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

ISSN 1550-1736

Cover design by Patrick Foley.

A Word from the Editor

The editorial board is pleased to present our latest issue of the *Mathematics Exchange*, a collection of six articles of interest to a broad audience at the undergraduate level. We appreciate how authors inspire and motivate our readership to follow their example in sharing their love of mathematics, and we hope you enjoy the fruits of their labor. We believe that getting students involved in publishing mathematics is a true milestone in helping them find their (permanent) place in the mathematical community, and we are honored and proud to be a part of that endeavor.

Graph coloring is an important subfield of graph theory. The smallest number of colors needed to color the vertices of a graph so that no two adjacent vertices share the same color is referred to as the chromatic number of the graph. The first article reviews the history of 4-chromatic unit-distance graph, and extends the ideas of the Fish Graph (O'Donnell and Hochberg, 1996) to construct a 4-chromatic unit-distance graph containing only 21 vertices, which is an improvement over a construction of Hochberg and O'Donnell's graph of the same type of 23 vertices.

The second article is a very clear, understandable, and well-written expository paper on the famous $\check{C}ebotarev$ Density Theorem. It provides all the details needed to prove the density theorem, and lists important applications showing that the density theorem has significant implications for primes in arithmetic progressions and binary quadratic forms. In addition, the explanations of difficult number theory concepts contained in this article, including L-series and density statements using Galois theory, are masterfully written. Reading this article will be enjoyable for both advanced undergraduate students and experts on $\check{C}ebotarev$ densities.

The abundancy index of a positive integer is the ratio of the sum of its divisors and itself. The third article is an interesting expository article on the abundancy index. This is an accessible topic in elementary number theory, and it has some surprising connection to the Riemann hypothesis.

John H. Conway's Base-13 function is a nowhere-continuous, real-valued function on $\mathbb{R}$. It provides a counterexample to the converse of Intermediate Value Theorem on any interval of finite length. The value of the Base 13-function $f(x)$ can be described by manipulating the digits of x' s base-13 expansion. Although $f(x)$ can be easily defined in plain language, it is not trivial to formulate it in arithmetic alone. In the fourth article, the author constructs a closed-form function comprising only of arithmetic and proves that it is equivalent to the Base-13 function on integers.

The fifth article presents the application of mathematical modeling in financial problems. To study the optimal balance between savings and consumption, the authors apply Dynamic Programming and Optimal Control Theory in optimization models.

With the numerical simulation using the past data, they create an optimal monthly savings and consumption plan for the given financial goal.

The final article is another article dealing with an accessible topic in elementary number theory. It introduces a generalized definition of amicable numbers, discusses some related questions, and shows that some integers are not feebly amicable with any other integer. This article provides a good example of how an undergraduate research project can be devised.

We hope that you will enjoy reading this issue of the *Mathematics Exchange*. As always, we welcome and encourage ideas on how we can better serve our readers.

Yayuan Xiao

10.30.2021

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 (including your advisor's name and contact information) in PDF form, usually prepared by LaTeX (preferred) or Microsoft Word, to the editor you selected. 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

A Word from the Editor

Editorial Committee and Call for Papers

Articles

Ball State Undergraduate Mathematics Exchange
https:// digitalresearch.bsu.edu/mathexchange
Vol. 15, No. 1 (Fall 2021)
Pages 2 – 8

A 21-Vertex 4-Chromatic Unit-Distance Graph of Girth 4

Daniel Kiteck , Kourtney Payne*

Daniel Kiteck received his Ph.D. in mathematics from the University of Kentucky in 2008. He has since enjoyed teaching at Indiana Wesleyan University, where he loves working with undergraduates in math research.

Kourtney Payne graduated Summa Cum Laude from Indiana Wesleyan University in 2017 with a Bachelor's degree in mathematics. She currently works as a senior physician compensation analyst at Sullivan Cotter in Southfield, Michigan.

Abstract The race to find the smallest 4-chromatic unit-distance graph of girth 4 stalled at 23 vertices in 1996. Using similar ideas to the 23-vertex graph, we constructed a 21-vertex graph. Unknown to us, the smallest possible of 17 vertices had already been created, but using a different approach. This paper carefully constructs our novel 21-vertex graph, while also comparing it to the 1996 23-vertex graph. We also give an overview of the construction of the 17-vertex graph.

1 Introduction

What is the smallest number of colors needed to color the points on the plane so that no two points at a unit-distance from each other have the same color? This smallest number is referred to as the *chromatic number of the plane*. Finding its value is a prominent open problem over a half-century old (search for *Hadwiger-Nelson Problem*). The only possibilities are five, six or seven colors. Four colors was eliminated as an option in 2018 when a unit-distance graph was found that is 5-chromatic, or, in other words, requires 5 colors to keep two adjacent vertices from having the same color [1]. Seven colors can be seen as an upper bound by coloring a regular-hexagon tiling of the plane in the following manner.

***Corresponding author**: daniel.kiteck@indwes.edu

Take a hexagon and its surrounding six hexagons and color them with seven different colors. Cover the plane by repeating this seven-color block. Depending on the given unit-distance, one can scale the colored tiling so that any two points a unit-distance apart will be different colors.

In thinking on the chromatic number of the plane, in 1975, Paul Erdős (of "Erdős number" fame) wondered if 4-chromatic unit-distance graphs without 3-cycles (or "triangle-free" or "of girth 4") exist:

> Let $\mathscr{S}$ be a subset of the plane which contains no equilateral triangle of size 1. Join two points of $\mathscr{S}$ if their distance is 1. Does this graph have chromatic number three? [2]

When Erdős republished the problem in 1979, he said the "...chromatic number is probably at most 3, but I do not see how to prove this." [3] Uncharacteristic for Erdős, he predicted incorrectly. In 1979, Nicholas Wormald showed that such a graph does indeed exist by publishing a 4-chromatic unit-distance 6448-vertex graph without 3-cycles (and, in fact, without 4-cycles) [4].

Alexander Soifer felt like 6448 vertices were a lot. So, in 1992, Soifer informally asked for the smallest example of a 4-chromatic unit-distance graph without 3-cycles [5] (p. 41, 110). From 1994 to 1996, three mathematicians accepted the challenge. Soifer says "A true World Series played out on the pages of *Geombinatorics...* ," a new journal Soifer had recently started [5] (p. 41, 42). First, the size was greatly reduced to 56 vertices by Paul O'Donnell in 1994 [6]. O'Donnell put two 5-pointed stars on a regular decagon, and then carefully connected seven 5-cycles in his construction. Next, Kiran Chilakamarri further reduced the size to 47 vertices in January 1995 [7], using a very different graph than O'Donnell's; Soifer called this the "Moth Graph" [5] (p. 118). Rob Hochberg improved the record by one (to 46), but he did not publish this since he heard about an even better result about to be published [5] (p. 125). The better result was, in July 1995, O'Donnell regaining first place with 40 vertices [8]. This graph has five-fold rotational symmetry. But then, O'Donnell and Hochberg combined forces to make the "Fish Graph" with an impressive 23 vertices in April 1996 [9] (Figure 1). This was the record for two decades. The details and pictures of the graphs of this "World Series" are recorded in chapter 15 of Soifer's book *The Mathematical Coloring Book: Mathematics of Coloring and the Colorful Life of Its Creators* [5]. (This book also states many related open problems.)

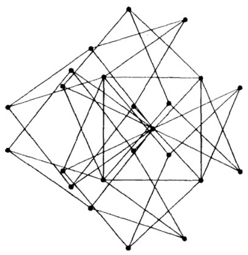


Figure 1: 23-Vertex "Fish Graph" (1996)

We, the authors of the article you are currently reading, thought we had found the first improvement since the Fish Graph. But only after we had finished our research, we realized that a 17-vertex graph (the smallest possible) had been found in 2016 by Geoffrey Exoo and Dan Ismailescu [10] (Figure 4). The following presents a novel approach to lower the 1996 record by two by extending the ideas of the Fish Graph. The construction of the Fish Graph will also be reviewed. An overview of the smallest graph possible, a 17-vertex graph, is given near the end of this article.

2 Construction of the 21-Vertex Graph

We found a 22-vertex graph (Figure 2) but realized we could coincide the two highlighted vertices to make a 21-vertex graph. The construction of the 21-vertex graph also shows how the 22-vertex graph is constructed.

Begin with a unit square WXYZ (Figure 3a). Add on V1, S and T so two rhombuses extend from the unit square (Figure 3b). Note that S has freedom to move while still keeping the graph unit-distance.

Construct pentagon Q1-Q2-Q3-Q4-Q5 where all edges are unit-distance, except possibly Q1-Q5, along with unit attachments Q1-V1, Q1-T, Q2-X, Q3-Z, Q4-X, and Q5-Z (Figure 3c). Moving S keeps everything unit-distance while taking Q1-Q5 from less than one unit to more than one unit. Fix S so that Q1-Q5 is unit-distance. This uses the argument of the Intermediate Value Theorem; this argument can be found in greater detail in [4] and [8].

The notation "X-Y→Z" is used to mean that Z is constructed to be unit-distance from both X and Y, such that X-Y-Z-X is counter-clockwise.

Select U2 unit-distance from Y. Then do the following construction, where the U's are vertices of the first pentagon of the Fish Graph (see below): U2-W→U3, Y-U3→U4, W-U2→U1, and U4-U1→U5 (Figure 3d). This construction gives U2 lots of freedom: the pentagon U1-U2-U3-U4-U5 is guaranteed to be unit-length with unit attachments for a range of choices for U2.

