. Certain theorems on solvable Lie algebras. *Izv. Vysš. Učebn. Zaved. Matematika*, 1966(6(55)):95-98, 1966.

[11] G. M. Mubarakzjanov. Classification of real structures of lie algebras of fifth order. *Izv. Vysshikh Uchebn. Zavedenii Mat.*, 3(34):99-106, 1964

[12] G. M. Mubarakzjanov. Classification of solvable lie algebras of sixth order with a non-nilpotent basis element. *Izv. Vysshikh Uchebn. Zavedenii Mat.*, 4(35):104-116, 1963.

[13] G. M. Mubarakzjanov. On solvable lie algebras. *Izv. Vysshikh Uchebn. Zavedenii Mat.*, 1(32):114-123, 1963.

[14] Jean-Claude Ndogmo. *Sur les fonctions invariantes sous l'action coadjointe d'une algebre de Lie resoluble avec nilradical abelien.* ProQuestLLC, Ann Arbor, MI, 1994. Thesis (Ph.D.)–Universite de Montreal (Canada).

[15] J. C. Ndogmo and P. Winternitz. Solvable Lie algebras with abelian nilradicals. *J. Phys. A*, 27(2):405–423, 1994.

[16] Alan Parry. *A Classification of real indecomposable solvable Lie algebras of small dimension with codimension one nilradical.* 2007. Thesis (M.Sc.)–Utah State University.

[17] J. Patera, R.T. Sharp, P. Winternitz, and H. Zassenhaus. Invariants of real low dimension lie algebras. *J. Math. Phys.*, 17:986–994, 1976.

[18] Craig Seeley. 7-dimensional nilpotent Lie algebras. *Trans. Amer. Math. Soc.*, 335(2):479–496, 1993.

[19] Anastasia Shabanskaya. Classification of six dimensional solvable indecomposable Lie algebras with a codimension one nilradical over R. ProQuestLLC, Ann Arbor, MI, 2011. Thesis(Ph.D.)–The University of Toledo.

[20] Anastasia Shabanskaya and Gerard Thompson. Six-dimensional Lie algebras with a five-dimensional nilradical. *J. Lie Theory*, 23(2):313–355, 2013.

[21] P. Turkowski. Solvable lie algebras of dimension six. *J. Math. Phys.*, 31(6):1344–1350, 1990.

[22] K. A. Umlauf. *Über die Zusammensetzung der endlichen continuierliche Transformation gruppen insbesondere der Gruppen von Rang null.* 1891. Thesis(Ph.D.)–University of Leipzig.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 34 – 38

The group of monomial matrices

*Martin F. Martinez, Pietro Paparella**

Martin Martinez worked on this paper as a dual-enrolled high school student. He is now majoring in mathematics at the University of Washington Bothell. In his spare time, he enjoys learning about and working with electronic music production..

Pietro Paparella received the Ph.D. degree in mathematics from Washington State University in 2013 and is currently an Associate Professor of mathematics in the Division of Engineering and Mathematics at the University of Washington Bothell. His research interests are in nonnegative matrix theory, combinatorial matrix theory, discrete geometry, and the geometry of polynomials. His hobbies include guitar playing, charcoal drawing, and oil painting.

Abstract

A recent result is used to give a brief proof of the well-known fact that the set of *monomial* matrices forms a subgroup of the set of invertible matrices. In addition, another proof is given of the result that the inverse of an invertible nonnegative matrix is nonnegative if and only if the matrix is monomial.

1 Introduction

In this note, we utilize a recent result [3, Lemma 3.3] to give a brief proof that the set of *monomial* matrices forms a subgroup of the set of invertible matrices. The result is well-known, but, to the best of our knowledge, a proof is not readily available in the literature and deserves wider circulation. In addition, we give an elementary proof that

***Corresponding author:** pietrop@uw.edu

the inverse of an invertible nonnegative matrix is nonnegative if and only if the matrix is monomial.

2 Notation & Background

In this work, '$\mathbb{F}$' stands for $\mathbb{C}$ or $\mathbb{R}$. The algebra of n-by-n matrices with entries over $\mathbb{F}$ is denoted by $\mathsf{M}_n = \mathsf{M}_n(\mathbb{F})$ and the subset of invertible n-by-n matrices with entries from $\mathbb{F}$ is denoted by $\mathsf{GL}_n = \mathsf{GL}_n(\mathbb{F})$. The set of all n-by-1 column vectors is identified with the set of all ordered n-tuples with entries in $\mathbb{F}$ and thus denoted by $\mathbb{F}^n$. If $x \in \mathbb{F}^n$, then D_x denotes the diagonal matrix such that $d_{ii} = x_i$.

For $n \in \mathbb{N}$, denote by S_n the *symmetric group* of degree n. Given $\sigma \in S_n$, the *permutation matrix* with respect to σ, denoted by $P = P_\sigma \in \mathsf{M}_n$, is the n-by-n matrix such that $p_{ij} = \delta_{\sigma(i),j}$, where δ denotes the *Kronecker delta function*. The following facts concerning permutation matrices are well-known:

Proposition 1. *If σ, $\gamma \in S_n$, then:*

(i) $P_\sigma P_\gamma = P_{\gamma \circ \sigma}$;

(ii) $(P_\sigma)^{-1} = P_{\sigma^{-1}} = (P_\sigma)^\top$; *and*

(iii) P is a permutation matrix if and only if P is a matrix with entries from $\{0,1\}$ and every row and every column of P contains exactly one nonzero entry.

3 Monomial matrices

Definition 1. *If $A \in \mathsf{M}_n$, then A is called* monomial, *a* monomial matrix, *or a* generalized permutation matrix *if there is an invertible diagonal matrix D and a permutation matrix P such that $A = DP$. The set of all n-by-n monomial matrices is denoted by* $\mathsf{GP}_n = \mathsf{GP}_n(\mathbb{F})$

Remark 2. *If A is monomial with $A = DP$, then $a_{ij} = d_{ii}\delta_{\sigma(i),j}$. Following part (iii) of Proposition 1, A is monomial if and only if every row and every column of A contains exactly one nonzero entry.*

If $S \in \mathsf{GL}_n$, then the *relative gain array (RGA) of S*, denoted by $\Phi(S)$, is defined by $\Phi(S) = S \circ S^{-\top}$, where '$\circ$' denotes the *Hadamard* or entrywise product and $S^{-\top} := (S^{-1})^\top = (S^\top)^{-1}$. Johnson and Shapiro [4] showed that if $A = SD_xS^{-1}$, then

$$\Phi(S)x = \begin{bmatrix} a_{11} & \cdots & a_{nn} \end{bmatrix}^\top. \tag{1}$$

The following result, stated in slightly different terms, was established by Johnson and Paparella [3, Lemma 3.3] via the RGA.

Lemma 1. *If P is a permutation matrix and $x \in \mathbb{F}^n$, then $P^\top D_x P = D_y$, where $y := P^\top x$.*

Proof. Because a permutation similarity effects a simultaneous permutation of the rows and columns of a matrix, it follows that $P^\top D_x P$ is a diagonal matrix—say D_y.

Following (1) and part (ii) of Proposition 1,

$$y = \Phi(P^\top)x = [P^\top \circ (P^\top)^{-\top}]x = \left[P^\top \circ P^\top\right]x = P^\top x. \qquad \square$$

The following characterization is immediate from Lemma 1.

Corollary 1. *If $A \in \mathsf{M}_n$, then A is monomial if and only if $A = PD$, where D is an invertible diagonal matrix and P is a permutation matrix. Furthermore, if $A = D_x P$, where $x \in \mathbb{F}^n$, then $A = PD_y$, where $y := P^\top x$.*

Recall that if $A \in \mathsf{M}_n(\mathbb{R})$, then A is called *(entrywise) nonnegative* (respectively, *positive*), denoted by $A \geq 0$ (respectively, $A > 0$), if $a_{ij} \geq 0, 1 \leq i, j \leq n$ (respectively, $a_{ij} > 0, 1 \leq i, j \leq n$).

Lemma 2. *If A is monomial, then A is invertible and A^{-1} is monomial. Furthermore, if $A \geq 0$, then $A^{-1} \geq 0$.*

Proof. If A is monomial, then there is a vector $x \in \mathbb{F}^n$ with no zero entries and a permutation matrix P such that $A = D_x P$. By Corollary 1, $A = PD_y$, where $y = P^\top x$. The matrix A is invertible as it is the product of invertible matrices and

$$A^{-1} = (PD_y)^{-1} = (D_y)^{-1} P^{-1} = D_{y^{-1}} P^\top,$$

where $y^{-1} := \begin{bmatrix} x_1{}^{-1} & \cdots & x_n{}^{-1} \end{bmatrix}^\top$. By Definition 1, A^{-1} is a monomial matrix.

Notice that $A \geq 0$ if and only if $y > 0$. Thus, if $A \geq 0$, then $y^{-1} > 0$ and $A^{-1} \geq 0$ as it is the product of nonnegative matrices. □

Theorem 3. *GP_n is a subgroup of GL_n.*

Proof. The identity matrix is clearly monomial, so GP_n is nonempty. In view of Lemma 2, it suffices to demonstrate closure. To this end, if $A, B \in \mathsf{GP}_n(\mathbb{F})$, then there are permutation matrices P and Q such that $A = D_x P$ and $B = D_y Q$. Thus,

$$\begin{aligned} AB &= (D_x P)(D_y Q) \\ &= D_x((PD_y)Q) && \text{(associativity)} \\ &= D_x((D_z P)Q) && \text{(Lemma 1 with } z := Py) \\ &= (D_x D_z)(PQ) && \text{(associativity)} \\ &= D_{x \circ z}(PQ), \end{aligned}$$

where '$\circ$' denotes the *Hadamard product*. □

4 Nonnegative subgroups of Invertible Matrices

In 2013, Ding and Rhee [1] proved that an invertible matrix and its inverse are *stochastic* (i.e., entrywise nonnegative with rows summing to unity) if and only if the invertible matrix is a permutation matrix. In a subsequent work [2], they gave another proof of this result and used the result to show that an invertible matrix and its inverse are entrywise nonnegative if and only if the invertible matrix is monomial.

The import of the second result above can be gleaned from the following context. Recall that the set of invertible nonnegative matrices with matrix multiplication forms a *monoid*, i.e., it satisfies the *closure*, *associativity*, and *identity* group-axioms. However, as can be readily seen with two-by-two matrices, the inverse of an invertible nonnegative matrix need not be nonnegative.

The set of permutation matrices forms a nonnegative multiplicative subgroup of the set of invertible matrices, and it is natural to ask whether there are other nontrivial subsets of invertible nonnegative matrices that form a subgroup.

Theorem 3 above and Theorem 4 below provide the answer.

Theorem 4. *If A is nonnegative and invertible, then $A^{-1} \geq 0$ if and only if A is monomial.*

Proof. The sufficiency of this condition was shown in Lemma 2.

To demonstrate necessity, we modify the elementary argument given by Ding and Rhee [1] for stochastic matrices.

To this end, suppose that A is a nonnegative invertible matrix and that $A^{-1} \geq 0$. For convenience, write $B = A^{-1}$. Since $AB = I$, it follows that

$$\sum_{k=1}^{n} a_{ik} b_{kj} = \delta_{ij}.$$

In particular, if $i \neq j$, then

$$\sum_{k=1}^{n} a_{ik} b_{kj} = 0. \tag{2}$$

Fix $i \in \{1, \ldots, n\}$. Because A is invertible, the ith row of A must possess at least one positive entry—say a_{ir}. The nonnegativity of both matrices ensures that each summand on the left-hand side of (2) equals zero, i.e., $a_{ik} b_{kj} = 0$, $\forall k \in \{1, \ldots, n\}$. Since $a_{ir} > 0$, it follows that $b_{rj} = 0$ whenever $j \neq i$. Since the rth row of B cannot be zero, it must be the case that $b_{ri} > 0$.

Next, we show that a_{ir} is the only nonzero entry in the ith row of A. To the contrary, if $a_{is} > 0$, with $r \neq s$, then the argument above implies that $b_{sj} = 0$ whenever $j \neq i$ and $b_{si} > 0$. Thus, the rth and sth rows of B are (positive) multiples of each other, contradicting the invertibility of B.

Since $A^\top B^\top = I$, another application of the argument above with respect to the rth row of $A^\top$ demonstrates that a_{ir} is the only nonzero entry in the rth column of A. As i was arbitrary, the result applies to every row of A and because A is invertible, it must be the case that every row and every column of A contains exactly one nonzero entry, i.e., A is monomial. □

Corollary 2. *Any subgroup of invertible matrices in which every matrix is nonnegative must be a subgroup of the set of nonnegative momonial matrices.*

Bibliography

[1] J. Ding and N. H. Rhee. Teaching tip: when a matrix and its inverse are stochastic. *College Math. J.*, 44(2):108–109, 2013.

[2] J. Ding and N. H. Rhee. When a matrix and its inverse are nonnegative. *Missouri J. Math. Sci.*, 26(1):98–103, 2014.

[3] C. R. Johnson and P. Paparella. Perron spectratopes and the real nonnegative inverse eigenvalue problem. *Linear Algebra Appl.*, 493:281–300, 2016.

[4] C. R. Johnson and H. M. Shapiro. Mathematical aspects of the relative gain array $(A \circ A^{-T})$. *SIAM J. Algebraic Discrete Methods*, 7(4):627–644, 1986.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 39 – 48

Cram with Square Polyominoes

Michael Fraboni, Emma Miller*

Michael Fraboni is a professor of mathematics at Moravian University. He is interested in involving undergraduates in meaningful research experiences on combinatorial game theory related topics..

Emma Miller graduated from Moravian University in 2022 with a bachelor's of science in physics and mathematics. This paper was grew from her senior honors thesis which focused on the expansion of common combinatorial games. Emma is currently employed full-time and is currently considering continuing her education in either mathematics or science management.

Abstract

We will consider expansions of CRAM, a game frequently studied in the area of combinatorial game theory. The game of CRAM is classically played with dominoes, a type of polyomino. We will define CRAM WITH HIGHER POLYOMINOES and use efficient packing results to establish the outcome classes for several board shapes and choices of polyominoes.

1 Introduction and Background

Combinatorial Game theory studies finite strategy games of perfect information.

Definition 1.1 [from [1]]. *A* combinatorial game

(i) *is* finite*: the game must end and cannot end in a tie;*

(ii) *is* based on pure strategy*: it has no elements of chance such as coin tosses, dice rolls, or randomly drawn cards; nor of skill, such as darts or hockey;*

(iii) *is* sequential*: has alternating players who take turns moving;*

***Corresponding author:** frabonim@moravian.edu

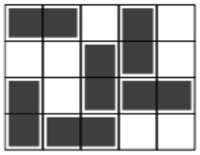


Figure 1: A game position of CRAM on a 5×4 board.

(iv) is a game of perfect information*: all players know all possible moves they can make as well as all moves the other players can make.*

Examples of combinatorial games include chess and go. Tic-tac-toe does not meet this definition since the game might end in a tie, while any games with secret information, such as most card games, are not included.

Our discussion focuses on games with only two players, a Left Player and a Right Player. We will assume the Left Player plays first, followed by the Right Player.

While analyzing these games, we assume both players *play perfectly*. In other words, if a player has a move or strategy that will allow the player to win, the player will play it.

A common example of a combinatorial game is CRAM. We define this game below, but it is well studied in, for example, [2] and [1].

Definition 1.2 *The game of* CRAM *is played on a board of square tiles. These boards can be of any size and arrangement. The two players alternately play a domino on two vertically or horizontally adjacent tiles. Any tile on the board can only hold a single cell of a domino. The game ends when the next player cannot play another domino. The last player to play wins.*

An example of a CRAM position can be seen in Figure 1. As each domino played is indistinguishable from all other dominoes and as both players have the same choice of possible moves, CRAM is an *impartial game*.

In Section 2 we introduce the idea of arbitrary polyominoes, and in Section 3 we extend the definition of CRAM to use these polyominoes instead of just dominoes. Then in Section 4 we study one version of this extended CRAM game using square polyominoes and are able to characterize the outcome of most games played on square boards. If the board is sufficiently small relative to the size of the piece then the first player to play will always have a winning strategy. If the board is square and exactly one less than three times the size of the piece then the second player has a winning strategy. Finally, if the board is square and three times the size of the piece or more, and the piece and board size are congruent mod 2 then the first player will have a winning strategy.

1.1 Outcome Classes

One final idea from combinatorial game theory is that of the outcome class of a game. The Fundamental Theorem of Combinatorial Games tells us how potential outcomes from these games are limited.

Theorem 1.3 (from [1]) *For any game played between two players, either the Left Player can force a win moving first or the Right Player can force a win moving second. Both cannot be true.*

This theorem implies that any game has one of four potential outcomes, Left Player will win regardless if she moves first or second, Right Player will win regardless if he moves first or second, the next player to play will win regardless of who that player is, or the second player to play will win. A second player win is also known as a previous player win since at any given moment in a game, the player who just moved is going to be the second player to play.

These outcomes are known as *outcome classes* and we will classify game positions according to their outcome classes. For a game position $\mathscr{G}$, the outcome class is denoted as $o(\mathscr{G})$. Because CRAM is an impartial game, any winning strategy for Left Player will work for Right Player, so our study focuses on the Next, $\mathscr{N}$, and Previous, $\mathscr{P}$, outcomes.

Definition 1.4 *A game $\mathscr{G}$ where the next player to play can force a win has the outcome class* Next*, denoted $o(\mathscr{G}) = \mathscr{N}$.*

A game $\mathscr{G}$ where the previous player who played can force a win has the outcome class Previous*, denoted $o(\mathscr{G}) = \mathscr{P}$.*

An example of a game with outcome class $\mathscr{N}$ is CRAM on a 1×4 board since the next player to play may place their domino on the middle two tiles. This ends the game as there are no more valid moves available.

An example of a game with the outcome class $\mathscr{P}$ is CRAM on a 2×2 board. Exactly two dominoes must be played before the game ends. As the first player cannot block the second player from playing a piece, the game has the outcome class $\mathscr{P}$.

2 Polyomino Preliminaries

Our goal is to extend CRAM to games using other pieces, so we take a moment to generalize the idea of a domino.

Definition 2.1. *A* polyomino, *P, is a finite collection of square cells such that each cell is vertically or horizontally adjacent to another cell.*

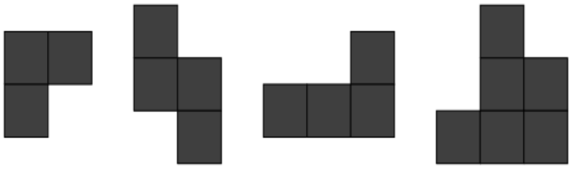


Figure 2: Examples of polyominoes.

See Figure 2 for examples of polyominoes. In this paper we will focus in particular on rectangular polyominoes.

Definition 2.2. *A* rectangular polyomino *$R_{u,v}$ is made of u columns of cells and v rows of cells.*

A square polyomino, *denoted R_u, is a rectangular polyomino with $u = v$.*

Examples are shown in Figure 3. We focus our attention specifically on square polyominoes, R_u.

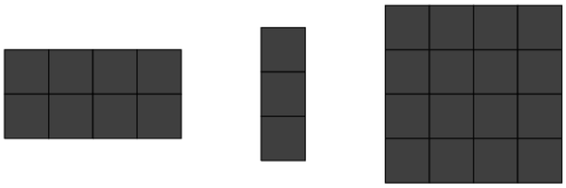


Figure 3: Examples of $R_{4,2}$, $R_{1,3}$, and R_4.

One common area of study for polyominoes is the study of packings and we will use some of these ideas in our results.

Definition 2.3 *A* packing *is an arrangement of polyominoes on a board such that no other polyominoes can be placed on the board.*

A packing number *is the number of polyominoes in a packing. An* efficient packing *on some board $\mathscr{B}$ of a polyomino P is a packing which uses the most possible copies of P. The number of copies used is the* efficient packing number *and is denoted* $\mathrm{p}_{\mathscr{B}}(P)$.

Similarly, a clumsy packing *on some board $\mathscr{B}$ of a polyomino P is a packing which uses the fewest possible copies of P. The number of copies used is the* clumsy packing number *and is denoted* $\mathrm{cp}_{\mathscr{B}}(P)$.

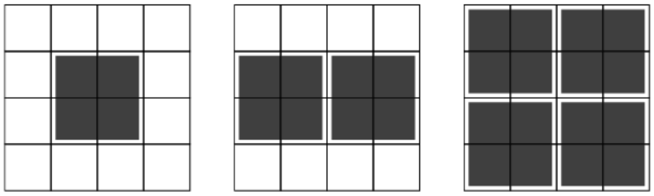


Figure 4: Examples of possible packings of a R_2 on a 4 × 4 board B. In this case, $\mathrm{cp}_{\mathscr{B}}(P)(R_2) = 1$ and $\mathrm{cp}_{\mathscr{B}}(P)(R_2) = 4$.

Additional resources that highlight packings include [5], [6], and [7]. Previous work on clumsy packings on finite boards specifically can be seen in [4]. An example of packings can be seen in Figure 4.

As an example, note that the efficient packing of $R_{u,v}$ on a rectangular board is trivial.

Observation 2.4. *For any $R_{u,v}$ efficient packing on a rectangular $n \times m$ board B, the* $\mathrm{p}_{\mathscr{B}}(R_{u,v}) = \lfloor \frac{n}{u} \rfloor \times \lfloor \frac{m}{v} \rfloor$*. In particular, if $n = m$ and $u = v$,* $\mathrm{p}_{\mathscr{B}}(R_u) = \lfloor \frac{n}{u} \rfloor^2$.

3 CRAM with Higher Polyominoes

We now extend the classical rules of CRAM to include general polyominoes and consider when we may determine the outcome class of such a game.

Definition 3.1. *For any polyomino P, we define the game of* CRAM WITH *P as played on a board B. Two players alternately play a free P on Bsuch that any tile of Bcan hold at most one cell of a P. The game ends when the next player cannot play another polyomino. The last player to play wins. In general, we call this class of games* CRAM WITH HIGHER POLYOMINOES.

CRAM WITH HIGHER POLYOMINOES was previously defined in [3] where the authors refer to the game as CRAMOMINOES.

We begin with some results relating game play in CRAM WITH HIGHER POLYOMINOES to packings of polyominoes. We then use these results to determine the outcome class of several games played on square boards.

3.1 CRAM and Packings

Since a game of CRAM WITH HIGHER POLYOMINOES ends when there is a packing, we will consider the packing of polyominoes on finite boards to determine the outcome of a game.

Lemma 3.2. *For a completed game position $\mathscr{G}$ of* CRAM WITH *P on a board $\mathscr{B}$, the number of moves to reach the position $\mathscr{G}$, $n(\mathscr{G})$, satisfies*

$$\mathrm{cp}_{\mathscr{B}}(P) \leq n(\mathscr{G}) \leq \mathrm{p}_{\mathscr{B}}(P).$$

Proof. As a game $\mathscr{G}$ of CRAM WITH P on a board Bmust end with a packing, the smallest possible number of moves is the clumsy packing number, $\mathrm{cp}_{\mathscr{B}}(P)$. The largest possible number of moves is the packing number, $\mathrm{p}_{\mathscr{B}}(P)$. Thus,

$$\mathrm{cp}_{\mathscr{B}}(P) \leq n(\mathscr{G}) \leq \mathrm{p}_{\mathscr{B}}(P).$$

□

We cannot simply use the clumsy packing number to determine the outcome of a game. For example, on a 9×2 board B, $\mathrm{cp}_{\mathscr{B}}(R_2) = 3$. However, if Left Player plays the move from Figure 5 the board must have a packing number of 4, which is greater than the clumsy packing number of the entire board.

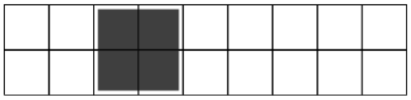


Figure 5: An example of a move which forces the number of moves in the finished game to be greater than the clumsy packing number.

However, for CRAM WITH P on a board $\mathscr{B}$, if $\mathrm{cp}_{\mathscr{B}}(P) = \mathrm{p}_{\mathscr{B}}(P)$, then the number of moves in a completed game must satisfy $n(\mathscr{G}) = \mathrm{cp}_{\mathscr{B}}(P)$ by Lemma 3.2. In this case then, the outcome class of a game comes down to whether the packing numbers are odd or even.

Corollary 3.3. *For a game $\mathscr{G}$ of* CRAM WITH P *on a board $\mathscr{B}$, where $\mathrm{cp}_{\mathscr{B}}(P) = \mathrm{p}_{\mathscr{B}}(P)$, if $\mathrm{cp}_{\mathscr{B}}(P)$ is even, the game has the outcome class $\mathscr{P}$ and if $\mathrm{cp}_{\mathscr{B}}(P)$ is odd, the game has the outcome class $\mathscr{N}$.*

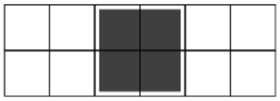


Figure 6: A game of CRAM with R_2 on a 6×2 board has the clumsy packing of 2, however, the game has the outcome class $\mathscr{N}$ since the first player may play as shown above. This leaves exactly two moves left and a win for the first player.

Unfortunately, if the $\mathrm{cp}_{\mathscr{B}}(P) \neq \mathrm{p}_{\mathscr{B}}(P)$ and if $\mathrm{cp}_{\mathscr{B}}(P)$ is odd, the game does not necessarily have the outcome class $\mathscr{N}$ as discussed in the previous example with Figure 5. Similarly, if the $\mathrm{cp}_{\mathscr{B}}(P) \neq \mathrm{p}_{\mathscr{B}}(P)$ and if $\mathrm{cp}_{\mathscr{B}}(P)$ is even, the game does not necessarily have the outcome class $\mathscr{P}$. An example of this can be seen in Figure 6.

The following observation is a trivial alternative for establishing specific cases of games with the outcome class $\mathscr{N}$ and $\mathscr{P}$. Typically, if one piece can be played to win the game, that game has the outcome class $\mathscr{N}$. The following observation provides a method for determining the outcome class of a game using the clumsy packing number to prove that one piece can win the game. This means we do not need to consider all other possible moves to determine the outcome. Similarly, if no pieces can be played to start the game, that game has the outcome class $\mathscr{P}$.

Observation 3.4. Let $\mathscr{G}$ be CRAM WITH P on a board $\mathscr{B}$, if $\mathrm{cp}_{\mathscr{B}}(P) = 1$ then, $o(\mathscr{G}) = \mathscr{N}$ and if $\mathrm{cp}_{\mathscr{B}}(P) = 0$ then, $o(\mathscr{G}) = \mathscr{P}$.

In the following section we will consider CRAM WITH P specifically when playing with square polyominoes.

4 $u \times u$ Square Cram

We consider CRAM WITH P using $u \times u$ square polyominoes, R_u.

In many of the following proofs, for a game of CRAM WITH R_u it will be useful to consider the center of a rectangular $a \times b$ board. For a square polyomino R_u, when u is odd, we will consider the single middlemost cell of the polyomino to be the center cell. When u is even, we will consider the middlemost 2×2 set of cells to be the center cells. An analogous idea used to define the center tile(s) on an $a \times b$ board when both a and b are either odd or even. If a is odd and b is even then the two tiles in the intersection of the center column with the two middlemost rows will be the center tiles. The case when b is odd and a is even is similar.

Definition 4.1. *We will say a player plays* centerish *if the number of center cells of a polyomino placed in the center tiles of the board is maximized.*

Note if both the board and polyomino have odd dimensions then there is only one play that would be considered centerish. Similarly, if both have even dimensions there is only one centerish play as all four center cells of the polyomino must be on all four center tiles of the board. When one has odd dimensions and the other has even dimensions there would be four plays that are considered centerish. See Figure 7 for an example.

In the following theorems, we establish the outcome for CRAM WITH R_u played on boards with sides less then or equal to $3u-1$ and square boards greater than $3u-1$ if the board and piece are both even or odd. This covers all games of CRAM WITH R_u played on square boards except when the size of the square board is even and u is odd or vice-versa.

Theorem 4.2. *Let $n, u \in \mathbb{Z}^+$, $u > 1$. If $u \leq n \leq 3u-2$, then* CRAM WITH R_u *on an $n \times n$ board has the outcome class $\mathcal{N}$.*

Proof. Assume $n, u \in \mathbb{Z}^+$. Let $\mathcal{G}$ be CRAM WITH R_u on a rectangular $n \times n$ board Bwhere $u \leq n \leq 3u-2$. If the Left Player plays centerish, the tiles without a cell create at most a $u-1$ gap between the played piece and one or two of the edges of Bas shown in Figure 7. As a $u-1$ gap cannot hold another polyomino, the $\text{cp}_{\mathcal{B}}(R_n) = 1$. Therefore by Observation 3.4, $o(\mathcal{G}) = \mathcal{N}$. □

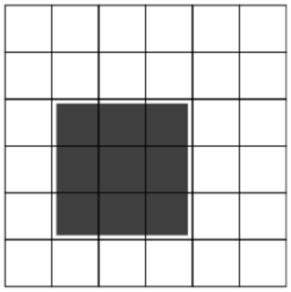


Figure 7: An example of CRAM WITH R_3 on a 6×6 board showing one of the ways for Left Player to play centerish.

Theorem 4.3. Let $n, u \in \mathbb{Z}^+$, $u > 1$. On an $n \times (3u-1)$ board, CRAM WITH R_u has the outcome class $\mathcal{P}$.

Proof. Let $\mathscr{G}$ be CRAM WITH R_u on a $n \times (3u-1)$ board B. If $n < u$ then no moves are possible and the game is a previous player win, so assume $n \geq u$. Let $\mathscr{C}$ be a $u \times (3u-1)$ portion of the board. Then $\mathrm{cp}_{\mathscr{C}}(R_u) = \mathrm{p}_{\mathscr{C}}(R_u) = 2$ and so any complete packing on $\mathscr{C}$ must contain 2 copies of R_u. As any set of u adjacent nonempty columns is analogous to $\mathscr{C}$, if the Left Player plays, the Right Player will always be able to play directly above or below that piece ending play in these columns. Therefore, for any move the Left Player makes, the Right Player has a response, and so $o(\mathscr{G}) = \mathscr{P}$. □

While Corollary 4.4 below follows directly from Theorem 4.2., we present an interesting alternate proof here.

Corollary 4.4. *Let $u \in \mathbb{Z}^+$, $u > 1$. Any game of* CRAM WITH R_u *on a* $(3u-1) \times (3u-1)$ *board has the outcome class* $\mathscr{P}$.

Proof. Let $\mathscr{G}$ be CRAM WITH R_u on a rectangular $(3u-1) \times (3u-1)$ board B.

Consider any packing of R_u on B. Note a $u \times u$ region at every corner of Bmust contain a cell of an R_u, an example is shown in Figure 8. If there is not a cell in this region, an R_u may be played in that region. Since no one R_u may overlap more than one of these regions, we need at least 4 copies of R_u to pack B. Thus, $\mathrm{cp}_{\mathscr{B}}(R_u) \geq 4$.

By Observation 2.4, $\mathrm{p}_{\mathscr{B}}(R_u) = \left\lfloor \frac{3u-1}{u} \right\rfloor^2 = 4$.

As $\mathrm{cp}_{\mathscr{B}}(R_u) = \mathrm{p}_{\mathscr{B}}(R_u) = 4$ and 4 is an even number, by Corollary 3.3, $o(\mathscr{G}) = \mathscr{P}$. □

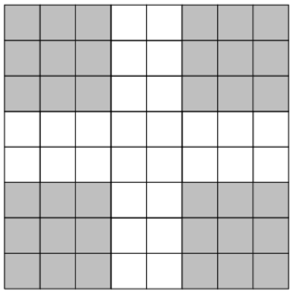


Figure 8: For a game of CRAM WITH R_3 on a 8×8 board, the shaded regions represent the 4 corner regions that must contain a piece of R_3.

In the following proofs we will discuss boards where the polyomino and length of the board are both even or both odd. In this situation, the piece can be played in the true center of the board so that the edge of the piece is the same number of cells away from the edge of the board on the top and bottom as well as the left and right. We will use this idea to discuss a mirroring strategy for playing the game.

Theorem 4.5. Let $u \in \mathbb{Z}^+$, $u > 1$, and let k be a non-negative, even integer. If $n = 3u + k$ then CRAM WITH R_u on an $n \times n$ board has the outcome class $\mathscr{N}$.

Proof. Given u, k as above, note u, n are both even or both odd.

Assume $u, n, \in \mathbb{Z}^+$. Let $n \geq 3u$ and u, n be even. Define $\mathscr{G}$ as CRAM WITH R_u on an $n \times n$ board B. As the size of Band R_u are both even, the Left Player may play in the true center of B. Then for any move Right Player makes, the Left Player has a response using rotational symmetry as follows: if a piece is played in one location, rotate the board 180° and place the polyomino in the same location. Thus after playing the first polyomino, the Left Player has a response to every Right Player move resulting in $o(\mathscr{G}) = \mathscr{N}$.

A similar argument holds when u, n are both odd. □

If the dimensions of the piece and the board are not both odd or even, then the first piece can sometimes be played centerish followed by a rotational mirroring strategy similar to that discussed in the proof of Theorem 4.6.

Theorem 4.6 Given $u, n, a \in \mathbb{Z}^+$. If u and n are both even or both odd with $u \leq n$ and $u \leq a \leq 3u - 2$ then CRAM WITH R_u on a $n \times a$ board has the outcome class $\mathscr{N}$.

Proof. Assume $u, n, a \in \mathbb{Z}^+$. Let u and n be both even or both odd, $u \leq n$ and let $u \leq a \leq 3u - 2$. Let $\mathscr{G}$ be CRAM WITH R_u on a $n \times a$ board. If the Left Player plays centerish, she splits the board into two disjoint $\frac{n-u}{2} \times a$ boards. As these are identical boards, through a mirroring strategy we see that $o(\mathscr{G}) = \mathscr{N}$. □

5 Conclusion and Future Work

Moving forward, there are a number other board sizes (including boards which are not rectangular) and different types of polyominoes to consider in the study of CRAM WITH HIGHER POLYOMINOES. In addition, expansions that put restrictions on the packings of the boards such as the ability for the polyomino/boards to rotate, adding a component of gravity, changing the type of board (such as a torus or Klein Bottle), and many more parameters could easily motivate future work in this area.

Bibliography

[1] Michael Albert, Richard J. Nowakowski, and David Wolfe, *Lessons in play*, A K Peters/CRC Press, 2007.

[2] Elwyn R. Berlekamp, John H. Conway, and Richard K. Guy, *Winning ways for your mathematical plays*, 2nd ed., Vol. 1, A K Peters/CRC Press, 2001.

[3] Emma Miller, Gabrielle Demchak, Victoria Samuels, Jacob Freeh, and Jacob Smith, *Fixed cram with higher polyominoes* **Unpublished Work** (2021).

[4] Emma Miller, Mitchel O'Connor, and Nathan Shank, *Clumsy packing of polyominoes in finite space*, arXiv, 2022.

[5] Joseph O'Rourke, Jacob E. Goodman, and Csaba D. Tóth, *Handbook of discrete and computational geometry*, 3rd ed., Chapman and Hall/CRC, 2017.

[6] Bill Sands, *The gunpost problem*, Mathematics Magazine **44** (1971), 193–196.

[7] Stefan Walzer, *Clumsy packings in the grid*, Bachelor's Thesis, Karlsruhe Institute of Technology (2012).

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 49 – 58

Characterizing distances between points in the level sets of a class of continuous functions on a closed interval

Henry Riely, Yuanming Luo*

Henry Riely received his Ph.D. from the Washington State University in 2019. He is a lecturer at Kennesaw State University. His main mathematical interests lie in analysis, especially stochastic processes and harmonic analysis.