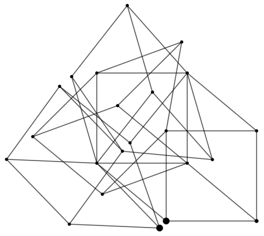


Figure 2: 22-Vertex

We next construct the second pentagon of the Fish Graph (see below), indicated with V's. Start with V1 that is already constructed. Then do V1-Y→V5, V5-U5→V4, W-V4→V3, and U5-V3→V2. Finally, connect V2 to V1 to complete the pentagon, but V2-V1 may not be unit-distance (Figure 3e). Thus, we need to make V2-V1 unit-distance. As U2 varies, the length of V2-V1 goes from below one unit to above one unit. Thus, fix U2 so that V2-V1 is one unit (again using the Intermediate Value Theorem). This finishes the construction of the 21-vertex graph. (Figure 3f).

Now a note on the construction of the Fish Graph. The Fish Graph has the starting square, but it does not have the two rhombuses or the Q-pentagon (the first pentagon constructed). Instead, immediately from the starting square, the U-pentagon and V-pentagon are constructed identically as given (except the vertex V1 will only be in the V-pentagon, and not overlap a vertex since the Fish Graph does not have the rhombuses). Then, to complete the Fish Graph, copies of the two pentagons (U and V) are flipped about the horizontal line through the center of the starting square, and the two pentagon copies are connected in the same way to the starting square as the two pentagon originals, just now "upside-down." This completes the construction of the Fish Graph and explains its horizontal line of symmetry (Figure 1) [9]. One vertex coincides with both pairs of pentagons, making 23 vertices instead of 24.

Here is a proof that the new 21-vertex graph is 4-chromatic. The proof uses elements

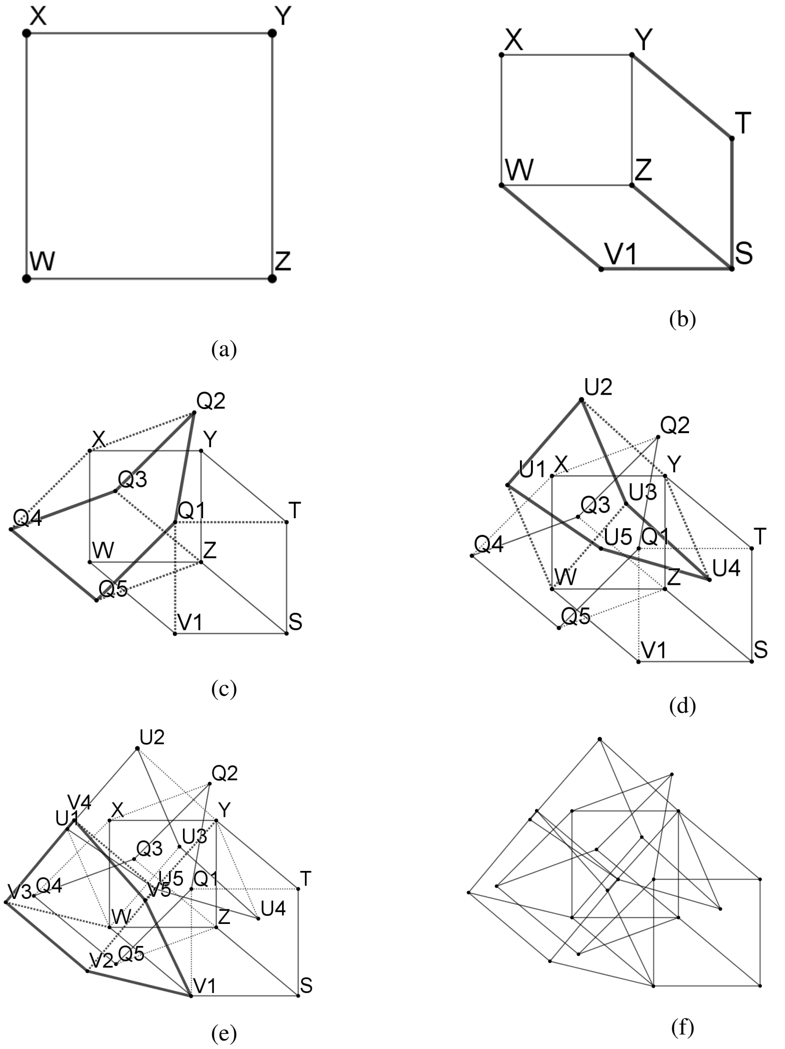


Figure 3: Constructing 21-Vertex Graph

from the proof for the Fish Graph (see below). In any attempted 3-coloring of the 21-vertex graph, one pair of diagonal vertices of the starting unit square must be the same color. Suppose W and Y are the same color, say green. It follows that U5 must also be green. Regardless of how the remaining vertices are colored, every vertex of the V-pentagon is now attached to a green vertex, so the V-pentagon can only be colored with two colors different than green, say blue and red. But pentagons are 3-chromatic, so a fourth color must be introduced.

Now suppose X and Z are the same color, say green. Let W be red and Y be blue. It follows that at least one of V1 and T must be green since otherwise S would be adjacent to three different colors with V1, Z, and T, so S would be forced to be a fourth color.

Then all five vertices of the Q-pentagon are attached to green vertices, so green cannot be used to color the Q-pentagon. But since pentagons are 3-chromatic, a fourth color must be introduced.

For the Fish Graph, the two pairs of pentagons (original U and V and the copies of U and V) work the same way on the vertices of the starting square as the U/V pair and Q pentagon do in the 21-vertex graph, making the Fish Graph 4-chromatic as well.

3 The Smallest Possible: a 17-Vertex Graph Found by Exoo and Ismailescu (2016)

As mentioned, we, the authors, were initially unaware that the 1996 record of 23 vertices in the Fish Graph had already been bettered in 2016. In fact, the new record of 17 vertices has been shown to be the smallest possible 4-chromatic unit-distance graph without 3-cycles (Figure 4). The full construction by Geoffrey Exoo and Dan Ismailescu is in [10]. The following gives an overview.

In a graph, a set of vertices is called *independent* if no two vertices in the set are adjacent. For their starting strategy, Exoo and Ismailescu say "The crucial idea of our approach is summarized in the two paragraphs below."

> Let G be a triangle-free 3-chromatic unit distance graph. For a given proper 3-coloring of the vertices, and a given independent set I, we say that I is *monochromatic* if all vertices of I receive the same color.
>
> Let $\mathscr{I}$ be a collection of independent sets of size 3 such that for every proper 3-coloring of G there exists a set $I \in \mathscr{I}$ which is monochromatic. It is then sufficient to attach 5-cycles **only** to the independent sets from $\mathscr{I}$, and the resulting graph will still be 4-chromatic. [10](p. 52)

This is a generalization of the technique used in the Fish Graph and this paper's 21 vertex graph: A set of pentagon(s) ("5-cycles") is attached in a manner where any proper coloring would have them attached to a monochromatic set, thus forcing another color, so pushing the graph from 3-chromatic to 4-chromatic.

Exoo and Ismailescu then use this method to construct a desired 21-vertex graph (distinctly different from the 21-vertex graph given in this paper– the 21-vertex graph by Exoo and Ismailescu has no connection to the Fish Graph). They start with an 11-vertex graph, and using a computer program, show that there is the desired collection of two independent sets (that at least one of the sets is monochromatic for any proper coloring). Then, laying the 11-vertex graph on a coordinate plane, they use computation technology to solve a system of non-linear equations that come from the restrictions of the graph to find the proper 5-cycles to attach, and where to attach them. Since two 5-cycles are attached, the graph is pushed up from 11 to 21 vertices. They then immediately better the 21 vertices. They show that one can find a starting graph with only **one** independent set that must be monochromatic regardless of coloring. The starting graph has 14 vertices, so when the 5-cycle is attached, the resulting final graph has only 19 vertices. This can be shown to be the smallest such graph using the strategy of independent sets and attaching 5-cycles. Exoo and Imailescu then wondered if they could be close enough to use computation technology to find any smaller, even the smallest.

They searched for graphs of order *n* that satisfy the following properties:

- 4-chromatic and *edge critical*, that is removal of any edge produces a graph which is 3-colorable.
- triangle-free and contain no forbidden subgraph of order up to 7 inclusive [10] (p. 61)

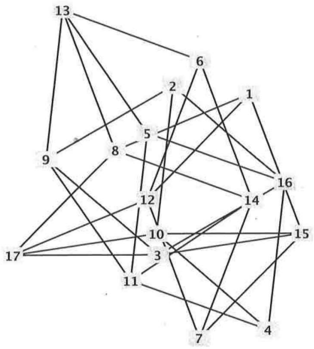


Figure 4: 17-Vertex: a smallest possible triangle-free 4-chromatic unit-distance graph, by Exoo and Ismailescu in 2016; picture from [10] p. 63

They did not elaborate on how they searched for these graphs. From these graphs, they then determined which could be unit-distance. They found none with less than 16 vertices. They found one with 16 vertices, but it did not work because there were two places where there was a unit-distance between a pair of vertices, and when the edges were filled in, it caused triangles. There were no others with 16 vertices. But they found one with 17 vertices. To show that a unit-distance embedding existed, they solved a non-linear polynomial system of six equations and six unknowns. “This system has 48 real solutions...which translates into 12 different embeddings discounting symmetries” [10] (p. 61). Since any smaller graphs would have shown up in their exhaustive list, 17 vertices must be the smallest possible.

The question still remains of what is the smallest 4-chromatic unit-distance graph with no 3-cycles *and* no 4-cycles. The smallest known 4-chromatic unit-distance graph with no 3- or 4-cycles is with 45 vertices [9]. But is this the smallest? This can be extended for larger cycles. It is hoped that considerations along these lines might help solve the chromatic number of the plane problem.

The authors thank Robert Hochberg for pointing us in the right direction going from 22 vertices to 21 vertices.

Bibliography

[1] Aubrey D.N.J. de Grey. "The Chromatic Number of the Plane Is at Least 5". Geombinatorics,XXVIII 1, 18-31, 2018

[2] Paul Erdős. "Unsolved Problems". Congressus Numerantium XV: Proceedings of the 5th British Combinatorics Conference 1975,681, 1976

[3] Paul Erdős "Combinatorial problems in geometry and number theory". Proceedings of Symposia in Pure Mathematics, XXXIV, American Mathematical Society: Relations between Combinatorics and Other Parts of Mathematics. Providence, RI, 149-162, 1979

[4] Nicholas C. Wormald. "A 4-Chromatic Graph with a Special Plane Drawing". Journal of the Australian Mathematical Society (Series A),28,1-8, 1979.

[5] Alexander Soifer. "The Mathematical Coloring Book: Mathematics of Coloring and the Colorful Life of Its Creators". New York, NY: Springer New York,2009.

[6] Paul O'Donnell. "A Triangle-Free 4-Chromatic Graph in the Plane". Geombinatorics IV 1, 23-29, 1994.

[7] K. B. Chilakamarri. "A 4-Chromatic Unit Distance Graph with no triangles". Geombinatorics,IV 3, 64-76, 1995

[8] Paul O'Donnell. "A 40-Vertex 4-Chromatic Triangle-Free Unit Distance Graph". Geombinatorics V 1, 30-34, 1995.

[9] Robert Hochberg, Paul O'Donnell. "Some 4-Chromatic Unit-Distance Graphs Without Small Cycles". Geombinatorics V 4, 134-141, 1996.

[10] Geoffrey Exoo and Dan Ismailescu. "Small Order Triangle-Free 4-Chromatic Unit Distance Graphs". Geombinatorics XXVI 2, 49-64, 2016.

Ball State Undergraduate Mathematics Exchange
`https:// digitalresearch.bsu.edu/mathexchange`
Vol. 15, No. 1 (Fall 2021)
Pages 9 – 27

On the Density Theorem of Čebotarev

*Shaunak Bhandarkar**

Shaunak Bhandarkar wrote the following expository paper when he was a junior at Homestead High School in California. He has currently begun his undergraduate studies at Stanford University and is excited to explore fields like algebraic number theory in greater depth.

Abstract

In this paper, we do exactly what the title implies: prove the *Čebotarev* Density Theorem. This is an extremely valuable theorem because it is a vast generalization of Dirichlet's Theorem on primes in an arithmetic progression, which states that for any $a, n \in_{+}$ relatively prime, there are infinitely many primes that are $\equiv a \pmod{n}$. Our theorem goes even further to the case of other number fields; we will show that the prime ideals in an imaginary quadratic field K are virtually equidistributed among the conjugacy classes of Artin symbols in the Galois group of a Galois extension L over K. Note that L need not be abelian over K.

1 Introduction

We start by introducing the L-functions. This will familiarize us with the most basic definitions as well as important functions. Then, we talk about convergence of L-functions, which will be especially important in later sections.

Next, we briefly visit some character theory. Specifically, the study of Dirichlet characters will help us prove important statements regarding partial zeta functions that will aid us in our journey to the Density Theorem. We then return to our study of L-functions and incorporate some of the theory that we have built up to this point. In particular, we derive an important theorem regarding where an L-function is analytic.

At this point, we introduce the notion of density. Starting with polar density, we explore various density-related properties and go on to prove some powerful results, such as the Artin map being surjective.

***Corresponding author**:shaunakb@stanford.edu

Then, we move on to Dirichlet density (and briefly introduce natural density). We prove that if polar density exists, then so does Dirichlet density, and that the two are equal. This nicely connects these two forms of density. We also explore some properties of Dirichlet density.

In the next section, we deepen our treatment of L-functions. We introduce several of the concepts in class field theory that allow us to derive preliminary density results. Most importantly, we prove that for a nontrivial Dirichlet character of the ray class group, the corresponding L-function does not vanish at $s = 1$.

By generalizing our arguments in the study of L-functions, we establish the theory needed to prove the main theorem in the case of an abelian extension $L \supset K$. At this point, we finally arrive at the main theorem, and prove it in the case of non-abelian extensions by cleverly connecting it to the abelian case.

Finally, we come to what is arguably the most important section: applications of the Čebotarev Density Theorem. This theorem has prolific applications, ranging from the theory of binary quadratic forms to the first main theorem of complex multiplication, although we just list a few. We then part with some concluding remarks.

2 A Review of L-Series

In this paper, we follow largely follow arguments presented in Milne [5] (i.e. most of the definitions, lemmas, propositions, and theorems we prove originate from there). For the enthusiastic reader interested in learning class field theory, Milne's notes are an excellent resource! We start by introducing basic notions needed to prove the main theorem. The first big topic is L-Series. These sums carry valuable information pertaining to prime density, which we will see later on. We assume prior knowledge of group theory (see Artin [3]) as well as basic knowledge of number fields and algebraic number theory. For a refresher of some of the assumed knowledge, check out Bhandarkar[7].

Definition 1. A Dirichlet series is a sum of the form

$$f(s) = \sum_{n \geq 1} \frac{a(n)}{n^s}$$

where $a(n) \in$ and $s = \sigma + it \in$. An Euler product belonging to a number field K is a product of the form

$$g(s) = \prod_{\mathfrak{p}} \frac{1}{(1-\theta_1(\mathfrak{p})N\mathfrak{p}^{-s})\cdots(1-\theta_d(\mathfrak{p})N\mathfrak{p}^{-s})}$$

where $\theta_i(\mathfrak{p}) \in$, $s \in$, and $\mathfrak{p}$ runs over all but finitely many prime ideals of the ring of integers, $\mathscr{O}_K$. Also, N over here denotes the norm function.

Let us look at two important examples of Dirichlet series.

1. The Riemann zeta function is

$$\zeta(s) = \sum_{n \geq 1} \frac{1}{n^s} = \prod_{p} \frac{1}{1-p^{-s}}$$

Notice that the sum is equal to the product because of unique factorization in .

2. More importantly, we will explore the Dedekind zeta function,

$$\zeta_K(s) = \sum_{\mathfrak{a}\geq 0} \frac{1}{N\mathfrak{a}^s} = \prod_{\mathfrak{p}} \frac{1}{1-N\mathfrak{p}^{-s}}$$

The sum is over the integral ideals of $\mathscr{O}_K$ while the product is over the prime ideals of $\mathscr{O}_K$. Furthermore, the sum above is equal to the product because of unique factorization of ideals into prime ideals in the ring of integers $\mathscr{O}_K$ (because it is a Dedekind domain).

Definition 2. Let $I_K^{\mathfrak{m}}$ denote the set of fractional ideals in $\mathscr{O}_K$ that are coprime to the modulus $\mathfrak{m}$. Define a Dirichlet character χ to be a homomorphism

$$\chi : I_K^{\mathfrak{m}} \longrightarrow^{\times}$$

that is trivial over the principal class $P_{K,1}$ of the ray class group $C_{\mathfrak{m}} = I_K^{\mathfrak{m}}/P_{K,1}$. In other words, χ is a character over the ray class group.

Notice that χ somewhat resembles the Artin map (which we will explicitly characterize in Theorem 20), though it is not quite the same. Still, characters are especially useful when dealing with L-functions.

Definition 3. A Dirichlet L-series for a given character χ is

$$L(s,\chi) = \sum_{\mathfrak{a}\subset\mathscr{O}_K,(\mathfrak{a},\mathfrak{m})=1} \frac{\chi(\mathfrak{a})}{N\mathfrak{a}^s} = \prod_{(\mathfrak{p},\mathfrak{m})=1} \frac{1}{1-\chi(\mathfrak{p})N\mathfrak{p}^{-s}}$$

Once again, we can turn the sum into the product because of unique factorization of ideals in $\mathscr{O}_K$.

3 Convergence of L-series

In this section, we list some analytic statements regarding the convergence of Dirichlet series. We omit the proof of most theorems in this section; they generally reduce to extensive computation. Still, they make good exercises for the reader.

Let

$$f(s) = \sum_{n\geq 1} \frac{a(n)}{n^s}$$

be a Dirichlet series and let $S(x) = \sum_{n\leq x} a(n)$, and suppose there exist constants a and b such that $|S(x)| \leq ax^b$ for all large x. Then, $f(s)$ converges uniformly for s in

$$D(b,\delta,\varepsilon) = \{\Re(s) \geq b+\delta, \arg(s-b) \leq \pi/2-\varepsilon\}$$

for all $\delta,\varepsilon \geq 0$, and it converges to an analytic function on the half plane $\Re(s) > b$. (Note that $\Re(s)$ denotes the real part of s.)

Lemma 4. *The Riemann zeta function $\zeta(s)$ has a meromorphic continuation to the half plane $\Re(s) > 0$ with a simple pole at $s = 1$.*

Lemma 5. *For s real and $s > 1$,*

$$\frac{1}{s-1} \leq \zeta(s) \leq 1 + \frac{1}{s-1}$$

Hence, $\zeta(s)$ has a simple pole at $s = 1$ and

$$\zeta(s) = \frac{1}{s-1} + \textit{function holomorphic near } 1$$

Proof. This is left as an exercise to the reader. (Hint: Look at the graph of $y = x^{-s}$ and relate $\zeta(s)$ to the area under the curve.) □

Armed with this fact, we can look at other interesting Dirichlet series.