Yuanming Luo is a junior undergraduate student at Georgia Institute of Technology. He is working toward both math and computer science degrees. His primary research interests are partial differential equations and numerical methods..

Abstract

Given a continuous function $f : [a,b] \to \mathbb{R}$ such that $f(a) = f(b)$, we investigate the set of distances $|x-y|$ where $f(x) = f(y)$. In particular, we show that the only distances this set must contain are ones which evenly divide $[a,b]$. Additionally, we show that it must contain at least one third of the interval $[0, b-a]$. Lastly, we explore some higher dimensional generalizations.

1 Introduction

Imagine waking up on a crisp fall morning and deciding to use the day for a hike. You drive to the trailhead and begin to chart a route. Since you must return to your car, your elevation will be the same at the beginning and the end of your walk. Are there other elevations through which you will pass twice? Clearly there are. If you

***Corresponding author:** hriely@kennesaw.edu

begin on an ascent, you must descend to return to the trailhead, and if you begin with a decent, you must eventually ascend. If the trail is perfectly flat, then at every moment your elevation is shared by every other moment. This intuition is often given in an introductory calculus course to illustrate the intermediate value theorem.

A natural follow up question: Can anything be said about the time elapsed between two points of equal elevation? For instance, if your hike lasts an hour, we know that there are two instants, separated by an hour, of equal elevation, namely the start and the finish. Need there be two such instances separated by a half hour? The answer, it turns out, is yes. Separated by 25 minutes? No, it's possible to design a hike with no 25 minute time interval leaving you at the same elevation that you started. So what is special about 30 minutes? Can we characterize all such durations? In this paper, we answer this and related questions.

More abstractly, we will investigate the level sets of real-valued continuous functions on closed intervals, whose endpoints get sent to the same real number. To map these functions onto our hiking analogy, given $f : [a,b] \to \mathbb{R}$, we can think of a and b as the start and end times of our hike, and we can think of $f(x)$ as our elevation at time x. Then f must be continuous to rule out teleportation and $f(a) = f(b)$ so that we begin and end at the same elevation. Each level set of f can be thought of as a collection of times of equal elevation (we will call these times *isopoints*). In this paper we will study the distances between all isopoints. In our framing, the distance between isopoints should be thought of as a time-distance; however, thinking of a and b as locations in space is an equally valid interpretation.

In Theorem 1, we will show that every duration of time that evenly divides the total length of the hike is a distance between isopoints. In Theorem 2, we will prove that that Theorem 1 gives the *only* such distances common to all possible hikes. Lastly, we will prove a sequence of Lemmas to build up to Theorem 3, our main result, which states that at least one third of the time-lengths between 0 and the total duration of the hike are distances between isopoints.

2 Notation

Given a nonempty closed interval $[a,b] \subset \mathbb{R}$ and a real number λ, we will use $C_\lambda([a,b])$ to represent the set of continuous functions on $[a,b]$ mapping both endpoints to λ. More precisely,

$$C_\lambda([a,b]) := \{f : [a,b] \to \mathbb{R} \mid f \text{ is continuous and } f(a) = f(b) = \lambda\}$$

We will use $C_\mathbb{R}([a,b])$ to refer to all functions in some $C_\lambda([a,b])$:

$$C_\mathbb{R}([a,b]) := \bigcup_{\lambda \in \mathbb{R}} C_\lambda([a,b]) = \{f : [a,b] \to \mathbb{R} \mid f \text{ is continuous and } f(a) = f(b)\}$$

Given a function $f \in C_\mathbb{R}([a,b])$, and a subset $X \subseteq [a,b]$, let

$$D_f(X) := \{d > 0 : |x - y| = d \text{ and } f(x) = f(y) \text{ for some } x, y \in X\}$$

If X is absent as in D_f, assume $X = [a,b]$.

The contents of this paper are motivated by Exercise 5.4.6. in [1].

A^{o}, $\overline{A}$, and ∂A will be used to denote the topological interior, closure, and boundary of A respectively. $\mu(A)$ will be used for the Lebesgue measure of A.

3 Main Results

We begin by proving that, for every $f \in C_{\mathbb{R}}([a,b])$, D_f contains the sequence $b-a$, $\frac{b-a}{2}$, $\frac{b-a}{3}$, $\ldots$

Theorem 1. *Let f be a real-valued, continuous function on the closed interval $[a,b]$ such that $f(a) = f(b)$. Given any $n \in \mathbb{N}$, there exist x and y in $[a,b]$ such that $|x-y| = \frac{b-a}{n}$ and $f(x) = f(y)$.*

Proof. We may assume without loss of generality that $[a,b] = [0,1]$. If not, just apply the result to $f(a+(b-a)x)$.

Define $g(x) = f(x+\frac{1}{n}) - f(x)$ and consider the sum

$$g(0) + g\left(\frac{1}{n}\right) + g\left(\frac{2}{n}\right) + \cdots + g\left(\frac{n-1}{n}\right) \tag{1}$$

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

$$= f(1) - f(0) = 0 \tag{3}$$

where (3) follows from (2) due to cancellation.

If every term in (1) is 0, then the result follows immediately because $f(\frac{k+1}{n}) = f(\frac{k}{n})$ for $k = 0,1,\ldots,n-1$. If (1) contains one or more nonzero terms, then there must be at least one positive and one negative term in order for the sum to be zero. That is, $g\left(\frac{k_1}{n}\right) < 0$ and $g\left(\frac{k_2}{n}\right) > 0$ for some integers k_1 and k_2 between 0 and $n-1$. Thus, by the intermediate value theorem, $g(c) = 0$ for some c between $\frac{k_1}{n}$ and $\frac{k_2}{n}$ (the continuity of g follows from the continuity of f). Therefore, we have $f(c+\frac{1}{n}) - f(c) = 0$. □

Theorem 1 provides a partial answer to the question posed in the introduction. If we hike for an hour, there will be two instants, 30 minutes apart, of equal elevation because 30 minutes is half of an hour. The same is true for 20 minutes, 15 minutes, etc. We are not done, however, because we haven't ruled out other durations. Our next result shows that no other duration is *guaranteed* to separated two equipoints.

Theorem 2. *Given a closed interval $[a,b]$, let $0 < d < b-a$. If d is not of the form $\frac{b-a}{n}$ for some $n \in \mathbb{N}$, then there exists a continuous function $f : [a,b] \to \mathbb{R}$ with $f(a) = f(b)$ such that $d \notin D_f$.*

Proof. Once again, we can assume without loss of generality that $[a,b] = [0,1]$. First, let $p(x)$ be any continuous d-periodic function with $p(0) \neq p(1)$. Note that the existence of such functions hinges on the fact that $d \neq \frac{1}{n}$. Next, let $m(x)$ be any strictly monotone continuous function such that $m(0) = 0$ and $m(1) = p(0) - p(1)$. We can insist on strict monotonicity since $m(0) = 0 \neq p(0) - p(1) = m(1)$. Then $p+m$ is

The definitions of $C_{\mathbb{R}}([a,b])$ and D_f are given in Section 2: Notation.

continuous as the sum of continuous functions. Furthermore, $(p+m)(0) = p(0) = p(1) + p(0) - p(1) = (p+m)(1)$.

To finish, we must show that $d \notin D_{p+m}$. Indeed, for all $x \in [0, 1-d]$, we have

$$\begin{aligned}(p+m)(x+d) - (p+m)(x) &= p(x+d) - p(x) + m(x+d) - m(x)\\ &= 0 + m(x+d) - m(x) \neq 0\end{aligned}$$

using the monotonicity of m and the periodicity of p. □

Taken together, Theorem 1 and Theorem 2 tell us that, on a hike that begins and ends at the same height, the only durations we know, a priori, will separate times of equal elevation, must evenly divide that total time of the hike. This is expressed formally in the following corollary:

Corollary 1.

$$\bigcap_{f \in C_{\mathbb{R}}([a,b])} D_f = \left\{ \frac{b-a}{n} : n \in \mathbb{N} \right\}.$$

Proof. Theorem 1 gives one inclusion and Theorem 2 gives the other. □

Corollary 1 characterizes the distances which are common to all functions in $C_{\mathbb{R}}([a,b])$. One then might wonder whether this represents a small intersection of large overlapping sets or there is a particular $f \in C_{\mathbb{R}}([a,b])$ such that $D_f = \left\{ \frac{b-a}{n} : n \in \mathbb{N} \right\}$. It turns out to be the former. Each D_f is considerably larger than the set of divisors of $b-a$. In fact, we show in Theorem 3 that each D_f contains at least a third of the numbers between 0 and $b-a$. Before we prove it, we need to develop a series of lemmas about D_f. We will start with results about the size of D_f for very simple functions, and generalize until we can analyze D_f for arbitrary $f \in C_{\mathbb{R}}([a,b])$. We begin by showing that shrinking the domain of f shrinks D_f.

Lemma 1. *If $A \subseteq B$, then $D_f(A) \subseteq D_f(B)$.*

Proof. Assume $d \in D_f(A)$. Then there are points $x, y \in A$ such that $|x-y| = d$ and $f(x) = f(y)$. But $A \subseteq B$, so x and y are also in B. Thus, $d \in D_f(B)$. □

Next, we will show that for constant functions f, D_f is at least as big as the domain of f.

Lemma 2. *Let f be a constant function on a bounded set $A \subset \mathbb{R}$. Assume A has a maximum value m. Then $\mu(D_f(A)) \geq \mu(A)$.*

Proof. Notice that $D_f(A)$ contains the set $m - A = \{m - a \mid a \in A\}$. Therefore $\mu(D_f(A)) \geq \mu(m-A) = \mu(A)$. □

In subsequent lemmas, it will be convenient to make assumptions like $f(x) > \lambda$ for all $x \in A$ or $\max_{A_1}(f) \leq \max_{A_2}(f)$. To help ensure we don't lose any generality, we will prove that certain transformations of f preserve D_f. More precisely, we will prove that D_f is invariant with repsect to horizontal and vertical reflections and translations of the graph of f. Since it's no extra work, we will prove a more general fact: that applying injective functions to the range of f and isometric functions to the domain of f does not affect D_f.

Lemma 3. *Let f be any real valued function on $A \subseteq \mathbb{R}$. If $g : f(A) \to \mathbb{R}$ is injective and $T : A \to \mathbb{R}$ is isometric, then $D_f(A) = D_{g\circ f}(A) = D_{f\circ T^{-1}}(T(A))$.*

Proof. Since g is injective, $f(x) = f(y)$ if and only if $g(f(x)) = g(f(y))$. Hence $d \in D_f(A)$ if and only if $d \in D_{g\circ f}(A)$, and so $D_f(A) = D_{g\circ f}(A)$.

Since all isometries are invertible, we have $f(x) = f(y)$ if and only if $f(T^{-1}Tx) = f(T^{-1}Ty)$. Furthermore, $|Tx - Ty| = |x - y|$ because T is isometric. Therefore, given any $d \geq 0$ there exist points $x, y \in A$ such that $d = |x - y|$ and $f(x) = f(y)$ if and only if there exists points $x', y' \in T(A)$ such that $f(T^{-1}x') = f(T^{-1}y')$ and $d = |x' - y'|$. Indeed, this correspondance is given by $x' = Tx$ and $y' = Ty$. Therefore, $D_f(A) = D_{f\circ T^{-1}}(T(A))$. □

In the next lemma, we will consider the case of continuous functions $f : [a,b] \to \mathbb{R}$ where $f(a)$ and $f(b)$ are both either global minima or global maxima. In other words, we will look at functions $f \in C_\lambda([a,b])$ where a and b are the *only* points where $f = \lambda$. It becomes quite easy to calculate D_f in this case.

Lemma 4. *Let $f \in C_\lambda([a,b])$ and suppose either $f(x) > \lambda$ for all $a < x < b$ or $f(x) < \lambda$ for all $a < x < b$. Then $D_f([a,b]) = (0, b-a]$.*

Proof. We may assume without loss of generality that $f(x) > \lambda$ for all $a < x < b$ because $D_f([a,b])$ does not change when the graph of f is reflected over the line $y = \lambda$, i.e., $D_f([a,b]) = D_{-f+2\lambda}([a,b])$, as we established in Lemma 3.

It is clear that $b - a \in D_f([a,b])$ since $f(a) = f(b)$, so we will let $d \in (0, b-a)$ and show that $d \in D_f([a,b])$. Define $g(x) = f(x+d) - f(x)$. Note that $g(a) = f(a+d) - f(a) = f(a+d) - \lambda > 0$ because $f(a+d) > \lambda$. Also, $g(b-d) = f(b) - f(b-d) = \lambda - f(b-d) < 0$ because $f(b-d) > \lambda$.

The intermediate value theorem guarantees the existence of a $c \in (a, b-d)$ such that $g(c) = f(c+d) - f(c) = 0$, i.e., $f(c+d) = f(c)$. Therefore $d \in D_f([a,b])$. □

Having settled the case where the global minima or maxima of f are located at the endpoints of a single closed interval, we will now ask the same question when f is defined on the union of two intervals. Once again, assuming that f has global minima at every endpoint point or global maxima at every endpoint, what does D_f look like? By Lemma 4, we already know how each interval will contribute to D_f when considered separately. In the following lemma, we characterize the ”interactions” between the two intervals. For convenience, we will assume that every endpoint is the location of a global minimum and we will make an assumption about where the global maximum is located. After proving the lemma, we will discuss how those assumptions can be discarded using Lemma 3.

Lemma 5. *Given any $a_1 < a_2 \leq a_3 < a_4$, define $A = [a_1, a_2] \cup [a_3, a_4]$, and let $f : A \to \mathbb{R}$ be a continuous function such that $f(a_k) = \lambda$ for $1 \leq k \leq 4$. If $f(x) > \lambda$ for all $x \in A^{\text{o}}$ and $\max_{[a_1,a_2]}(f) \geq \max_{[a_3,a_4]}(f)$, then $D_f(A) \supseteq [a_3 - a_1, a_4 - a_1]$.*

Proof. It is clear that $a_3 - a_1, a_4 - a_1 \in D_f(A)$ since $f(a_1) = f(a_3) = f(a_4)$, so we will let $d \in (a_3 - a_1, a_4 - a_1)$ and show that $d \in D_f(A)$. We will do this in three cases, depending on whether d is greater than, less than, or equal to $a_4 - a_2$. Define $g(x) = f(x+d) - f(x)$.

Case 1: $d > a_4 - a_2$.

In this case, we compute $g(a_1) = f(a_1+d) - f(a_1) = f(a_1+d) - \lambda > 0$ and $g(a_4 - d) = f(a_4) - f(a_4 - d) = \lambda - f(a_4 - d) < 0$. Here, we've used that $a_1 + d \in (a_3, a_4)$ and $a_4 - d \in (a_1, a_2)$ and $f > \lambda$ on these two open intervals. The intermediate value theorem then guarantees a $c \in (a_1, a_4 - d)$ such that $g(c) = f(c+d) - f(c) = 0$. Hence $d \in D_f(A)$.

Case 2: $d < a_4 - a_2$.

In this case, once again we compute $g(a_1) = f(a_1+d) - f(a_1) = f(a_1+d) - \lambda > 0$. This time, however, we observe that $g(t) \leq 0$ for some $t \in (a_1, a_2)$. Otherwise, we would have $f(t+d) > f(t)$ for all $t \in (a_1, a_2)$, contradicting the assumption $\max_{[a_1,a_2]}(f) \geq \max_{[a_3,a_4]}(f)$.

If $g(t) = 0$, we have $f(t+d) = f(t)$. If $g(t) < 0$, then the intermediate value theorem gives a $c \in (t, a_2)$ such that $g(c) = f(c+d) - f(c) = 0$. In either case, $d \in D_f(A)$.

Case 3: $d = a_4 - a_2$.

This case is trivial as $f(a_2) = \lambda = f(a_4) = f(a_2 + d)$. □

The hypotheses we've inserted into Lemma 5 impose significant constraints on the scope of the result, so it's worth pausing to consider how these can be relaxed, beginning with the assumption that $f(x) > \lambda$ for all $x \in A^{\text{o}}$. It's not so simple as stating that D_f is invariant with respect to vertical reflections since the inequality $\max_{[a_1,a_2]}(f) \geq \max_{[a_3,a_4]}(f)$ becomes $\min_{[a_1,a_2]}(-f) \leq \min_{[a_3,a_4]}(-f)$. However, as a corollary to Lemma 5, we will show that we can combine the two cases by insisting that $\max_{[a_1,a_2]}(|f-\lambda|) \geq \max_{[a_3,a_4]}(|f-\lambda|)$.

Corollary 2. *Given any $a_1 < a_2 \leq a_3 < a_4$, define $A = [a_1, a_2] \cup [a_3, a_4]$, and let $f : A \to \mathbb{R}$ be a continuous function such that $f(a_k) = \lambda$ for $1 \leq k \leq 4$. Suppose either $f(x) > \lambda$ for all $x \in A^{\text{o}}$ or $f(x) < \lambda$ for all $x \in A^{\text{o}}$. If $\max_{[a_1,a_2]}(|f-\lambda|) \geq \max_{[a_3,a_4]}(|f-\lambda|)$, then $D_f(A) \supseteq [a_3 - a_1, a_4 - a_1]$.*

Proof. If $f(x) > \lambda$ for all $x \in A^{\text{o}}$ then $|f - \lambda| = f - \lambda$, and adding λ to both sides of $\max_{[a_1,a_2]}(f-\lambda) \geq \max_{[a_3,a_4]}(f-\lambda)$ gives $\max_{[a_1,a_2]}(f) \geq \max_{[a_3,a_4]}(f)$, so $D_f(A) \supseteq [a_3 - a_1, a_4 - a_1]$ by Lemma 5.

On the other hand, if $f(x) < \lambda$ for all $x \in A^{\text{o}}$ then $|f - \lambda| = -f + \lambda$ and subtracting λ from both sides of $\max_{[a_1,a_2]}(-f+\lambda) \geq \max_{[a_3,a_4]}(-f+\lambda)$ gives $\max_{[a_1,a_2]}(-f) \geq \max_{[a_3,a_4]}(-f)$. Applying Lemma 5 to $-f$ and invoking the invariance proven in Lemma 3, we have $D_f(A) = D_{-f}(A) \supseteq [a_3 - a_1, a_4 - a_1]$. □

We cannot easily discard the $\max_{[a_1,a_2]}(|f-\lambda|) \geq \max_{[a_3,a_4]}(|f-\lambda|)$ hypothesis of Corollary 2. It's tempting to say that we do not lose generality due to Lemma 3, however, we must be careful. Indeed, Lemma 3 says that $D_f(A) = D_{f\circ T^{-1}}(T(A))$ for all isometries $T : A \to \mathbb{R}$. However, given $f : [a_1,a_2] \cup [a_3,a_4] \to \mathbb{R}$ such that $\max_{[a_1,a_2]}(|f-\lambda|) < \max_{[a_3,a_4]}(|f-\lambda|)$, there is no isometry T mapping $[a_1,a_2] \cup [a_3,a_4]$ to itself such that $\max_{[a_1,a_2]}(|f\circ T^{-1}-\lambda|) \geq \max_{[a_3,a_4]}(|f\circ T^{-1}-\lambda|)$ unless $[a_1,a_2]$ and $[a_3,a_4]$ are the same length.

Now that we've studied D_f for functions defined on a single interval and the union of two intervals, we will generalize to functions defined on n intervals. Locating specific intervals becomes very complicated due to interactions among the intervals, so we will return to our goal of lower bounding the size of D_f.

Lemma 6. *Given any $a_1 < b_1 \leq a_2 < b_2 \leq \cdots \leq a_n < b_n$, define $A = \cup_{k=1}^{n}[a_k,b_k]$ and let $f : A \to \mathbb{R}$ be a continuous function such that $f(a_k) = f(b_k) = \lambda$ for $1 \leq k \leq n$. Suppose either $f(x) > \lambda$ for all $x \in A^{\mathrm{o}}$ or $f(x) < \lambda$ for all $x \in A^{\mathrm{o}}$. Then $\mu(D_f(A)) \geq \mu(A)$.*

Proof. We will use proof by induction on n, the number of intervals.

Base case (n=1):

The base case is covered by Lemma 4, which gives us $D_f([a_1,b_1]) = (0,b_1-a_1]$. Therefore, $\mu(D_f([a_1,b_1])) = \mu([a_1,b_1]) = b_1-a_1$.

Induction Step:

Our goal is to prove that $\mu(D_f(\cup_{k=1}^{n+1}[a_k,b_k])) \geq \mu(\cup_{k=1}^{n+1}[a_k,b_k])$. Assume, without loss of generality, that $\max_{[a_1,b_1]}(|f-\lambda|) \geq \max_{[a_{n+1},b_{n+1}]}(|f-\lambda|)$. We do not lose generality because D_f is invariant with respect to horizontal reflections, i.e., $D_{f(x)} = D_{f(-x)}$. Then, by Corollary 2, $D_f([a_1,b_1] \cup [a_{n+1},b_{n+1}]) \supseteq (a_{n+1}-a_1, b_{n+1}-a_1)$. Combining this fact with Lemma 1 gives

$$\begin{aligned} D_f(\cup_{k=1}^{n+1}[a_k,b_k]) &\supseteq D_f(\cup_{k=1}^{n}[a_k,b_k]) \cup D_f([a_1,b_2] \cup [a_{n+1},b_{n+1}]) \\ &\supseteq D_f(\cup_{k=1}^{n}[a_k,b_k]) \cup (a_{n+1}-a_1, b_{n+1}-a_1). \end{aligned}$$

Next, observe that $(a_{n+1}-a_1, b_{n+1}-a_1)$ and $D_f(\cup_{k=1}^{n}[a_k,b_k])$ are disjoint. Indeed, if $d \in D_f(\cup_{k=1}^{n}[a_k,b_k])$, then $d \leq b_n - a_1 \leq a_{n+1} - a_1$. Computing the length of both sides and applying the induction hypothesis, we get

$$\begin{aligned} \mu(D_f(\cup_{k=1}^{n+1}[a_k,b_k])) &\geq \mu(D_f(\cup_{k=1}^{n}[a_k,b_k]) \cup (a_{n+1}-a_1, b_{n+1}-a_1)) \\ &= \mu(D_f(\cup_{k=1}^{n}[a_k,b_k])) + \mu((a_{n+1}-a_1, b_{n+1}-a_1)) \\ &\geq \mu(\cup_{k=1}^{n}[a_k,b_k]) + \mu((a_{n+1}-a_1, b_{n+1}-a_1)) \\ &= \mu(\cup_{k=1}^{n+1}[a_k,b_k]) \end{aligned}$$

□

Having established a lower bound on the size of D_f when f is defined on the finite union of closed intervals, we will now use a simple limiting argument to generalize to functions defined on the countable union of closed intervals.

Lemma 7. *Let $\{I_n\}$ be a countable collection of closed intervals and define $A = \cup_{n=1}^{\infty} I_n$. Assume that A is bounded and $\{I_n\}$ have disjoint interiors. Let f be a continuous function on A such that $f(x) = \lambda$ on the endpoints of each I_n and either $f(x) > \lambda$ for all $x \in A^{\mathrm{o}}$ or $f(x) < \lambda$ for all $x \in A^{\mathrm{o}}$. Then $\mu(D_f(A)) \geq \mu(A)$.*

Proof. Fix $\varepsilon > 0$. Since A is bounded and $\{I_n\}$ have disjoint interiors, we know that $\lim_{n\to\infty} \mu\left(\cup_{k=n}^{\infty} I_k\right) = 0$. Thus there exists some $N \in \mathbb{N}$ such that $\mu\left(\cup_{k=N}^{\infty} I_k\right) < \varepsilon$. Applying Lemma 6 and Lemma 1 yields

$$\begin{aligned} \mu(D_f(A)) &\geq \mu\left(D_f\left(\cup_{k=1}^{N} I_k\right)\right) \\ &\geq \mu\left(\cup_{k=1}^{N} I_k\right) \\ &= \mu(A) - \mu\left(\cup_{k=N}^{\infty} I_k\right) \\ &> \mu(A) - \varepsilon \end{aligned}$$

Therefore, $\mu(D_f(A)) \geq \mu(A)$ because ε was arbitrary. □

With access to these lemmas, we are now prepared to prove that $D_f([a,b])$ must contain at least a third of the distances in $(0, b-a]$.

Theorem 3. *If $f \in C_\lambda([a,b])$ then $\mu(D_f) \geq \frac{b-a}{3}$.*

Proof. Let $A_>$, $A_<$, and $A_=$ be the subsets of $[a,b]$ on which f is greater than, less than, and equal to λ respectively.

$A_>$ and $A_<$ are the preimages of open sets under a continuous function and are thus open. Therefore, each is a countable union of open intervals. Applying Lemma 7 to the closure of each tells us that $\mu(D_f(\overline{A_>})) \geq \mu(\overline{A_>}) = \mu(A_>)$ and $\mu(D_f(\overline{A_<})) \geq \mu(\overline{A_<}) = \mu(A_<)$[1]. Applying Lemma 2 to $A_=$ gives $\mu(D_f(A_=)) \geq \mu(A_=)$. Combining these three inequalities with Lemma 1, we have

$$\begin{aligned} \mu(D_f([a,b])) &\geq \max\left(\mu(D_f(\overline{A_>})), \mu(D_f(\overline{A_<})), \mu(D_f(A_=))\right) \\ &\geq \max\left(\mu(A_>), \mu(A_<), \mu(A_=)\right) \\ &\geq \frac{b-a}{3} \end{aligned}$$

where the last line follows from $\mu(A_>) + \mu(A_<) + \mu(A_=) = b-a$. □

4 Future Work

4.1 Is $\frac{b-a}{3}$ a minimum?

Theorem 3 establishes a lower bound on D_f for functions in $C_\lambda([a,b])$. The key was to restrict our attention to $A_> = \{x : f(x) > \lambda\}$ because if $f(x) = f(y)$, then either both x

[1] Dropping the closure doesn't change the length because the union of countably many intervals has a countable boundary.

and y are in $A_>$ or neither are. The same holds for $A_<$ and $A_=$. In other words, points in D_f cannot arise due to "interactions" among $A_>$, $A_<$, and $A_=$. With this in mind, the bound in Theorem 3 seems tight: simply define a function which is positive on the first third of $[a,b]$, negative on the second third, and zero on the last third. Then each of $A_>$, $A_<$, and $A_=$ should contribute $(0,\frac{b-a}{3}]$ to D_f. For example, let

$$f(x) = \begin{cases} \sin x & \text{if } 0 \leq x \leq 2\pi \\ 0 & \text{if } 2\pi \leq x \leq 3\pi \end{cases}.$$

The reason this strategy doesn't work is $A_=$. Indeed, $D_f(A_>) = D_f(A_<) = (0,\pi]$. However, $A_= = \{0,\pi\} \cup [2\pi,3\pi]$ and $D_f(A_=) = (0,3\pi]$. This makes D_f as large as possible due to interactions between the points 0 and π and the interval $[2\pi,3\pi]$.

The trouble with the previous example is the presence of isolated points 0 and π in $A_=$. The former is unavoidable, but we can eliminate the latter by making f zero *between* the intervals on which it is positive and negative. Let

$$f(x) = \begin{cases} \sin x & \text{if } 0 \leq x \leq \pi \\ 0 & \text{if } \pi \leq x \leq 2\pi \\ -\sin x & \text{if } 2\pi \leq x \leq 3\pi \end{cases}.$$

Now $A_= = \{0,3\pi\} \cup [\pi,2\pi]$ and $D_f(A_=) = (0,2\pi]$, but D_f is still strictly greater than the bound established in Theorem 3.

Had we defined D_f slightly differently to ignore the endpoints of the domain of f, the previous example would prove Theorem 3 is sharp. More precisely, if we instead define $D_f = \{d > 0 : |x-y| = d \text{ and } f(x) = f(y) \text{ for some } a < x < y < b\}$, then $D_f = (0,\pi]$ in the previous example.

However, if we stick to our original definition, is there an $f \in C_\lambda([a,b])$ with $\mu(D_f) = \frac{b-a}{3}$? If not, what is the infimum of D_f over all such f?

4.2 Generalization

What does $D_f(X)$ look like when X is not a closed interval? We could broaden the class of functions we look at by defining

$$C_\lambda(X) = \left\{ f : \overline{X} \to \mathbb{R} \mid f \text{ is continuous and } f(x) = \lambda \text{ for all } x \in \partial X \right\}.$$

What is $\bigcap D_f$ over all such f and what is the infimum of $\mu(D_f)$?

We could also explore functions with an n-dimensional domain and/or m-dimensional codomain.

What does $D_f(X)$ look like when X is n-dimensional? The more general definition of $C_\lambda(X)$ proposed above works just fine in this case. For simplicity, we might want to start with cubes or spheres, and slowly relax the constraints on X. Additionally, as in Section 4.1 we should amend the definition D_f to ignore the boundary of X. Otherwise $D_f = (0,(X)]$ always (unless X is disconnected).

What does $D_f([a,b])$ look like when the codomain of f is m-dimensional? If $m > 1$ the minimum $D_f([a,b])$ becomes $\{b-a\}$. Consider, for example, $f : [0,2\pi] \to \mathbb{R}^2$ defined by $f : x \mapsto (\cos x, \sin x)$. The only pair of points in $[0, 2\pi]$ which get mapped to the same output are 0 and 2π, so $D_f = \{2\pi\}$.

To construct an interesting generalization, we must then restrict our attention to functions mapping a closed interval to some subset $A \subset \mathbb{R}^m$. If A contains any "loops," the minimum $D_f([a,b])$ becomes $\{b-a\}$, so A should be a one-dimensional "loop-free" set.

Lastly, if the previous questions are settled, perhaps we could define

$$C_\lambda(X,Y) = \left\{ f : \overline{X} \to Y \mid f \text{ is continuous and } f(x) = \lambda \text{ for all } x \in \partial X \right\}$$

and classify D_f in terms of $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$.

Bibliography

[1] Abbott, Stephen, *Understanding Analysis,* Undergraduate Texts in Mathematics, Springer, New York, 2015.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 59 – 75

ICA can consistently bin similar sources together: The case with 3 sinusoidal sources separated into 2 components

Erin Munro Krull, Breanna Ollech, Kayley Grabowski*

Erin Munro Krull is an assistant professor in the department of mathematical sciences at Ripon College. Her research interests are in computational neuroscience, particularly data analysis and mathematical modeling applied to sleep and epilepsy.

Breanna Ollech worked on this paper as a senior at Ripon College studying Mathematics, Secondary Education and Exercise science. She is currently a first year mathematics teacher at OshKosh West High School. Along with teaching, she is also the softball coach and source of strength student mentor.

Kayley Grabowski worked on this paper as a senior at Ripon College studying Mathematics, Economics, and Business Management. She is currently a PhD candidate in Economics at the University of Illinois at Chicago. Outside of school, Kayley enjoys baking cookies, watching and playing sports, and being outdoors.

***Corresponding author:** munrokrulle@ripon.edu

Abstract

Independent Component Analysis (ICA) is a blind-source separation method, meaning that it takes in a recording with multiple sensors and attempts to unmix it into the original sources. For example, suppose there are 4 people (sources) speaking in a room with 4 microphones (sensors), then ICA unmixes the recording from the 4 microphones to give tracks of the individual people called ICA components. ICA is currently used to decompose a variety of signals with many sensors, including fMRI and EEG data. However, its use in interpreting data with fewer sensors, such as the local field potential (LFP), is limited because of concerns about how it handles over-complete data (data with more sources than sensors). While there has been some success in enhancing ICA so that it can extract more sources than sensors, we focus on how ICA handles over-complete data. In this paper, we show that ICA consistently bins sources with similar spatial maps together when there are 3 sinusoidal sources and 2 sensors.

1 Introduction

Many neurophysiological recordings of the brain used to study micro-circuits include the local field potential (LFP). There are a wide variety of LFP recordings freely available [45], as well as recent technological developments in recording the LFP [19, 22]. One of the benefits of field potential recordings is that they often simultaneously record nearby action potentials along with field potentials reflecting the summation of many cells acting at once, some possibly from far away [26, 46]. While these recordings may have limited spatial resolution, they tend to have very high time resolution. Standard methods of decomposing the LFP for further analysis include spectral analysis, which addresses frequency content, and current source density, which uses field potential physical properties to derive current sources and sinks. Another decomposition method is independent component analysis (ICA), which can be used to separate overlapping sources that contributed to the recording. For example, suppose there are two voices recorded over two microphones, so that their amplitude differs on each microphone. By taking advantage of the distinct voices and their spatial differences in amplitude, ICA attempts to separate the two-sensor recording so that each component contains an individual voice. (See figure 1 for an illustration.)

ICA is one of many methods of blind-source separation. While there are many resources describing this method in detail [6, 24, 25], we will briefly describe the framework here using the LFP as an example. Suppose there are n sources $\vec{s}(t) = \begin{bmatrix} s_0(t) & s_1(t) & \dots & s_{n-1}(t) \end{bmatrix}$ affecting the LFP, and the LFP is recorded by an electrode with n sensors $\vec{x}(t) = \begin{bmatrix} x_0(t) & x_1(t) & \dots & x_{n-1}(t) \end{bmatrix}$. ICA will take the data from the sensors, and separate them into n components $\vec{c}(t) = \begin{bmatrix} c_0(t) & c_1(t) & \dots & c_{n-1}(t) \end{bmatrix}$ so that the time series of the components are as statistically independent as possible. ICA assumes that sources are mixed linearly onto each sensor. That means ICA assumes there is some mixing matrix M so that $\vec{x} = M\vec{s}$, where the columns of M represent the source spatial maps or relative amplitude across sensors. Likewise, ICA decomposes signals by returning an unmixing matrix U, so that $\vec{c} = U\vec{x}$. Assuming all the original sources are independent, then $\vec{c}(t)$ is a linear estimate of $\vec{s}(t)$. Unlike the other common decomposition methods principal component analysis (PCA) or factor analysis (FA), ICA does not necessarily make restrictions on whether the spatial maps of the components are oriented at right-angles, and doesn't necessarily favor high-amplitude directions. This makes ICA a relatively flexible decomposition method, which is used in many fields besides neuroscience [36], including analytical chemistry [33], cancer omics datasets [43], gravity and magnetic signal processing [49], and image processing [4, 9, 10].

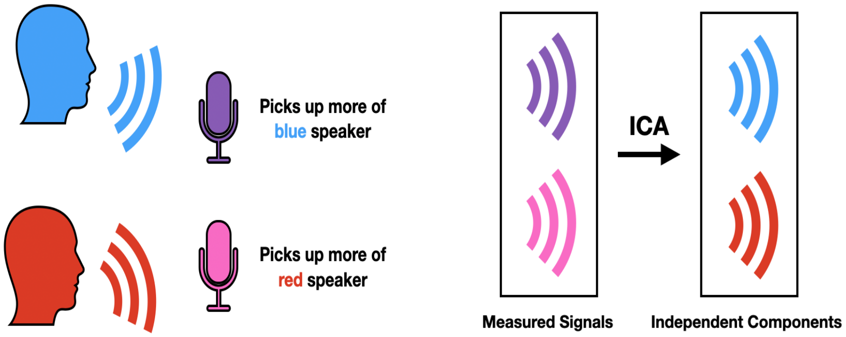


Figure 1: Illustration of ICA A) Different sources have differing amplitudes on each sensor. Therefore, each sensor will have its own mixture of all sources. In this illustration, the top microphone picks up more of the blue speaker than the red speaker, so its voice mixture is represented as deep purple. At the same time, the bottom microphone picks up more of the red speaker than the blue speaker, so its mixture is represented as fuchsia. B) ICA attempts to separate sources so they are as independent as possible. In this case, ICA would take the recorded mixtures and attempt to separate them into components containing the original voices. (Figure originally published in [44], reprinted with permission.)