Let $f(s)$ be a Dirichlet series for which there exists constants C, a, and $b < 1$ such that $|S(x) - ax| \leq Cx^b$. Then, f extends to a meromorphic function on $\Re(s) > b$ with a simple pole at $s = 1$ with residue a.

Proof. For the Dirichlet series $f(s) - a\zeta(s)$, $|S(x)| \leq Cx^b$, so by Proposition , this series converges for $\Re(s) > b$. The result readily follows. □

Before we move on, we encounter one last lemma that will prove to be useful soon.

Lemma 6. *Let $u_1, u_2, \cdots$ be a sequence of real numbers ≥ 2 for which*

$$f(s) = \prod_{j=1}^{\infty} \frac{1}{1 - u_j^{-s}}$$

is uniformly convergent on each region $D(1, \delta, \varepsilon)$ (with $\delta, \varepsilon > 0$). Then,

$$\log f(s) \sim \sum \frac{1}{u_j^s}$$

as $s \to 1^+$ (i.e. $s \to 1$ with $\Re(s) > 1$).

Proof. This is a simple exercise in manipulating sums. (Hint: use the Maclaurin series for $\log(1-x)$ and then break the double sum apart.) □

4 haracters and Partial Zeta Functions

Now, we introduce some basic character theory. In particular, knowing certain statements about characters - namely, the orthogonality relations - will aid us in our study of L-functions.

Definition 7. A one-dimensional representation of a group G, i.e. $\chi : G \longrightarrow^{\times}$ is a character of G. Note that this map is a homomorphism.

For a character χ of G, we have that $\sum_{a\in G}\chi(a) = \begin{cases} |G| & \text{if } \chi = \chi_0 \text{ (the trivial character)} \\ 0 & \text{otherwise} \end{cases}$

Proof. The first part is obvious. If we have a nontrivial character χ, then for some $g \in G$, $\chi(g) \neq 1$. Then,

$$\chi(g)\sum_{a\in G}\chi(a) = \sum_{a\in G}\chi(ga) = \sum_{a\in G}\chi(a),$$

meaning $\sum_{a\in G}\chi(a) = 0$, as desired. □

Suppose the group G is abelian. Fix some $a \in G$. Then,

$$\sum_{\chi\in\hat{G}}\chi(a) = \begin{cases} |G| & \text{if } a = 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, $\hat{G} = (G, C^\times)$ is the character group of G.

Proof. Using the fact that G is noncanonically isomorphic to $\hat{G}$, this proof becomes identical to that of the previous proposition. □

Before we introduce some new tools, let us provide some motivation to our treatment of L-functions. Let K be a number field and $\mathfrak{m}$ be some modulus. Begin with the Dedekind zeta function, $\zeta_K(s)$. For some class $\mathfrak{k} \in C_\mathfrak{m}$ (i.e., the class group), define the partial zeta function to be

$$\zeta(s,\mathfrak{k}) = \sum_{\mathfrak{a}\neq(0),\mathfrak{a}\in\mathfrak{k}} \frac{1}{N\mathfrak{a}^s}$$

Note that for every character χ of the class group,

$$\zeta_K(s) = \sum_{\mathfrak{k}\in C_\mathfrak{m}} \zeta(s,\mathfrak{k}) \text{ and}$$

$$L(s,\chi) = \sum_{\mathfrak{k}\in C_\mathfrak{m}} \chi(\mathfrak{k})\zeta(s,\mathfrak{k})$$

In other words, knowing about $\zeta(s,\mathfrak{k})$ can tell us about the Dedekind zeta function as well as the corresponding L-function.

Theorem 8. *The partial zeta function $\zeta(s,\mathfrak{k})$ is analytic for $\Re(s) > 1 - \frac{1}{[K:]}$ except for a simple pole at $s = 1$. If we let $g_\mathfrak{m}$ denote the residue at $s = 1$, then $g_\mathfrak{m}$ is independent of $\mathfrak{k}$.*

Proof. We omit the proof of this theorem, mainly because it relies on the famous class number formula; to see a detailed derivation, refer to Janusz [4]. It allows us to determine exactly what $g_\mathfrak{m}$ is. □

Corollary 9. *If χ is not the trivial character, the L-function $L(s,\chi)$ is analytic for $\Re(s) > 1 - \frac{1}{[K:]}$.*

Proof. Near $s = 1$,

$$L(s,\chi) = \sum_{\mathfrak{t}\in C_{\mathfrak{m}}} \chi(\mathfrak{t})\zeta(s,\mathfrak{t}) = \frac{\sum_{\mathfrak{t}\in C_{\mathfrak{m}}} \chi(\mathfrak{t})g_{\mathfrak{m}}}{s-1} + \text{holomorphic function}$$

and Proposition shows us that the numerator of the first term is 0. □

5 Polar Density

At last, we come across one type of density. We assume the reader is familiar with notions such as the inertial degree and ramification index that are used to study the decomposition of prime ideals over number fields; to learn these topics in algebraic number theory (and more), see Marcus [2]. For a set T of prime ideals of K, we define $\zeta_{K,T}(s) = \prod_{\mathfrak{p}\in T} \frac{1}{1-N\mathfrak{p}^{-s}}$.

Definition 10. If some positive integral power $\zeta_{K,T}(s)^n$ of $\zeta_{K,T}(s)$ extends to a meromorphic function on a neighborhood of 1 having a pole of order m at 1, we say that T has polar density $\delta(T) = \frac{m}{n}$.

[Properties of Polar Density]

We have the following assertions:

1. The set of all prime ideals of K has polar density 1.
2. The polar density of every set is nonnegative.
3. If T is the disjoint union of T_1 and T_2, and two of the three polar densities exist, then so does the third, and we have $\delta(T) = \delta(T_1) + \delta(T_2)$.
4. If $T \subset T'$, then $\delta(T) \leq \delta(T')$.
5. A finite set has density zero.

Proof.

1. In this case, the set T is the set of all prime ideals of K, so $\zeta_{K,T}(s) = \zeta_K(s)$, which extends to a neighborhood of 1, where it has a simple pole. Thus $\frac{m}{n} = 1$, as desired.
2. Having a negative polar density means $m < 0$, i.e., $\zeta_{K,T}(s)$ is holomorphic in a neighborhood of $s = 1$ and zero there. However, $\zeta_{K,T}(1) = \prod_{\mathfrak{p}\in T} \frac{1}{1-N\mathfrak{p}^{-1}} > 0$, meaning polar density is nonnegative.
3. Observe that $\zeta_{K,T}(s) = \zeta_{K,T_1}(s) \cdot \zeta_{K,T_2}(s)$. Suppose $\zeta_{K,T}(s)^n$ and $\zeta_{K,T_1}(s)^{n_1}$ extend to meromorphic functions with poles of order m and m_1, respectively; the other two cases are identical. Then

$$\zeta_{K,T_2}(s)^{nn_1} = \frac{\zeta_{K,T}(s)^{nn_1}}{\zeta_{K,T_1}(s)^{nn_1}}$$

extends to a meromorphic function in a neighborhood of $s = 1$ and has a pole there of order $mn_1 - m_1 n$. Thus, $\delta(T_2) = \frac{mn_1 - m_1 n}{nn_1} = \frac{m}{n} - \frac{m_1}{n_1} = \delta(T) - \delta(T_1)$, as desired.

4. This follows readily from 3.
5. This is obvious; $m = 0$ because $\zeta_{K,T}(s)$ is finite and positive. Moreover, there is no pole at $s = 1$.

□

If T contains no primes $\mathfrak{p}$ for which $N\mathfrak{p}$ is prime (in), then $\delta(T) = 0$.

Proof. Let $\mathfrak{p}$ be a prime in T. Since $N\mathfrak{p} = p^f$ (where p lies under $\mathfrak{p}$ in and f denotes the inertial degree of $\mathfrak{p}$), we must have $f \geq 2$; if $f = 1$, $N\mathfrak{p}$ would be prime. Moreover, for any given prime $p \in$, there are at most $[K :]$ primes of K lying over p. Thus, $\zeta_{K,T}(s)$ can be decomposed into a product $\prod_{1 \leq i \leq [K:]} g_i(s)$ of $[K :]$ infinite products over the prime numbers, with each factor of g_i being either a 1 or a $\frac{1}{1-p^{-fs}}$ (for every prime p). Thus, for any i, $g_i(1) \leq \prod_p \frac{1}{1-p^{-fp}} \leq \prod_p \frac{1}{1-p^{-2}} = \zeta(2) = \frac{\pi^2}{6}$. Thus, $g_i(s)$ is holomorphic at $s = 1$, meaning that the order of the pole there must be 0 (recall that polar density cannot be negative). We conclude that $\delta(T) = 0$. □

Corollary 11. *Let T_1 and T_2 be sets of prime ideals in K. If the sets differ only by primes $\mathfrak{p}$ for which $N\mathfrak{p}$ is not prime and one of the two sets has polar density, then so does the other, and the densities are equal.*

At last, the time has come to exploit the power of polar density. It turns out we can derive some important analytic results.