ICA is relatively straightforward to interpret if the recording involves a lot of sensors. If there are more sensors than sources, then higher amplitude components may summarize major sources (above the signal-to-noise ratio), while smaller amplitude components may represent noise in the recording. Currently, ICA is used to analyze a variety of neurophysiological data, including fMRI [8], MRI [50], MEG [5], EEG [3, 36], voltage sensitive dye [1, 14, 21, 39], and PET [48]. Many of these data sets have an abundant number of sensors that are assumed to be greater than the number of relevant sources, and there is research on how to choose a subset of ICA components so that they are reliable and match biologically plausible or known sources [11, 13, 31, 51]. There are several studies that use ICA to interpret LFP data, even though these recordings tend to have fewer sensors [20]. However, these studies cannot necessarily assume that there are fewer relevant sources than sensors.

While ICA is thought to work well if there are no more sources than sensors, it is still unclear how ICA handles over-complete or under-determined data where there are more sources than sensors. Several methods try to address this issue by modifying the ICA algorithm so that more sources are extracted, possibly taking advantage of sparseness or other features in the data [36]. There are other studies which demonstrate that ICA can produce consistent results, across multiple ICA runs and subjects, even when data is over-complete [2, 13, 15, 25, 27, 28, 32, 35]. While some of these studies compare ICA components by measuring similarity between spatial maps, none of these address how and why certain sources may be combined into a single component. The uncertainty in how to interpret ICA components extracted from over-complete data, along with some instability of ICA components, may be the reason why ICA is not always recommended as an analysis tool for decomposing LFP. Reviews may instead point to decompositions which rely more on spectral analysis or on forward models of known biophysical structures [12, 17, 19, 29, 38, 41, 47]. On the other hand, ICA is more readily used in studies involving EEG [2, 3, 30] and fMRI [48, 42], where recordings tend to contain many more sensors.

We focus on describing how ICA handles over-complete data. In particular, we ran simulations to test how ICA separates 3 noisy sinusoidal sources recorded on 2 sensors into 2 components. Our results show that ICA can separate sources in a predictable manner, namely sources with similar spatial maps across the sensors are binned together. There is very little variation in how sources are binned together across different ICA runs, as long as 2 of the 3 sources are closer together in terms of their spatial map. If all 3 sources have equidistant spatial maps, then we see more variation between ICA runs. Our results indicate we may be able to determine how ICA bins original sources together by looking at the reliability and spatial maps of ICA components over the sensors. Moreover, viewing ICA components as binned sources may have advantages in interpreting the original data. For example, in a recording of a 4-part chorus, we may be more interested in components that contain the 4 voice parts, not the individual voices.

2 Methods

We ran simulations using Google Colaboratory[1], which used Python version 3.6.9. For each simulation, we used the same 3 sources:

$$\begin{aligned} s_0 &= \sin(2\cdot 2\pi t) + \text{random noise} \\ s_1 &= \sin(3\cdot 2\pi t) + \text{random noise} \\ s_2 &= \sin(5\cdot 2\pi t) + \text{random noise} \end{aligned}$$

where the random noise is uniformly distributed over $[-0.5, 0.5]$. Our 3 sources are illustrated in figure 2. We used sinusoidal functions with relatively prime frequencies so we could easily distinguish which sources were separated into which components using Fourier analysis. We added non-Gaussian noise at the same amplitude as the sinusoidal function to help satisfy the conditions of ICA, which are that original sources are independent and non-Gaussian. Without noise, we may see a pattern in the data since the combined signals repeat every $2\cdot 3\cdot 5$ seconds. With the added noise, we will see that these signals appear fairly independent when plotted against each other. For our analysis, all sources were sampled at 1000 Hz, and recorded for 100 s.

We mixed the sources onto two sensors using a mixing matrix M of the following

[1] Google Research, `https://colab.research.google.com`, accessed June 2022

form:

$$M = \left[\begin{array}{ccc} \vec{m}_0 & \vec{m}_1 & \vec{m}_2 \end{array}\right] = \left[\begin{array}{ccc} \cos(\alpha_0) & \cos(\alpha_1) & \cos(\theta) \\ \sin(\alpha_0) & \sin(\alpha_1) & \sin(\theta) \end{array}\right]$$

so that the recorded data mixtures are $\vec{x}(t) = M\vec{s}(t)$. The angles α_0, α_1, and θ in our mixing vectors represent the spatial map or relative amplitude of the original sources on each sensor. For instance, an angle of 0° means the source is recorded entirely on the first sensor, while an angle of 90° means the source is recorded entirely on the second sensor. All other angles represent how the source is distributed across both sensors. Angles that are $> 90°$ or $< -0°$ represent mixtures where the source sign is flipped from one sensor to another, which can frequently occur in voltage recordings.

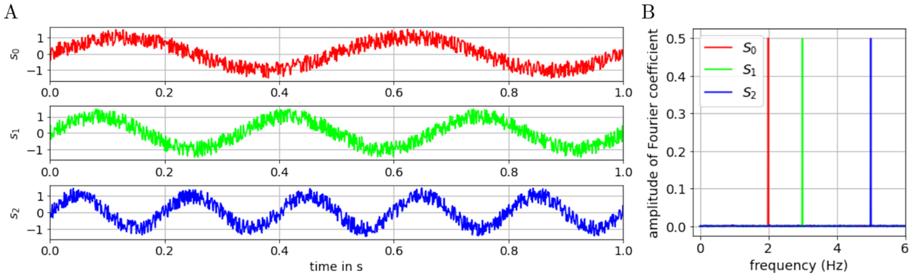


Figure 2: Figure 2: Original sources A) Time series of the 3 original sources. Each source is a sine wave at a different frequency combined with uniform noise. We used the same sources in all simulations. B) The amplitude of the Fourier transform of each source.

While the mixing vectors in M all have an amplitude of 1, the same data mixture can be produced by any reciprocal pair of mixing vectors and sources. For example, the negative mixing vector and matching source would produce the same mixture:

$$\left[\begin{array}{ccc} -\vec{m}_0 & \vec{m}_1 & \vec{m}_2 \end{array}\right] \left[\begin{array}{c} -s_0 \\ s_1 \\ s_2 \end{array}\right] = -\vec{m}_0(-s_0) + \vec{m}_1(s_1) + \vec{m}_2(s_2) = \vec{x}$$

Similarly, if we scale one mixing vector by a constant k and its source by $1/k$, we would also get the same data mixture. For this reason, the scaled mixing vector $k\vec{m}_i$ is considered equivalent to the unit vector $\vec{m}_i$ and we can use the angle to represent the spatial map for both vectors. Moreover, since $-\vec{m}_i$ is equivalent to $\vec{m}_i$, then we only consider angles from $-45°$ to $135°$.

We ran several cases, where we focus on a fixed α_0 and α_1 and vary θ from $-45°$ to $135°$. For each value of θ, we apply ICA to the data mixture 10 times, using both the FastICA algorithm [37] and the extended infomax algorithm ([16], Version 1.2.1: 10.5281/zenodo.7314185). The algorithm returns the un-mixing matrix U and estimated mixing matrix $M_{\text{est}} = U^{-1}$. Note that M_{est} is both the mixing matrix of the ICA components and the algorithm's best attempt at estimating M with a 2×2 matrix. We calculated the angles of the estimated mixing vectors within M_{est}, and compared them to the angles of the original mixing vectors.

We also applied the discrete Fourier transform to the ICA components [18], and compared the frequency content of the components with the original sources by calculating the relative amplitude as follows: Let $|y_i(t)|$ be the root mean square (RMS) amplitude of a signal $y_i(t)$, and $Y_i(\omega)$ be the Discrete Fourier transform coefficient of the signal at frequency ω Hz. Then the relative amplitude for frequency ω in component $c_i(t)$ compared to the original source $s_j(t)$ is $\frac{|C_i(\omega)|/|c_i(t)|}{|S_j(\omega)|/|s_j(t)|}$.

3 Results

3.1 A detailed case: spatial maps of the first two sources are 90° apart.

For our first case, we show what happens in detail with 2 sources and 2 sensors when $\alpha_0 = 0^\circ$ and $\alpha_1 = 90^\circ$, and with 3 sources and 2 sensors where $\theta = -30^\circ$ for the third source. We used the FastICA algorithm for all results shown below. The extended infomax algorithm yielded nearly identical results. Figure 3 shows the data mixtures presented to ICA, along with the time series for the components $U\vec{x}(t) = \vec{c}(t)$ and their Fourier transforms. Note that, while we give the ICA algorithm the entire data mixture, ICA disregards all time information. For example, in the 2-source case, the algorithm only takes the data set shown in figure 3Ai into account. ICA appears to separate the 2 sources mixed onto 2 sensors perfectly. On the other hand, the 2-dimensional data mixture of the 3 sources appears to be separated so that source s_1 is contained in c_0 while sources s_0 and s_2 are mostly contained in c_1. It appears that s_1 is separated out of c_1 entirely, while c_0 contains some amount of the frequencies found in s_0 and s_2, though not enough to cloud s_1.

To compare the ICA components with the original sources further, we also looked at the estimated mixing matrix $M_{\text{est}} = U^{-1}$, whose columns are mixing vectors which represent the ICA component's spatial map or relative amplitude across the sensors. Both the original mixing vectors and the estimated mixing vectors are shown in figure 4. Since the scale of the mixing vectors doesn't affect the spatial maps, as explained in Methods, we can represent them with an angle between -45° and 135°.

We combine the frequency and angle comparisons, as shown in figure 5. The angles illustrate how the spatial map of the original sources may be binned in the ICA components. Frequency content is represented by color, where the relative amplitude of each frequency is represented by color intensity. With these comparisons, we see that the separation for 2 sources is very close to the original sources. Likewise, for 3 sources one ICA component is almost entirely dedicated to the original source with angle 90°, while the other two sources are binned in terms of both angle and frequency content. In fact, figure 5 shows results for 10 ICA runs, overlaid on top of each other. These results show that there is very little variation between ICA runs for these data mixtures.

We now expand our analysis so that θ varies from -45° to 135°. Figure 6 shows three cases where α_0 and α_1 are 90° apart. We see in all of these cases that the ICA components follow the original sources in both angle and frequency content. The ICA components tend to bin together whichever sources have spatial maps with closer

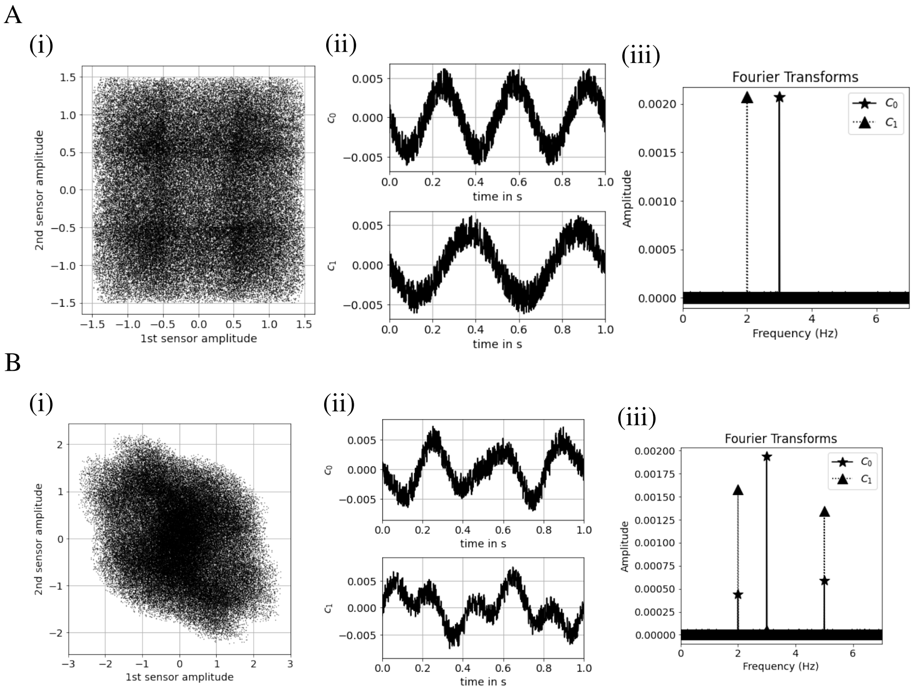


Figure 3: ICA applied to mixtures with 2 or 3 sources A) (i) The data mixture with 2 sources, $\alpha_0 = 0°$, and $\alpha_1 = 90°$. This is the same data where s_0 is recorded entirely on the first sensor and s_1 is recorded entirely on the second sensor. (ii) Example ICA components. Components can be returned in either order, and may be scaled differently than the original sources. Both of these components are negative relative to the original sources. (iii) The amplitude of the component Fourier coefficients confirm that the sources are separated very well, with each component almost entirely containing a single frequency. B) (i) The sources s_0, s_1, and s_2 mixed onto the two sensors using M with $\alpha_1 = 0°$, $\alpha_2 = 90°$, and $\theta = -30°$. (ii) Example ICA components. The first component c_0 appears to contain mostly s_1 (which oscillates at 3 Hz), while the second component c_1 appears to be a mixture of s_0 and s_2. (iii) The amplitude of the component Fourier coefficients confirm that c_0 is predominantly composed of s_1, but also contains some of the other sources. We also see that c_1 contains a fairly even amount of s_0 and s_2 since their frequency amplitude is about the same.

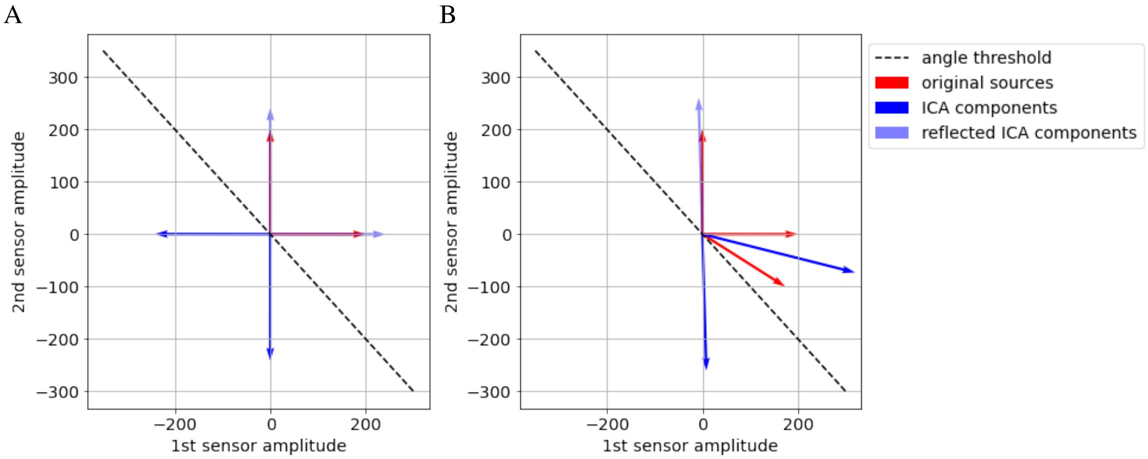


Figure 4: Angles represent the spatial map or relative amplitude over sensors. A) Mixing vectors representing the spatial maps of sources s_0 and s_1 on each sensor (red, magnified 200 times), along with vectors representing the spatial maps of ICA components as reported in M_{est} (shown in blue). We use these vectors to calculate the angles for the original sources and the ICA components. In our analysis, if the component amplitude vector has angle $< -45°$ or $> 135°$ then we reflect the vector across the origin. B) Mixing vectors representing the spatial maps of the original 3 sources on each sensor (red, magnified 200 times), along with the ICA components (blue). From these vectors, it appears that ICA separated s_1 into a single component, and s_0 and s_2 into the other component, similar to figure 3.

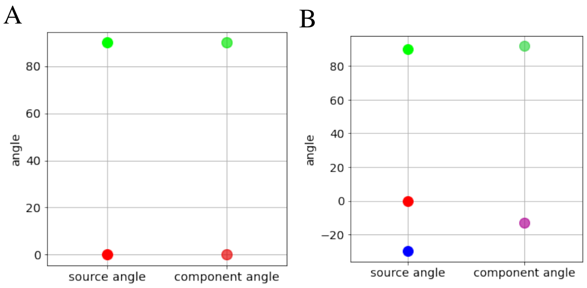


Figure 5: Comparison of angle and frequency distribution for $\alpha_1 = 0$, $\alpha_2 = 90$, $\theta = -30$ degrees. Angles for the spatial maps of the original sources are on the left, while angles for the ICA component spatial maps are on the right. Color indicates the relative amplitude of different frequencies: red for 2 Hz, green for 3 Hz, and blue for 5 Hz. Both panels show trials for 10 ICA decompositions, where the marker for each run has opacity set to 1/10. A) Two sources and two sensors. B) Three sources and two sensors.

angles. Since -45° is considered equivalent to 135°, then angles may be closer across this angle threshold. There appears to be almost no variation between ICA runs. The only instance with some variation is where θ is half-way between α_0 and α_1.

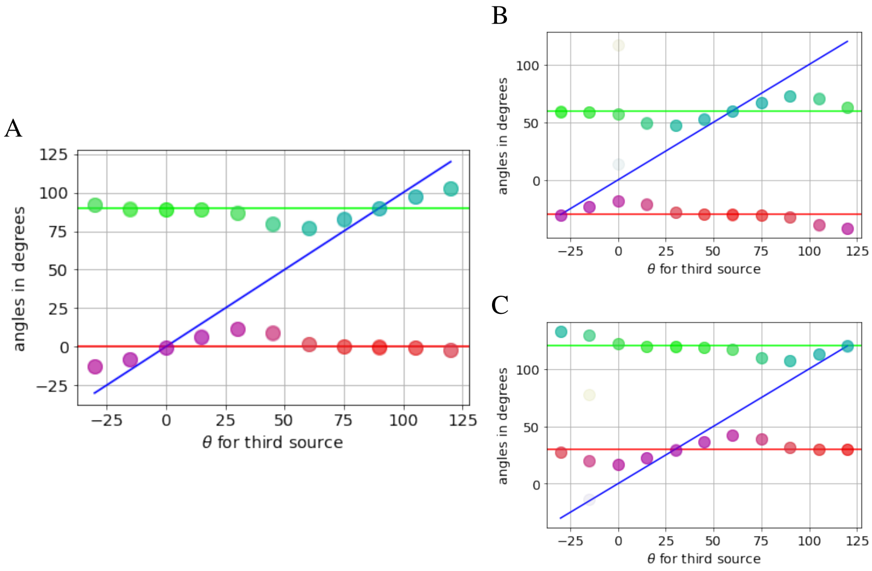


Figure 6: Comparison of angle and frequency distribution: base angles are 90° **apart.** Solid lines represent the spatial map angle of the original sources, along with their color-coded frequency: red for 2 Hz, green for 3 Hz, and blue for 5 Hz. Dots represent the angle and frequency content for the ICA components, similar to figure 5. A) Results for ICA runs with θ ranging from -30° to 135° when $\alpha_0 = 0$ and $\alpha_1 = 90^\circ$. B) Results for $\alpha_0 = -30^\circ$ and $\alpha_1 = 60^\circ$. C) Results for $\alpha_0 = 30^\circ$ and $\alpha_1 = 120^\circ$.

3.2 Cases where spatial maps of the first two sources are less than 90° apart.

Figure 7 shows results for cases where α_0 and α_1 are 60° apart. The ICA components still follow the original sources in most cases. However, now that α_0 and α_1 are 60° apart, it is possible for all three angles to be equidistant over the 180° range. When they are equidistant, we see a lot more variability between ICA runs using the FastICA algorithm. This makes sense since there is no true optimal separation of sources if they are equidistant. The results for the extended infomax algorithm didn't show as much variety, but instead showed some discontinuity in results when the sources are equidistant. (See figure 1.1 in the appendix.) There may be more variability in the results for values of θ in between the ones chosen.

Figures 8 and 9 show results where α_0 and α_1 are 45° apart and 30° apart, respectively. When α_0 and α_1 are 45° apart, we still see some variability between ICA runs when the angles are more equidistant, but the majority of the ICA runs return components that bin together the original sources with the closest angle. When α_0 and α_1 are 30° apart, we see almost no variability, similar to the case where α_0 and α_1 are 90° apart.

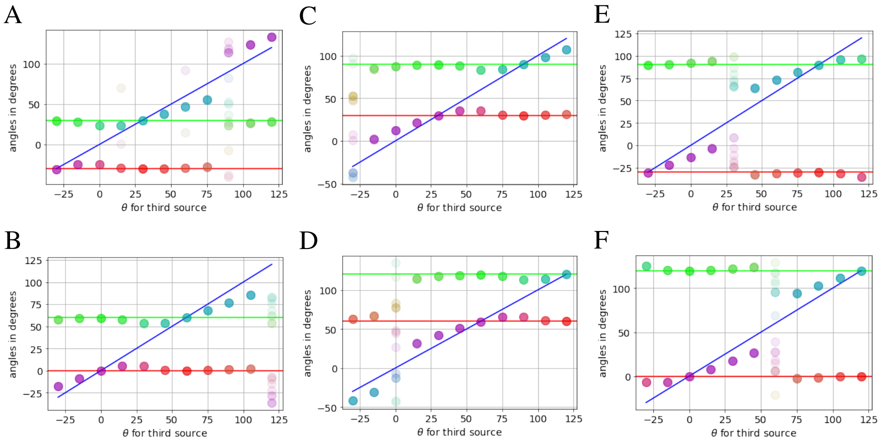


Figure 7: Comparison of angle and frequency distribution: base angles are 60° **apart.** Results where all combinations of α_0 and α_1 are 60° apart, starting with $\alpha_0 = -30^\circ$ and $\alpha_1 = 30^\circ$. Since angles that are 180 degrees apart are equivalent, then we also consider cases where the angle is 60° apart across the boundary: $-45^\circ = 135^\circ$. For instance $\alpha_0 = -30^\circ = 150^\circ$ and $\alpha_1 = 90^\circ$ are also 60° apart. Since all possible angles span 180°, we see the greatest instability when the 3 sources are all equidistant at 60°.

In line with previous examples where ICA bins sources with the closest angle together, the original sources with spatial map angles α_0 and α_1 are binned together until θ is within 30° of either angle.

4 Discussion

We applied the ICA algorithm to 2-sensor data mixtures composed of 3 noisy sinusoidal sources. The spatial map of each source across the sensors is characterized by the angle of its mixing vector. We found that ICA systematically binned sources with the closest angle together. ICA would evenly split sources if one source's angle was equidistant to the other two. The ICA components were stable across multiple ICA runs in most cases. The largest variability was seen when all 3 source angles were equidistant. Our results give evidence that ICA predictably separates sources and that ICA components can be interpreted as estimated *groups* of original sources.

While we examined the stability of ICA components systematically based on their spatial maps, several other studies demonstrated the stability of ICA components with biologically plausible data [2, 25, 13, 15, 27, 28, 32]. Also, while we examined how ICA components bin sources together by their spatial maps, there are several methods that compare ICA spatial maps to choose the ideal number of components when there are an abundant number of sensors [11, 23].

Understanding how ICA bins sources together may shed light on how best to use ICA to decompose data. Researchers can run ICA multiple times to see whether ICA

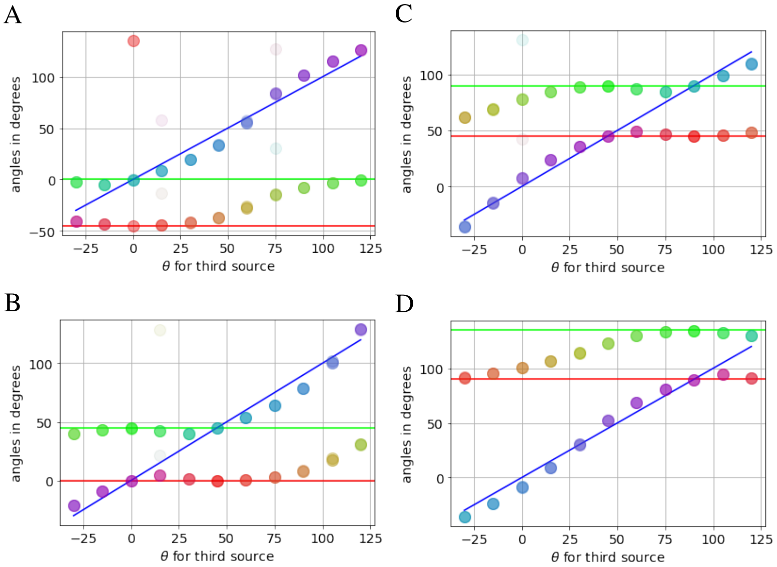


Figure 8: Comparison of angle and frequency distribution: base angles are 45° **apart.** We see more stability when the base angles are closer together. However, there are some cases with instability when all 3 sources are close to equidistant.

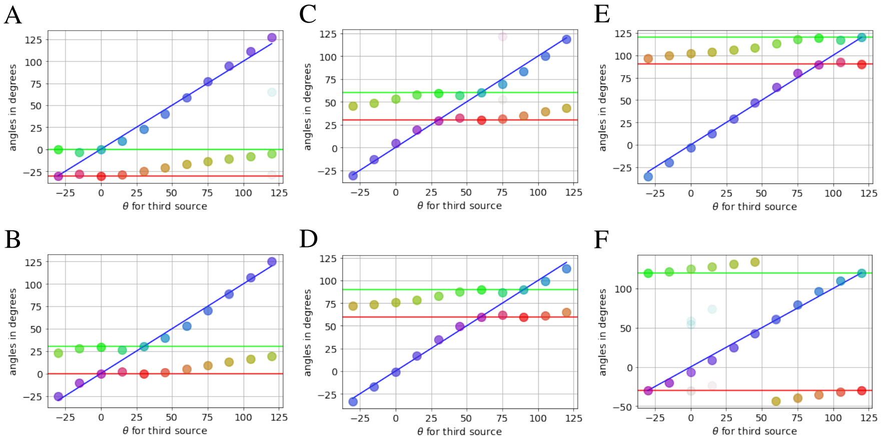


Figure 9: Comparison of angle and frequency distribution: base angles are 30° **apart.** Source separation appears very stable when two sources are significantly closer together.

components are consistent across ICA runs [32]. If the ICA components have a high degree of variability, then the underlying original sources may have equidistant spatial maps. While we defined distance using the dot product, other studies have used a variety of different measures to compare spatial maps [11, 27]. If there are equidistant sources, decreasing the total number of components may bin the equidistant sources together so that the ICA components are more stable. ICA components may also be more stable by increasing the total number of components enough to resolve equidistant sources. If there are fewer sensors than desired ICA components, this may be done with a reliable method to extract more sources than sensors.

The idea that ICA bins sources together can help address the consistency of ICA components across different recordings. For instance, suppose we have a set of 10 LFP recordings, each taken from a different animal under the same context, so we have the same number of neurophysiological sources. Suppose further that each recording used 32 sensors, but that some of the sensors are faulty in 5 of the recordings. The faulty sensors still pick up the LFP, but their baseline voltage drifts - which can happen often since neural electrodes are highly sensitive. These faulty sensors add an extra source, which ICA may pick out as a single component since the spatial map is concentrated on one sensor - making it very distinct from other spatial maps. This leaves fewer components for the neurophysiological sources. So some of the neurophysiological sources that ICA separated in recordings without faulty sensors may be binned together in recordings with faulty sensors. We may be able to tell which sources are binned together by looking at their spatial maps. Most current studies that use ICA on LFP focus on large-amplitude, easily replicable components. Unlocking which components are binned together may allow a better interpretation of smaller-amplitude components.

Seeing how ICA bins sources together can help ICA components be seen as relevant functional groups of sources. These functional groups may not necessarily be neuronal populations, but may represent afferent synapses, active cell parts such as dendrites, glia, or cell assemblies [7, 34, 35, 40, 41]. We may even be able to quantify how well sources are separated from each other based on the relative distance between the spatial maps of the ICA components.

Our results represent an initial study in how ICA treats over-complete data, where we focus on 3 sinusoidal sources with the same amplitude separated into 2 components. We used the same 3 sources for all of our simulations. Future work in this area could consider many different types of sources that vary in frequency spectrum and amplitude, along with different combinations in the number of sources and components. In particular, natural sources can include a whole range of frequencies at varying amplitudes. If two sources share some of the same frequencies, then ICA may have a harder time distinguishing between the two sources. Also, ICA normalizes the data given to it so that amplitudes in all directions are the same. This means lower-amplitude sources may not be normalized if they are not mixed in a distinct direction. Therefore, ICA may bin sources differently in over-complete data if some of the sources have much higher amplitude than others. We used the FastICA algorithm and the extended infomax algorithm in all our simulations. While the results using both of these algorithms were nearly identical, we did note some differences in ICA component stability when spatial maps are equidistant. Other ICA algorithms may yield different results in the amount of variability between ICA runs as well as distribution of angles between components.

ICA is designed to separate sources so that components are as independent as possible. The idea that ICA bins similar sources together not only follows this goal, but can allow for more insightful interpretation of separated sources. Indeed, binning sources together may be the most desirable outcome. For example, in the LFP we may be more interested in separating functional groups of neurons than many individual neurons with similar activity. While current studies that use ICA in interpreting the LFP tend to focus on just a few replicable components, ICA separation allowed them to have insights that may not have been possible by other means. Therefore, using ICA may allow immediate insights into micro-circuits in the brain.

1 Results for the extended infomax algorithm when base angles are 60° apart.

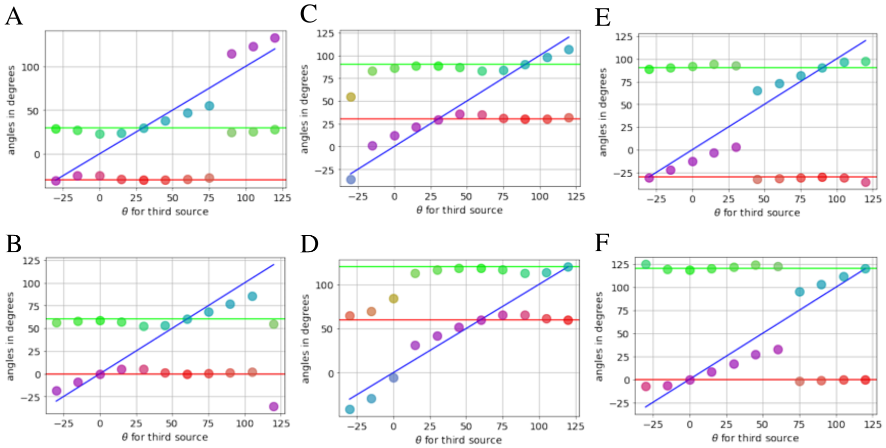


Figure 1.1: Comparison of angle and frequency distribution using the extended infomax algorithm: base angles are 60° apart. Instead of the instability we see with the FastICA algorithm, results tend to jump discontinuously. However, we may see some instability if we try finer-grained values for θ.

Bibliography

[1] Aimon, S., Katsuki, T., Jia, T., Grosenick, L., Broxton, M., Deisseroth, K., Sejnowski, T. J., and Greenspan, R. J. (2019). Fast near-whole–brain imaging in adult Drosophila during responses to stimuli and behavior. *PLoS Biology,* 17(2):e2006732.

[2] Artoni, F., Menicucci, D., Delorme, A., Makeig, S., and Micera, S. (2014). RELICA: A method for estimating the reliability of independent components. *NeuroImage,* 103:391–400.

[3] Babiloni, C., Barry, R. J., Başar, E., Blinowska, K. J., Cichocki, A., Drinkenburg, W. H., Klimesch, W., Knight, R. T., Silva, F. L. d., Nunez, P., Oostenveld, R., Jeong, J., Pascual-Marqui, R., Valdes-Sosa, P., and Hallett, M. (2019). International Federation of Clinical Neurophysiology (IFCN)– EEG research workgroup: Recommendations on frequency and topographic analysis of resting state EEG rhythms. Part1: Applications in clinical research studies. *Clinical Neurophysiology,* 131(1):285–307.

[4] Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face Recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks*, 13(6):1450.

[5] Bénar, C.-G., Velmurugan, J., López-Madrona, V., Pizzo, F., and Badier, J. (2021). Detection and localization of deep sources in magnetoencephalography: a review. *Current Opinion in Biomedical Engineering*, 18:100285.

[6] Brown, G. D., Yamada, S., and Sejnowski, T. J. (2001). Independent component analysis at the neural cocktail party. *Trends in Neurosciences,* 24(1):54–63.

[7] Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extra cellular fields and currents-EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience,* 13(6):407–420.

[8] Calhoun, V. D. and Lacy, N.d. (2017). Ten Key Observations on the Analysis of Resting-state Functional MR Imaging Data Using Independent Component Analysis. *Neuroimaging Clinics of North America,* 27(4):561–579.

[9] Calinon, S. and Billard, A. (2005). Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML'05, page105–112, NewYork, NY, USA. Association for Computing Machinery.

[10] Cvejic, N., Bull, D., and Canagarajah, N. (2007). Improving Fusion of Surveillance Images in Sensor Networks Using Independent Component Analysis. *IEEE Transactions on Consumer Electronics,* 53(3):1029–1035.

[11] Du, Y., He, X., and Calhoun, V. D. (2021). Smart (splitting-merging assisted reliable) inde pendent component analysis for brain functional networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC),* pages 3263–3266.

[12] Einevoll, G. T., Kayser, C., Logothetis, N. K., and Panzeri, S. (2013). Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nature Reviews Neurscience,* 14(11):770–785.

[13] Esposito, F. and Goebel, R. (2011). Extracting functional networks with spatial independent component analysis. *Current Opinion in Neurology,* 24(4):378–385.

[14] Frost, W. N., Brandon, C. J., Bruno, A. M., Humphries, M. D., Moore-Kochlacs, C., Sejnowski, T. J., Wang, J., and Hill, E. S. (2015). Monitoring Spiking Activity of Many Individual Neurons in Invertebrate Ganglia. *Advances in Experimental Medicine and Biology,* 859:127–145.

[15] Głąbska, H., Potworowski, J., Łęski, S., and Wójcik, D. K. (2014). Independent Components of Neural Activity Carry Information on Individual Populations. *PLoS ONE,* 9(8):e105071.

[16] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M.S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience,* 7(267):1–13.

[17] Gratiy, S. L., Devor, A., Einevoll, G. T., and Dale, A. M. (2011). On the estimation of population-specific synaptic currents from laminar multielectrode recordings. *Frontiers in neuroinformatics,* 5:32.

[18] Harris, C. R., Millman, K. J., Walt, S. J. v. d., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. v., Brett, M., Haldane, A., Río, J. F. d., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature,* 585(7825):357–362.

[19] Harris Bozer, A. L., Uhelski, M. L., and Li, A.-L. (2017). Extrapolating meaning from local field potential recordings. *Journal of Integrative Neuroscience,* 16(1):107–126.

[20] Herreras, O., Makarova, J., and Makarov, V. A. (2015). New uses of LFPs: Pathway-specific threads obtained through spatial discrimination. *Neuroscience,* 310:486–503.

[21] Hill, E. S., Moore-Kochlacs, C., Vasireddi, S. K., Sejnowski, T. J., and Frost, W. N. (2010). Validation of Independent Component Analysis for Rapid Spike Sorting of Optical Recording Data. *Journal of Neurophysiology,* 104(6):3721–3731.

[22] Hong, G. and Lieber, C. M. (2019). Novel electrode technologies for neural recordings. *Nature Reviews Neuroscience,* 20(6):330–345.

[23] Hu, G., Waters, A. B., Aslan, S., Frederick, B., Cong, F., and Nickerson, L. D. (2020). Snowball ICA: A Model Order Free Independent Component Analysis Strategy for Functional Magnetic Resonance Imaging Data. *Frontiers in Neuroscience,* 14:569657.