Theorem 12. *Let $L \supset K$ be a field extension of finite degree and let M be its Galois closure. Then the set of prime ideals of K that split completely in L has density $\frac{1}{[M:K]}$.*

Proof. The first thing to notice is that a prime ideal $\mathfrak{p}$ of K splits completely in L if and only if it splits completely in M. One direction is easy: if it splits completely in M, it must split completely in the subfield L. If it splits completely in L, then it also splits completely in every conjugate field L'. All of these conjugate fields must lie under the decomposition field (the fixed field of the decomposition group of (M/K)), and so their compositum is a field lying under the decomposition field as well. This field is just M! $\mathfrak{p}$ splits completely only up to and including the decomposition field, so we conclude that it splits completely in M as well.

Thus, it suffices to prove this theorem with the assumption that L is Galois over K. Let S be the set of prime ideals of K that split completely in L and let T be the primes of L lying over a prime ideal in S. For each $\mathfrak{p} \in S$, there are exactly $[L : K]$ prime ideals $\mathfrak{P} \in T$, and for each of them, $N_K^L(\mathfrak{P}) = \mathfrak{p}$ (where N_K^L denotes relative norm). Thus, $N\mathfrak{P} = N\mathfrak{p}$ (where N denotes norm over). This tells us that $\zeta_{L,T}(s) = \zeta_{K,S}(s)^{[L:K]}$. Also, T contains every prime ideal of L that is unramified over K and for which $N\mathfrak{P}$ is prime (in). Thus, T differs from the set of all prime ideals in L by a set of polar density 0 (using Corollary 11), and so T has density 1. Moreover, this shows that $\zeta_{K,S}$ has the property signifying that S is a set of polar density $\frac{1}{[L:K]}$, as desired. □

Corollary 13. *If $f(x) \in K[x]$ splits into linear factors modulo $\mathfrak{p}$ for all but finitely many prime ideals $\mathfrak{p}$ of K, then f splits into linear factors in K.*

Proof. If L is the splitting field of f, then L is Galois over K. Now, use Theorem 12 on L/K. For more interesting details, see Bhandarkar[7], Section 4. □

Corollary 14. *For every abelian extension L/K and every finite set S of primes of K including those that ramify in L, let I_K^S denote the fractional ideals that are prime to all ideals in S. Then, the Artin map*

$$\left(\frac{L/K}{\cdot}\right) : I_K^S \longrightarrow (L/K)$$

is surjective.

Proof. Let H be the image of the Artin map; it is some subgroup of (L/K). If its fixed field is L^H, then we see that $H = (L/L^H)$ is the image. For all $\mathfrak{p} \notin S$, $\left(\frac{L^H/K}{\mathfrak{p}}\right) = \left(\frac{L/K}{\mathfrak{p}}\right)|_{L^H} = 1$, which implies that $\mathfrak{p}$ splits completely in L^H. Thus, all but finitely many prime ideals of $\mathscr{O}_K$ split completely in L^H, so Theorem 12 tells us that $[L^H : K] = 1$; in other words, the Artin map is surjective. □

6 Dirichlet Density

Define two functions $f(s)$ and $g(s)$ for $s > 1$ and real. We write $f(s) \sim g(s)$ as $s \to 1^+$ if $\lim_{s\to 1^+} \frac{f(s)}{g(s)} = 1$. Then, $f(s) \sim \delta \log \frac{1}{s-1}$ as $s \to 1^+$ means

$$\lim_{s\to 1^+} \frac{f(s)}{\log \frac{1}{s-1}} = \delta.$$

When f and g are holomorphic in a neighborhood of $s = 1$ except for possibly poles at $s = 1$, then $f \sim g$ if and only if f and g differ by a function that is holomorphic in a neighborhood of $s = 1$.

Definition 15. Let T be a set of primes of K. If there exists a δ such that

$$\sum_{\mathfrak{p}\in T} \frac{1}{N\mathfrak{p}^s} \sim \delta \log \frac{1}{s-1} \text{ as } s \to 1^+$$

then we say that T has Dirichlet density δ.

Definition 16. If the limit

$$\lim_{x\to\infty} \frac{\text{number of } \mathfrak{p} \in T \text{ with } N\mathfrak{p} \le x}{\text{number of } \mathfrak{p} \text{ with } N\mathfrak{p} \le x}$$

exists, then we call it the natural density of T.

Natural density is much more intuitive than the other types of density, and one might wonder if at all natural density is ever equal to Dirichlet density or polar density. The answer, though reassuring, is somewhat surprising:

1. If polar density exists, then so does Dirichlet density, and the two are equal.
2. If natural density exists, then so does Dirichlet density, and the two are equal.

Proof. We only prove the first part. If T has polar density $\frac{m}{n}$, then

$$\zeta_{K,T}(s)^n = \frac{a}{(s-1)^m} + \frac{g(s)}{(s-1)^{m-1}}$$

where g is holomorphic near $s=1$. Furthermore, $a>0$ because $\zeta_{K,T}(s)>0$ for $s>1$ and real. Taking logs and applying Lemma 6 gives us

$$n\sum_{\mathfrak{p}\in T}\frac{1}{N\mathfrak{p}^s} = m\log\frac{1}{s-1}$$

In other words, T has Dirichlet density $\frac{m}{n}$, as desired. □

A set can have a Dirichlet density without having a natural density. For example, let T be the set of prime numbers with leading digit 1. Then, T does not have a natural density, but it has a Dirichlet density, namely $\log_{10} 2$. Thus, it is a stronger statement to say that a set has natural density.

Also, notice that polar densities are rational numbers. Thus, every set having a natural density that is irrational will not have a polar density! For a more detailed discussion on natural and Dirichlet density, check out Conrad [9].

Now, we shall see that Dirichlet density has similar properties to those of polar density:

[Properties of Dirichlet Density]

1. The set of all prime ideals of K has Dirichlet density 1.
2. The Dirichlet density of any set is nonnegative.
3. If T is the disjoint union of T_1 and T_2, and two of the three Dirichlet densities exist, then so does the third, and $\delta(T) = \delta(T_1) + \delta(T_2)$.
4. If $T \subset T'$, then $\delta(T) \leq \delta(T')$.
5. If T is finite, then $\delta(T) = 0$.

Proof.

1. The set of prime ideals of K even has polar density 1, which is stronger.
2. For $s>0$ and real, $\frac{1}{N\mathfrak{p}^s} > 0$ and for $s \to 1^+$, $\log\frac{1}{s-1} > 0$, so Dirichlet density must be nonnegative.

3. Clearly,
$$\sum_{\mathfrak{p}\in T}\frac{1}{N\mathfrak{p}^s}=\sum_{\mathfrak{p}\in T_1}\frac{1}{N\mathfrak{p}^s}+\sum_{\mathfrak{p}\in T_2}\frac{1}{N\mathfrak{p}^s}$$
so long as $\Re(s)>1$. Thus, if
$$\sum_{\mathfrak{p}\in T_1}\frac{1}{N\mathfrak{p}^s}\sim\delta_1\log\frac{1}{s-1}\text{ and }\sum_{\mathfrak{p}\in T_2}\frac{1}{N\mathfrak{p}^s}\sim\delta_2\log\frac{1}{s-1}$$
then
$$\sum_{\mathfrak{p}\in T}\frac{1}{N\mathfrak{p}^s}\sim(\delta_1+\delta_2)\log\frac{1}{s-1}$$
The other two cases are virtually identical to this one.

4. This readily follows from 3.

5. When T is finite, $\sum_{\mathfrak{p}\in T}\frac{1}{N\mathfrak{p}^s}$ is holomorphic for all s and thus bounded near any point. In particular, as $s\to 1^+$, the Dirichlet density must go to 0.

□

Let T be the set of prime ideals of K having degree 1 over , i.e., for which the inertial degree $f(\mathfrak{p}|p)=1$. Then, $\delta(T)=1$.

Proof. Proposition tells us that the complement of T has polar density equal to 0, and thus, Dirichlet density equal to 0 as well. □

Corollary 17. *Let T be as in the proposition. Then, for every set S of primes in K having Dirichlet density,*
$$\delta(T\cap S)=\delta(S)$$

Proof. The complement T' of T has Dirichlet density 0, so $\delta(S)=\delta(S\cap T)+\delta(S\cap T')=\delta(S\cap T)$, since $\delta(S\cap T')\le\delta(T')=0$. □

7 Making Magic out of L-functions

At last, it is time to put together some of our basic results. We can do this by playing around with L-functions. The value of L-functions, especially as $s\to 1^+$ is crucial to our discussion surrounding the Čebotarev Density Theorem.

Definition 18. Recall that for a number field K and a modulus $\mathfrak{m}$, we say that a subgroup $H\subset I_K^{\mathfrak{m}}$ is a congruence subgroup for $\mathfrak{m}$ if it satisfies $P_{K,1}\subset H\subset I_K^{\mathfrak{m}}$. In this case, the quotient $I_K^{\mathfrak{m}}/H$ is called a generalized ideal class group for $\mathfrak{m}$.

Let $\mathfrak{m}$ be a modulus for K and let H be a congruence subgroup for $\mathfrak{m}$:
$$P_{K,1}\subset H\subset I_K^{\mathfrak{m}}$$

Then, if $L(1,\chi)$ is nonzero for all nontrivial characters χ of the ray class group $I_K^{\mathfrak{m}}/H$, $\delta(\{\mathfrak{p}\in H\})=\frac{1}{(I_K^{\mathfrak{m}}:H)}$; otherwise, it is 0.

Proof. Let $h = (I_K^{\mathfrak{m}} : H)$ and χ be a character of $I_K^{\mathfrak{m}}$ trivial on H, and as usual, let

$$L(s,\chi) = \prod_{\mathfrak{p}\nmid\mathfrak{m}} \frac{1}{1-\chi(\mathfrak{p})N\mathfrak{p}^{-s}}$$

Lemma 6 tells us that

$$\log L(s,\chi) \sim \sum_{\mathfrak{p}\nmid\mathfrak{m}} \frac{\chi(\mathfrak{p})}{N\mathfrak{p}^s} \text{ as } s \to 1^+$$

But Proposition (note that $I_K^{\mathfrak{m}}/H$ is abelian) gives us

$$\sum_{\chi} \chi(\mathfrak{p}) = \begin{cases} h \text{ if } \mathfrak{p} \in H \\ 0 \text{ if } \mathfrak{p} \notin H \end{cases}$$

Thus, summing over all χ, we get

$$\sum_{\chi} \log L(s,\chi) \sim h \sum_{\mathfrak{p}\in H} \frac{1}{N\mathfrak{p}^s} \text{ as } s \to 1^+$$

Now, if $\chi \neq \chi_0$, then $L(s,\chi)$ is holomorphic near $s = 1$, ie. $L(s,\chi) = (s-1)^{m(\chi)}(g(s))$, where $m(\chi) \geq 0$, $g(1) \neq 0$, and $g(s)$ is holomorphic near $s = 1$. Thus, $\log L(s,\chi) \sim m(\chi)\log(s-1) = -m(\chi)\log\frac{1}{s-1}$. If $\chi = \chi_0$, then

$$L(s,\chi) = \frac{\zeta_K(s)}{\prod_{\mathfrak{p}|\mathfrak{m}} \frac{1}{1-N\mathfrak{p}^{-s}}}$$

which means that

$$\log L(s,\chi_0) \sim \log \zeta_K(s) \sim \log \frac{1}{s-1}$$

Thus, we find that

$$h \sum_{\mathfrak{p}\in H} \frac{1}{N\mathfrak{p}^s} \sim (1 - \sum_{\chi\neq\chi_0} m(\chi)) \log \frac{1}{s-1}$$

and hence

$$\delta(\{\mathfrak{p} \in H\}) = \frac{1 - \sum_{\chi\neq\chi_0} m(\chi)}{h}$$

This shows that $\delta(\{\mathfrak{p} \in H\}) = \frac{1}{h}$ if $L(1,\chi) \neq 0$ for every $\chi \neq \chi_0$; otherwise, the density must be 0 (i.e. exactly one of the $m(\chi)$ must be equal to 1, meaning at most one $L(s,\chi)$ can have a zero at $s = 1$ since Dirichlet density is nonnegative, and it must be a simple zero). □

Now, we visit an inequality that will give us useful information about L-functions:

Theorem 19 (The Second Inequality)**.** *For every Galois extension L of K and modulus* $\mathfrak{m}$ *of K,*

$$(I_K^{\mathfrak{m}} : P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})) \leq [L : K]$$

Note that here, $I_L^{\mathfrak{m}}$ denotes the set of fractional ideals of L (lying above ideals of $I_K^{\mathfrak{m}}$) prime to $\mathfrak{m}$.

Proof. Let $H = P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})$. If $\mathfrak{p}$ splits in L, then $f(\mathfrak{P}|\mathfrak{p}) = 1$ for all $\mathfrak{P} \subset \mathscr{O}_L$ lying over $\mathfrak{p} \subset \mathscr{O}_K$, in which case $\mathfrak{p}$ is the norm of any prime ideal of $\mathscr{O}_L$ lying over it. Thus, $\{\mathfrak{p} \in H\}$ contains the set of prime ideals splitting completely in L. Then, Theorem 12 tells us that

$$\delta(\{\mathfrak{p} \in H\}) \geq [L:K]^{-1} > 0$$

Moreover, Proposition tells us that if $\delta(\{\mathfrak{p} \in H\}) > 0$, it must be equal to $(I_K^{\mathfrak{m}} : H)^{-1}$. This only occurs if for all nontrivial characters χ of $I_K^{\mathfrak{m}}/H$, $L(1,\chi) \neq 0$. Finally, we have

$$(I_K^{\mathfrak{m}} : H) = \delta(\{\mathfrak{p} \in H\})^{-1} \leq [L:K]$$

□

This theorem is particularly important because it tells us that if H is of the form as in Proposition , then $L(1,\chi) \neq 0$ for all nontrivial characters χ of $I_K^{\mathfrak{m}}/H$. But when we are given a Galois extension $L \supset K$, how do we know this hypothesis is satisfied? Lucky for us, Artin Reciprocity comes to the rescue!

Theorem 20 (Reciprocity Law)**.** *Let L be a finite Abelian extension of K, and let S be the set of primes of K ramifying in L. Then, the Artin map*

$$\left(\frac{L/K}{\cdot}\right) : I_K^S \longrightarrow (L/K)$$

admits a modulus $\mathfrak{m}$ *such that a prime of K (finite or infinite) ramifies if and only if it divides* $\mathfrak{m}$ *and induces the isomorphism*

$$I_K^{\mathfrak{m}}/(P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})) \xrightarrow{\sim} (L/K)$$

This theorem is literally the very foundation of class field theory. To use this theorem, we also introduce another important theorem of class field theory: the Existence Theorem.

Theorem 21 (Existence Theorem)**.** *For every congruence subgroup H modulo* $\mathfrak{m}$*, there exists a finite Abelian extension* L/K *such that* $H = P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})$.

This theorem is nice because it complements Artin Reciprocity in a way that allows us to construct an important bijection. Notice that for H and L as in the theorem, Artin Reciprocity allows us to construct the isomorphism

$$I_K^{\mathfrak{m}}/H \xrightarrow{\sim} (L/K)$$

In particular, there is a field $L_{\mathfrak{m}}$ known as the ray class field modulo $\mathfrak{m}$ for which the Artin map defines an isomorphism

$$C_{\mathfrak{m}} = I_K^{\mathfrak{m}}/(P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})) \xrightarrow{\sim} (L_{\mathfrak{m}}/K)$$

For a field $L \subset L_{\mathfrak{m}}$, set

$$N_K^L(C_{\mathfrak{m},L}) = (P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})) \pmod{P_{K,1}}$$

Thus, the Existence Theorem provides the following beautiful corollary:

Corollary 22. *For a modulus $\mathfrak{m}$, the map $L \mapsto N_K^L(C_{\mathfrak{m},L})$ is a bijection from the set of Abelian extensions of K contained in $L_{\mathfrak{m}}$ to the set of subgroups of $C_{\mathfrak{m}}$.*

Proof. This is a rather neat result of applying the Galois correspondence. □

Thus, class field theory shows us that the hypothesis of Proposition is satisfied: every congruence subgroup H is of the form $P_{K,1} \cdot N_K^L(I_L^{\mathfrak{m}})$ for a unique Abelian extension $L \supset K$. For our particular discussion, we obtain the following corollary:

Corollary 23. *For any modulus $\mathfrak{m}$ of K and any nontrivial Dirichlet character $\chi : C_{\mathfrak{m}} \longrightarrow^{\times}$, $L(1,\chi) \neq 0$.*

8 Proof of the Cebotarev Density Theorem

At last, we have the tools necessary to prove our main theorem. We will start by handling the abelian case and cleverly use that to tackle the nonabelian case.

Theorem 24. *Let $\mathfrak{m}$ be a modulus for K, and let H be a congruence subgroup for $\mathfrak{m}$. For any class $\mathfrak{t} \in I_K^{\mathfrak{m}}/H$, the set of prime ideals in $\mathfrak{t}$ has Dirichlet density $\frac{1}{(I_K^{\mathfrak{m}}:H)}$.*

Proof. It suffices to prove a more general version of Proposition . Consider some class $\mathfrak{t} \in I_K^{\mathfrak{m}}/H$ and let $\mathfrak{a}$ be a coset representative of this class. Also, let $h = (I_K^{\mathfrak{m}} : H)$. Much like we considered the sum $\sum_\chi \log L(s,\chi)$, we now consider the sum

$$\sum_\chi \chi(\mathfrak{a})^{-1} \log L(s,\chi) \sim \sum_\chi \chi(\mathfrak{a})^{-1} \sum_{\mathfrak{p} \nmid \mathfrak{m}} \frac{\chi(\mathfrak{p})}{N\mathfrak{p}^s} = \sum_{\mathfrak{p} \nmid \mathfrak{m}} \sum_\chi \frac{\chi(\mathfrak{a}^{-1}\mathfrak{p})}{N\mathfrak{p}^s} = h \sum_{\mathfrak{p} \in \mathfrak{t}} \frac{1}{N\mathfrak{p}^s}$$

where we obtain the last equality by applying our character orthogonality relations.

Now, Corollary 23 shows us that $L(1,\chi) \neq 0$ for any nontrivial χ. Thus, using the terminology of Proposition , we see that if $L(s,\chi) = (s-1)^{m(\chi)} g(s)$ near $s = 1$, then in fact $m(\chi) = 0$. Thus, density-wise, $\log L(s,\chi) \sim -m(\chi) \log \frac{1}{s-1} = 0$ as $s \to 1^+$, so $L(s,\chi)$ for nontrivial characters χ do not contribute to the Dirichlet density.