[24] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks,* 13(4-5):411–430.

[25] Jung, T.-P., Makeig, S., McKeown, M. J., Bell, A. J., Lee, T.-W., and Sejnowski, T. J. (2001). Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE,* 89(7):1107–1122.

[26] Kajikawa, Y. and Schroeder, C. E. (2011). How local is the local field potential? *Neuron,* 72(5):847–858.

[27] Makarov, V. A., Makarova, J., and Herreras, O. (2010). Disentanglement of local field potential sources by independent component analysis. *Journal of Computational Neuroscience,* 29(3):445 301–457.

[28] Makarova, J., Ibarz, J. M., Makarov, V. A., Benito, N., and Herreras, O. (2011). Parallel Readout of Pathway-Specific Inputs to Laminated Brain Structures. *Frontiers in Systems Neuroscience,* 5:77.

[29] Martínez-Cañada, P., Noei, S., and Panzeri, S. (2021). Methods for inferring neural circuit interactions and neuromodulation from local field potential and electroencephalogram measures. *Brain Informatics,* 8(1):27.

[30] Martínez-Cancino, R., Delorme, A., Truong, D., Artoni, F., Kreutz-Delgado, K., Sivagnanam, S., Yoshimoto, K., Majumdar, A., and Makeig, S. (2021). The open EEGLAB portal Interface: High-Performance computing with EEGLAB. *NeuroImage,* 224:116778.

[31] McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O., and Sastry, A. V. (2021). Optimal dimensionality selection for independent component analysis of transcriptomic data. *BMC Bioinformatics,* 22(1):584.

[32] Meinecke, F., Ziehe, A., Kawanabe, M., and Müller, K.-R. (2002). A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. *IEEE transactions on bio-medical engineering,* 49(12Pt2):1514–1525.

[33] Monakhova, Y. B. and Rutledge, D. N. (2019). Independent components analysis (ICA) at the “cocktail-party” in analytical chemistry. *Talanta,* 208:120451.

[34] Munro, E. and Kopell, N. (2012). Subthreshold somatic voltage in neocortical pyramidal cells can control whether spikes propagate from the axonal plexus to axon terminals: a model study. *Journal of Neurophysiology,* 107(10):2833–2852.

[35] Munro Krull, E., Sakata, S., and Toyoizumi, T. (2019). Theta Oscillations Alternate With High Amplitude Neocortical Population Within Synchronized States. *Frontiers in neuroscience,* 13:316.

[36] Naik, G. R. and Kumar, D. K. (2011). An Overview of Independent Component Analysis and Its Applications. *Informatica,* 35(1):63–81.

[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12:2825–2830.

[38] Pesaran, B., Vinck, M., Einevoll, G. T., Sirota, A., Fries, P. ,Siegel, M., Truccolo, W., Schroeder, C. E., and Srinivasan, R. (2018). Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nature Neuroscience,* 21(7):903–919.

[39] Reidl, J., Starke, J., Omer, D. B., Grinvald, A., and Spors, H. (2007). Independent component analysis of high-resolution imaging data identifies distinct functional domains. *NeuroImage,* 34(1):94–108.

[40] Sakurai, Y., Osako, Y., Tanisumi, Y., Ishihara, E., Hirokawa, J., and Manabe, H. (2018). Multiple Approaches to the Investigation of Cell Assembly in Memory Research—Present and Future. *Frontiers in Systems Neuroscience,* 12:21.

[41] Sinha, M. and Narayanan, R. (2022). Active Dendrites and Local Field Potentials: Biophysical Mechanisms and Computational Explorations. *Neuroscience,* 489:111–142.

[42] Smitha, K., Raja, K. A., Arun, K., Rajesh, P., Thomas, B., Kapilamoorthy, T., and Kesavadas, C. (2017). Resting state fMRI: A review on methods in resting state connectivity analysis and resting state networks. *The Neuroradiology Journal,* 30(4):305–317.

[43] Sompairac, N., Nazarov, P. V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., Kairov, U., and Zinovyev, A. (2019). Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *International Journal of Molecular Sciences,* 20(18):4414.

[44] Talebi, S. (2021). Independent Component Analysis (ICA): Finding hidden factors in data. *Towards Data Science,* `https://towardsdatascience.com/independent-component-analysis-ica-a3eba0ccec35`. Accessed 27 October 2023.

[45] Teeters, J. L., Harris, K. D., Millman, K. J., Olshausen, B. A., and Sommer, F. T. (2008). Data Sharing for Computational Neuroscience. *Neuroinformatics,* 6(1):47–55.

[46] Torres, D., Makarova, J., Ortuño, T., Benito, N., Makarov, V. A., and Herreras, O. (2019). Local and Volume-Conducted Contributions to Cortical Field Potentials. *Cerebral Cortex*, 29(12):5234–5254.

[47] Unakafova, V. A. and Gail, A. (2019). Comparing Open-Source Tool boxes for Processing and Analysis of Spike and Local Field Potentials Data. *Frontiers in Neuroinformatics,* 13:57.

[48] Yakushev, I., Drzezga, A., and Habeck, C. (2017). Metabolic connectivity: methods and application. *Current Opinion in Neurology,* 30(6):677–685.

[49] Zhang, N. and Nie, J. (2015). Independent Component Analysis Based Blind Source Separation Algorithm and its Application in the Gravity and Magnetic Signal Processing. 2015 *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 269–273.

[50] Zhang, Q., Hu, G., Tian, L., Ristaniemi, T., Wang, H., Chen, H., Wu, J., and Cong, F. (2018). Examining stability of independent component analysis based on coefficient and component matrices for voxel-based morphometry of structural magnetic resonance imaging. *Cognitive Neurodynamics,* 12(5):461–470.

[51] Zhao, W., Li, H., Hu, G., Hao, Y., Zhang, Q., Wu, J., Frederick, B. B., and Cong, F. 367 (2021). Consistency of independent component analysis for FMRI. *Journal of Neuroscience Methods,* 351:109013.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 76 – 92

Irrationality of the Riemann-Zeta function at the positive integers

*Yoochan Noh**

Yoochan Noh is a high school student at Korea International School, with a serious focus on college-level mathematics subjects. After graduation, he plans to pursue degrees in Applied Mathematics and Computer Science, driven by his passion for problem-solving and the transformative potential of technology. Inspired by previous small individual research projects, Yoochan sought to undertake a more substantial and profound research endeavor in mathematics. This led him to explore the complex and deep theme of the Riemann-Zeta function. Yoochan's aspiration extends beyond this research, as he hopes to engage in various types of research in the future, further expanding his understanding and contributing to the advancement of mathematical knowledge.

Abstract

The Riemann Zeta function, usually denoted by the Greek letter ζ, was defined in 1737 by a Swiss mathematician Leonhard Euler. This function is an infinite converging sum of powers of natural numbers, and it has explicit expressions in terms of π at positive even integers. In this paper we will discuss various irrationality proofs, focusing on irrationality of certain values of the Zeta function.

1 Introduction

We start with the definition of the Riemann-Zeta function (that we will just call Zeta function from now on).

***Corresponding author:** ycnoh0211@gmail.com

Definition 1.1. *On the complex half-plane $\{(z) > 1 \mid z \in \mathbb{C}\}$ the Riemann-Zeta function is defined by the following expression:*

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z} = \frac{1}{1^z} + \frac{1}{2^z} + \frac{1}{3^z} + \dots + \frac{1}{n^z} + \dots$$

It is easy to show that the sum converges in this region.

Definition 1.1 will suffice for our purposes, but $\zeta(s)$ can be extended to the whole complex plane by [2]. Leonhard Euler did some basic computations with $\zeta(s)$. In particular, he famously solved the Basel Problem which is the question of determining the precise value of $\zeta(2)$. We cover his proof in the modern language in Section . Euler also generalized the computation to all positive even integers. One of the main results of this paper is a different proof of this formula that we give in Section 6.

The rest of the paper is organized as follows:

In Section 2 we discuss preliminaries needed to solve the Basel problem. In Section 3 we prove that certain radical expressions, and π, are irrational. In Section 4 we solve the Basel problem and compute $\zeta(4)$. In Section 5 we define Bernoulli numbers, an important preliminary for computing the Zeta function at the even integers. In Section 6 we discuss how the Zeta function at the even integers can be expressed in terms of Bernoulli numbers. In Section 7 we introduce the notion of being transcendental and explain that transcendentality of π [10] implies that $\zeta(2k)$, $k \geq 1$ is irrational. In Section 8 we prove that $\zeta(3)$ is irrational following [3, Theorem 2]. In Section 9 we show some advanced results, generalizations, and conjectures of the irrationality of the Zeta function.

2 Preliminaries

In order to prove irrationality of the Zeta function at certain values, it is necessary to understand certain preliminaries such as the infinite product formula for the sine function.

2.1 Logarithms of infinite products

Lemma 2.1. *For an infinite convergent product*

$$S = \prod_{n=1}^{\infty} a_n$$

it is always the case that

$$\log(S) = \sum_{n=1}^{\infty} \log(a_n)$$

Proof. If the infinite product converges to a positive number, continuity of the logarithm function permits the interchange of the limit and the logarithm. So,

$$\begin{aligned}\log\prod_{n=1}^{\infty}a_n &= \log\left(\lim_{k\to\infty}\prod_{n=1}^{k}a_n\right)\\ &= \lim_{k\to\infty}\log\left(\prod_{n=1}^{k}a_n\right)\\ &= \lim_{k\to\infty}\sum_{n=1}^{k}\log(a_n)\\ &= \sum_{n=1}^{\infty}\log(a_n)\end{aligned}$$

□

2.2 Derivatives of infinite sums

Lemma 2.2. *Suppose we have a sequence of functions f_n differentiable on $[a,b]$. If we have the series $\sum_{n=1}^{\infty} f_n(x)$ converging to $f(x)$ on $[a,b]$:*

$$f(x)=\sum_{n=1}^{\infty}f_n(x)$$

and the series of derivatives $\sum_{n=1}^{\infty} f_n'(x)$ converges uniformly on $[a,b]$, then we have

$$f'(x)=\sum_{n=1}^{\infty}f_n'(x) \qquad (a\le x\le b)$$

Proof. This is a standard result, see e.g. [11, Theorem 7.17]. □

2.3 Infinite product of the sine function

Theorem 2.3. *We have the equalities*

$$\sin(x)=x\left(\prod_{k=1}^{\infty}\left(1+\frac{x}{k\pi}\right)\left(1-\frac{x}{k\pi}\right)\right)=x\left(\prod_{k=1}^{\infty}\left(1-\frac{x^2}{k^2\pi^2}\right)\right)$$

and

$$\frac{\sin(x)}{x}=\prod_{k=1}^{\infty}\left(1-\frac{x^2}{k^2\pi^2}\right)$$

that may be understood as expressing $\sin(x)$ *as an infinite product over its roots at $n\pi$ for $n\in\mathbb{Z}$.*

We refer to [8] for the proof.

3 Irrationality of radicals, and of π

In this section we give some elementary irrationality proofs. In particular, we prove that π is irrational using integral techniques. Somewhat similar, but much more advanced, methods will be used in Section 8 to prove that $\zeta(3)$ is irrational, a famous result due to Apéry [1].

Proposition 3.1. *For a prime number k, $\sqrt[n]{k}$ is an irrational number, for any $n \in \mathbb{Z}_{\geq 2}$.*

Proof. We first observe that $\sqrt[n]{k}$ is a root of the polynomial $x^n - k$. According to the rational root theorem, which uses ratios of the factors of the leading coefficient and the constant of a polynomial to determine its integer roots, a rational root of a polynomial with integer coefficients that is written in lowest terms $\frac{p}{q}$ must have denominator q that divides the leading coefficient, and numerator p that divides the constant coefficient. Here, $1 \equiv 0 \pmod{q}$, so q has to be 1, while $k \equiv 0 \pmod{p}$, so p is either k or 1. Therefore, any rational root of $x^n - k$ must be an integer.

The only way for $\sqrt[n]{k}$ to be an integer is if k is an n-th power of an integer, where $n \geq 2$. Since k is a prime number, it can only be expressed as p^1 when factorized. However, $1 \not\equiv 0\ (\bmod n)$, so $\sqrt[n]{k}$ cannot be an integer, and hence not a rational number. □

Corollary 3.2. For prime numbers k and l, $\sqrt[n]{k + \sqrt[m]{l}}$ is an irrational number for any $n, m \in \mathbb{Z}_{>0}$

Proof. We can first assume that $\sqrt[n]{k + \sqrt[m]{l}}$ is rational, thereby stating that

$$\sqrt[n]{k + \sqrt[m]{l}} = \frac{p}{q}, \text{ where } p, q \in \mathbb{Z}$$

This gives

$$k + \sqrt[m]{l} = \frac{p^n}{q^n}$$

$$\sqrt[m]{l} = \frac{p^n}{q^n} - k$$

Since both $\frac{p^n}{q^n}$ and k are rational, it can be concluded that $\sqrt[m]{l}$ is also rational. However, have already shown that $\sqrt[m]{l}$ is irrational in Proposition 3.1, which contradicts the initial hypothesis. Therefore, $\sqrt[n]{k + \sqrt[m]{l}}$ has to be irrational. □

Theorem 3.3 (Irrationality of π). *The number π is irrational.*

Proof. For any integrable function $f(x)$ by integration by parts we have:

$$\int f(x) \sin x \, dx = -f(x) \cos x + f'(x) \sin x - \int f''(x) \sin x \, dx$$

By using the values of $\sin(0) = 0$, $\cos(0) = 1$, $\sin(\pi) = 0$, and $\cos(\pi) = -1$,

$$\int_0^{\pi} f(x) \sin x \, dx = f(\pi) + f(0) - \int_0^{\pi} f''(x) \sin x \, dx$$

If $f(x)$ is a polynomial of degree $2n$, $n \in \mathbb{Z}_{>0}$, then repeating the calculation $n+1$ times would give

$$\int_0^{\pi} f(x) \sin x \, dx = F(\pi) + F(0) + \int_0^{\pi} f^{(2n+2)}(x) \sin x \, dx = F(\pi) + F(0) \quad (1)$$

where $F(x) = f(x) - f''(x) + f^{(4)}(x) - \cdots + (-1)^n f^{(2n)}(x)$ and the last equality is from $f^{(2n+2)}(x) = 0$ (here $f^{(k)}(x)$ stands for the k-th derivative of $f(x)$).

Assume that π is rational, that is $\pi = \frac{p}{q}$ with $p, q \in \mathbb{Z}$ and $q \neq 0$. We will choose a particular polynomial $f(x)$ such that $F(0) + F(\pi)$ is an integer. Then, we will also show that $\int_0^\pi f(x) \sin x \, dx$ lies between 0 and 1, exclusively. Since no such integer can exist, this will obtain contradiction and π has to be irrational.
For $n \in \mathbb{Z}_{>0}$, let

$$f(x) = \frac{x^n (p - qx)^n}{n!}. \tag{2}$$

For $F(\pi) + F(0)$ to be an integer, we need to show that both $f^{(2n)}(\pi)$ and $f^{(2n)}(0)$ are integers.
For the chosen function $f(x)$,

$$f(\pi - x) = f\left(\frac{p}{q} - x\right) = \frac{\left(\frac{p}{q} - x\right)^n \left(p - q\left(\frac{p}{q} - x\right)\right)^n}{n!} = \frac{\left(\frac{p}{q} - x\right)^n (qx)^n}{n!} = \frac{x^n (p - qx)^n}{n!} = f(x)$$

Also, using the chain rule we see that for any $k \in \mathbb{Z}_{>0}$ we have

$$f^{(k)}(x) = (-1)^n f^{(k)}(\pi - x)$$

and

$$f^{(2n)}(0) = (-1)^{2n} f^{(2n)}(\pi) = f^{(2n)}(\pi)$$

So, if we show that $f^{(2n)}(0)$ is an integer, $f^{(2n)}(\pi)$ would also be an integer. We can express the function $f(x)$ in 2 ways:

$$f(x) = \frac{x^n (p - qx)^n}{n!} = \sum_{j=0}^{2n} \frac{c_j}{n!} x^j$$

for some $c_j \in \mathbb{Z}$. Also (according to the Taylor series),

$$f(x) = \frac{f(0)}{0!} + \frac{f'(0)}{1!} x + \frac{f''(0)}{2!} x^2 + \cdots + \frac{f^{(2n)}(0)}{(2n)!} x^{2n}.$$

The coefficients at x^j for both equations should be equal.

$$\frac{c_{2n}}{n!} = \frac{f^{(2n)}(0)}{(2n)!}$$

Thus,

$$\frac{(2n)!}{n!} c_{2n} = f^{(2n)}(0)$$

Since $\frac{(2n)!}{n!} c_{2n}$ is an integer, then $f^{(2n)}(0)$ would also be an integer, which proves that $\int_0^\pi f(x) \sin x \, dx$ is an integer.
The next step is to show that (1) equates to a value strictly between 0 and 1.

$$f(x) = \frac{x^n (p - qx)^n}{n!} = \frac{x^n}{n!} (p - qx)^n$$

For $0 < x < \pi$, $\frac{x^n}{n!} > 0$, and $(p - qx)^n > 0$, so $f(x) > 0$ for $0 < x < \pi$. Also, since $\sin x > 0$ for $0 < x < \pi$ too, $f(x) \sin x > 0$ for the same domain, and therefore $\int_0^\pi f(x) \sin x \, dx > 0$. If the domain for x is $0 < x < \pi$, it can also be written as $0 < \pi - x < \pi$. By multiplying the 2 together, we get $0 < x(\pi - x) < \pi^2$.

Then,

$$0 < f(x) = \frac{x^n(p-qx)}{n!} = q^n \frac{x^n(\pi - x)^n}{n!} < q^n \frac{\pi^{2n}}{n!}$$

Since we are free to choose the value of n, we just need to show that

$$\lim_{n\to\infty} q^n \frac{\pi^{2n}}{n!} < \frac{1}{2} \tag{3}$$

Indeed, then we have $f(x) < \frac{1}{2}$ so

$$\int_0^\pi f(x)\sin x\, dx < \frac{1}{2}\int_0^\pi \sin x\, dx = 1.$$

To show (3), we just need to look at the Taylor series of the value of $e^{q\pi^2}$

$$e^{q\pi^2} = 1 + \frac{q\pi^2}{1!} + \frac{q^2\pi^4}{2!} + \frac{q^3\pi^6}{3!} + \dots$$

So, this infinite series converges to $e^{q\pi^2}$, a real number. However, if a particular infinite series $\sum_{n=0}^{\infty} a_n$ converges and $a_n > 0$, then

$$\lim_{n\to\infty} a_n = 0 < \frac{1}{2}$$

In our case,

$$\lim_{n\to\infty} q^n \frac{\pi^{2n}}{n!} = 0 < \frac{1}{2}$$

Therefore, using proof by contradiction, this shows that π is irrational. □

4 Elementary computation of $\zeta(2)$ and $\zeta(4)$.

A famous question, known as the Basel problem, is computing $\zeta(2)$. This result demonstrates that the infinite sum of the squares of the inverses of positive natural numbers is equal to the square of the number π divided by 6. We can write this as:

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots + \frac{1}{n^2} + \dots = \frac{\pi^2}{6}$$

The proof that we give below goes back to Euler [12, Theorem 1]. We also generalize the computation to calculate $\zeta(4)$.

4.1 Solving the Basel problem

Through the infinite product of the sine function formula seen in 2.3, we have concluded that:

$$\frac{\sin x}{x} = \prod_{k=1}^{\infty}\left(1 - \frac{x^2}{k^2\pi^2}\right) = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{2^2\pi^2}\right)\left(1 - \frac{x^2}{3^2\pi^2}\right)\left(1 - \frac{x^2}{4^2\pi^2}\right)\cdots \tag{4}$$

In addition to this, we can use the Taylor series to achieve the following equation:

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

$$\frac{\sin x}{x} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \frac{x^8}{9!} - \dots \tag{5}$$

Theorem 4.1 (Basel problem)**.** *We have*

$$\zeta(2) = \frac{\pi^2}{6} \tag{6}$$

Proof. The idea is to compare the coefficients at x^2 obtained using formulas (5) and (4). The coefficient at x^2 using (5) is $-\frac{1}{6}$. Let us compute the coefficient using (4). We have

$$\frac{\sin x}{x} = \lim_{n\to\infty} \left(\frac{\sin x}{x}\right)_n \tag{7}$$

where

$$\left(\frac{\sin x}{x}\right)_n := \prod_{k=1}^{n} \left(1 - \frac{x^2}{k^2\pi^2}\right) \tag{8}$$

We can investigate the finite order terms $\left(\frac{\sin x}{x}\right)_n$ starting with $n = 3$. Note that we have

$$\left(\frac{\sin x}{x}\right)_3 = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{2^2\pi^2}\right)\left(1 - \frac{x^2}{3^2\pi^2}\right) =$$

$$= \left(1 - \frac{x^2}{\pi^2} - \frac{x^2}{2^2\pi^2} + T_4(2)x^4\right)\left(1 - \frac{x^2}{3^2\pi^2}\right) = 1 - \frac{x^2}{\pi^2} - \frac{x^2}{2^2\pi^2} - \frac{x^2}{3^2\pi^2} + T_4(3)x^4 + T_6(3)x^6 =$$

$$= 1 - \frac{x^2}{\pi^2} - \frac{x^2}{2^2\pi^2} - \frac{x^2}{3^2\pi^2} + T(3)$$

where we denote by $T_m(n)$ the coefficient at x^m in the expansion of $\left(\frac{\sin x}{x}\right)_n$ and by $T(n)$ the sum $\sum_{m=4}^{2n} T_m(n)x^m$.

Similarly one can compute

$$\left(\frac{\sin x}{x}\right)_4 = 1 - \frac{x^2}{\pi^2} - \frac{x^2}{2^2\pi^2} - \frac{x^2}{3^2\pi^2} - \frac{x^2}{4^2\pi^2} + T(4)$$

$$\left(\frac{\sin x}{x}\right)_5 = 1 - \frac{x^2}{\pi^2} - \frac{x^2}{2^2\pi^2} - \frac{x^2}{3^2\pi^2} - \frac{x^2}{4^2\pi^2} - \frac{x^2}{5^2\pi^2} + T(5)$$

Using the same pattern, for an arbitrary $n \geq 3$ we get

$$\left(\frac{\sin x}{x}\right)_n = 1 - \frac{x^2}{\pi^2} - \frac{x^2}{2^2\pi^2} - \frac{x^2}{3^2\pi^2} - \dots - \frac{x^2}{n^2\pi^2} + T(n) \tag{9}$$

$$= T(n) + 1 - \left(\frac{1}{\pi^2} + \frac{1}{2^2\pi^2} + \frac{1}{3^2\pi^2} + \dots + \frac{1}{n^2\pi^2}\right)(x^2). \tag{10}$$

Now due to (7), by taking the limits of both sides, the coefficient at x^2 in (4) is equal to

$$-\left(\frac{1}{\pi^2} + \frac{1}{2^2\pi^2} + \frac{1}{3^2\pi^2} + \dots + \frac{1}{n^2\pi^2} + \cdots\right)$$

Therefore, we can equate the obtained coefficients to get

$$\frac{1}{\pi^2} + \frac{1}{2^2\pi^2} + \frac{1}{3^2\pi^2} + \dots + \frac{1}{n^2\pi^2} + \dots = \frac{1}{3!}.$$

Multiplying by π^2 both sides of the equation, we get:

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots+\frac{1}{n^2}+\cdots=\frac{\pi^2}{6}$$

which proves the result.

□

4.2 Computing $\zeta(4)$

It has previously been stated that:

$$\zeta(2)=\sum_{i=1}^{\infty}\frac{1}{i^2} \tag{11}$$

Also, by following on with the definition of the Riemann Zeta function:

$$\zeta(4)=\sum_{i=1}^{\infty}\frac{1}{i^4} \tag{12}$$

Theorem 4.2. *We have*

$$\zeta(4)=\frac{\pi^2}{90} \tag{13}$$

Proof. Here, we are trying to modify equation (11) by squaring it, and also modify equation (7) so that it can be compared with the original equation and find the coefficient of x^4. Both of the modified equations will then be used to find $\zeta(4)$.

$$\begin{aligned}(\zeta(2))^2&=\left(\sum_{i=1}^{\infty}\frac{1}{i^2}\right)\left(\sum_{j=1}^{\infty}\frac{1}{j^2}\right)\\&=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\frac{1}{i^2}\frac{1}{j^2}=\sum_{i=1}^{\infty}\frac{1}{i^4}+\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\frac{1}{i^2}\frac{1}{j^2}+\sum_{j=1}^{\infty}\sum_{i=j+1}^{\infty}\frac{1}{i^2}\frac{1}{j^2}\\&=\zeta(4)+2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\frac{1}{i^2}\frac{1}{j^2}\end{aligned}$$

Recall equation (8). From here, through the equation (and the proof of the Basel theorem above), we can see that the number of terms in the sum determining the coefficient at x^2 is $\binom{n}{1}$, and that the number of terms in the sum determining the coefficient at x^4 is $\binom{n}{2}$ For example,

$$\begin{aligned}\left(\frac{\sin x}{x}\right)_3&=1-\frac{x^2}{\pi^2}\left(\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}\right)+\frac{x^4}{\pi^4}\left(\frac{1}{1^2}\frac{1}{2^2}+\frac{1}{1^2}\frac{1}{3^2}+\frac{1}{2^2}\frac{1}{3^2}\right)-\cdots\\&=1-\frac{x^2}{\pi^2}\left(\sum_{i=1}^{3}\frac{1}{i^2}\right)+\frac{x^4}{\pi^4}\left(\sum_{1\le i<j\le 3}\frac{1}{i^2}\frac{1}{j^2}\right)-\cdots\end{aligned}$$

So, in general:

$$\left(\frac{\sin x}{x}\right)_n=1-\frac{x^2}{\pi^2}\left(\sum_{i=1}^{n}\frac{1}{i^2}\right)+\frac{x^4}{\pi^4}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{i^2}\frac{1}{j^2}\right)-\cdots$$

In equation (7) we have established that:

$$\lim_{n\to\infty}\left(\frac{\sin x}{x}\right)_n = \frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \ldots$$

Therefore,

$$\lim_{n\to\infty}\frac{1}{\pi^4}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{i^2}\frac{1}{j^2} = \frac{1}{5!}$$

and

$$\lim_{n\to\infty}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{i^2}\frac{1}{j^2} = \frac{\pi^4}{5!}. \tag{14}$$

It was previously shown that:

$$(\zeta(2))^2 = \zeta(4) + 2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\frac{1}{i^2}\frac{1}{j^2}$$

So, applying equation (14) we get

$$\zeta(4) = (\zeta(2))^2 - 2\frac{\pi^4}{5!}$$

But then

$$\zeta(4) = (\zeta(2))^2 - 2\frac{\pi^4}{5!} = \left(\frac{\pi^2}{6}\right)^2 - \frac{\pi^4}{60} = \frac{\pi^4}{36} - \frac{\pi^4}{60} = \frac{\pi^4}{90}$$

□

5 Bernoulli numbers

Definition 5.1. *Bernoulli numbers, often denoted B_n, are set of rational numbers that are often used in analysis. They are defined via the equation:*

$$\frac{t}{e^t - 1} = \sum_{k=0}^{\infty}\frac{B_k}{k!}t^k$$

One can modify the equation

$$\left(\frac{e^t - 1}{t}\right)\left(\sum_{k=0}^{\infty}\frac{B_k}{k}t^k\right) = 1$$

by using the Taylor series expansion for $\frac{e^t-1}{t}$:

$$\left(\frac{1}{1!} + \frac{t}{2!} + \frac{t^2}{3!} + \ldots\right)\left(\frac{B_0}{0!} + \frac{B_1}{1!}t + \ldots\right) = 1$$

Here, we can see that $B_0 = 1$, and the coefficient of t^k becomes:

$$\frac{B_k}{k!}\frac{1}{1!} + \frac{B_{k-1}}{(k-1)!}\frac{1}{2!} + \frac{B_{k-2}}{(k-2)!}\frac{1}{3!} + \cdots + \frac{B_0}{0!}\frac{1}{(k+1)!} = 0$$

Or,

$$B_0 = 1, \binom{k+1}{k}B_k + \binom{k+1}{k-1}B_{k-1} + \cdots + \binom{k+1}{0}B_0 = 0$$

This gives a recursive way to compute B_k.

6 Computing $\zeta(2k)$ for $k \geq 1$

According to equation (4), we have:

$$\frac{\sin x}{x} = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{2^2\pi^2}\right)\left(1 - \frac{x^2}{3^2\pi^2}\right)\left(1 - \frac{x^2}{4^2\pi^2}\right)\dots$$

Or, we can rewrite this equation by putting the natural log on both sides:

$$\log\left(\frac{\sin x}{x}\right) = \log\left(\left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{2^2\pi^2}\right)\left(1 - \frac{x^2}{3^2\pi^2}\right)\left(1 - \frac{x^2}{4^2\pi^2}\right)\dots\right)$$

$$\log(\sin x) - \log x = \sum_{k=1}^{\infty} \log\left(1 - \frac{x^2}{k^2\pi^2}\right)$$

So, if we take the derivative on both sides of the equation in terms of x, we get (as long as $|x| < \pi$):

$$\cot x - \frac{1}{x} = \sum_{k=1}^{\infty}\left(-\frac{2x}{k^2\pi^2}\right)\frac{1}{1 - \frac{x^2}{k^2\pi^2}}$$

Therefore

$$\cot x = \frac{1}{x} + \sum_{k=1}^{\infty}\left(-\frac{2x}{k^2\pi^2}\right)\frac{1}{1 - \frac{x^2}{k^2\pi^2}} =$$

$$= \frac{1}{x} - 2\sum_{k=1}^{\infty}\left(-\frac{x}{k^2\pi^2}\right)\left(1 + \frac{x^2}{k^2\pi^2} + \frac{x^4}{k^4\pi^4} + \dots\right) =$$

$$= \frac{1}{x} - 2\left(\frac{\zeta(2)}{\pi^2}x + \frac{\zeta(4)}{\pi^4}x^3 + \frac{\zeta(6)}{\pi^6}x^5 + \dots\right)$$

and so

$$\frac{\cos x}{\sin x} = \frac{1}{x} - 2\sum_{k=1}^{\infty}\frac{\zeta(2k)}{\pi^{2k}}x^{2k-1} \tag{15}$$

According to Euler's formula, we have the equation:

$$e^{ix} = i\sin x + \cos x$$

If we substitute $-x$ instead of x into the equation, we get:

$$e^{-ix} = i\sin(-x) + \cos(-x) = -i\sin x + \cos x$$

Therefore,

$$\frac{e^{ix} + e^{-ix}}{2} = \cos x$$

And,

$$\frac{e^{ix} - e^{-ix}}{2i} = \sin x$$

So, using these two equations we have

$$\frac{\cos x}{\sin x} = \frac{\frac{e^{ix}+e^{-ix}}{2}}{\frac{e^{ix}-e^{-ix}}{2i}} = i\left(\frac{e^{ix}+e^{-ix}}{e^{ix}-e^{-ix}}\right) = i\left(\frac{e^{2ix}+1}{e^{2ix}-1}\right) = i\left(1 + \frac{2}{e^{2ix}-1}\right) = i + \frac{1}{x}\left(\frac{2ix}{e^{2ix}-1}\right)$$

Substituting into (15) gives

$$i+\frac{1}{x}\left(\frac{2ix}{e^{2ix}-1}\right)-\frac{1}{x}=-2\sum_{k=1}^{\infty}\frac{\zeta(2k)}{\pi^{2k}}x^{2k-1}$$

According to Definition 5.1

$$i+\frac{1}{x}\sum_{k=0}^{\infty}\frac{B_k}{k!}(2ix)^k-\frac{1}{x}=-2\sum_{k=1}^{\infty}\frac{\zeta(2k)}{\pi^{2k}}x^{2k-1}$$

This equation implies that

$$\sum_{k=2}^{\infty}\frac{B_k}{k!}(2ix)^k=-2\sum_{k=1}^{\infty}\frac{\zeta(2k)}{\pi^{2k}}x^{2k}$$

since the other coefficients in the sum on the left hand side cancel out. If we compare the coefficients of x^{2k} on both sides, we get

$$\frac{B_{2k}}{(2k)!}(2i)^{2k}=-2\frac{\zeta(2k)}{\pi^{2k}}$$

$$\zeta(2k)=\frac{B_{2k}}{(2k)!}(2i)^{2k}\frac{-\pi^{2k}}{2}$$

This simplifies to a general expression

$$\zeta(2k)=\frac{(-1)^{k+1}2^{2k-1}\pi^{2k}B_{2k}}{(2k)!} \tag{16}$$

generalizing (6) and (13).

7 Transcendentality of π and irrationality of $\zeta(2k)$, $k\geq 1$

We now explain that (16) implies that the numbers $\zeta(2k)$, $k\geq 1$ are irrational. Indeed, it is well-known that π is not just irrational but transcendental (refer to [10]).

Definition 7.1. *A number $\alpha\in\mathbb{R}$ is transcendental if it is not a root of any polynomial*

$$a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$$

with integer coefficients $a_i\in\mathbb{Z}$.

Proposition 7.2. *If $\alpha\in\mathbb{R}$ is transcendental, it is also irrational.*

Proof. Suppose that α is rational that is we have $\alpha=\frac{p}{q}$ for $p,q\in\mathbb{Z}$ with $q\neq 0$. Then α is a root of $qx-p=0$ contradicting Definition 7.1. □

According to (16), the numbers $\zeta(2k)$, $k\geq 1$ are of the form $a\pi^{2k}$ for $a\in\mathbb{Q}$ (since $B_{2k}\in\mathbb{Q}$ for any $k\geq 1$). Therefore, irrationality of $\zeta(2k)$, $k\geq 1$ follows from transcendentality of π and the following fact that generalizes Proposition 7.2.

Proposition 7.3. *If $\alpha\in\mathbb{R}$ is transcendental then α^k is irrational for any $k\geq 1$.*

Proof. The proof is similar to the proof of Proposition 7.2. Fix a $k\geq 1$ and suppose that α^k is rational that is we have $\alpha^k=\frac{p}{q}$ for $p,q\in\mathbb{Z}$ with $q\neq 0$. Then α is a root of qx^k-p contradicting Definition 7.1. □

8 Irrationality of $\zeta(3)$

In this section, we are going to integrate certain expressions involving logarithms and polynomials to show that $\zeta(3)$ is irrational. This is a famous result of Apéry [1]. We are going to follow the more elementary exposition of Beukers [3] while trying to give more details and some motivation.

First of all, we choose a certain polynomial and prove that it has integer coefficients. The choice of this polynomial is akin to the particular choice of the function $f(x)$ of (2) in the proof of Theorem 3.3 that π is irrational. We use the notation $f^{(n)}(x)$ for the n-th derivative of $f(x)$.

Lemma 8.1. *The polynomial $P_n(x) = \frac{1}{n!}(x^n(1-x)^n)^{(n)}$ has integer coefficients.*

Proof.

$$P_n(x) = \frac{1}{n!}\,(x^n(1-x)^n)^{(n)} = \frac{1}{n!}((x-x^2)^n)^{(n)}$$

According to the binomial theorem, this becomes

$$\frac{1}{n!}\left(\binom{n}{0}x^n(-x^2)^0 + \binom{n}{1}x^{n-1}(-x^2)^1 + \binom{n}{2}x^{n-2}(-x^2)^2 + \cdots + \binom{n}{n}x^0(-x^2)^n\right)^{(n)}$$

Let $a_i := (-1)^i\binom{n}{i}$, which is an integer. Then,

$$\begin{aligned} P_n(x) &= \left(\frac{a_0}{n!}x^n + \frac{a_1}{n!}x^{n+1} + \cdots + \frac{a_n}{n!}x^{2n}\right)^{(n)} \\ &= \frac{n!a_0}{n!0!} + \frac{(n+1)!a_1}{n!1!}x^1 + \cdots + \frac{(2n)!a_n}{n!n!}x^n \\ &= \binom{n}{0}a_0 + \binom{n+1}{1}a_1x^1 + \cdots + \binom{2n}{n}a_nx^n \end{aligned}$$

Since $\binom{p}{q} \in \mathbb{Z}$, $P_n(x)$ is an polynomial with integer coefficients.