However, if $\chi = \chi_0$, then as we found before, $\log L(s,\chi_0) \sim \log \frac{1}{s-1}$. Thus, by summing $\log L(s,\chi)$ across all χ in the character group, we see that

$$h \sum_{\mathfrak{p} \in \mathfrak{t}} \frac{1}{N\mathfrak{p}^s} \sim \log \frac{1}{s-1} \text{ or } \delta(\{\mathfrak{p} \in \mathfrak{t}\}) = \frac{\sum_{\mathfrak{p} \in \mathfrak{t}} \frac{1}{N\mathfrak{p}^s}}{\log \frac{1}{s-1}} = \frac{1}{h}$$

as desired. □

Corollary 25. *Let $L \supset K$ be a finite Abelian extension and let $\sigma \in (L/K)$. Then, the set of prime ideals $\mathfrak{p}$ of K that are unramified in L and for which $\left(\frac{L/K}{\mathfrak{p}}\right) = \sigma$ has Dirichlet density $\frac{1}{[L:K]}$.*

Proof. Artin Reciprocity gives us the isomorphism $I_K^{\mathfrak{m}}/H \xrightarrow{\sim} (L/K)$ for some modulus $\mathfrak{m}$ and congruence subgroup H. Thus, the inverse image of σ is one entire class $\mathfrak{k}$ of $I_K^{\mathfrak{m}}/H$. At this point, we may apply Theorem 24 to obtain the result. □

Voilà! We have just proven the Čebotarev Density Theorem for Abelian extensions $L \supset K$! At this point, we may extend to the general (not necessarily abelian) case:

Theorem 26 (Čebotarev)**.** *Let L be a finite Galois extension of the field K and suppose $\sigma \in (L/K)$. Moreover, denote C by the conjugacy class of σ in (L/K). Then, the set*

$$T = \{\mathfrak{p} \textit{ a prime ideal in } \mathscr{O}_K \mid \mathfrak{p} \textit{ unramified in } L, \left(\frac{L/K}{\mathfrak{p}}\right) = C\}$$

has Dirichlet density

$$\delta(T) = \frac{|C|}{|(L/K)|} = \frac{|C|}{[L:K]}.$$

Proof. Since (L/K) is not necessarily abelian, we try to cleverly reduce to this case. Let $\sigma \in (L/K)$ have order f and let $M = L^{\langle\sigma\rangle}$ be the fixed field of the set of automorphisms $\langle\sigma\rangle$ (subgroup of automorphisms generated by σ). Then, L is a cyclic extension of M of degree f, and the Artin map gives us an isomorphism

$$C_{\mathfrak{m}}/H \xrightarrow{\sim} \langle\sigma\rangle$$

for some modulus $\mathfrak{m}$ of M and $H = P_{M,1} \cdot N_M^L(I_L^{\mathfrak{m}})$.

Now, let $\mathfrak{p}$ be a prime of $\mathscr{O}_K$, $\mathfrak{q}$ be prime lying above $\mathfrak{p}$ in $\mathscr{O}_M$, and $\mathfrak{P}$ be a prime lying above $\mathfrak{q}$ in $\mathscr{O}_L$. If we let $c = |C|$ and $d = [L:K]$, we must show that $\delta(T) = \frac{c}{d}$. Also, we must note that in this proof, we ignore the finitely many primes that are not prime to $\mathfrak{m}$ (i.e. primes that ramify).

Let

$$T_{M,\sigma} = \{\mathfrak{q} \subset \mathscr{O}_M \mid \left(\frac{L/M}{\mathfrak{q}}\right) = \sigma, f(\mathfrak{q}|\mathfrak{p}) = 1\}$$

By Corollary 25, we know that the set of primes satisfying the first condition (i.e. $\left(\frac{L/M}{\mathfrak{q}}\right) = \sigma$) of $T_{M,\sigma}$ has density $\frac{1}{f}$, and thus, $T_{M,\sigma}$ has density $\frac{1}{f}$ (using Corollary 17).

Now, let

$$T_{L,\sigma} = \{\mathfrak{P} \subset \mathscr{O}_L \mid \left(\frac{L/K}{\mathfrak{P}}\right) = \sigma\}$$

We aim to relate $T_{M,\sigma}$ and $T_{L,\sigma}$.

Lemma 27. *We have the following two assertions:*

1. *The map $\mathfrak{P} \mapsto \mathfrak{q} = \mathfrak{P} \cap \mathscr{O}_M$ defines a bijection $T_{L,\sigma} \to T_{M,\sigma}$.*
2. *The map $\mathfrak{P} \mapsto \mathfrak{p} = \mathfrak{P} \cap \mathscr{O}_K : T_{L,\sigma} \to T$ sends exactly $\frac{d}{cf}$ primes of $T_{L,\sigma}$ to each prime of T.*

Proof.

1. Take some $\mathfrak{P} \in T_{L,\sigma}$ and let $\mathfrak{q} = \mathfrak{P} \cap \mathscr{O}_M$ and $\mathfrak{p} = \mathfrak{P} \cap \mathscr{O}_K$. Then, the Decomposition Group $D(\mathfrak{P}|\mathfrak{p}) \cong (\mathscr{O}_L/\mathfrak{P} \;/\; \mathscr{O}_K/\mathfrak{p})$ is generated by σ but σ fixes the residue field $\mathscr{O}_M/\mathfrak{q}$ (because it fixes M). Thus, $\mathscr{O}_M/\mathfrak{q} = \mathscr{O}_K/\mathfrak{p}$, meaning that $f(\mathfrak{q}|\mathfrak{p}) = 1$. This means that $\mathfrak{q} \in T_{M,\sigma}$, so we have a map

$$\mathfrak{P} \mapsto \mathfrak{q} = \mathfrak{P} \cap \mathscr{O}_M : T_{L,\sigma} \to T_{M,\sigma}$$

This map is injective because $f(\mathfrak{P}|\mathfrak{q}) = f(\mathfrak{q}|\mathfrak{p})^{-1} f(\mathfrak{P}|\mathfrak{p}) = 1 \cdot f = f$, so $\mathfrak{P}$ is the only prime of $\mathscr{O}_L$ lying over $\mathfrak{q}$. Moreover, this map is surjective because for any prime $\mathfrak{P}$ lying over $\mathfrak{q} \in T_{M,\sigma}$,

$$\left(\frac{L/K}{\mathfrak{P}}\right) = \left(\frac{L/K}{\mathfrak{P}}\right)^{f(\mathfrak{q}|\mathfrak{p})} = \left(\frac{L/M}{\mathfrak{q}}\right) = \sigma$$

and so $\mathfrak{P}$ lies in $T_{L,\sigma}$. Thus, our map is a bijection.

2. Fix a $\mathfrak{p}_0 \in T$ and let $\mathfrak{P}_0 \in T_{L,\sigma}$ lie over $\mathfrak{p}_0$. Then, for $\tau \in (L/K)$,

$$\left(\frac{L/K}{\tau\mathfrak{P}_0}\right) = \tau \left(\frac{L/K}{\mathfrak{P}_0}\right) \tau^{-1}$$

and so

$$\tau \left(\frac{L/K}{\mathfrak{P}_0}\right) \tau^{-1} = \sigma \iff \tau \in C_G(\sigma)$$

where $C_G(\sigma)$ denotes the centralizer of σ in (L/K). Therefore, the map $\tau \mapsto \tau\mathfrak{P}_0$ gives us a bijection

$$C(\sigma)/D(\mathfrak{P}_0|\mathfrak{p}_0) \longrightarrow \{\mathfrak{P} \in T_{L,\sigma} \mid \mathfrak{P} \cap \mathscr{O}_K = \mathfrak{p}_0\}$$

where $D(\mathfrak{P}_0|\mathfrak{p}_0)$ denotes decomposition group. The decomposition group is $\langle\sigma\rangle$, which has order f and $C_G(\sigma)$ has order $\frac{d}{c}$ because there is a bijection

$$\tau \mapsto \tau\sigma\tau^{-1} : (L/K)/C_G(\sigma) \to C$$

Therefore, $(C_G(\sigma) : D(\mathfrak{P}_0|\mathfrak{p}_0)) = \frac{d}{cf}$. Thus, we have shown that for each $\mathfrak{p} \in T$, there are exactly $\frac{d}{cf}$ primes $\mathfrak{P} \in T_{L,\sigma}$ lying over $\mathfrak{p}$. This completes part 2.

□

Returning to our proof, we can combine statements 1 and 2 to obtain the map

$$\mathfrak{q} \mapsto \mathfrak{p} = \mathfrak{q} \cap \mathscr{O}_K$$

which is a $\frac{d}{cf} : 1$ map $T_{M,\sigma} \to T$. For such a $\mathfrak{q}$, $N_K^M(\mathfrak{q}) = \mathfrak{p}$, so $N\mathfrak{q} = N\mathfrak{p}$. Hence

$$\sum_{\mathfrak{p}\in T} \frac{1}{N\mathfrak{p}^s} = \frac{cf}{d} \sum_{\mathfrak{q}\in T_{M,\sigma}} \frac{1}{N\mathfrak{q}^s} \sim \frac{cf}{d} \cdot \frac{1}{d} \log \frac{1}{s-1} = \frac{c}{d} \log \frac{1}{s-1}$$

which completes the proof of the Čebotarev Density Theorem. □

Interestingly enough, the prime number theorem generalizes nicely to general number fields; it is called the Landau Prime Ideal Theorem. Using this theorem and keeping the notation we used above, if we set

$$\pi_C(x) = \{\mathfrak{p} \text{ is a finite, unramified prime ideal of } \mathscr{O}_K \mid \left(\frac{L/K}{\mathfrak{p}}\right) = C, N\mathfrak{p} \leq x\}$$

then we can obtain the following effective form of the Density Theorem:

$$\pi_C(x) \sim \frac{c}{d}\frac{x}{\log x}.$$

9 Applications of the Density Theorem

The Density Theorem has many applications throughout number theory. By no means do we provide a full treatment of its applications; rather, we focus on a few rather elegant examples. We start by pointing out a simple yet special case: Dirichlet's Theorem on Primes in Arithmetic Progression (we follow the argument presented in Triantafillou [6]).