□

Now, we prove that a certain double integral is a rational expression in terms of 1 and $\zeta(3)$, and provide a bound on the denominator of this expression. A more complicated integral of a similar form (with x^r and y^s replaced by $P_n(x)$ and $P_n(y)$) will be used later in the proof. This later integral may be regarded as the analog of $\int_0^\pi f(x)\sin x\,dx$ in the proof of Theorem 3.3 that π is irrational and the lemma below is analagous to proving that $\int_0^\pi f(x)\sin x\,dx$ is an integer.

Lemma 8.2. *Fix an $n \in \mathbb{Z}_{>0}$. Then for any $0 \le s \le r \le n$, we have*

$$\int_0^1\int_0^1 -\frac{\log(xy)}{1-xy}x^ry^s\,dx\,dy = \frac{A+B\cdot\zeta(3)}{(1,\ldots,n)^3}$$

for some $A,B \in \mathbb{Z}$

Proof. Consider the integral:

$$\int_0^1\int_0^1 -\frac{\log xy}{1-xy}x^ry^s\,dx\,dy$$

Note that the integral is proper since the integrand is bounded near $xy = 1$ and convergent near $xy = 0$ (even if $r = s = 0$). Since for $|xy| < 1$ we have

$$\frac{1}{1-xy} = 1 + xy + x^2y^2 + x^3y^3 + \dots \tag{17}$$

the integral is equal to

$$-\int_0^1\int_0^1 \log(xy)x^ry^s(1+xy+x^2y^2+\dots)\,dx\,dy =$$

$$= -\int_0^1\int_0^1 \sum_{k=0}^{\infty} \log(xy)x^ry^sx^ky^k\,dx\,dy$$

$$= -\int_0^1 \sum_{k=0}^{\infty}\int_0^1 \log(xy)x^{r+k}y^{s+k}\,dx\,dy$$

(the interchange of the sum and the integral is justified since for any fixed $y \in (0,1)$ the convergence in (17) is uniform in $x \in [0,1]$). Integrating with respect to x and using integration by parts, it is easy to deduce:

$$-\int_0^1 \sum_{k=0}^{\infty}\int_0^1 \log(xy)x^{r+k}y^{s+k}\,dx\,dy = -\sum_{k=0}^{\infty}\left(\int_0^1 \frac{y^{s+k}\log(y)}{r+k+1} - \int_0^1 \frac{y^{s+k}}{(r+k+1)^2}dy\right)$$

By integrating with respect to y in a similar way, we get:

$$-\sum_{k=0}^{\infty}\left(\int_0^1 \frac{y^{s+k}\log(y)}{r+k+1} - \int_0^1 \frac{y^{s+k}}{(r+k+1)^2}dy\right) = \sum_{k=0}^{\infty}\left(\frac{1}{(r+k+1)(s+k+1)^2} + \frac{1}{(r+k+1)^2(s+k+1)}\right)$$

$$= \sum_{k=0}^{\infty}\left(\frac{r+s+2k+2}{(r+k+1)^2(s+k+1)^2}\right)$$

Since

$$\frac{1}{(s+k+1)^2} - \frac{1}{(r+k+1)^2} = \frac{(r-s)(r+s+2k+2)}{(r+k+1)^2(s+k+1)^2}$$

and

$$\sum_{k=0}^{\infty}\left(\frac{1}{(s+k+1)^2} - \frac{1}{(r+k+1)^2}\right) = \frac{1}{(s+1)^2} - \frac{1}{(r+1)^2} + \frac{1}{(s+2)^2} - \frac{1}{(r+2)^2} + \dots$$

We can conclude that for $r > s$:

$$\sum_{k=0}^{\infty}\left(\frac{r+s+2k+2}{(r+k+1)^2(s+k+1)^2}\right) = \frac{1}{r-s}\left(\sum_{k=0}^{\infty}\frac{1}{(s+k+1)^2} - \sum_{k=0}^{\infty}\frac{1}{(r+k+1)^2}\right) = \frac{1}{r-s}\sum_{k=1}^{r-s}\cdot\frac{1}{(s+k)^2}$$

where the last equality follows by cancelling out the terms of the two series. Also, since $r-s < n$, $(s+k)^2(r-s)$ would be a divisor of $(1,\dots,n)^3$ since $r < n$. Therefore,

$$\int_0^1\int_0^1 -\frac{\log xy}{1-xy}x^ry^s\,dx\,dy$$

is a rational number with denominator dividing $(1,\dots,n)^3$ for $r > s$.
On the other hand, for $r = s$, the equation becomes:

$$\sum_{k=0}^{\infty}\left(\frac{1}{(r+k+1)(s+k+1)^2} + \frac{1}{(r+k+1)^2(s+k+1)}\right) = 2\sum_{k=0}^{\infty}\frac{1}{(r+k+1)^3}$$

Since $\zeta(3)$ is equal to $\sum_{k=1}^{\infty} \frac{1}{k^3}$,

$$2\sum_{k=0}^{\infty} \frac{1}{(r+k+1)^3} = 2\left(\zeta(3) - \sum_{k=1}^{r} \frac{1}{k^3}\right) \tag{18}$$

This implies the result since every k^3 in the expression $\sum_{k=1}^{r} \frac{1}{k^3}$ divides $(1,\dots,n)^3$. □

To continue the analogy with the proof of Theorem 3.3 that π is irrational, we also need to be able to bound above the integral that we use for the irrationality proof. Recall that in that proof, we just wanted to show that $\int_0^{\pi} f(x)\sin x\,dx$ lies between 0 and 1; here the argument will be more complicated.

Lemma 8.3. *Fix an $n \in \mathbb{Z}_{>0}$. Let $P_n(x)$ be as in Lemma 8.1. Then we have*

$$\int_0^1\int_0^1 -\frac{\log(xy)}{1-xy} P_n(x)P_n(y)\,dx\,dy \leq 2\left(\frac{1}{27}\right)^n \zeta(3)$$

Proof. Consider the integral:

$$\int_0^1\int_0^1 -\frac{\log(xy)}{1-xy} P_n(x)P_n(y)\,dx\,dy$$

Since

$$-\frac{\log(xy)}{1-xy} = \int_0^1 \frac{1}{1-(1-xy)z}\,dz \tag{19}$$

we can rewrite the integral as a triple integral:

$$\int_0^1\int_0^1\int_0^1 \frac{1}{1-(1-xy)z}\,dz P_n(x)P_n(y)\,dx\,dy$$

We have

$$\int_0^1\int_0^1\int_0^1 \frac{1}{1-(1-xy)z} P_n(x)P_n(y)\,dz\,dx\,dy = \frac{1}{n!}\int_0^1\int_0^1\int_0^1 \frac{1}{1-(1-xy)z} P_n(y)\,d\left((x^n(1-x)^n)^{(n-1)}\right)dy\,dz$$

Swapping the order of integration and integrating by parts with respect to x, we get

$$\frac{1}{n!}\int_0^1\int_0^1\int_0^1 yz\cdot\left(\frac{1}{1-(1-xy)z}\right)^2 (x^n(1-x)^n)^{(n-1)} P_n(y)\,dx\,dy\,dz$$

using the fact that $(x^n(1-x)^n)^{(n-1)}$ is 0 at $x=0$ and $x=1$. Integrating with respect to x by parts $n-1$ more times in a similar fashion, we get

$$\frac{1}{n!}\int_0^1\int_0^1\int_0^1 n!\frac{x^n y^n z^n (1-x)^n P_n(y)}{(1-(1-xy)z)^{n+1}}\,dx\,dy\,dz = \int_0^1\int_0^1\int_0^1 \frac{x^n y^n z^n (1-x)^n P_n(y)}{(1-(1-xy)z)^{n+1}}\,dx\,dy\,dz$$

We now make a change of variables $x=u,\ y=v,\ z=\frac{1-w}{1-(1-uv)w}$. One can check that this defines a differentiable bijective map $[0,1]^3 \to [0,1]^3$ with Jacobian

$$(u,v,w) = \frac{-uv}{(1-(1-uv)w)^2}$$

Let

$$f(x,y,z) = \frac{x^n y^n z^n (1-x)^n P_n(y)}{(1-(1-xy)z)^{n+1}}$$

By changing the variables in the integral (see [11]) we have (this requires a bit of computation)

$$\int_0^1\int_0^1\int_0^1 f(x,y,z)\,dx\,dy\,dz = \int_0^1\int_0^1\int_0^1 f(x(u,v,w),y(u,v,w),z(u,v,w))|(u,v,w)|\,du\,dv\,dw =$$

$$= \int_0^1\int_0^1\int_0^1 (1-w)^n(1-u)^n\frac{P_n(v)}{1-(1-uv)w}\,du\,dv\,dw$$

Integrating with respect to v by parts n times (and switching the order of integration) similarly to before, this integral is equal to

$$\int_0^1\int_0^1\int_0^1 \frac{u^n(1-u)^n v^n(1-v)^n w^n(1-w)^n}{(1-(1-uv)w)^{n+1}}\,du\,dv\,dw$$

The integrand expression is easy to estimate. Indeed, we have

$$1-(1-uv)w = (1-w)+uvw \geq 2\sqrt{1-w}\sqrt{uvw}$$

on $[0,1]^3$ by arithmetic-geometric mean inequality. Therefore, we have

$$\frac{u(1-u)v(1-v)w(1-w)}{1-(1-uv)w} \leq \frac{1}{2}\sqrt{u}(1-u)\sqrt{v}(1-v)\sqrt{w(1-w)}$$

on $[0,1]^3$. The maximum of $g(t)=\sqrt{t}(1-t)$ for $t\in[0,1]$ occurs at $t=\frac{1}{3}$ and the maximum of $h(t)=t(1-t)$ for $t\in[0,1]$ occurs at $t=\frac{1}{2}$. This implies that

$$\frac{u(1-u)v(1-v)w(1-w)}{1-(1-uv)w} \leq \frac{1}{27}$$

Therefore, we have

$$\begin{aligned}
&\int_0^1\int_0^1\int_0^1 \frac{u^n(1-u)^n v^n(1-v)^n w^n(1-w)^n}{(1-(1-uv)w)^{n+1}}\,du\,dv\,dw = \\
&\quad\int_0^1\int_0^1\int_0^1 \left(\frac{u(1-u)v(1-v)w(1-w)}{1-(1-uv)w}\right)^n \frac{1}{1-(1-uv)w}\,du\,dv\,dw \leq \\
&\qquad\left(\frac{1}{27}\right)^n \int_0^1\int_0^1\int_0^1 \frac{1}{1-(1-uv)w}\,du\,dv\,dw \leq \\
&\qquad\quad\left(\frac{1}{27}\right)^n \int_0^1\int_0^1 -\frac{\log(uv)}{(1-uv)}\,du\,dv = 2\left(\frac{1}{27}\right)^n\zeta(3)
\end{aligned}$$

where the penultimate inequality is by (19) and the last equality follows from (18) in the proof of Lemma 8.2. □

Finally, we are ready to show that $\zeta(3)$ is irrational.

Theorem 8.4. *$\zeta(3)$ is irrational.*

Proof. Consider the integral

$$I_n := \int_0^1\int_0^1 -\frac{\log(xy)}{1-xy}P_n(x)P_n(y)\,dx\,dy$$

from Lemma 8.3. Then there exist some $A',B'\in\mathbb{Z}$ such that

$$I_n = \frac{A'+B'\cdot\zeta(3)}{(1,\ldots,n)^3}$$

Indeed, if $P_n(x) = \sum_{i=0}^{n} b_i x^i$, then $b_i \in \mathbb{Z}$ by Lemma 8.1. But then we have

$$I_n = \sum_{r=0}^{n} \sum_{s=0}^{n} b_r b_s \int_0^1 \int_0^1 -\frac{\log(xy)}{1-xy} x^r y^s \, dx \, dy$$

by linearity and the claim follows from Lemma 8.2.

Suppose that $\zeta(3)$ is rational. Then we have $\zeta(3) = \frac{p}{q}$ for some $p, q \in \mathbb{Z}$ with $q > 0$. Consider $|A' + B'\zeta(3)|$. Note that all the terms in the integrand expression of I_n are positive on $(0,1)$ (one can check that $P_n(x)$ is a polynomial in $x(1-x)$ with positive coefficients) so $I_n \neq 0$. On one hand, we have

$$|A' + B'\zeta(3)| = |A' + B'\frac{p}{q}| = \frac{|A'q + B'p|}{q} \geq \frac{1}{q} \tag{20}$$

On the other hand,

$$|A' + B'\zeta(3)| = I_n (1, \dots, n)^3 \leq 2 \left(\frac{1}{27} \right)^n \zeta(3)(1, \dots, n)^3 \tag{21}$$

It is enough to show that $\lim_{n\to\infty} 2\left(\frac{1}{27}\right)^n \zeta(3)(1, \dots, n)^3 = 0$. Indeed, then we have

$$2 \left(\frac{1}{27} \right)^n \zeta(3)(1, \dots, n)^3 < \frac{1}{q}$$

for n large enough which is a contradiction with (20) and (21). However, it is well-known that the Prime Number Theorem [5] implies that $\lim_{n\to\infty} \sqrt[n]{(1, \dots, n)} = e$. But then

$$\lim_{n\to\infty} 2 \left(\frac{1}{27} \right)^n \zeta(3)(1, \dots, n)^3 = \lim_{n\to\infty} 2 \left(\frac{1}{27} \right)^n \zeta(3) e^{3n} = 2 \left(\frac{e^3}{27} \right)^n \zeta(3) = 0$$

where the last equality is since $\frac{e^3}{27} < 1$. □

9 Advanced results

In this paper we have shown that the the Riemann-Zeta function is irrational at the even positive integers and gave an exposition of Beukers' proof [3] that $\zeta(3)$ is irrational. Not much is known about irrationality of the Riemann-Zeta function at odd integers $\zeta(2k+1)$, for $k > 1$. We finish this paper by listing some known results:

(i) Infinitely many of $\zeta(2k+1)$, for $k \geq 1$ are irrational, see [7].

(ii) It was shown in [13] that one of $\zeta(5), \zeta(7), \dots, \zeta(17), \zeta(19)$ is irrational. The same author also proved a stronger result that one of $\zeta(5), \zeta(7), \zeta(9), \zeta(11)$ is irrational.

These partial result motivate the following conjecture, widely believed to be true but inaccessible with the current tools.

Conjecture 9.1. *All the $\zeta(2k+1), k \geq 1$ are irrational. Moreover, they are transcendental and algebraically independent from powers of π.*

Here being *algebraically independent* from powers of π means that no $\zeta(2k+1), k \geq 1$ is a root of any polynomial with coefficients of the form

$$a_0 + a_1\pi + a_2\pi^2 + \dots + a_n\pi^n$$

for some $n \in \mathbb{N}$ and $a_i \in \mathbb{Z},\ 1 \leq i \leq n$.

Bibliography

[1] R. Apéry, *Irrationalité de $\zeta 2$ et $\zeta 3$*, Journées Arithmétiques de Luminy, Astérisque, no. 61 (1979), pp. 11-13.

[2] T.M. Apostol, *Introduction to analytic number theory*, Springer Science & Business Media (1998).

[3] F. Beukers, *A note on the irrationality of $\zeta(2)$ and $\zeta(3)$*, In: Pi: A Source Book. Springer, New York, NY (2004).

[4] K. Conrad, *Irrationality of π and e* (2009), K.Conrad's website: https://kconrad.math.uconn.edu/blurbs/analysis/irrational.pdf.

[5] G.H. Hardy, E.M. Wright, *An introduction to the theory of numbers*, Oxford university press (1979).

[6] A.J. Hildebrand, *Introduction to analytic number theory*, Math 531 lecture notes (2005), University of Illinois website: https://faculty.math.illinois.edu/ hildebr/ant/.

[7] S. Johannes, *Infinitely many odd zeta values are irrational. By elementary means*, arXiv preprint arXiv:1802.09410 (2018).

[8] V.J. Kanovei, *The correctness of Euler's method for the factorization of the sine function into an infinite product*, Russian Mathematical Surveys Volume 43 (1998).

[9] D.H. Mayer, *On a ζ function related to the continued fraction transformation*, Bulletin de la Société Mathématique de France 104 (1976).

[10] I. Niven, *The transcendence of π*, The American Mathematical Monthly 46.8 (1939): 469-471.

[11] W. Rudin, *Principles of mathematical analysis*, McGraw-Hill New York (1976).

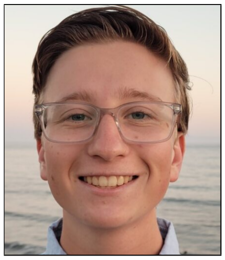
[12] W.R. Sullivan, *Numerous proofs of $\zeta(2) = \frac{\pi^2}{6}$* (2013).

[13] V.V. Zudilin, *One of the eight numbers $\zeta(5)$, $\zeta(7)$,..., $\zeta(17)$, $\zeta(19)$ is irrational*, Mathematical Notes 70.3 (2001): 426-431.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 93 – 117

On the Cauchy Transform of the Complex Power Function

Benjamin Faktor, Michael Kuhn, Gahl Shemy*



Benjamin Faktor is an undergraduate student in mathematics at the University of California Santa Barbara. His interests are in analysis and PDE, which he hopes to continue studying in graduate school. His hobbies include body surfing and hiking.

Michael Kuhn worked on this paper with fellow undergraduate classmates after they finished their first quarter of complex analysis. He recently graduated with majors in Computer Science and Mathematics, and plans to spend some time working as a software engineer. Outside of those disciplines he enjoys surfing and swimming.

Gahl Shemy worked on this paper as an undergraduate studying mathematics at UCSB's College of Creative Studies. She is now a first year math Ph.D. student at the University of Michigan, hoping to do research in representation theory or number theory.

Abstract

The integral $\int_{|z|=1} \frac{z^\beta}{z-\alpha} dz$ for $\beta = \frac{1}{2}$ has been comprehensively studied by Mortini and Rupp for pedagogical purposes. We write for a similar purpose, elaborating on their work with the more general consideration $\beta \in \mathbb{C}$. This culminates in an explicit solution

***Corresponding author:** mkuhn@ucsb.edu

in terms of the hypergeometric function for $|\alpha| \neq 1$ and any $\beta \in \mathbb{C}$. For rational β, the integral is reduced to a finite sum. A differential equation in α is derived for this integral, which we show has similar properties to the hypergeometric equation.

1 Introduction

The purpose of this paper is to investigate integrals of the form

$$\int_{|z|=1} \frac{z^\beta}{z-\alpha}\, dz. \tag{1}$$

Our personal interest in this type of integral stems from a recent paper due to Mortini and Rupp [1], in which the authors evaluate (1) for $\beta = \frac{1}{2}$ using various methods.

Initially we note that the function z^β must be defined, for general $\beta \in \mathbb{C}$, in terms of some branch of the complex logarithm. In our notation, for $0 < \theta < 2\pi$, $\log_\theta(z)$ will represent the branch of the complex logarithm with branch cut $\{re^{i\theta} : r \geq 0\}$; it is defined on the simply connected domain $\mathbb{C} \setminus \{re^{i\theta} : r \geq 0\}$, and we fix $\log_\theta(1) = 0$. Under these conditions our branch is

$$\log_\theta(z) = \ln|z| + i\arg_\theta(z)$$

where $\arg_\theta$ is the argument function with values in $(\theta - 2\pi, \theta)$.
This branch can be related to the branch of the square root discussed in [1] by taking $t_0 = \theta - 2\pi$.
We denote by $\mathrm{Arg}(z)$ the argument of z falling in the range $[0, 2\pi)$, and by $\arg(z)$ the equivalence class (modulo 2π) of all possible values for the argument of z. Any condition with $\arg(z)$ is considered satisfied if one representative satisfies the condition.

The implications of using a branch of the complex logarithm to define the complex power are that even when we choose $|\alpha| \neq 1$, the meromorphic function

$$m_{\alpha,\beta,\theta}(z) := \frac{z^\beta}{z-\alpha} = \frac{e^{\beta \log_\theta(z)}}{z-\alpha} \tag{2}$$

will not be analytic, or even *continuous*, on the boundary of the unit disk. This is due to the branch cut necessary for the $\log_\theta$ function used in (2). The discontinuity at the branch cut, although merely a jump, prevents a simple evaluation with direct application of Cauchy's Residue Theorem. Rather, one must proceed using different methods.

The main results of the paper are explicit expressions of (1) in the two cases of $|\alpha| > 1$ and $|\alpha| < 1$. Specifically, we prove:

Theorem 1. *When $|\alpha| > 1$,*

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = \begin{cases} -2\pi i \alpha^\beta & \beta \in \mathbb{Z}_{<0}, \\ 0 & \beta = 0, \\ e^{i\beta\theta}\left(1 - e^{-2\pi i\beta}\right)\frac{1}{\beta}\left[1 - {}_2F_1(1,\beta;1+\beta;\alpha^{-1}e^{i\theta})\right] & \beta \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0}. \end{cases}$$

When $|\alpha| < 1$*,*

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = \begin{cases} 2\pi i\alpha^{\beta} & \beta \in \mathbb{Z}_{\geq 0}, \\ e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\frac{1}{\beta}\,{}_2F_1(1,-\beta;1-\beta;\alpha e^{-i\theta}) & \beta \in \mathbb{C}\setminus\mathbb{Z}_{\geq 0}. \end{cases}$$

In §2, the unit circle is approximated with a contour of integration which avoids the branch cut in order to derive an equation involving (1). The connection between (1) and the hypergeometric function, ${}_2F_1$, is made in §3 through the identification of a core integral in §3.1. In §4, series manipulation leads to the proof of Theorem 1. The particular case when $\beta \in \mathbb{Q}\setminus\mathbb{Z}$ is further simplified in §5, and in §6 we include a derivation of a differential equation for which (1) is a solution. Provided in §7, the appendix, is a discussion of measure theory topics leading up to the statement of Lebesgue's Dominated Convergence Theorem; adequate references are cited there for the curious reader.

2 Contour Method

We first extend §1 in [1], evaluating (1) via contour integration. For this section alone (§2) it is additionally assumed that $\mathrm{Arg}(\alpha) \neq \theta$ and $\alpha \neq 0$, so that α does not lie on the branch cut. Furthermore, we assume that $\Re(\beta) > 0$, as this condition will be necessary for certain bounds. The purpose of this section is to prove the following lemma:

Lemma 1. *If* $\mathrm{Arg}(\alpha) \neq \theta$*, and* $\Re(\beta) > 0$*, then for* $0 < |\alpha| < 1$*,*

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = 2\pi i\alpha^{\beta} + e^{i\beta\theta}(1-e^{-2\pi i\beta})\int_0^1 \frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt,$$

and for $|\alpha| > 1$*,*

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = e^{i\beta\theta}(1-e^{-2\pi i\beta})\int_0^1 \frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt.$$

Proof. There are 3 main steps:

§2.1) constructing a proper contour;

§2.2) finding singularities and computing their residues;

§2.2) using limits to derive a useful equation.

The lemma follows from plugging (11), (35), (13), (38), and (22) all back into (9). □

2.1 Constructing the Contour

Take a branch of the complex logarithm $\log_\theta$ in the definition of z^β, and let the contour of integration $\Gamma_{\varepsilon,\theta,\rho}$ consist of:

a) the line segment $L_{\varepsilon,\theta,\rho} := \{z \in \mathbb{C} : \rho \leq |z| \leq 1, \arg z = \theta + \varepsilon\}$,

b) the arc $C_{\varepsilon,\theta} := \{z \in \mathbb{C} : |z| = 1, \theta + \varepsilon \leq \arg z \leq \theta + 2\pi - \varepsilon\}$,

c) the line segment $M_{\varepsilon,\theta,\rho} := \{z \in \mathbb{C} : 1 > |z| > \rho, \arg z = \theta + 2\pi - \varepsilon\}$,

d) the arc $D_{\varepsilon,\theta,\rho} := \{z \in \mathbb{C} : |z| = \rho, \theta + 2\pi - \varepsilon \geq \arg z \geq \theta + \varepsilon\}$, oriented as usual, with the bounded region enclosed on the left as we trace the contour. For example, for the principal branch of log ($\log_\pi$ in our notation), the contour is as in Figure 2.1. Under this definition, we have

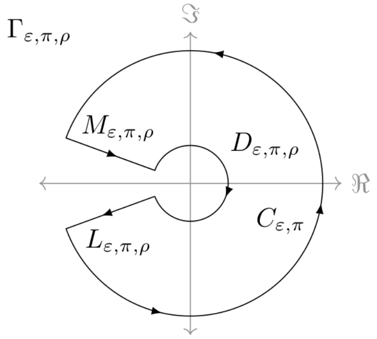

$\Gamma_{\varepsilon,\pi,\rho}$
$\Im$
$M_{\varepsilon,\pi,\rho}$
$D_{\varepsilon,\pi,\rho}$
$\Re$
$C_{\varepsilon,\pi}$
$L_{\varepsilon,\pi,\rho}$

Figure 2.1: Contour for $\theta = \pi$.

$$\int_{\Gamma_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} = \int_{C_{\varepsilon,\theta}} m_{\alpha,\beta,\theta} + \int_{D_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} + \int_{L_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} + \int_{M_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}. \tag{3}$$

One can choose any parameterization of the four curves, noting that smooth equivalence of parameterizations will guarantee generality. In particular, we choose

a) $L_{\varepsilon,\theta,\rho}$: $z(t) = te^{i(\theta+\varepsilon)}$ for $\rho \le t \le 1$,
$$\int_{L_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}(z)\,dz = \int_{\rho}^{1} m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)}\,dt; \tag{4}$$

b) $C_{\varepsilon,\theta}$: $z(t) = e^{it}$ for $\theta+\varepsilon \le t \le \theta+2\pi-\varepsilon$,
$$\int_{C_{\varepsilon,\theta}} m_{\alpha,\beta,\theta}(z)\,dz = \int_{\theta+\varepsilon}^{\theta+2\pi-\varepsilon} m_{\alpha,\beta,\theta}(e^{it})\,ie^{it}dt; \tag{5}$$

c) $M_{\varepsilon,\theta,\rho}$: $z(t) = te^{i(\theta+2\pi-\varepsilon)}$ for $1 \ge t \ge \rho$,
$$\int_{M_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}(z)\,dz = \int_{1}^{\rho} m_{\alpha,\beta,\theta}(te^{i(\theta+2\pi-\varepsilon)})e^{i(\theta+2\pi-\varepsilon)}\,dt; \tag{6}$$

d) $D_{\varepsilon,\theta,\rho}$: $z(t) = \rho e^{it}$ for $\theta+2\pi-\varepsilon \ge t \ge \theta+\varepsilon$,
$$\int_{D_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}(z)\,dz = \int_{\theta+2\pi-\varepsilon}^{\theta+\varepsilon} m_{\alpha,\beta,\theta}(\rho e^{it})\,i\rho e^{it}dt. \tag{7}$$

2.2 Applying the Residue Theorem

Applying Cauchy's Residue Theorem requires computing residues for singularities contained within the contour. To compute the residues of the meromorphic function $m_{\alpha,\beta,\theta}(z)$ defined in (2), first note that $e^{\beta \log_\theta(z)}$ is analytic in $\mathbb{C} \setminus \{re^{i\theta} \in \mathbb{C} : r \geq 0\}$, so the only singularity of $m_{\alpha,\beta,\theta}$ is at α, and this singularity only becomes relevant when $|\alpha| \leq 1$. This singularity is a simple pole, since

$$\lim_{z\to\alpha}(z-\alpha)m_{\alpha,\beta,\theta}(z) = \lim_{z\to\alpha} e^{\beta \log_\theta(z)} = \alpha^\beta \neq 0 \tag{8}$$

but

$$\lim\nolimits_{z\to\alpha}(z-\alpha)^2 m_{\alpha,\beta,\theta}(z) = \lim\nolimits_{z\to\alpha}(z-\alpha)e^{\beta \log_\theta(z)} = 0.$$

Evaluating as in (8), the residue at α is found to be α^β. In order to derive an equation involving (1), one might consider first taking the limit $\varepsilon \to 0^+$ and then $\rho \to 0^+$ in (3):

$$\lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_{\Gamma_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} = \lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\left[\int_{C_{\varepsilon,\theta}} m_{\alpha,\beta,\theta} + \int_{D_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} + \int_{L_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} + \int_{M_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}\right]. \tag{9}$$

Since the contour $\Gamma_{\varepsilon,\theta,\rho}$ in (9) lies in the interior of the simply connected domain of $\log_\theta$ whenever $\varepsilon, \rho > 0$, $m_{\alpha,\beta,\theta}$ is analytic on the path of integration so long as α does not lie on $\Gamma_{\varepsilon,\theta,\rho}$. In this case, Cauchy's Residue Theorem applies and so

$$\int_{\Gamma_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} = 2\pi i\, n(\Gamma_{\varepsilon,\theta,\rho},\alpha)\,\mathrm{Res}(m_{\alpha,\beta,\theta},\alpha) = 2\pi i \alpha^\beta\; n(\Gamma_{\varepsilon,\theta,\rho},\alpha)$$

where $n(\Gamma_{\varepsilon,\theta,\rho},\alpha)$ is the winding number of $\Gamma_{\varepsilon,\theta,\rho}$ around α. Note that by definition of the contour, and because $\mathrm{Arg}(\alpha) \neq \theta$ by assumption, we have

$$n(\Gamma_{\varepsilon,\theta,\rho},\alpha) = \begin{cases} 1 & \text{if } 0 < \varepsilon < \min_{\arg(\alpha)}\{|\arg(\alpha) - \theta|\} \text{ and } 0 < \rho < |\alpha| < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where the notation $\min_{\arg(\alpha)}$ in (10) denotes that the minimum is taken over all possible representatives of $\arg(\alpha)$. It follows that

$$\lim_{\varepsilon\to 0^+} n(\Gamma_{\varepsilon,\theta,\rho},\alpha) = \begin{cases} 1 & \text{if } 0 < \rho < |\alpha| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

since ε can certainly be made smaller than $|\arg(\alpha) - \theta| > 0$, and that

$$\lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+} n(\Gamma_{\varepsilon,\theta,\rho},\alpha) = \begin{cases} 1 & \text{if } 0 < |\alpha| < 1, \\ 0 & \text{otherwise} \end{cases}$$

since ρ can certainly be made smaller than $|\alpha| > 0$. Therefore

$$\lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_{\Gamma_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta} = \begin{cases} 2\pi i \alpha^\beta & \text{if } 0 < |\alpha| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

In evaluating $\lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_{C_{\varepsilon,\theta}} m_{\alpha,\beta,\theta}$, we use (5) to express

$$\begin{aligned}\lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_{C_{\varepsilon,\theta}} m_{\alpha,\beta,\theta} &= \lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_{\theta+\varepsilon}^{\theta+2\pi-\varepsilon} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt,\\ &= \lim_{\varepsilon\to 0^+}\int_{\theta+\varepsilon}^{\theta+2\pi-\varepsilon} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt,\\ &= \mathscr{P}\mathscr{V}\int_{\theta}^{\theta+2\pi} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt,\\ &= \int_{\theta}^{\theta+2\pi} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt, &(12)\\ &= \int_{|z|=1} m_{\alpha,\beta,\theta}(z)\, dz. &(13)\end{aligned}$$

To see that the value of the improper integral in (12) is the same as its principal value, note that whenever an improper integral converges, its principal value converges as well (and to the same value). By convention, (12) is evaluated as

$$\int_{\theta}^{\theta+2\pi} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt = \lim_{\varepsilon\to 0}\int_{\theta+\varepsilon}^{\theta+\pi} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt + \lim_{\varepsilon'\to 0}\int_{\theta+\pi}^{\theta+2\pi-\varepsilon'} m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}dt. \tag{14}$$

It suffices to show that $g(t) = m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}$ is bounded on $[\theta,\theta+2\pi]$ in order for the right hand side of (14) to converge, and thus for the desired improper integral to converge. We first bound the real part of $\beta\log_\theta(z)$, noting that

$$\begin{aligned}\beta\log_\theta(z) &= [\Re(\beta)+i\Im(\beta)]\,[\ln|z|+i\arg_\theta(z)],\\ &= [\Re(\beta)\ln|z|-\Im(\beta)\arg_\theta(z)]+i[\Re(\beta)\arg_\theta(z)+\Im(\beta)\ln|z|], &(15)\end{aligned}$$

where $\arg_\theta := \Im(\log_\theta)$. Since we fix $\log_\theta(1)=0$ for every $0<\theta<2\pi$, the continuity of $\log_\theta$ on its simply connected domain implies that

$$-2\pi < \arg_\theta(z) < 2\pi$$

for all z in the domain and for all θ. Further, continuity also implies that even as z approaches the branch cut (in a limiting sense),

$$-2\pi \le \arg_\theta(z) \le 2\pi. \tag{16}$$

Equations (15) and (16) along with the assumption $\Re(\beta)>0$ give the bound

$$\Re(\beta\log_\theta(z)) = \Re(\beta)\ln|z| - \Im(\beta)\arg_\theta(z) \le \Re(\beta)\ln|z| + 2\pi|\Im(\beta)|. \tag{17}$$

Since $|e^z| = e^{\Re(z)}$, we can now bound

$$\begin{aligned}|m_{\alpha,\beta,\theta}(z)| &= \frac{|e^{\beta\log_\theta(z)}|}{|z-\alpha|},\\ &= \frac{e^{\Re(\beta\log_\theta(z))}}{|z-\alpha|}, &(18)\\ &\le \frac{e^{\Re(\beta)\ln|z|+2\pi|\Im(\beta)|}}{||z|-|\alpha||}. &(19)\end{aligned}$$

For $t\in[\theta,\theta+2\pi]$, the bound (19) immediately gives

$$|m_{\alpha,\beta,\theta}(e^{it})\, ie^{it}| \le \frac{e^{\Re(\beta)\ln|e^{it}|+2\pi|\Im(\beta)|}}{||e^{it}|-|\alpha||} = \frac{e^{2\pi|\Im(\beta)|}}{|1-|\alpha||}. \tag{20}$$

Thus both limits on the right hand side of (14) converge, and hence the equality in (12) is justified.

Now we show that the portion of the integral over the contour $D_{\varepsilon,\theta,\rho}$ approaches 0 as $\varepsilon \to 0^+, \rho \to 0^+$. Using (19) and applying an ML-bound to (7) yields

$$\begin{aligned}\left|\int_{D_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}(z)\,dz\right| &= \left|\int_{\theta+2\pi-\varepsilon}^{\theta+\varepsilon} m_{\alpha,\beta,\theta}(\rho e^{it})\; i\rho e^{it} dt\right|,\\ &\le \rho\,\frac{e^{\Re(\beta)\ln|\rho|+2\pi|\Im(\beta)|}}{|\rho-|\alpha||}(2\pi-2\varepsilon),\\ &\le 2\pi e^{2\pi|\Im(\beta)|}\frac{\rho^{\Re(\beta)}}{|\frac{|\alpha|}{\rho}-1|}. \end{aligned} \tag{21}$$

Since $|\alpha|>0$, (21) gives

$$\begin{aligned}\left|\lim_{\rho\to0^+}\lim_{\varepsilon\to0^+}\int_{D_{\varepsilon,\theta}} m_{\alpha,\beta,\theta}(z)\,dz\right| &\le \lim_{\rho\to0^+}\lim_{\varepsilon\to0^+}\left[2\pi e^{2\pi|\Im(\beta)|}\frac{\rho^{\Re(\beta)}}{|\frac{|\alpha|}{\rho}-1|}\right],\\ &= \lim_{\rho\to0^+}\left[2\pi e^{2\pi|\Im(\beta)|}\frac{\rho^{\Re(\beta)}}{|\frac{|\alpha|}{\rho}-1|}\right],\\ &= 0. \end{aligned} \tag{22}$$

We now consider the limiting value of the integral along $L_{\varepsilon,\theta,\rho}$. The core difficulty of this part of the contour integral is in evaluating

$$\lim_{\varepsilon\to0^+}\int_{\rho}^{1} m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)}\,dt. \tag{23}$$

The strategy is to use Lebesgue's Dominated Convergence Theorem (see §7, specifically Theorem 2). Take the family of functions defined on $t\in[0,1]$:

$$\mathscr{F}_L := \left\{ f_\varepsilon(t) = m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)} \;\middle|\; 0<\varepsilon<\pi \right\}. \tag{24}$$

The function $te^{i(\theta+\varepsilon)}$ is continuous in the positive real variable t, and $m_{\alpha,\beta,\theta}$ is continuous on its simply connected domain except at the point α; this singularity is undesirable. To achieve continuity of the functions in question, we restrict our view to the following collection instead:

Fix r with $0<r<\min_{\arg(\alpha)}|\arg(\alpha)-\theta|$, and define

$$\mathscr{F}_L^* := \left\{ f_\varepsilon(t) = m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)} \;\middle|\; 0<\varepsilon<r \right\}. \tag{25}$$

The point α is outside the sector between $\theta-r$ and $\theta+r$ (see Figure 2.3 for illustration), so now the functions are continuous. This allows us to conclude that $\mathscr{F}_L^*$ is a set of Lebesgue measurable functions.

Moreover, in $\mathscr{F}_L^*$, we have

$$\begin{aligned}\lim_{\varepsilon\to 0^+} f_\varepsilon(t) &= \lim_{\varepsilon\to 0^+} m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)},\\ &= \lim_{\varepsilon\to 0^+} \frac{e^{\beta\log_\theta(te^{i(\theta+\varepsilon)})}}{te^{i(\theta+\varepsilon)}-\alpha}e^{i(\theta+\varepsilon)},\\ &= \lim_{\varepsilon\to 0^+} \frac{e^{\beta(\ln t+i(\theta-2\pi+\varepsilon))}}{te^{i(\theta+\varepsilon)}-\alpha}e^{i(\theta+\varepsilon)},\\ &= \frac{e^{\beta(\ln t+i(\theta-2\pi))}}{te^{i\theta}-\alpha}e^{i\theta} =: g_\theta(t). \qquad (26)\end{aligned}$$

Note that $t\in[\rho,1]\subseteq(0,\infty)$ above, and since $\operatorname{Arg}(\alpha)\neq\theta$ we have that $te^{i\theta}-\alpha\neq 0$ in the limit. The limit above is evaluated using this fact along with continuity of the exponential.
Equivalently, this means that for any sequence $\varepsilon_n\to 0^+$, f_{ε_n} converges pointwise to g_θ as $n\to\infty$.

In fact, the conditions for Lebesgue's Dominated Convergence Theorem above can be shown for $\mathscr{F}_L$ rather than $\mathscr{F}_L^*$ using a slightly more advanced argument. For the last condition however, which requires us to bound functions in the family by a Lebesgue integrable function, it is much easier to consider only $\mathscr{F}_L^*$. For a given $f_\varepsilon\in\mathscr{F}_L^*$, let $\delta := \min_{\arg(\alpha)}\{|\arg(\alpha)-(\theta+\varepsilon)|\}$. That is, δ gives the minimum difference in angle between $\theta+\varepsilon$ and the vector from the origin out to α.

If $\delta\geq\frac{\pi}{2}$, simple geometry gives that α is at least a distance of $|\alpha|$ away from the segment $te^{i(\theta+\varepsilon)}$ for $t\in[\rho,1]$. To see this, consider Figure 2.2 and note that the side of the triangle opposite the angle of size δ is the longest side of the triangle (since δ is either right or obtuse). Thus the shortest distance d from α to the line segment $L_{\varepsilon,\theta,\rho}$ is bounded below:

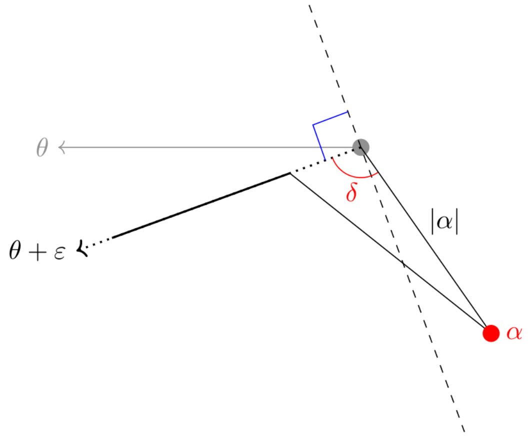


Figure 2.2: Illustration of the case $\delta\geq\frac{\pi}{2}$.

$$d > |\alpha|. \qquad (27)$$

If instead $\delta < \frac{\pi}{2}$, then α is at least a distance of $|\alpha|\sin(\delta)$ away from the segment $te^{i(\theta+\varepsilon)}$ for $t \in [\rho,1]$. To see this consider the similar picture in Figure 2.3 and note that the altitude dropped from α to the line containing $L_{\varepsilon,\theta,\rho}$ is precisely of length $|\alpha|\sin(\delta)$ (although the distance will be greater if $|\alpha|$ is so small or so large that the altitude dropped onto the line does not strike within the segment parameterized by $t \in [\rho,1]$).

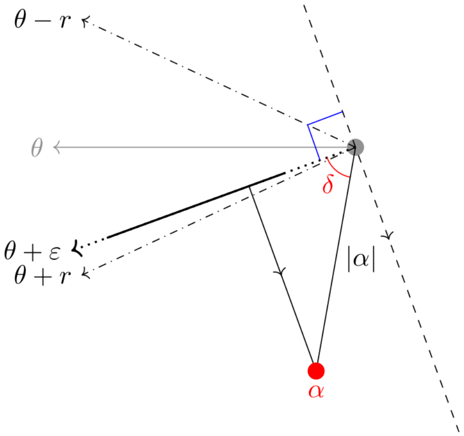


Figure 2.3: Illustration of the case $\delta < \frac{\pi}{2}$.

Now in our consideration of $\mathscr{F}_L^*$, we have $\varepsilon < r < \delta$ and so

$$|\alpha|\sin(\delta) > |\alpha|\sin\left(\min_{\arg(\alpha)}\left\{|\arg(\alpha) - (\theta \pm r)|\right\}\right) > k > 0,$$

for a fixed constant k dependent on r, θ, and α. Thus in this case as well, the shortest distance d from α to the line segment $L_{\varepsilon,\theta,\rho}$ is bounded below:

$$d > k. \tag{28}$$

Consequently, for $K := \max\{\frac{1}{|\alpha|}, \frac{1}{k}\}$, we have that every function $f_\varepsilon \in \mathscr{F}_L^*$ has for all $t \in [\rho,1]$ that

$$\begin{aligned} |f_\varepsilon(t)| &= |m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)}|, \\ &= |m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})|, \\ &= \frac{e^{\Re(\beta\log_\theta(te^{i(\theta+\varepsilon)}))}}{|te^{i(\theta+\varepsilon)} - \alpha|}, \qquad (29) \\ &\leq Ke^{\Re(\beta)\ln|te^{i(\theta+\varepsilon)}|+2\pi|\Im(\beta)|}, \\ &\leq Ke^{\Re(\beta)|t|+2\pi|\Im(\beta)|} =: h_r(t). \qquad (30) \end{aligned}$$

(The third equality holds by (18); the first inequality holds by (17) and the reasoning which led to (27) and (28); the final inequality holds since $|t| > \ln|t|$ for all $t \in \mathbb{R}$ and the because exponential is strictly increasing on $\mathbb{R}$.)

Clearly this h_r is integrable on $[\rho,1]$ for all $\rho > 0$, since it is simply a scaled exponential.

Now, consider any arbitrary sequence $(\varepsilon_n) \to 0^+$ with $\varepsilon_n < r$, and define a sequence of functions (f_{ε_n}); note $f_{\varepsilon_n} \in \mathscr{F}_L^*$ for all n. From (26) we have $(f_{\varepsilon_n}) \to g_\theta$ pointwise, and $|f_{\varepsilon_n}(t)| \leq h_r(t)$ for all n and for all $t \in [0,1]$, as shown in (30). Therefore Lebesgue's Dominated Convergence Theorem implies that

$$\lim_{n\to\infty} \int_\rho^1 f_{\varepsilon_n}(t)\, dt = \int_\rho^1 g_\theta(t)\, dt. \tag{31}$$

for all $\rho > 0$.

Dispensing with the condition $\varepsilon_n < r$, it is still true for any arbitrary sequence $(\varepsilon_n) \to 0^+$ that

$$\lim_{n\to\infty} \int_\rho^1 f_{\varepsilon_n}(t)\, dt = \int_\rho^1 g_\theta(t)\, dt \tag{32}$$

since $(\varepsilon_n) \to 0^+$ has a tail which is completely bounded above by r, and thus convergence of the tail shown in (31) implies convergence of the whole sequence. Since (32) holds for arbitrary (ε_n), this implies

$$\lim_{\varepsilon\to 0^+} \int_\rho^1 f_\varepsilon(t)\, dt = \int_\rho^1 g_\theta(t)\, dt. \tag{33}$$

for $f_\varepsilon \in \mathscr{F}_L$.

Hence we evaluate (23) and find

$$\begin{aligned}\lim_{\varepsilon\to 0^+} \int_\rho^1 m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)}\, dt &= \int_\rho^1 \lim_{\varepsilon\to 0^+} m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)} dt,\\ &= \int_\rho^1 \frac{e^{\beta(\ln t + i(\theta-2\pi))}}{te^{i\theta}-\alpha} e^{i\theta}\, dt,\\ &= e^{i\beta(\theta-2\pi)} \int_\rho^1 \frac{e^{\beta \ln t}}{t-\alpha e^{-i\theta}}\, dt. \end{aligned} \tag{34}$$

But g_θ is continuous for $t \in (0,1]$ and bounded as $t \to 0$. Therefore, allowing improper integrals, and drawing from equations (4) and (34) it is straightforward to compute

$$\begin{aligned}\lim_{\rho\to 0^+} \lim_{\varepsilon\to 0^+} \int_{L_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}(z)\, dz &= \lim_{\rho\to 0^+} \lim_{\varepsilon\to 0^+} \int_\rho^1 m_{\alpha,\beta,\theta}(te^{i(\theta+\varepsilon)})e^{i(\theta+\varepsilon)}\, dt,\\ &= \lim_{\rho\to 0^+} \left[e^{i\beta(\theta-2\pi)} \int_\rho^1 \frac{e^{\beta \ln t}}{t-\alpha e^{-i\theta}}\, dt \right],\\ &= e^{i\beta(\theta-2\pi)} \int_0^1 \frac{e^{\beta \ln t}}{t-\alpha e^{-i\theta}}\, dt. \end{aligned} \tag{35}$$

Finally we take limits in the last integral on the right hand side of (9) along $M_{\varepsilon,\theta,\rho}$. Similarly the difficulty in this case is evaluating

$$\lim_{\varepsilon\to 0^+} \int_1^\rho m_{\alpha,\beta,\theta}(te^{i(\theta+2\pi-\varepsilon)})e^{i(\theta+2\pi-\varepsilon)}\, dt \tag{36}$$

using Lebesgue's Dominated Convergence Theorem. Analogous steps as those used for $L_{\varepsilon,\theta,\rho}$ can be applied to the $M_{\varepsilon,\theta,\rho}$ case to show that

$$\begin{aligned}\lim_{\varepsilon\to 0^+}\int_1^{\rho} m_{\alpha,\beta,\theta}(te^{i(\theta+2\pi-\varepsilon)})e^{i(\theta+2\pi-\varepsilon)}\,dt &= -\int_{\rho}^{1}\lim_{\varepsilon\to 0^+} m_{\alpha,\beta,\theta}(te^{i(\theta+2\pi-\varepsilon)})e^{i(\theta+2\pi-\varepsilon)}\,dt,\\ &= -\int_{\rho}^{1}\frac{e^{\beta(\ln(t)+i\theta)}}{te^{i(\theta+2\pi)}-\alpha}e^{i(\theta+2\pi)}\,dt,\\ &= -e^{i\beta\theta}\int_{\rho}^{1}\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt. \qquad (37)\end{aligned}$$

Just as before the integrand is bounded on $[0,1]$. Using (37) there is no issue writing

$$\begin{aligned}\lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_{M_{\varepsilon,\theta,\rho}} m_{\alpha,\beta,\theta}(z)\,dz &= \lim_{\rho\to 0^+}\lim_{\varepsilon\to 0^+}\int_1^{\rho} m_{\alpha,\beta,\theta}(te^{i(\theta+2\pi-\varepsilon)})e^{i(\theta+2\pi-\varepsilon)}\,dt,\\ &= \lim_{\rho\to 0^+}\left[-e^{i\beta\theta}\int_{\rho}^{1}\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt\right],\\ &= -e^{i\beta\theta}\int_{0}^{1}\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt. \qquad (38)\end{aligned}$$

3 The Hypergeometric Function Connection

3.1 A Core Integral

In order to fully evaluate (1) using the contour method outlined in §2, the following integral from Lemma 1 must be evaluated:

$$\int_0^1 \frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt, \qquad (39)$$

which exists for $\Re(\beta) > -1$. The integral in (39) is in fact an improper integral and can be written

$$\lim_{\rho\to 0}\int_{\rho}^{1}\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt.$$

Algebraic manipulations give

$$\begin{aligned}\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}} &= \frac{e^{\ln t}}{t-\alpha e^{-i\theta}}e^{(\beta-1)\ln t},\\ &= \frac{t-\alpha e^{-i\theta}+\alpha e^{-i\theta}}{t-\alpha e^{-i\theta}}e^{(\beta-1)\ln t},\\ &= \left(1+\frac{\alpha e^{-i\theta}}{t-\alpha e^{-i\theta}}\right)e^{(\beta-1)\ln t}. \qquad (40)\end{aligned}$$

Integrating first over the interval $[\rho,1]$ and using (40) yields

$$\int_{\rho}^{1}\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt = \int_{\rho}^{1} e^{(\beta-1)\ln t}\,dt + \alpha e^{-i\theta}\int_{\rho}^{1}\frac{e^{(\beta-1)\ln t}}{t-\alpha e^{-i\theta}}\,dt. \qquad (41)$$

Notice that the function $e^{(\beta-1)\log(t)}$ is the derivative of $\frac{1}{\beta}e^{\beta\log(t)}$, which is analytic on $[\rho,1]$. Thus

$$\int_\rho^1 e^{(\beta-1)\ln t}\,dt = \frac{1}{\beta}e^{\beta\ln(1)} - \frac{1}{\beta}e^{\beta\ln(\rho)} = \frac{1}{\beta} - \frac{1}{\beta}e^{\beta\ln(\rho)}. \tag{42}$$

Substituting (42) into (41) and taking limits gives

$$\lim_{\rho\to0}\int_\rho^1 \frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt = \lim_{\rho\to0}\left[\frac{1}{\beta}-\frac{1}{\beta}e^{\beta\ln(\rho)}\right] + \alpha e^{-i\theta}\lim_{\rho\to0}\int_\rho^1 \frac{e^{(\beta-1)\ln t}}{t-\alpha e^{-i\theta}}\,dt,$$

$$\int_0^1 \frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt = \frac{1}{\beta} + \alpha e^{-i\theta}\int_0^1 \frac{e^{(\beta-1)\ln t}}{t-\alpha e^{-i\theta}}\,dt; \tag{43}$$

the integral on the right hand side of (43) exists for $\Re(\beta) > 0$. Again the convergence of improper integrals follows from the boundedness of the integrands. Moving the constant inside the integral in (43) gives

$$\int_0^1 \frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt = \frac{1}{\beta} - \int_0^1 \frac{e^{(\beta-1)\ln t}}{1-\left(\frac{1}{\alpha}e^{i\theta}\right)t}\,dt. \tag{44}$$

Therefore finding a solution to (1) using the contour integration method necessitates working with the following "core integral" for $z = \frac{1}{\alpha}e^{i\theta}$:

$$\int_0^1 t^{\beta-1}(1-zt)^{-1}\,dt. \tag{45}$$

The choice to write $t^{\beta-1}$ rather than $e^{(\beta-1)\ln t}$ in (45) is intentional, since generality is not lost when any branch $\log_\theta$ for $\theta \not\equiv 0$ is used to define this complex power of $t \in [0,1]$.

3.2 Definition & Relevant Identities

We investigate the integral in (45) by making use of the well-studied hypergeometric function ${}_2F_1(a,b,c;z)$. For $|z| < 1$, this function is defined as the infinite series

$${}_2F_1(a,b,c;z) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n n!}z^n, \quad c \in \mathbb{C}\setminus\mathbb{Z}_{\leq 0} \tag{46}$$

where $(x)_n = \frac{\Gamma(x+n)}{\Gamma(x)}$ is the rising Pochhammer symbol.

The hypergeometric series generalizes the geometric series, and is prominent in the study of linear differential equations with three regular singular points. The hypergeometric function is notably a solution to the hypergeometric equation, discussed in §6.

A comprehensive collection of identities involving ${}_2F_1$ can be found in [2]. The most notable for our purposes is the following:

For $|z| < 1$ and $\Re(c) > \Re(b) > 0$,

$${}_2F_1(a,b,c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\,dt. \tag{47}$$

Letting $a = 1$, $b = \beta$, and $c = 1 + \beta$ under the conditions for (47) gives

$$\begin{aligned} {}_2F_1(1,\beta,1+\beta;z) &= \frac{\Gamma(1+\beta)}{\Gamma(\beta)\Gamma(1)} \int_0^1 t^{\beta-1}(1-t)^0(1-tz)^{-1}\,dt, \\ &= \beta \int_0^1 t^{\beta-1}(1-zt)^{-1}\,dt, \end{aligned} \tag{48}$$

such that the integral above is exactly the integral in (45), only scaled.

3.3 Final Steps of the Contour Method

We conclude the contour method for $|\alpha| > 1$ by proving the following statement, making use of the hypergeometric identity (48).

Proposition 1. *When* $|\alpha| > 1$*,* $\mathrm{Arg}(\alpha) \neq \theta$*,* $\theta \neq 0$ *(mod* 2π*), and* $\Re(\beta) > 0$*,*

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = e^{i\beta\theta}(1-e^{-2\pi i\beta})\frac{1}{\beta}\left[1 - {}_2F_1(1,\beta;1+\beta;\alpha^{-1}e^{i\theta})\right]. \tag{49}$$

Proof. Since $|\alpha| > 1$, the last argument in the hypergeometric function satisfies

$$\left|\frac{1}{\alpha}e^{i\theta}\right| = \frac{1}{|\alpha|} < 1.$$

Under the assumption $\Re(\beta) > 0$, one can apply the identity (48) and find

$$\int_0^1 \frac{t^{\beta-1}}{1-(\alpha^{-1}e^{i\theta})t}\,dt = \frac{1}{\beta}\,{}_2F_1(1,\beta;1+\beta;\alpha^{-1}e^{i\theta}). \tag{50}$$

With this expression for the core integral, an application of Lemma 1 and (44) completes the proof. □

The case $0 < |\alpha| < 1$ cannot be approached in the same manner. While the initial steps in the contour method still hold, the integral identity from (48) does not apply since the last argument in the hypergeometric function now satisfies

$$\left|\frac{1}{\alpha}e^{i\theta}\right| = \frac{1}{|\alpha|} > 1;$$

which is outside the domain of (47).

4 Series Method

Fortunately there exist methods outside of contour integration which allow us to express (1) in terms of the hypergeometric function in all cases. Rather than dealing with integral identities of the hypergeometric function, one can work with series to produce a term of the form (46).

4.1 Proof of the main result

Consider first $|\alpha| > 1$. Recall from (12) and (13) that

$$\begin{aligned}\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta}(z)\,dz &= \int_{|z|=1} \frac{e^{\beta \log_\theta(z)}}{z-\alpha}\,dz, \\ &= \lim_{\varepsilon\to 0^+}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon} m_{\alpha,\beta,\theta}(e^{it})ie^{it}\,dt. \qquad (51)\end{aligned}$$

For $\theta - 2\pi < t < \theta$, $\log_\theta(e^{it}) = it$, so one can then rewrite the integrand as

$$\begin{aligned} m_{\alpha,\beta,\theta}(e^{it})ie^{it} &= \frac{e^{\beta(\log_\theta(e^{it}))}}{e^{it}-\alpha} ie^{\log_\theta(e^{it})}, \\ &= i\frac{e^{(\beta+1)(\log_\theta(e^{it}))}}{e^{it}-\alpha}, \\ &= i\frac{e^{i(\beta+1)t}}{e^{it}-\alpha}, \\ &= -\frac{ie^{i(\beta+1)t}}{\alpha}\cdot\frac{1}{1-\frac{1}{\alpha}e^{it}}, \\ &= -\frac{ie^{i(\beta+1)t}}{\alpha}\sum_{k=0}^{\infty}\alpha^{-k}e^{ikt}; \qquad (52)\end{aligned}$$

where (52) follows by rewriting in terms of a convergent geometric series. Pulling the factor of e^{it} inside the series yields

$$\begin{aligned} m_{\alpha,\beta,\theta}(e^{it})ie^{it} &= -ie^{i\beta t}\sum_{k=0}^{\infty}\alpha^{-(k+1)}e^{i(k+1)t}, \\ &= -ie^{i\beta t}\sum_{k=1}^{\infty}\alpha^{-k}e^{ikt}, \\ &= -i\sum_{k=1}^{\infty}\alpha^{-k}e^{i(\beta+k)t}. \qquad (53)\end{aligned}$$

For a fixed $|\alpha| > 1$ we have that $|\alpha|^{-1} < 1$, so

$$\left|\sum_{k=1}^{\infty}\alpha^{-k}e^{ikt}\right| \leq \sum_{k=1}^{\infty}|\alpha^{-k}e^{ikt}| = \sum_{k=1}^{\infty}|\alpha|^{-k} =: K_\alpha < \infty.$$

We define a sequence of functions (f_n), where $f_n : [\theta-2\pi,\theta] \to \mathbb{C}$ are given by

$$f_n(t) := \sum_{k=1}^{n}\alpha^{-k}e^{i(\beta+k)t} = e^{i\beta t}\sum_{k=1}^{n}\alpha^{-k}e^{ikt}.$$

Each function in the sequence is bounded via

$$\begin{aligned}|f_n(t)| &= \left|e^{i\beta t}\sum_{k=1}^{n}\alpha^{-k}e^{ikt}\right|, \\ &\leq K_\alpha|e^{i\beta t}|, \\ &= K_\alpha e^{-\Im(\beta)t} =: g_\alpha(t).\end{aligned}$$

Note that g_α is integrable on $[\theta-2\pi+\varepsilon,\theta-\varepsilon]$ since it is simply a scaled exponential. It is clear that each f_n is continuous as a finite sum of analytic functions, so again this continuity means the functions are measurable. Since their pointwise limit is the expression in (53), Lebesgue's Dominated Convergence Theorem (see §7, Theorem 2 specifically) implies

$$\begin{aligned}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}\sum_{k=1}^{\infty}\alpha^{-k}e^{i(\beta+k)t}\,dt &= \lim_{n\to\infty}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}\sum_{k=1}^{n}\alpha^{-k}e^{i(\beta+k)t}\,dt,\\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}\alpha^{-k}e^{i(\beta+k)t}\,dt,\\ &= \sum_{k=1}^{\infty}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}\alpha^{-k}e^{i(\beta+k)t}\,dt, \qquad (54)\end{aligned}$$

where the second equality holds since it is merely the interchange of an integral and finite sum.

Using (54) along with (53) yields

$$\begin{aligned}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon} m_{\alpha,\beta,\theta}(e^{it})ie^{it}\,dt &= -i\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}\sum_{k=1}^{\infty}\alpha^{-k}e^{i(\beta+k)t}\,dt,\\ &= -i\sum_{k=1}^{\infty}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}\alpha^{-k}e^{i(\beta+k)t}\,dt,\\ &= -i\sum_{k=1}^{\infty}\alpha^{-k}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}e^{i(\beta+k)t}\,dt. \qquad (55)\end{aligned}$$

An individual summand of (55) consists of an α^{-k} term multiplied by an integral. The integrand, $e^{i(\beta+k)t}$, is an entire function of t and thus is bounded on the compact set $t\in[\theta-2\pi,\theta]$ by some M. Note that this bound M can be chosen independent of k by letting

$$M > e^{-\Im(\beta)t} = |e^{i(\beta+k)t}| \qquad \forall t\in[\theta-2\pi,\theta].$$

The length of the curve being integrated over is at most

$$(\theta-\varepsilon)-(\theta-2\pi+\varepsilon) = 2\pi-2\varepsilon < 2\pi =: L,$$

where L does not depend on ε. Because the integrand is entire it must be continuous on the path of integration, and so the ML-bound gives that

$$\left|\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}e^{i(\beta+k)t}\,dt\right| \le ML,$$

where M and L are given above and independent of ε and k. Thus each term of the series in (55) is bounded in modulus by $ML|\alpha|^{-k}$, so that

$$\left|\sum_{k=1}^{\infty}\alpha^{-k}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon}e^{i(\beta+k)t}\,dt\right| \le \sum_{k=1}^{\infty}ML|\alpha|^{-k}. \qquad (56)$$

Since $|\alpha|>1$, the right hand side of (56) converges, and the Weierstrass M-test implies that the series in (55) is uniformly convergent. Substituting the expression from (55) back into (51) allows the exchange of limit and infinite sum in (57) to find that

$$\begin{aligned}\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta}(z)\,dz &= \lim_{\varepsilon\to 0^+}\left[-i\sum_{k=1}^{\infty}\alpha^{-k}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon} e^{i(\beta+k)t}\,dt\right],\\ &= -i\sum_{k=1}^{\infty}\alpha^{-k}\lim_{\varepsilon\to 0^+}\int_{\theta-2\pi+\varepsilon}^{\theta-\varepsilon} e^{i(\beta+k)t}\,dt, \qquad (57)\\ &= -i\sum_{k=1}^{\infty}\alpha^{-k}\int_{\theta-2\pi}^{\theta} e^{i(\beta+k)t}\,dt. \qquad (58)\end{aligned}$$

The integrand in (58) is entire, and it has an antiderivative $\frac{e^{i(\beta+k)t}}{i(\beta+k)}$ when $\beta+k\neq 0$; this antiderivative is also entire. This fact not only ensures the equality between (57) and (58), but it also allows the use of the Complex Fundamental Theorem of Calculus to conclude that for $\beta\notin\mathbb{Z}_{<0}$,

$$\int_{\theta-2\pi}^{\theta} e^{i(\beta+k)t}\,dt = \left[\frac{e^{i(\beta+k)t}}{i(\beta+k)}\right]_{\theta-2\pi}^{\theta} = \frac{e^{i(\beta+k)\theta}}{i(\beta+k)}\left(1-e^{-2\pi i\beta}\right). \qquad (59)$$

Using definition (46) as well as our intermediates (58) and (59) we find

$$\begin{aligned}\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta}(z)\,dz &= -i\sum_{k=1}^{\infty}\alpha^{-k}\frac{e^{i(\beta+k)\theta}}{i(\beta+k)}\left(1-e^{-2\pi i\beta}\right),\\ &= -e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\sum_{k=1}^{\infty}\alpha^{-k}\frac{e^{ik\theta}}{\beta+k},\\ &= -e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\frac{1}{\beta}\left[\left(\sum_{k=0}^{\infty}\frac{\beta}{\beta+k}(\alpha^{-1}e^{i\theta})^k\right)-1\right],\\ &= -e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\frac{1}{\beta}\left[\left(\sum_{k=0}^{\infty}\frac{(1)_k(\beta)_k}{(1+\beta)_k k!}(\alpha^{-1}e^{i\theta})^k\right)-1\right],\\ &= -e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\frac{1}{\beta}\left[{}_2F_1(1,\beta;1+\beta;\alpha^{-1}e^{i\theta})-1\right],\\ &= e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\frac{1}{\beta}\left[1-{}_2F_1(1,\beta;1+\beta;\alpha^{-1}e^{i\theta})\right], \qquad (60)\end{aligned}$$

so long as $\beta\neq 0$. This completes the proof of Theorem 1 in the case where $\beta\in\mathbb{C}\setminus\mathbb{Z}_{\leq 0}$. To handle the cases when $\beta\in\mathbb{Z}_{\leq 0}$, note that

$$\int_{\theta-2\pi}^{\theta} e^{i(\beta+k)t}\,dt = \begin{cases}2\pi & \beta+k=0,\\ 0 & \beta+k\in\mathbb{Z}\setminus\{0\},\end{cases} \qquad (61)$$

since the bounds of integration align with the period of the exponential unless the exponent is 0. Thus when $\beta\in\mathbb{Z}_{\leq 0}$,

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta}(z)\,dz = -i\sum_{k=1}^{\infty}\alpha^{-k}2\pi\delta_{\beta,-k} \qquad (62)$$

where $\delta_{\beta,-k}$ is the classical Kronecker delta function. This completes the proof of Theorem 1 in the case where $\beta\in\mathbb{Z}_{\leq 0}$, the first part of the main result.

Next consider $|\alpha|<1$. We proceed in a manner analogous to that of the $|\alpha|>1$ case, omitting details for the sake of brevity. It holds that

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = \begin{cases} e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\sum_{k=0}^{\infty}\frac{\alpha^k}{\beta-k}e^{-ik\theta} & \beta\notin\mathbb{Z}_{\geq 0}, \\ 2\pi i\alpha^{\beta} & \beta\in\mathbb{Z}_{\geq 0}. \end{cases} \tag{63}$$

An application of (46) shows that for $\beta\neq 0$,

$$\begin{aligned}\sum_{k=0}^{\infty}\frac{\alpha^k}{\beta-k}e^{-ik\theta} &= \frac{1}{\beta}\sum_{k=0}^{\infty}\frac{-\beta}{-\beta+k}\left(\alpha e^{-i\theta}\right)^k, \\ &= \frac{1}{\beta}\sum_{k=0}^{\infty}\frac{(1)_k(-\beta)_k}{(1-\beta)_k k!}\left(\alpha e^{-i\theta}\right)^k, \\ &= \frac{1}{\beta}\,{}_2F_1(1,-\beta;1-\beta;\alpha e^{-i\theta}). \end{aligned} \tag{64}$$

Combining (63) and (64) completes the proof of Theorem 1.

4.2 Reconciling Methods

Note that the steps of the contour method described in §2 and the simplifications in §3.1 still hold so long as $\mathrm{Arg}(\alpha)\neq\theta$, $\Re(\beta)>0$, $\theta\neq 0 \pmod{2\pi}$, and $|\alpha|\neq 0,1$. The nontrivial equations of the series method hold so long as $|\alpha|\neq 1$ and $\beta\notin\mathbb{Z}_{\geq 0}$. Thus under all these conditions one can write an identity for (45) in the case where $0<|\alpha|<1$:

$$e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\frac{1}{\beta}\,{}_2F_1(1,-\beta;1-\beta;\alpha e^{-i\theta}) = 2\pi i\alpha^{\beta} + e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)\int_0^1\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt, \tag{65}$$

$$\frac{1}{\beta}\,{}_2F_1(1,-\beta;1-\beta;\alpha e^{-i\theta}) = \frac{2\pi i\alpha^{\beta}}{e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)} + \int_0^1\frac{e^{\beta\ln t}}{t-\alpha e^{-i\theta}}\,dt,$$

$$\frac{1}{\beta}\,{}_2F_1(1,-\beta;1-\beta;\alpha e^{-i\theta}) = \frac{2\pi i\alpha^{\beta}}{e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)} + \frac{1}{\beta} - \int_0^1\frac{e^{(\beta-1)\ln t}}{1-\left(\frac{1}{\alpha}e^{i\theta}\right)t}\,dt, \tag{66}$$

$$\int_0^1\frac{e^{(\beta-1)\ln t}}{1-\left(\frac{1}{\alpha}e^{i\theta}\right)t}\,dt = \frac{2\pi i\alpha^{\beta}}{e^{i\beta\theta}\left(1-e^{-2\pi i\beta}\right)} + \frac{1}{\beta}\left[1-{}_2F_1(1,-\beta;1-\beta;\alpha e^{-i\theta})\right] \tag{67}$$

where the equality in (65) follows from Lemma 1 and (64), and the equality in (66) follows from (44).

5 Computing the Example $\beta=\frac{m}{n}$

Since the hypergeometric function gives the value of (1) as an infinite series which is still difficult to explicitly evaluate, it is desirable to compute examples for which the hypergeometric function can be simplified more. We show this is the case when $\beta=\frac{m}{n}\in\mathbb{Q}\setminus\mathbb{Z}$, with $m\in\mathbb{Z}$, $n\in\mathbb{N}$. We ignore the cases $\beta\in\mathbb{Z}$ since these are easily evaluated without need of the hypergeometric function.

Corollary 1. *Let $\beta = \frac{m}{n} \in \mathbb{Q}\setminus\mathbb{Z}$. When $|\alpha| < 1$,*

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = \alpha^{-\frac{m}{n}}(1-e^{-2\pi i\frac{m}{n}})\sum_{j=0}^{n-1} e^{\frac{2\pi ijm}{n}}\log\left(1-e^{\frac{2\pi ij}{n}}\sqrt[n]{\alpha e^{-i\theta}}\right)$$

and when $|\alpha| > 1$

$$\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta} = (1-e^{-2\pi i\frac{m}{n}})\left(\frac{n}{m}e^{i\frac{m}{n}\theta}+\alpha^{\frac{m}{n}}\sum_{j=0}^{n-1} e^{-\frac{2\pi ijm}{n}}\log\left(1-e^{\frac{2\pi ij}{n}}\sqrt[n]{\alpha^{-1}e^{i\theta}}\right)\right).$$

Proof. From Theorem 1, one has for $\beta \notin \mathbb{Z}$ that

$$\int_{\partial\mathbb{D}} m_{\alpha,\frac{m}{n},\theta} = \begin{cases} e^{i\frac{m}{n}\theta}\left(1-e^{-2\pi i\frac{m}{n}}\right)\frac{n}{m}\,{}_2F_1(1,-\frac{m}{n};1-\frac{m}{n};\alpha e^{-i\theta}) & |\alpha| < 1, \\ e^{i\frac{m}{n}\theta}\left(1-e^{-2\pi i\frac{m}{n}}\right)\frac{n}{m}\left[1-{}_2F_1(1,+\frac{m}{n};1+\frac{m}{n};\alpha^{-1}e^{i\theta})\right] & |\alpha| > 1. \end{cases} \tag{68}$$

Therefore the main difficulty in evaluating (68) lies in computing

$${}_2F_1\left(1,\frac{m}{n};1+\frac{m}{n};z\right) \tag{69}$$

for non-integral $\frac{m}{n}$ and for $0 < |z| < 1$.

Since $\frac{m}{n} \notin \mathbb{Z}$, it is never the case that the parameter $c = 1+\frac{m}{n}$ in (69) is 0 or a negative integer. The hypergeometric series is therefore well defined, and using the definition (46) yields

$$\begin{aligned} {}_2F_1\left(1,\frac{m}{n};1+\frac{m}{n};z\right) &:= \sum_{k=0}^{\infty}\frac{(1)_k(\frac{m}{n})_k}{k!(1+\frac{m}{n})_k}z^k. \\ &= \sum_{k=0}^{\infty}\frac{\frac{m}{n}}{k+\frac{m}{n}}z^k \\ &= m\sum_{k=0}^{\infty}\frac{1}{m+nk}z^k \\ &= \frac{m}{z^{\frac{m}{n}}}\sum_{k=0}^{\infty}\frac{1}{m+nk}z^{\frac{m}{n}+k} =: G(z). \end{aligned} \tag{70}$$

Note also that since $0 < |z|$, division by a fractional power of z causes no issue. The particular choice of branch for defining the n^{th} root does not matter so long as the choice is consistent across the fractional powers (see remarks 1 and 2).

Notice that the expression for $G(z)$ produces an even simpler expression for $G(z^n)$, given by

$$G(z^n) = \frac{m}{z^m}\sum_{k=0}^{\infty}\frac{1}{m+nk}z^{m+nk} \tag{71}$$

On the other hand, for any branch of the logarithm with $\log(1) = 0$ analytic in a ball of radius 1 at $z = 1$, we have

$$\log(1-z) = \sum_{k=1}^{\infty} -\frac{1}{k}z^k$$

whenever $|z| < 1$. The difference between the above expression and that in (71) is that only terms of the form z^{m+nk} for $k \in \mathbb{N}\cup\{0\}$ appear in (71), whereas a z^k term appears in the series for $\log(1-z)$ for every $k \in \mathbb{N}$. To rectify this we express $G(z^n)$ as some series $\sum_{s=1}^{\infty} \frac{\delta_s}{s} z^s$, possibly with leading factors, where δ_s takes on the value 1 whenever n divides $s-m$ (so that s is of the form $m+nk$ for some $k \in \mathbb{N}$) and is 0 otherwise.

To find a suitable function δ_s, recall that the sum of all of the n^{th} roots of unity is 0 for $n > 1$. It is natural then that

$$\delta_s = \frac{1}{n} \sum_{j=0}^{n-1} e^{\frac{2\pi i j(s-m)}{n}} = \begin{cases} 1 & \text{if } n \mid (s-m) \\ 0 & \text{otherwise} \end{cases} \tag{72}$$

is the desired function. To see the validity of this claim, we first consider when $n \mid (s-m)$. In this case, we have $\frac{s-m}{n} \in \mathbb{Z}$, and so

$$\frac{1}{n} \sum_{j=0}^{n-1} e^{\frac{2\pi i j(s-m)}{n}} = \frac{1}{n} \sum_{j=0}^{n-1} 1 = 1 \tag{73}$$

since $\frac{j(s-m)}{n} \in \mathbb{Z}$. On the other hand, when $n \nmid (s-m)$, let $d := \gcd(n, s-m)$ and define $\eta := \frac{n}{d}$. Note that $d < n$ else we have $n \mid (s-m)$, and thus $\eta > 1$. Now

$$e^{\frac{2\pi i(s-m)}{n}} = \zeta_\eta^{\frac{s-m}{d}} \tag{74}$$

where ζ_η is the first primitive η^{th} root of unity. Note also that because d is the greatest common divisor of $s-m$ and n, then $\frac{s-m}{d}$ is coprime to η. This implies that $\zeta_\eta^{\frac{s-m}{d}}$ is another primitive η^{th} root of unity. Now

$$\begin{aligned}
\delta_s &= \frac{1}{n} \sum_{j=0}^{n-1} e^{\frac{2\pi i j(s-m)}{n}}, \\
&= \frac{1}{d\eta} \sum_{j=0}^{d\eta-1} \zeta_\eta^{\frac{s-m}{d} j}, \\
&= \frac{1}{d\eta} \left(\sum_{j=0}^{\eta-1} \zeta_\eta^{\frac{s-m}{d} j} + \sum_{j=\eta}^{2\eta-1} \zeta_\eta^{\frac{s-m}{d} j} + \cdots + \sum_{j=(d-1)\eta}^{d\eta-1} \zeta_\eta^{\frac{s-m}{d} j} \right), && (75) \\
&= \frac{1}{\eta} \sum_{j=0}^{\eta-1} \zeta_\eta^{\frac{s-m}{d} j}, && (76) \\
&= 0. && (77)
\end{aligned}$$

The equality between (75) and (76) holds since $\zeta_\eta^{\frac{s-m}{d} j} = \zeta_\eta^{\frac{s-m}{d} j'}$ when $j \equiv j' \pmod{\eta}$. The final equality, (77), holds since the sum over every η^{th} root of 1 is 0.

One can therefore express

$$G(z^n) = \frac{m}{z^m} \sum_{s=1}^{\infty} \frac{\delta_s}{s} z^s \tag{78}$$

since this series gives the same terms as the series in (71). To evaluate this series, note that for each j in $\{0, \ldots, n-1\}$, the series $\sum_{s=1}^{\infty} \left| \frac{e^{\frac{2\pi i j(s-m)}{n}}}{s} \right| z^s$ is a power series which converges absolutely for $|z| < 1$. Letting the value of this series be denoted b_j, we also note that $\sum_{j=0}^{n-1} b_j$ converges since it is a finite sum. We may therefore exchange the order of summation to get

$$\sum_{j=0}^{n-1}\sum_{s=1}^{\infty}\frac{e^{\frac{2\pi ij(s-m)}{n}}}{s}z^s=\sum_{s=1}^{\infty}\sum_{j=0}^{n-1}\frac{e^{\frac{2\pi ij(s-m)}{n}}}{s}z^s=G(z^n) \tag{79}$$

which allows evaluation of $G(z^n)$ by simplifying the left hand side of (79):

$$\begin{aligned}G(z^n)=\frac{m}{z^m}\sum_{j=0}^{n-1}\sum_{s=1}^{\infty}\frac{e^{\frac{2\pi ij(s-m)}{n}}}{s}z^s &= \frac{m}{z^m}\sum_{j=0}^{n-1}e^{-\frac{2\pi ijm}{n}}\sum_{s=1}^{\infty}\frac{e^{\frac{2\pi ijs}{n}}}{s}z^s,\\ &= \frac{m}{z^m}\sum_{j=0}^{n-1}-e^{-\frac{2\pi ijm}{n}}\sum_{s=1}^{\infty}-\frac{1}{s}(e^{\frac{2\pi ij}{n}}z)^s,\\ &= \frac{m}{z^m}\sum_{j=0}^{n-1}-e^{-\frac{2\pi ijm}{n}}\log(1-e^{\frac{2\pi ij}{n}}z). \end{aligned}\tag{80}$$

Remark 1. Notice that the only requirement of the branch of log we choose is that it is analytic in the ball of radius 1 at $z=1$, and that $\log(1)=0$.

Finally, to come up with an expression for $G(z)$ as opposed to $G(z^n)$, simply substitute $z^{\frac{1}{n}}$ in the expression above, yielding

$$G(z)=-\frac{m}{n}z^{-\frac{m}{n}}\sum_{j=0}^{n-1}e^{-\frac{2\pi ijm}{n}}\log(1-e^{\frac{2\pi ij}{n}}\sqrt[n]{z}) \tag{81}$$

whenever $|z|<1$.

Remark 2. The choice of branch for $\sqrt[n]{\cdot}$ does not matter, so long as the choice is consistent across the expression for $G(z)$. To see this more clearly, rewrite

$$G(z)=-\frac{m}{n}\sum_{j=0}^{n-1}\left(e^{-\frac{2\pi ij}{n}}\frac{1}{\sqrt[n]{z}}\right)^m\log\left(1-e^{\frac{2\pi ij}{n}}\sqrt[n]{z}\right). \tag{82}$$

This sum is symmetric over the n^{th} roots of z. Any branch of $\sqrt[n]{\cdot}$ must map an input z to one of the n possible roots ω of $\omega^n=z$. The symmetry in (82) implies that, no matter the branch chosen, this sum will always have the same terms.

From (70) we know that $G(z)={}_2F_1\left(1,\frac{m}{n};1+\frac{m}{n};z\right)$, and hence (81) allows us to conclude that for $\beta=\frac{m}{n}\in\mathbb{Q}\setminus\mathbb{Z}$ with $m\in\mathbb{Z}$, $n\in\mathbb{N}$,

$${}_2F_1\left(1,\frac{m}{n};1+\frac{m}{n};z\right)=-\frac{m}{n}z^{-\frac{m}{n}}\sum_{j=0}^{n-1}e^{-\frac{2\pi ijm}{n}}\log(1-e^{\frac{2\pi ij}{n}}\sqrt[n]{z}) \tag{83}$$

for all $|z|<1$. Finally, when $|\alpha|<1$ we have $|\alpha e^{-i\theta}|<1$, so substituting $z=\alpha e^{-i\theta}$ into (83) proves the first conclusion of Corollary 1. Similarly, when $|\alpha|>1$, we have that $|\frac{1}{\alpha}e^{i\theta}|<1$, and hence substituting $z=\alpha^{-1}e^{i\theta}$ into (83) proves the second conclusion. □

6 Differential Equation

A key feature of the hypergeometric equation

$$z(1-z)\frac{d^2F}{dz^2}+(c-(a+b+1)z)\frac{dF}{dz}-abF=0 \tag{84}$$

is its regular singular points, and it is well-known that they are $0,1,\infty$. Hence, one might wish to derive a second-order ordinary differential equation in the variable α for which $I(\alpha)=\int_{\partial\mathbb{D}} m_{\alpha,\beta,\theta}$ is a solution, and determine its regular singular points.

6.1 The case $|\alpha|>1$

For $|\alpha|>1$, $\beta\notin\mathbb{Z}_{\leq 0}$, the desired equation follows by relating (1) to the hypergeometric function $F(z)={}_2F_1(a,b,c;z)$, which is famously a solution of (84).
Let $f(z)={}_2F_1(1,\beta,1+\beta;z)$. Then f solves the equation

$$z(z-1)\frac{d^2f}{dz^2}+((1+\beta)-(2+\beta)z)\frac{df}{dz}-\beta f=0. \tag{85}$$

Consider the change in variables $\alpha=\frac{1}{z}e^{i\theta}$ (equiv. $z=\frac{1}{\alpha}e^{i\theta}$), and make the following necessary calculations:

$$\begin{aligned}\frac{df}{dz}&=\frac{d\alpha}{dz}\frac{df}{d\alpha}=-\frac{1}{z^2}e^{i\theta}\frac{df}{d\alpha}=-\alpha^2e^{-i\theta}\frac{df}{d\alpha},\\ \frac{d^2f}{dz^2}&=\frac{d\alpha}{dz}\cdot\frac{d}{d\alpha}\frac{df}{dz}\\ &=-\alpha^2e^{-i\theta}\left(-\alpha^2e^{-i\theta}\frac{d^2f}{d\alpha^2}-2\alpha e^{-i\theta}\frac{df}{d\alpha}\right)\\ &=\alpha^4e^{-2i\theta}\frac{d^2f}{d\alpha^2}+2\alpha^3e^{-2i\theta}\frac{df}{d\alpha}.\end{aligned}$$

By substituting into (85), notice that $f_*(\alpha):=f\left(\frac{1}{\alpha}e^{i\theta}\right)$ solves

$$\alpha^{-1}e^{i\theta}\left(\alpha^{-1}e^{i\theta}-1\right)\left(\alpha^4e^{-2i\theta}\frac{d^2f_*}{d\alpha^2}+2\alpha^3e^{-2i\theta}\frac{df_*}{d\alpha}\right)+\left((1+\beta)-(2+\beta)(\alpha^{-1}e^{i\theta})\right)\left(-\alpha^2e^{-i\theta}\frac{df_*}{d\alpha}\right)-\beta f_*=0,$$

which after some simplification becomes

$$p_2(\alpha)\frac{d^2f_*}{d\alpha^2}+p_1(\alpha)\frac{df_*}{d\alpha}-\beta f_*=0, \tag{86}$$

where $p_2(\alpha)=\alpha^2-\alpha^3e^{-i\theta}, p_1(\alpha)=\alpha(\beta+4)-\alpha^2(\beta+3)e^{-i\theta}$.
From Theorem 1, $f_*(\alpha)=1-kI(\alpha)$, with the abbreviation $k=\frac{\beta}{e^{i\beta\theta}(1-e^{-2\pi i\beta})}$, and we calculate the derivatives to be

$$\begin{aligned}\frac{df_*}{d\alpha}&=\frac{df_*}{dI}\frac{dI}{d\alpha}=-k\frac{dI}{d\alpha},\\ \frac{d^2f_*}{d\alpha^2}&=-k\frac{d^2I}{d\alpha^2}.\end{aligned}$$

Substitution into (86) yields that $I(\alpha)$ solves the equation

$$p_2(\alpha)\left(-k\frac{d^2I}{d\alpha^2}\right)+p_1(\alpha)\left(-k\frac{dI}{d\alpha}\right)-\beta\left(1-kI\right)=0,$$

or rather,

$$p_2(\alpha)\frac{d^2I}{d\alpha^2}+p_1(\alpha)\frac{dI}{d\alpha}-\beta I=e^{i\beta\theta}(e^{-2\pi i\beta}-1). \tag{87}$$

From (87) it is clear that the normalized coefficients $\frac{p_1(\alpha)}{p_2(\alpha)}\alpha$ and $\frac{-\beta}{p_2(\alpha)}\alpha^2$ are analytic in a neighborhood of 0. Similarly, $\frac{p_1(\alpha)}{p_2(\alpha)}(\alpha-e^{i\theta})$ and $\frac{-\beta}{p_2(\alpha)}(\alpha-e^{i\theta})^2$ are analytic in a neighborhood of $e^{i\theta}$. These coefficients have poles at 0 and $e^{i\theta}$, so these are regular singular points. To classify the point at infinity, let $x=1/\alpha$ and rewrite (87) in x. Akin to a previous change of variables, one has

$$\frac{dI}{d\alpha}=-x^2\frac{dI}{dx},$$
$$\frac{d^2I}{d\alpha^2}=x^4\frac{d^2I}{dx^2}+2x^3\frac{dI}{dx},$$

so that (87) becomes

$$p_2\left(\frac{1}{x}\right)\left(x^4\frac{d^2I}{dx^2}+2x^3\frac{dI}{dx}\right)+p_1\left(\frac{1}{x}\right)\left(-x^2\frac{dI}{dx}\right)-\beta I=e^{i\beta\theta}(e^{-2\pi i\beta}-1),$$

or equivalently,

$$q_2(x)\frac{d^2I}{dx^2}+q_1(x)\frac{dI}{dx}-\beta I=e^{i\beta\theta}(e^{-2\pi i\beta}-1), \tag{88}$$

where $q_2(x)=x^2-xe^{-i\theta}$, $q_1(x)=-x(\beta+2)+(\beta+1)e^{-i\theta}$. By a similar line of reasoning, the regular singular points of (88) are $x=0$ and $x=e^{-i\theta}$, so $\alpha=\infty$ and $\alpha=e^{i\theta}$ are both regular singular points of (87).

Thus equation (87), for which (1) is a solution, has precisely three regular singular points at $0,e^{i\theta},\infty$, reminiscent of (84). Any function satisfying a differential equation with three regular singular points may be expressed using the hypergeometric function, so this result supports the validity of the relationship derived.

6.2 The case $|\alpha|<1$

When $|\alpha|<1$, $\beta\notin\mathbb{Z}_{\geq 0}$, one can proceed exactly as §6.1 and make use of Theorem 1. Let $g(z)={}_2F_1(1,-\beta,1-\beta,z)$. Then g solves the equation

$$z(z-1)\frac{d^2g}{dz^2}+((1-\beta)-(2-\beta)z)\frac{dg}{dz}+\beta g=0.$$

The change of variables $\alpha=ze^{i\theta}$ gives that $g_*(\alpha):=g(\alpha e^{-i\theta})$ solves

$$r_2(\alpha)\frac{d^2g_*}{d\alpha^2}+r_1(\alpha)\frac{dg_*}{d\alpha}+\beta g_*=0, \tag{89}$$

where $r_2(\alpha)=\alpha^2-\alpha e^{i\theta}$, $r_1(\alpha)=(1-\beta)e^{i\theta}-(2-\beta)\alpha$. Theorem 1 gives $g_*(\alpha)=kI(\alpha)$, where again $k=\frac{\beta}{e^{i\beta\theta}(1-e^{-2\pi i\beta})}$. This scaling does not change the equation, so $I(\alpha)$ is also a solution of (89), with g_* replaced with I. Just as before, one reasons that $\alpha=0,e^{i\theta}$ are regular singular points of this equation. In the variable $x=\frac{1}{\alpha}$, the equation (89) written for I is

$$s_2(x)\frac{d^2I}{dx^2}+s_1(x)\frac{dI}{dx}+\beta I=0,$$

where $s_2(x) = x^2 - x^3 e^{i\theta}, s_1(x) = \beta x - (1+\beta)e^{i\theta}x^2$, of which the regular singular points are $x = 0, e^{-i\theta}$. Finally, one concludes that the regular singular points of the hypergeometric-like differential equation that $I(\alpha)$ solves are $0, e^{i\theta}, \infty$.

7 Appendix

In order to use Lebesgue's Dominated Convergence Theorem, which is essential to the proofs in the paper, a short introduction to basic measure theory is needed. Careful treatment of the necessary measure theory topics is handled in Chapter 11 of Rudin's *Principles of Mathematical Analysis* (pgs. 300-315 of [3]). For more resources on the topic see the reviews in [4]. If the reader chooses to forego the short introduction to measure theory, it will at least help to understand the advantages of Lebesgue integration over the standard Riemann integration taught in introductory calculus.

For one, the Lebesgue integral extends to a much wider class of functions than the Riemann integral. A classic example is that of the Dirichlet function δ which takes value 1 on rationals and 0 on irrationals. This function is not Riemann integrable on the interval $[0,1]$, since no matter what partition P we pick for the interval, at least one irrational and one rational must lie in each interval of the partition. Thus for every partition P, $U(P,\delta) = 1$ and $L(P,\delta) = 0$, where U and L are the upper and lower sums of δ with partition P, respectively. This shows that δ is not Riemann integrable.

The Dirichlet function *is* Lebesgue integrable however, and its Lebesgue integral over the interval $[0,1]$ is simple to compute. Let λ be the Lebesgue measure on $\mathbb{R}$. By definition of the Lebesgue integral, we have

$$\int_{[0,1]} \delta \, d\lambda = 1 \cdot \lambda([0,1] \cap \mathbb{Q}) + 0 \cdot \lambda([0,1] \setminus \mathbb{Q}) = 0$$

since the rationals form a Lebesgue measure 0 subset of the reals. This example shows in particular that a "large" number of discontinuities (think uncountably many, as in the Dirichlet function) does not necessarily prevent a function from being Lebesgue integrable, while it does prevent it from being Riemann integrable.

Another limitation of Riemann integration is the difficulty in passing a limit under the integral sign. Given a sequence of Riemann-integrable functions $\{f_n\}$ which converge to some function f, it would be convenient if

$$\lim_{n\to\infty} \int_a^b f_n(x)\, dx = \int_a^b \lim_{n\to\infty} f_n(x)\, dx.$$

Indeed, there are cases where this holds – for instance, if $\{f_n\}$ converges to a function f *uniformly* (rather than just pointwise) on the finite interval $[a,b]$. This fact can be seen as a consequence of Lebesgue's Dominated Convergence Theorem, again demonstrating its utility (see the comment under Exercise 11.6 in [3]). It is mainly for this reason we use the integral of Lebesgue rather than that of Riemann, since in the Lebesgue context it is much easier to justify switching the limit and integral.

The reader should also notice that we discuss measurability and integrability of functions with codomain $\mathbb{C}$ rather than $\mathbb{R}$ (see pg. 325 of [3]). Recall that $f : \mathbb{R} \to \mathbb{C}$ may be written in terms of two component functions $u, v : \mathbb{R} \to \mathbb{R}$

$$f(t) = u(t) + iv(t),$$

and the integral of f over $A \subseteq \mathbb{R}$ is aptly defined

$$\int_A f(t)\, d\lambda := \int_A u(t)\, d\lambda + i \int_A v(t)\, d\lambda.$$

It is natural then that the function $f : \mathbb{R} \to \mathbb{C}$ be integrable over A so long as u and v are. This also corroborates the definition that $f : \mathbb{R} \to \mathbb{C}$ is measurable if u and v are.

The need to discuss complex valued functions stems from our use of line integrals (pgs. 101-102 in [5]). Let $\Omega \subseteq \mathbb{C}$, and recall that the integral of the function $f : \Omega \to \mathbb{C}$ over a piecewise differentiable arc γ parameterized by $z : [a,b] \to \mathbb{C}$ is defined as

$$\int_\gamma f(z)\, dz := \int_a^b f(z(t))z'(t)\, dt.$$

Defining $u_\gamma, v_\gamma : [a,b] \to \mathbb{R}$ such that

$$[(f \circ z)z'](t) = u_\gamma(t) + iv_\gamma(t),$$

the integral of f over γ can be written terms of two real integrals:

$$\int_\gamma f(z)\, dz = \int_a^b u_\gamma(t)\, dt + i \int_a^b v_\gamma(t)\, dt.$$

If f is continuous on the piecewise-continuously differentiable curve γ, then certainly $(f \circ z)z' : [a,b] \to \mathbb{C}$ is piecewise continuous and bounded. From here one concludes that u_γ, v_γ are also piecewise continuous, bounded functions; the component functions are in fact Riemann integrable. It is known that if a function $g : \mathbb{R} \to \mathbb{R}$ is Riemann integrable on $[a,b]$, then g is also Lebesgue integrable on $[a,b]$ and the two separate notions of integration yield the exact same result (Theorem 11.33 [3]). In this paper, every claim of integrability is justified through this sense.

Now we state Lebesgue's Dominated Convergence Theorem (Theorem 11.32 in [3]).

Theorem 2. *Suppose A is a measurable set with respect to some measure μ, and let $\{f_n\}$ be a sequence of measurable functions (with respect to the same measure) such that*

$$f_n(x) \to f(x) \quad \text{as} \quad n \to \infty \quad \forall x \in A.$$

If there exists a μ-integrable function g on A such that

$$|f_n(x)| \leq g(x) \quad \forall n \in \mathbb{N}, \forall x \in A,$$

then

$$\lim_{n\to\infty} \int_A f_n\, d\mu = \int_A f\, d\mu.$$

Remark 3. In our case, the sets we integrate over are always intervals. Since we choose $\mu = \lambda$ when applying Theorem 2, and since intervals are Lebesgue measurable, the first condition of Theorem 2 holds easily.

In the mainline discussion, we sweep under the rug the application of Theorem 2 to separate real and imaginary components. However it should be clear that the arguments laid out there justify the separate applications — the following points support this:

(a) a sequence of functions $f_n : \mathbb{R} \to \mathbb{C}$ are measurable in our sense, then by definition one had to have shown the component sequences u_n, v_n are measurable.

(b) $f_n \to f$ pointwise, then also the components converge $u_n \to u$ and $v_n \to v$ pointwise.

(c) $|f_n| \leq g$, then $|u_n|, |v_n| \leq |f_n| \leq g$.

Acknowledgements

The authors sincerely thank Professor Mihai Putinar for encouraging our work on the problem and for providing helpful discussions and edits.

Bibliography

[1] R. Mortini and R. Rupp. The Cauchy Transform of the Square Root Function on the Circle. Complex Analysis and Operator Theory, 2022.

[2] A. Erdélyi. Higher Transcendental Functions. McGraw-Hill, New York, NY, 1953.

[3] W. Rudin. Principles of Mathematical Analysis. McGraw-Hill, New York, NY, 1976.

[4] J. C. Oxtoby. Book review: Measure theory. Bullet in of the American Mathematical Society, 1953.

[5] Lars V. Ahlfors. Complex Analysis. McGraw-Hill (India), Chennai, Tamil Nadu, 1979.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 17, No. 1 (Fall 2023)
Pages 118 – 133

Matricial Frameworks for the Mandelbrot and Filled Julia Sets

*Eric Babcock, Dawson Brindle, Mitch Hamidi, Lara Ismert**

Eric Babcock earned his BS in Software Engineering from Embry-Riddle Aeronautical University in Prescott. He currently works as a Software Engineer at SpaceX working on the network topology team for Starlink. In his free time, he likes playing sports and spending time with his wife and dog.

Dawson Brindle earned a bachelor's degree in aerospace engineering from Embry-Riddle Aeronautical University Prescott in 2023 and is currently working as a specialty test engineer. During his free time, he likes to try making different types of coffee and learning new concepts.

Mitch Hamidi earned his Ph.D. in mathematics from the University of Nebraska-Lincoln in 2019 and is currently an assistant professor of mathematics at Embry-Riddle Aeronautical University in Prescott, Arizona. His research interests lie in operator algebras and quantum information theory. In his free time, he enjoys playing music and hanging out with his wife, 3-month old son, and his four cats.

Lara Ismert earned her Ph.D. in mathematics from the University of Nebraska-Lincoln in 2019 and is currently an assistant professor of mathematics at Embry-Riddle Aeronautical University in Prescott, Arizona. Her research interests lie in operator algebras and quantum information theory. When she finds spare time, she loves performing in musical theatre and spending time with her husband, 3-month old son, and her four cats.

***Corresponding author:** ISMERTL@erau.edu

Abstract

Both the Mandelbrot set and filled Julia sets are subsets in the complex plane derived by studying iterations of complex polynomials. We develop a matricial framework to establish an alternate form of iteration by complex polynomials using a sequence of affine transformations. Using this framework, we are able to check membership in a filled Julia set and the Mandelbrot set by studying boundedness of sequences of matrices. Specifically, we show that a complex number belongs to the Mandelbrot set if and only if a particular sequence of matrices is bounded in the operator norm, and a complex number belongs to a filled Julia set if and only if a particular sequence of matrices is bounded in operator norm.

1 Introduction

The complex plane (denoted $\mathbb{C}$) is comprised of values $x+yi$, where x and y belong to the real numbers (denoted $\mathbb{R}$) and i is the imaginary number $\sqrt{-1}$. We call x the real part of $x+yi$ and y the imaginary part of $x+yi$. When we don't need to display the real and imaginary parts of a complex number, we simply denote it by a single variable, like z, c, or w.

1.1 Visualizations of the Complex Plane

To the unfamiliar eye, the complex plane looks no different than the familiar xy-coordinate plane, denoted $\mathbb{R}^2$. Indeed, a complex number $x+yi$ in $\mathbb{C}$ is spatially in the exact same location as the coordinate pair (x,y) in $\mathbb{R}^2$.

The perspective of vector geometry further supports the notion that these two spaces are not distinct. To do this, we can view a complex number of the from $x+yi$ as a vector with head at the origin which points a value of y in the direction of the vertical axis (imaginary values) and a value of x in the horizontal axis. Simply put, this representation allows for the complex plane to be viewed as a two-dimensional real vector space. Now, let $x_1, y_1, x_2, y_2 \in \mathbb{R}$. If we consider $x_1+y_1i, x_2+y_2i \in \mathbb{C}$ and $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^2$ as 2-dimensional vectors emanating from their respective origins (in $\mathbb{C}$ and $\mathbb{R}^2$), vector addition will yield outputs also in the exact same locations:

$$\text{In } \mathbb{R}^2: \qquad (x_1,y_1)+(x_2,y_2) = (x_1+x_2, y_1+y_2).$$

$$\text{In } \mathbb{C}: \qquad (x_1+y_1i)+(x_2+y_2i) = x_1+x_2+(y_1+y_2)i.$$

In contrast, there is a natural multiplication on $\mathbb{C}$ that $\mathbb{R}^2$ does not have.

$$\begin{aligned}\text{In } \mathbb{C}: (x_1+y_1i)*(x_2+y_2i) &= x_2x_2+x_1y_2i+x_2y_1i+y_1y_2i^2\\ &= x_2x_2+x_1y_2i+x_2y_1i+y_1y_2(-1)\\ &= x_1x_2-y_1y_2+(x_1y_2+x_2y_1)i.\end{aligned}$$

Clearly, coordinate-version of the complex product, $(x_1x_2-y_1y_2, x_1y_2+x_2y_1)$ is not equal to the $\mathbb{R}^2$ product (x_1x_2, y_1y_2). An interesting way to interpret the strange multiplication on $\mathbb{C}$, is that we can also think of elements of $\mathbb{C}$ not only as 2-dimensional vectors, but also as 2×2-matrices with real entries:

$$x + yi \mapsto \begin{bmatrix} x & y \\ -y & x \end{bmatrix}$$

Matrix multiplication appropriately implements the multiplication on $\mathbb{C}$. Observe:

$$\begin{bmatrix} x_1 & y_1 \\ -y_1 & x_1 \end{bmatrix} \begin{bmatrix} x_2 & y_2 \\ -y_2 & x_2 \end{bmatrix} = \begin{bmatrix} x_1x_2 - y_1y_2 & x_1y_2 + x_2y_1 \\ -x_1y_2 - x_2y_1 & x_1x_2 - y_1y_2 \end{bmatrix}$$

Notice that the $(1,1)$-entry of the matrix product above is $x_1x_2 - y_1y_2$, which is the real part of the complex product $x_1x_2 - y_1y_2 + (x_1y_2 + x_2y_1)i$, and the $(1,2)$-entry of the matrix product above is $x_1y_2 + x_2y_1$, which is the imaginary part of the complex product $x_1x_2 - y_1y_2 + (x_1y_2 + x_2y_1)i$.

Functions like $f(z) = z^2 + i$, or more generally, $f(z) = z^2 + c$ for some complex constant c, play a central role in the definitions of the Mandelbrot and filled Julia sets. Throughout the paper, we study particular properties of functions like f, and these properties depend deeply on the constant c. So, instead of denoting the map $z \mapsto z^2 + c$ by just "f," we will denote it by f_c to emphasize the dependence on c.

The function $f_c(z) = z^2 + c$ does two sequential processes to an input w:

$$w \quad \xrightarrow{(1)} \quad w^2 \quad \xrightarrow{(2)} \quad w^2 + c$$

The first process is most easily visualized in yet another representation of complex numbers, called *polar form*. In this perspective, $w = re^{i\theta}$, where $r, \theta \in \mathbb{R}^+$, which is graphed by going a distance r along the positive real axis and then rotating that point an angle of θ counterclockwise. In polar form, $w \mapsto w^2$ is equivalent to $re^{i\theta} \mapsto (re^{i\theta})^2 = r^2e^{i(2\theta)}$. In the latter, the distance r that w was from the origin is squared, and the angle θ that w was from the origin is doubled. The second process sending w^2 to $w^2 + c$ is most easily visualized in the 2-dimensional vector representation of $\mathbb{C}$ as translation of the point w^2 along the vector representation of c.

We want to emphasize here that **neither of these two processes are linear**. Indeed, squaring a number is a *quadratic* transformation, while translating a number is an *affine* transformation. The main results of this paper nonetheless reframe the iterative process applying f_c to w into a *linear* transformation. Our goal in pursuing this unnatural framework is to leverage the robust tools of linear algebra.

1.2 Filled Julia Sets and the Mandelbrot Set

While the debate of who discovered Julia sets and, subsequently, the Mandelbrot set (named after mathematician Benoit Mandelbrot), is ongoing within the mathematical research community [4], there is no debate on the important role Julia sets and the Mandelbrot set play in seemingly disparate scientific fields. Within mathematics, the Mandelbrot set and most Julia sets are examples of *fractals*, which are geometric subsets of the complex plane that have properties similar to the frozen crystals on a snowflake when you zoom in under a microscope. These subsets arise not only in fractal geometry, but also in computer graphics, control theory [11], robotics, and even various methods of encryption. For example, strong QR codes often have a fractal embedded in them, such as a Julia set [7]. The strength of using a Julia set as means of creating a "password" comes from the complexity of drawing the Julia set itself.

Classically, the Mandelbrot set and filled Julia sets are constructed by repeatedly iterating a complex polynomial of the form $f_c(z) = z^2 + c$ at points in the complex plane. Each point iterated is called a *seed*. A seed belongs to a filled Julia set and/or the Mandelbrot set depending on the long run behavior of repeated iterations by f_c – in particular, whether or not repeated iterations form a bounded sequence.

The largest road block in studying these fractals is the computational complexity of this process of iteration. Below is an image of a Julia set created in C^{++} whose pixels are colored by how quickly they escape boundedness. The light green points escape very slowly and border the points inside the filled Julia set who do not escape. The dark points on the outside escape very quickly.

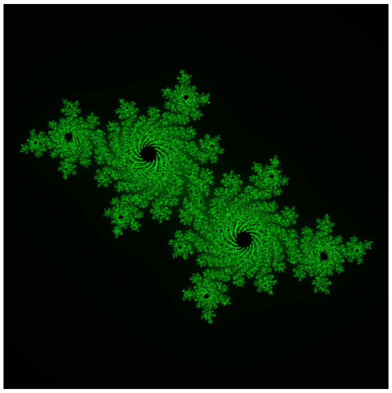


Figure 1.1: Julia Image for c = -.4 + -.59i

2 Preliminaries

For each $c \in \mathbb{C}$, define $f_c : \mathbb{C} \to \mathbb{C}$ by $f_c(z) = z^2 + c$.

2.1 Orbits and filled Julia sets

For $c \in \mathbb{C}$, the *orbit* of a complex number w under iterations of f_c, denoted $\mathscr{O}_c(w)$, is the sequence $\mathscr{O}_c(w) = \left\{ f_c^{(k)}(w) \right\}_{k=0}^{\infty}$, where $f_c^{(k)}(z)$ is the composition of k copies of f_c for $k \in \mathbb{N}$ and $f_c^{(0)}(z) = z$. We call w the *seed* of the orbit and c the *root* of the orbit. We say the orbit $\mathscr{O}_c(w)$ is *bounded* if there exists an $R > 0$ such that $\mathscr{O}_c(w)$ is a subset of $D(0,R) = \{z \in \mathbb{C} : |z| < R\}$. That is, an orbit is bounded if there exists an open disc centered at the origin in $\mathbb{C}$ that contains all elements of $\mathscr{O}_c(w)$.

Each orbit contains infinitely many complex numbers, and some orbits display their behavior relatively quickly (in the first few values), while other orbits take much longer (a lot of values) to show their true nature. Indeed, an orbit can produce two behaviors: the first being trending towards infinity and the second being bounded. When an orbit $\mathscr{O}_c(w)$ is bounded, it can have multiple sub behaviors–it could stay in a closed distance from a specific point, or it can produce a periodic orbit within a set of finitely many points.

Definition 2.1. *For $c \in \mathbb{C}$, the filled Julia set for c, denoted $\mathscr{J}_c$, is the set*

$$\mathscr{J}_c = \{w \in \mathbb{C} : \mathscr{O}_c(w) \text{ is bounded}\}.$$

Definition 2.2. *The Mandelbrot set, denoted $\mathfrak{M}$, is the set of all $c \in \mathbb{C}$ such that $\mathscr{J}_c$ is connected, i.e., $\mathscr{J}_c$ cannot be separated by two or more disjoint open subsets of $\mathbb{C}$.*

We will use an equivalent definition for this paper, which follows from Theorem 2.3 – see Proposition 3.1 in [2].

Theorem 2.3. *A complex number c is in the Mandelbrot set $\mathfrak{M}$ if and only if $0 \in \mathscr{J}_c$.*

For $c \in \mathbb{C}$, it is easy to see that $\mathscr{O}_c(0)$ is bounded if and only if $\mathscr{O}_c(c) = \mathscr{O}_c(0) \setminus \{0\}$ is bounded, which yields the following definition. The *Mandelbrot set* is the set of all complex numbers c such that $\mathscr{O}_c(c)$ is bounded, or equivalently, $c \in \mathscr{J}_c$, i.e., $\mathfrak{M} = \{c \in \mathbb{C} : c \in \mathscr{J}_c\}$.

2.2 Norms on Matrix Algebras

In our work, we will be interested determining the convergence of sequences of matrices. Given a matrix norm $\|\cdot\|$ on $M_n(\mathbb{R})$, a sequence $\{A_k\}_{n=1}^{\infty}$ of matrices in $M_n(\mathbb{R})$ converges to a matrix A with respect to $\|\cdot\|$ if

$$\lim_{k\to\infty} \|A_k - A\| = 0.$$

The two primary matrix norms we use are the *operator* and *Frobenius* norms, denoted $\|\cdot\|_{op}$ and $\|\cdot\|_F$, respectively. Given a matrix $A \in M_n(\mathbb{R})$, the *operator norm* of A is

$$\|A\|_{op} = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

and the *Frobenius norm* of A is

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij}|^2},$$

where A_{ij} is the ij-entry of A. It is well-known that the operator norm of a matrix A always equals its largest *singular value*, where the *singular values* of A are defined to be the square root of the eigenvalues of the matrix $A^T A$, and the Frobenius norm of A is always an upper bound for the operator norm of A, i.e., $\|A\|_{op} \leq \|A\|_F$.

2.3 Real and Imaginary Parts of an Iteration

Let $z = x + yi$ and $c = a + bi$. We denote the first iteration of z under f_c by z_1 – that is, $z_1 := f_c(z) = z^2 + c$. Observe

$$\begin{aligned} z_1 &= z^2 + c \\ &= (x+yi)^2 + (a+bi) \\ &= x^2 + 2xyi - y^2 + a + bi \\ &= \underbrace{(x^2 - y^2 + a)}_{x_1} + \underbrace{(2xy + b)}_{y_1} i. \end{aligned}$$

When we compute z_1 whilst keeping track of the real and imaginary parts of the seed z and root c, we are able to give a precise formula for the real and imaginary parts of $z_1 = x_1 + y_1 i$:

$$x_1 = x^2 - y^2 + a \qquad \text{and} \qquad y_1 = 2xy + b.$$

In fact, for all $k \in \mathbb{N}$, the real and imaginary parts of $z_k = x_k + y_k i$ are given recursively by the formulae

$$x_{k+1} = x_k^2 - y_k^2 + a \qquad \text{and} \qquad y_{k+1} = 2x_k y_k + b. \tag{1}$$

2.4 Affine Transformations

Our work aims to build a matricial framework for determining membership in the Mandelbrot and filled Julia sets. We do so by viewing $\mathbb{C}$ as a 2-dimensional real vector space and each consecutive element of an orbit under f_c as an *affine transformation* of the previous element. Thus, each orbit yields a sequence of *affine transformation matrices* from the algebra of 3×3 matrices with real entries, denoted $M_3(\mathbb{R})$.

Viewing $\mathbb{R}^2$ as a 2-dimensional real vector space, an *affine transformation* $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a transformation that can be decomposed as $T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, where A is a linear transformation and $\mathbf{b}$ is a vector along which you subsequently translate the vector $A\mathbf{x}$. In our work, we prefer to work solely with linear transformations, which requires us to utilize what are called *homogeneous coordinates.* Each vector $\mathbf{x}$ in $\mathbb{R}^2$ can be written in homogeneous coordinates as $\begin{bmatrix}\mathbf{x} & 1\end{bmatrix}^T$, and as such, an affine transformation T defined above can be realized as the projection of a matrix transformation $\tilde{T}$ on $\mathbb{R}^3$ given by

$$\tilde{T}\left(\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}\right) = \begin{bmatrix}A & \mathbf{b}\\ \mathbf{0}^T & 1\end{bmatrix}\begin{bmatrix}\mathbf{x}\\1\end{bmatrix} = \begin{bmatrix}A\mathbf{x}+\mathbf{b}\\1\end{bmatrix},$$

where the standard matrix for $\tilde{T}$ is called the *affine transformation matrix for T*.

3 Matricial Framework for the Mandelbrot Set

As $\mathbb{R}$-vector spaces, $\mathbb{R}^2$ and $\mathbb{C}$ are isomorphic: given $z = x + yi \in \mathbb{C}$, identify z with the vector $\begin{bmatrix}x\\y\end{bmatrix} \in \mathbb{R}^2$. Fix $c = a + bi \in \mathbb{C}$. Then the complex number $f_c(z)$ would be identified with the corresponding vector in $\mathbb{R}^2$:

$$\begin{bmatrix}x\\y\end{bmatrix} \sim z \mapsto f_c(z) \sim \begin{bmatrix}x^2 - y^2 + a\\ 2xy + b\end{bmatrix}.$$

Not surprisingly, this mapping on $\mathbb{R}^2$ is not linear, and thus cannot be implemented by matrix multiplication of a single 2×2-matrix with real entries. Instead, notice:

$$\begin{bmatrix}x^2 - y^2 + a\\ 2xy + b\end{bmatrix} = \begin{bmatrix}x & -y\\ 2y & 0\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} + \begin{bmatrix}a\\b\end{bmatrix}.$$

When the input z is the fixed complex number c, as is of interest when determining if c belongs to the Mandelbrot set, the transformation $c \mapsto f_c(c)$ is implemented in $\mathbb{R}^2$ by first applying a linear transformation to $\begin{bmatrix} a \\ b \end{bmatrix}$ and then translating by $\begin{bmatrix} a \\ b \end{bmatrix}$:

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim c \mapsto f_c(c) \sim \begin{bmatrix} a^2 - b^2 + a \\ 2ab + b \end{bmatrix} = \begin{bmatrix} a & -b \\ 2b & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix}.$$

As described in subsection 5.1, this is an affine transformation, and thus can be realized as a linear transformation when we “lift" to $\mathbb{R}^3$:

$$\begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \sim c \mapsto f_c(c) \sim \begin{bmatrix} a^2 - b^2 + a \\ 2ab + b \\ 1 \end{bmatrix} = \begin{bmatrix} a & -b & a \\ 2b & 0 & b \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ 1 \end{bmatrix}$$

Recall that c belongs to the Mandelbrot set if and only if $\mathscr{O}_c(c) = \{f_c^{(k)}(c) : k \in \mathbb{N}\}$ is bounded. Set $a_0 = a$ and $b_0 = b$. For $k \in \mathbb{N}$, define a_k and b_k to be the real and imaginary parts of $f_c^{(k)}(c)$, respectively, i.e.,

$$a_{k+1} + b_{k+1}i = f_c^{(k+1)}(a + bi) = f_c(a_k + b_k i).$$

For each $k \in \mathbb{N}$, consider the 3×3 matrix

$$\begin{bmatrix} a_k & -b_k & a \\ 2b_k & 0 & b \\ 0 & 0 & 1 \end{bmatrix},$$

and observe that

$$\begin{bmatrix} a_k & -b_k & a \\ 2b_k & 0 & b \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_k \\ b_k \\ 1 \end{bmatrix} = \begin{bmatrix} a_k^2 - b_k^2 + a \\ 2a_k b_k + b \\ 1 \end{bmatrix}$$

By Equation 1, this means that for each $k \in \mathbb{N}$ we have

$$\begin{bmatrix} a_{k+1} \\ b_{k+1} \\ 1 \end{bmatrix} = \begin{bmatrix} a_k & -b_k & a \\ 2b_k & 0 & b \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_k \\ b_k \\ 1 \end{bmatrix}. \tag{2}$$

Equation 2 shows that the orbit $\mathscr{O}_c(c)$ can be generated by recursively applying affine transformations to the real and imaginary parts of the previous image. In particular, for each $k \in \mathbb{N}$, define $A_k : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$A_k(\mathbf{x}) = \begin{bmatrix} a_{k-1} & -b_{k-1} \\ 2b_{k-1} & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} a \\ b \end{bmatrix}.$$

Then Equation 2 shows that the k^{th} iteration (written as an $\mathbb{R}^2$ vector in homogeneous coordinates) of f_c can be obtained by multiplying the $(k-1)^{\text{th}}$ iteration (written as an $\mathbb{R}^2$ vector in homogeneous coordinates) by the affine transformation matrix for A_{k-1}.

Define $M : \mathbb{R}^2 \to M_3(\mathbb{R})$ by $M(x,y) = \begin{bmatrix} x & -y & 0 \\ 2y & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Hence, we can rewrite Equation 2 as

$$\begin{bmatrix} a_k \\ b_k \\ 1 \end{bmatrix} = \left(M(a_{k-1}, b_{k-1}) + \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} a_{k-1} \\ b_{k-1} \\ 1 \end{bmatrix}. \tag{3}$$

We continue by studying properties of the sequence of matrices $\{M(a_k, b_k)\}_{k=0}^{\infty}$, which arise from the generation of $\mathscr{O}_c(c)$.

3.1 Properties of $M(x,y)$

Our work was motivated by determining if properties of $\{M(a_k, b_k)\}_{k=0}^{\infty}$ tells us anything about whether or not $c = a + bi$ is in the Mandelbrot set. In this section, we outline the spectral theory of $M(x,y)$.

Let $x, y \in \mathbb{R}$ be given. A quick computation shows that the eigenvalues for $M(x,y)$ are

$$\lambda(x,y) = \frac{x \pm \sqrt{x^2 - 8y^2}}{2} \quad \text{and} \quad \lambda = 0.$$

Hence, we can define functions $\lambda_+, \lambda_- : \mathbb{R}^2 \to \mathbb{C}$ by

$$\lambda_+(x,y) = \frac{x + \sqrt{x^2 - 8y^2}}{2} \qquad \text{and} \qquad \lambda_-(x,y) = \frac{x - \sqrt{x^2 - 8y^2}}{2}.$$

If $y \neq 0$, i.e., $x + yi \notin \mathbb{R}$, it's easy to check that the eigenspaces corresponding to $\lambda_\pm(a,b)$ are

$$E_{\lambda_+(x,y)} = \operatorname{span}\left\{ \begin{bmatrix} \frac{x+\sqrt{x^2-8y^2}}{4y} \\ 1 \\ 0 \end{bmatrix} \right\} \quad \text{and} \quad E_{\lambda_-(x,y)} = \operatorname{span}\left\{ \begin{bmatrix} \frac{x-\sqrt{x^2-8y^2}}{4y} \\ 1 \\ 0 \end{bmatrix} \right\}.$$

Likewise, if $y = 0$, i.e., $x + yi \in \mathbb{R}$, then $M(x,0)$ is diagonal with eigenspaces $E_x = \operatorname{span}\{\mathbf{e}_1\}$ and $E_0 = \operatorname{span}\{\mathbf{e}_2, \mathbf{e}_3\}$, where $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the standard basis for $\mathbb{R}^3$.

Lemma 3.1. *$M(x,y)$ has distinct eigenvalues if and only if $x \neq \pm 2\sqrt{2}y$.*

Proof. A quick computation yields $\lambda_+(x,y) = \lambda_-(x,y)$ if and only if $x = \pm 2\sqrt{2}y$, which proves the lemma. □

Proposition 3.2. *If $(x,y) \neq (0,0)$, then $M(x,y)$ is diagonalizable if and only if $x \neq \pm 2\sqrt{2}y$.*

Proof. Note that $M(0,0)$ is the 3×3 zero matrix, which is trivially diagonalizable. Suppose $(x,y) \neq (0,0)$. If $x \neq \pm 2\sqrt{2}y$, then $M(x,y)$ is diagonalizable since it has 3 distinct eigenvalues by the previous lemma.

It remains to show that if $M(x,y)$ is diagonalizable, then $x \neq \pm 2\sqrt{2}y$. We prove the contrapositive. If $x = \pm 2\sqrt{2}y$, then

$$\lambda_+(x,y) = \lambda_-(x,y)$$

by the previous lemma. Hence, we have $E_{\lambda_+(x,y)} = E_{\lambda_-(x,y)}$ is 1-dimensional, which implies $M(x,y)$ is not diagonalizable. □

Proposition 3.3. *$M(x,y)$ is singular for all $x,y \in \mathbb{R}$.*

Proof. Since $\lambda = 0$ is an eigenvalue, $M(x,y)$ is not invertible. □

It can be computed directly that the nonzero singular values of $M(x,y)$ are

$$\sigma_-(x,y) := \frac{1}{\sqrt{2}}\left(x^2+5y^2-\sqrt{x^4+10x^2y^2+9y^4}\right)^{1/2}$$
$$\sigma_+(x,y) := \frac{1}{\sqrt{2}}\left(x^2+5y^2+\sqrt{x^4+10x^2y^2+9y^4}\right)^{1/2}$$

Proposition 3.4. *The operator norm of $M(x,y)$ is $\sigma_+(x,y)$.*

Proof. The maximum singular value of $M(x,y)$ is $\sigma_+(x,y)$. □

3.2 Boundedness and Convergence of Iterations

We show that boundedness of an orbit $\mathscr{O}_{a+bi}(a+bi)$ is equivalent to boundedness of the sequence of matrices $\{M(a_k,b_k)\}_{k=0}^{\infty}$ with respect to the operator norm.

Proposition 3.5. *Let $c = a+bi$ and $A+Bi \in \mathbb{C}$ be given. Then $\{f_c^{(k)}(c)\}_{k=0}^{\infty}$ converges to $A+Bi$ if and only if $\{M(a_k,b_k)\}_{k=0}^{\infty}$ converges to $M(A,B)$ in operator (and Frobenius) norm.*

Proof. Observe that

$$M(a_k,b_k)-M(A,B) = \begin{bmatrix} a_k-A & -(b_k-B) & 0 \\ 2(b_k-B) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence, we have $\|M(a_k,b_k)-M(A,B)\|_F^2 = (a_k-A)^2+5(b_k-B)^2$. Since $(a_k-A)^2, (b_k-B)^2 \geq 0$ for all $k \in \mathbb{N}$, it follows that $\lim_{k\to\infty}\|M(a_k,b_k)-M(A,B)\|_F = 0$ if and only if

$$\lim_{k\to\infty}\left|f_c^{(k)}(c)-(A+Bi)\right| = \lim_{k\to\infty}\sqrt{(a_k-A)^2+(b_k-B)^2} = 0.$$

Therefore, $\{f_c^{(k)}(c)\}_{k=0}^{\infty}$ converges to $A+Bi$ if and only if $\{M(a_k,b_k)\}_{k=0}^{\infty}$ converges to $M(A,B)$ in Frobenius norm.

Since the operator and Frobenius norms are equivalent on $M_3(\mathbb{R})$, it follows that $\{M(a_k,b_k)\}_{k=0}^{\infty}$ converges to $M(A,B)$ in operator norm if and only if $\{M(a_k,b_k)\}_{k=0}^{\infty}$ converges to $M(A,B)$ in Frobenius norm, which completes the proof.

□

The following result justifies the study of our matricial framework in the context of the Mandelbrot set.

Theorem 3.6. *A complex number $c = a+bi$ is in the Mandelbrot set if and only if $\{M(a_k,b_k)\}_{k=0}^{\infty}$ is uniformly bounded in operator norm.*

Proof. Suppose $c = a + bi$ is in the Mandelbrot set. Then there exists an $R > 0$ such that $\left\| \begin{bmatrix} a_k \\ b_k \end{bmatrix} \right\|^2 = \left| f_c^{(k)}(c) \right|^2 < R$ for all $k \in \mathbb{N} \cup \{0\}$. It follows that for each $k \in \mathbb{N} \cup \{0\}$ we have

$$\|M(a_k, b_k)\|_F^2 = a_k^2 + 5b_k^2 \leq 5a_k^2 + 5b_k^2 < 5R.$$

Thus, $\sqrt{5R}$ is an upper bound for $\left\{ \|M(a_k, b_k)\|_{op} \right\}_{k=0}^{\infty}$ since the operator norm is bounded above by the Frobenius norm.

Conversely, suppose that $\{M(a_k, b_k)\}_{k=0}^{\infty}$ is bounded in operator norm. Then there exists a $C > 0$ such that $\|M(a_k, b_k)\|_{op} < C$ for all $k \in \mathbb{N} \cup \{0\}$. Hence, for each $k \in \mathbb{N} \cup \{0\}$ the definition of the operator norm yields

$$C > \|M(a_k, b_k)\|_{op} \geq \|M(b_k, a_k)\mathbf{e}_1\| = \left\| \begin{bmatrix} a_k \\ 2b_k \\ 0 \end{bmatrix} \right\| \geq \left\| \begin{bmatrix} a_k \\ b_k \\ 0 \end{bmatrix} \right\| = \left\| \begin{bmatrix} a_k \\ b_k \end{bmatrix} \right\|,$$

where $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the standard basis for $\mathbb{R}^3$. Thus, $\left\{ \left\| \begin{bmatrix} a_k \\ b_k \end{bmatrix} \right\| \right\}_{k=0}^{\infty}$ is bounded, which implies $c = a + bi$ is in the Mandelbrot set. □

As a consequence of Theorem 3.6, we obtain a bound on the sequence of eigenvalues arising from the sequence of affine transformation matrices implementing f_c.

Corollary 3.7. If $c = a + bi$ is in the Mandelbrot set, then $\{\lambda_{\pm}(a_k, b_k)\}_{k=0}^{\infty}$ is bounded.

Proof. The spectral radius of a matrix is bounded above by its operator norm. □

4 Matricial Framework for Filled Julia Sets

Definition 4.1. *Given $a, b, x, y \in \mathbb{R}$, let $c := a + bi$ and $z := x + yi$. Define*

$$J(c,z) := \begin{bmatrix} x & -y & a \\ 2y & 0 & b \\ 0 & 0 & 1 \end{bmatrix}$$

Note that $J(c,z)$ can be decomposed similar to $M(x,y)$:

$$J(c,z) = \begin{bmatrix} x & -y & a \\ 2y & 0 & b \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} x & -y & 0 \\ 2y & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ 0 & 0 & 1 \end{bmatrix} = M(x,y) + T_{\mathbf{c},1}$$

where $T_{\mathbf{c},1}$ denotes translation of a point in $\mathbb{R}^3$ along the vector $\begin{bmatrix} a & b & 1 \end{bmatrix}^T$. Just like the action of f_c on a seed $z = x + yi$, the action of the linear operator $J(c,z)$ moves the vector $\begin{bmatrix} x & y & 1 \end{bmatrix}^T$ to $\begin{bmatrix} x_1 & y_1 & 1 \end{bmatrix}^T$ via two processes, (1) $M(x,y)$ and (2) $T_{\mathbf{c},1}$.

Below we collect some properties of the matrix $J(c,z)$.

- The eigenvalues for $J(c,z)$ are $\{1, \lambda_+(c,z), \lambda_-(c,z)\}$, where

$$\lambda_{\pm}(c,z) = \frac{x \pm \sqrt{x^2 - 8y^2}}{2}.$$

At first glance, the eigenvalues for $J(c,z)$ appear to be identical to those for $M(x,y)$. This is strange because $J(c,z)$ is a matrix which has nonzero entries in the $(1,3)$ and $(2,3)$-entries, while $M(x,y)$ is 0 in these two places. When one computes the eigenvalues of a matrix T by solving the equation $\det(T-\lambda I)=\mathbf{0}$, the algorithm can indeed ignore values along which one is not expanding. What is important to note, however, is that the eigenvalues for $J(c,z_k)$ *will* include information about c because z_k includes information about c.

- The corresponding eigenspaces for $J(c,z)$ are spanned by the eigenvectors
$$\mathbf{e_1}=\begin{bmatrix}\frac{by-a}{1-x+2y^2}\\ \frac{b(x-1)-2ay}{1-x+2y^2}\\ 1\end{bmatrix}\qquad \mathbf{e}_+=\begin{bmatrix}\frac{x-\sqrt{x^2-8y^2}}{4y}\\ 1\\ 0\end{bmatrix}\qquad \mathbf{e}_-=\begin{bmatrix}\frac{x+\sqrt{x^2-8y^2}}{4y}\\ 1\\ 0\end{bmatrix}$$

- Information about the real and imaginary parts of z can be stripped off the matrix $J(c,z)$: the trace of $J(c,z)$ is $x+1$ and the determinant of $J(c,z)$ is $2y^2$.

- $J(c,z)^TJ(c,z)=\begin{bmatrix}x^2+4y^2 & -xy & ax+2by\\ -xy & y^2 & -ay\\ ax+2by & -ay & a^2+b^2+1\end{bmatrix}$ is symmetric.

In applications of linear algebra to quantum physics, mathematicians are particularly interested in *commutation relations* between pairs of matrices. Recall that for $x,y\in\mathbb{R}$, multiplication is commutative: $xy=yx$. However, given two matrices $A,B\in M_n(\mathbb{R})$, matrix multiplication is not always commutative: $AB\neq BA$. A commutation relation is simply the formula one finds when computing $AB-BA$, which, when A and B do not commute, will yield a nonzero matrix. Because of its frequency in literature, mathematicians denote $AB-BA$ by $[A,B]$. In the following two lemmas, we determine precisely when two matrices $J(c,z)$ and $J(c,w)$, where z might be different from w, will commute, and we also show when $J(c_1,z)$ and $J(c_2,z)$ commute, where c_1 and c_2 may also be different.

Lemma 4.2. *Let $c,z,w\in\mathbb{C}$ and set $c=a+bi$, $z:=x+yi$ and $w:=f+gi$. If $a\neq 0$, $[J(c,z),J(c,w)]=\mathbf{0}$ if and only if $x=f$ and $y=g$. If $a=0$, $[J(c,z),J(c,w)]=\mathbf{0}$ if and only if $fy=xg$.*

Proof. Observe
$$\begin{aligned}[J(c,z),J(c,w)]=\mathbf{0}\iff&\begin{bmatrix}xf-2yg & -xg & ax-bf+a\\ 2fy & -2yg & 2ay+b\\ 0 & 0 & 1\end{bmatrix}\\ &-\begin{bmatrix}fx-2yg & -fy & af-bg+a\\ 2xg & -2yg & 2ag+b\\ 0 & 0 & 1\end{bmatrix}\\ &=\mathbf{0}\\ \iff& fy_1\overset{(1)}{=}xg \text{ and } ax-by\overset{(2)}{=}af-g \text{ and } ay\overset{(3)}{=}ag.\end{aligned}$$

Case 1 ($a\neq 0$): If $a\neq 0$, (3) is equivalent to $y=g$, and if $y,g\neq 0$, this implies $x\overset{(1)}{=}f$. If $y=0$ ($g=0$) without loss of generality) then (3) implies $g=0$ ($y=0$). Finally, this yields $ax \overset{(2)}{=} af$, which implies $x = f$ when $a \neq 0$.

Case 2 ($a = 0$): If $a = 0$, then $-by \overset{(2)}{=} -bg$. If $b = 0$, equations (2) and (3) are vacuous and $xg \overset{(1)}{=} fy$ is equivalent to $[J(c,z), J(c,w)] = \mathbf{0}$. If $b \neq 0$, we get $y \overset{(2)}{=} g$. If $y_1 \neq 0$ (or, equivalently, $g \neq 0$) then $x = f$. If $a = y = g = 0$, then x and f can be distinct. □

Lemma 4.3. *Let $c_1, c_2, z \in \mathbb{C}$, and set $c_1 := a_1 + b_1 i$, $c_2 := a_2 + b_2 i$, and $z = x + yi$. Then $[J(c_1,z), J(c_2,z)] = \mathbf{0}$ if and only if either $a = c$ or $a \neq c$, $y = \frac{b-d}{2(a-c)}$ and $x = \frac{-(b-d)^2}{2(a-c)^2} - 1$.*

Proof. One can easily show that $[J(c_1,z), J(c_2,z)] = \mathbf{0}$ if and only if

$$(a_1 - a_2)x + (b_1 - b_2)y + (a_1 - a_2) \overset{(1)}{=} 0 \qquad \text{and} \qquad 2ya_2 + b_1 \overset{(2)}{=} 2ya_1 + b_2.$$

When $a_1 \neq a_2$, (2) can be written as $y = \frac{b_1 - b_2}{2(a_1 - a_2)}$. One can then plug this expression in for y in (1) and simplify to get $x = \frac{-(b_1 - b_2)^2}{2(a_1 - a_2)^2} - 1$. When $a_1 = a_2$, (2) implies $b_1 = b_2$, in which case $c_1 = c_2$. □

The last result in this section is a generalization of the previous section's main theorem. Specifically, in the previous section's main theorem we showed that a complex number $c = a + bi$ belongs to the Mandelbrot set if and only if the sequence of matrices $\{M(a_k, b_k)\}_{k=0}^{\infty}$ is uniformly bounded in the operator norm. Equivalently, $c \in \mathscr{J}_c$ if and only if $\{M(a_k, b_k)\}_{k=0}^{\infty}$ is uniformly bounded in the operator norm. Below, we allow for the seed z to differ from c and prove an analogous result.

Theorem 4.4. *A complex number z belongs to $\mathscr{J}_c$ if and only if the set of matrices $\{J(c, z_k) : n \in \mathbb{N}\}$ is bounded in operator norm.*

Proof. Fix $c = a + bi$ and suppose $z = x + yi \in \mathscr{J}_c$. Then there exists an $R > 0$ such that $\left\| \begin{bmatrix} x_k \\ y_k \end{bmatrix} \right\|^2 = \left| f_c^{(k)}(z_k) \right|^2 < R$ for all $k \in \mathbb{N} \cup \{0\}$. It follows that for each $k \in \mathbb{N} \cup \{0\}$ we have

$$\|J(c, z_k)\|_F^2 = x_k^2 + 5y_k^2 + \left\| \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \right\|^2 \leq 5x_k^2 + 5y_k^2 + \left\| \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \right\|^2 < 5R + \left\| \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \right\|^2.$$

Thus, $\sqrt{5R + a^2 + b^2 + 1}$ is an upper bound for $\left\{ \|J(c, z_k)\|_{op} \right\}_{k=0}^{\infty}$ since the operator norm is bounded above by the Frobenius norm.

Conversely, suppose that $\{J(c, z_k)\}_{k=0}^{\infty}$ is bounded in operator norm. Then there exists a $C > 0$ such that $\|J(c, z_k)\|_{op} < C$ for all $k \in \mathbb{N} \cup \{0\}$. Hence, for each $k \in \mathbb{N} \cup \{0\}$ the definition of the operator norm yields

$$C > \|J(c, z_k)\|_{op} \geq \|J(c, z_k)\mathbf{e}_1\| = \left\| \begin{bmatrix} x_k \\ 2y_k \\ 0 \end{bmatrix} \right\| \geq \left\| \begin{bmatrix} x_k \\ y_k \\ 0 \end{bmatrix} \right\| = \left\| \begin{bmatrix} x_k \\ y_k \end{bmatrix} \right\|,$$

where $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the standard basis for $\mathbb{R}^3$. Thus, $\left\{ \left\| \begin{bmatrix} x_k \\ y_k \end{bmatrix} \right\| \right\}_{k=0}^{\infty}$ is bounded, which implies $z = x + yi \in \mathscr{J}_c$.

□

5 Future Directions

5.1 Dynamical Systems and Markov Processes

In this section, we take $\mathbb{N}$ to denote $\{0,1,2,...\}$. Consider the set of countably infinite sequences of complex numbers:

$$\mathbb{C}^{\mathbb{N}} = \{(z_0, z_1, z_2, ...) : z_k \in \mathbb{C}\ \forall n \in \mathbb{N}\}.$$

We denote an element of $\mathbb{C}^{\mathbb{N}}$ by $(z_k)_{n=0}^{\infty}$. For any $k \in \mathbb{N}$, we write z_k to denote a single complex number within the sequence (z_k). Some of these sequences are related to the orbit of a seed $z \in \mathbb{C}$ under iterations of f_c for some root $c \in \mathbb{C}$. Indeed, if a sequence $(z_k)_{n=0}^{\infty} \in \mathbb{C}^{\mathbb{N}}$ satisfies $z_{n+1} = f_c(z_n)$ for all $n \in \mathbb{N}$, then $\mathscr{O}_c(z) = (z_k)_{n=0}^{\infty}$.

Example 5.1. Below we give some simple examples that show the presence of Julia set orbits inside the set $\mathbb{C}^{\mathbb{N}}$.

(i) If $c = 1$ and $z = 1$, then $\mathscr{O}_1(1) = (1, 2, 5, 26, ...)$.

(ii) If $c = i$ and $z = i$, then $\mathscr{O}_i(i) = (i, i-1, -i, i-1, ...)$.

(iii) If $c = i$ and $z = 2i$, then $\mathscr{O}_i(2i) = (2i, i-4, 15-7i, 176-209i, ...)$.

(iv) If $c = -1$ and $z = 1$, then $\mathscr{O}_{-1}(1) = (1, 0, -1, 0, ...)$.

To parallel our previous sections' work of re-framing iterations and orbits in terms of 2-dimensional real vectors, in this case, we would consider the set

$$(\mathbb{R}^2)^{\mathbb{N}} = \{(\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, ...) : \mathbf{z}_k \in \mathbb{R}^2\ \forall n \in \mathbb{N}\}.$$

Example 5.2. The analogue of the above examples for $\mathbb{C}^{\mathbb{N}}$ are easily translated into the $(\mathbb{R}^2)^{\mathbb{N}}$ picture.

(i) If $c = 1$ and $\mathbf{z} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, then $\mathscr{O}_1(\mathbf{z}) = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 26 \\ 0 \end{bmatrix}, ... \right)$.

(ii) If $c = i$ and $\mathbf{z} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, then $\mathscr{O}_i(\mathbf{z}) = \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, ... \right)$.

(iii) If $c = i$ and $\mathbf{z} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, then $\mathscr{O}_i(\mathbf{z}) = \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} -4 \\ 1 \end{bmatrix}, \begin{bmatrix} 15 \\ -7 \end{bmatrix}, \begin{bmatrix} 176 \\ -209 \end{bmatrix}, ... \right)$.

(iv) If $c = -1$ and $\mathbf{z} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, then $\mathscr{O}_{-1}(\mathbf{z}) = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, ... \right)$.

Just like in the $\mathbb{C}^{\mathbb{N}}$ setting, a sequence $(\mathbf{z}_k)_{n=0}^{\infty}$ of vectors in $(\mathbb{R}^2)^{\mathbb{N}}$ is the orbit $\mathscr{O}_{\mathbf{c}}(\mathbf{z})$ if

$$\text{for every } n \in \mathbb{N}, \qquad \mathbf{z}_{n+1} = J(\mathbf{z}_k, c)\mathbf{z}_k.$$

We can sort of think of the matrices $J(c, z_k)$ as detecting these orbits, but, of course, we have to check this equality holds at all $n \in \mathbb{N}$.

Given $w \in \mathbb{C}$, define a map $\mathsf{J}_{c,w} : (\mathbb{R}^2)^{\mathbb{N}} \to (\mathbb{R}^2)^{\mathbb{N}}$ by

$$\mathsf{J}_{c,w}(\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, ...) := (J(c,w_0)\mathbf{z}_0, J(c,w_1)\mathbf{z}_1, J(c,w_2)\mathbf{z}_2, ...).$$

When $z = w$, $\mathsf{J}_{c,w}$ is the *direct sum* of the matrices $J(c,w_k)$, commonly denoted $\oplus_{n=0}^{\infty} J(c,w_k)$. The notation and definition of direct sums will not be used in the rest of the paper, so we mention it only in case the reader is familiar with these operators.

Note that the action of $\mathsf{J}_{c,w}$ on the specific vector sequence $(\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, ...)$ just pushes each $\mathbf{w}_k$ to the left one position:

$$\mathsf{J}_{c,w}(\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, ...) = (J(c,w_0)\mathbf{w}_0, J(c,w_1)\mathbf{w}_1, J(c,w_2)\mathbf{w}_2, ...) = (\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, ...).$$

Since there's no place on the left for the $\vec{w}_0$ term to go, it just gets tossed off the left edge of the sequence; it walks the plank of the $\mathsf{J}_{c,w}$ pirate ship.

We will call $\mathsf{J}_{c,w}$ a "left shift operator," but we must note that, unlike classic left shift operators, $\mathsf{J}_{c,w}$ is only acting as a left shift on the subspace of vectors spanned by $(\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, ...)$ and images of $(\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, ...)$ under $\mathsf{J}_{c,w}$.

These operators have an interesting property. Recall that we found the eigenvalues for each $J(c,z_k)$ in a previous section: $\sigma(J(c,z_k)) = \{1, \lambda_+(c,z_k), \lambda_-(c,z_k)\}$. It's not hard to show that $\sigma(\mathsf{J}_{c,z})$ contains all of these eigenvalues, $\cup_{n=0}^{\infty} \sigma(J(c,z_k))$, and possibly more. We plan to study the operators $\mathsf{J}_{c,z}$ and their spectra in a future project and see how these relate to the filled Julia set $\mathscr{J}_c$.

6 Notation

- $c = a + bi$
- $z = x + yi$, $z_k = x_k + y_k i$
- $\mathscr{O}_c(z)$ - orbit of z under f_c
- $\mathscr{J}_c$ - filled Julia Set for c
- $\mathfrak{M}$ - Mandelbrot Set
- f_c - mapping of $z \mapsto z^2 + c$
- $M(x,y) = \begin{bmatrix} x & -y & 0 \\ 2y & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
- $J(c,z) = \begin{bmatrix} x & -y & a \\ 2y & 0 & b \\ 0 & 0 & 1 \end{bmatrix}$
- $\mathsf{J}_{c,z}$ - $\oplus_{k=0}^{\infty} J(c, z_k)$
- $M_n(\mathbb{R})$ - set of $n \times n$-matrices with real entries
- $\|\cdot\|_{op}$ - operator norm on $M_n(\mathbb{R})$
- $\|\cdot\|_F$ - Frobenius norm on $M_n(\mathbb{R})$
- $[A,B] = AB - BA$

Acknowledgments

We thank Dr. Anne Boettcher, Director of the Embry-Riddle Aeronautical Universty, prescott Campus Undergraduate Research Institute and Associate Dean of Research, as well as the Arizona/NASA Space Grant Consortium, for their support.

The material contained in this document is based upon work supported by a National Aeronautics and Space Administration (NASA) grant or cooperative agreement. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of NASA. This work was supported through a NASA grant awarded to the Arizona/NASA Space Grant Consortium.

Bibliography

[1] Neil J. Calkin, Eunice Y. S. Chan, Robert M. Corless, David J. Jeffrey, and Piers W. Lawrence. A Fractal Eigenvector. *The American Mathematical Monthly*, 129(6):503–523, 2022

[2] Adrien Douady, John Hamal Hubbard, and P. Lavaurs. Etude Dynamique des Polynômes Complexes. 1984.

[3] Estela A. Gavosto, James R. Miller, and John Sheu. Immersive 4D Visualization of Complex Dynamics. 1998.

[4] John Horgan. Who Discovered the Mandelbrot Set? *Scientific American*, 2009.

[5] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[6] David C. Lay, Steven R. Lay, and Judi J. McDonald. *Linear Algebra and its Applications.* Pearson, 5th edition, 2014.

[7] D. I. Magomedova and O. I. Sheluhin. Fractal models and algorithms for creating aprotective marking for integrity and authenticity bitmap images. In *2020 Systems of Signal Synchronization, Generating and Processing in Telecommunications*, pages 1–6,2020.

[8] Benoit Mandelbrot. *The fractal geometry of nature*. W. H. Freeman and Comp., New York, 3rd edition, 1983.

[9] Walter Rudin. *Real and Complex Analysis.* McGraw-Hill Education, 1987.

[10] Edward B. Saff and Arthur David Snider. *Fundamentals of Complex Analysis for Mathematics, Science, and Engineering.* Prentice Hall, 1993.

[11] Xin Zhang and Zhiqiang Xu. Implementation of Mandelbrot set and Julia set on SOPC platform. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 1494–1498, 2011.