Corollary 28 (Dirichlet)**.** *For any positive integers a and m, with* $\gcd(a,m) = 1$*, there are infinitely many primes p for which $p \equiv a \pmod{m}$.*

Proof. Using the Čebotarev Density Theorem, we will prove an even stronger result: that the set of primes $\equiv a \pmod{m}$ has Dirichlet density $\frac{1}{\phi(m)}$ in the set of primes (of), where ϕ denotes the totient function.

Now, let ζ_m be an m^{th} root of unity. Let $K =$ and $L = (\zeta_m)$ be a cyclotomic extension. We know that L/K is Galois and that $(L/K) \cong (/m)^\times$. This isomorphism can be made explicit by taking some $a \in (/m)^\times$ and mapping it to the unique automorphism that takes ζ_m^k to ζ_m^{ak}.

For a prime number $p \in$, $N(p) = p$. If $\mathfrak{P} \subset \mathscr{O}_L$ is a prime lying over p such that $\sigma \in (L/K)$ satisfies $\sigma(\alpha) \equiv \alpha^{N(p)} \pmod{\mathfrak{P}}$, we must have $\sigma(\zeta_m^k) = \zeta_m^{pk}$ for all k. As long as $p \nmid m$, p does not ramify in L, in which case $\left(\frac{L/K}{p}\right) = \bar{p} \in (L/K)$, where $\bar{p}$ is the class of p modulo m. Thus, $\left(\frac{L/K}{p}\right) = a$ if and only if $p \equiv a \pmod{m}$. At this point, the Density Theorem states that the density of primes p of such that $\left(\frac{L/K}{p}\right) = a$ is $\frac{1}{|(L/K)|} = \frac{1}{\phi(m)}$, as desired. □

So being able to prove Dirichlet's theorem with a snap of our fingers is a sign of just how powerful the Čebotarev Density Theorem is! Now, we move on to another interesting application, which explores primes that split completely in number fields. In particular, these primes can characterize a given extension $L \supset K$. First, we introduce some terminology.

Definition 29. Given two sets $\mathscr{S}$ and $\mathscr{T}$, we say $\mathscr{S} \dot{\subset} \mathscr{T}$ if $\mathscr{S} \subset \mathscr{T}$ up to a finite set of elements. We also say $\mathscr{S} \doteq \mathscr{T}$ if $\mathscr{S} \dot{\subset} \mathscr{T}$ and $\mathscr{T} \dot{\subset} \mathscr{S}$.

Definition 30. Given an extension $L \supset K$, we set

$$\mathscr{S}_{L/K} = \{\mathfrak{p} \text{ is a finite prime ideal of } K \mid \mathfrak{p} \text{ splits completely in } L\}.$$

Also, let

$\tilde{\mathscr{S}}_{L/K} = \{\mathfrak{p}$ is a finite prime ideal of $\mathscr{O}_K \mid \mathfrak{p}$ unramified in $L, f(\mathfrak{P}|\mathfrak{p}) = 1$ for some prime

$\mathfrak{P}$ of L lying over $\mathfrak{p}\}$.

Using this terminology, we can effectively state the following powerful theorem:

Theorem 31. *Let L and M be finite extensions of K. Then:*

1. *If M is Galois over K, then* $L \subset M \iff \mathscr{S}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$.
2. *If L is Galois over K, then* $L \subset M \iff \tilde{\mathscr{S}}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$

Proof. We begin with the proof of 2. When $L \subset M$, we easily have $\tilde{\mathscr{S}}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$; indeed, for $\mathfrak{p} \in \tilde{\mathscr{S}}_{M/K}$, $f(\mathfrak{P}|\mathfrak{p}) = 1$ for some $\mathfrak{P}$ lying over $\mathfrak{p}$ in $\mathscr{O}_M$. Thus, if $\mathfrak{q}$ is a prime of $\mathscr{O}_L$ lying over $\mathfrak{p}$ and under $\mathfrak{P}$, then we must have $f(\mathfrak{q}|\mathfrak{p}) = 1$. But since inertial degrees of all conjugates of a prime ideal are the same in a Galois extension, we conclude that $\mathfrak{p}$ has inertial degree $f = 1$ in L. Moreover, since it is unramified, we conclude that $\mathfrak{p}$ splits completely in L, and thus $\mathfrak{p} \in \mathscr{S}_{L/K}$ as well.

Conversely, suppose that $\tilde{\mathscr{S}}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$, and let N be a Galois extension of K containing both L and M; it suffices to show that $(N/M) \subset (N/L)$. Thus, given $\sigma \in (N/M)$, we need to prove that $\sigma |_L = 1$. By the Čebotarev Density Theorem, there is a prime $\mathfrak{p}$ in K, unramified in N such that $\left(\frac{N/K}{\mathfrak{p}}\right)$ is the conjugacy class of σ. Thus, there is some prime $\mathfrak{P}$ of N for which $\left(\frac{N/K}{\mathfrak{P}}\right) = \sigma$. We claim that $\mathfrak{p} \in \tilde{\mathscr{S}}_{M/K}$. To see this, let $\mathfrak{P}' = \mathfrak{P} \cap \mathscr{O}_M$. Then, for $\alpha \in \mathscr{O}_M$,

$$\alpha \equiv \sigma(\alpha) \equiv \alpha^{N(\mathfrak{p})} \pmod{\mathfrak{P}'}$$

where the first congruence follows from $\sigma |_M = 1$ and the second from the definition of the Artin symbol. Thus, the Artin symbol is trivial, meaning that $f(\mathfrak{P}'|\mathfrak{p}) = 1$ (since f is the order of the decomposition group generated by the Artin symbol, which is trivial). This means $\mathfrak{p} \in \tilde{\mathscr{S}}_{M/K}$, as desired. The Density Theorem implies that there are infinitely many such $\mathfrak{p}$'s. Thus, $\tilde{\mathscr{S}}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$ tells us that $\mathfrak{p} \in \mathscr{S}_{L/K}$, i.e., $\left(\frac{L/K}{\mathfrak{p}}\right) = 1$, meaning that $\sigma |_L = \left(\frac{N/K}{\mathfrak{P}}\right) |_L = \left(\frac{L/K}{\mathfrak{p}}\right) = 1$, as desired.

Now, to prove 1, note that $L \subset M$ easily implies $\mathscr{S}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$ using the exact same reasoning as in the proof of part 2 above. To show the other direction, let L' be the Galois closure of L over K. Using the reasoning from Theorem 12, we see that a prime of K splits completely in L if and only if it splits completely in L'. Thus, $\mathscr{S}_{L/K} = \mathscr{S}_{L'/K}$. Thus, our hypothesis $\mathscr{S}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$ may be rephrased as $\mathscr{S}_{M/K} \dot{\subset} \mathscr{S}_{L'/K}$. By part 2, we obtain $L' \subset M$, so $L \subset M$, and we are done. □

Why did we bother to prove all of that? For one, it tells us about the relationship between field extension and the prime ideals contained in them. Moreover, it allows us to formulate the following corollary:

Corollary 32. *Let L and M be Galois extensions of K. Then:*

1. $L \subset M \iff \mathscr{S}_{M/K} \dot{\subset} \mathscr{S}_{L/K}$.
2. $L = M \iff \mathscr{S}_{M/K} \dot{=} \mathscr{S}_{L/K}$.

Proof. Notice first that 1 immediately implies 2, so it suffices to prove just 1. Now, observe that if M is Galois over K, then $\tilde{\mathscr{S}}_{M/K}$ reduces to $\mathscr{S}_{M/K}$, so applying Theorem 31 immediately proves part 1 of this corollary. □

Now, we introduce one last application, which is to the theory of binary quadratic forms. Although we do not prove it here, it points out a beautiful interplay between binary quadratic forms and ideals in number fields.

Theorem 33. *Let $f(x,y) = ax^2 + bxy + cy^2$ be a primitive positive definite binary quadratic form of discriminant $D < 0$. Moreover, let $\mathscr{S}$ be the set of primes represented by f. Then, the Dirichlet density $\delta(\mathscr{S})$ exists and is equal to*

$$\delta(\mathscr{S}) = \begin{cases} \frac{1}{2h(D)} & \textit{if } f \textit{ is properly equivalent to its opposite} \\ \frac{1}{h(D)} & \textit{otherwise} \end{cases}$$

where $h(D)$ is the famous class number. In particular, f represents infinitely many prime numbers!

Proof. We omit the proof because it relies on developing a theory of ideals in orders of imaginary quadratic fields. Still, we refer the reader to Cox[1]. □

10 Some Parting Remarks

The Čebotarev Density Theorem is elegant and powerful. It is at a crossroads between algebraic and analytic number theory, and has various applications across number theory. In this paper, we presented a formulation and proof of the density theorem in the language of class field theory to illustrate the connection between the two. However, it is interesting to note that when Čebotarev first proved this theorem, he did not make use of class field theory; rather, it was the density theorem that motivated much of the formulation of class field theory! To learn about the history of the Čebotarev Density Theorem and read Čebotarev's original proof, we refer the reader to Lenstra and Stevenhagen [8]. We sincerely hope the reader has taken away something useful from this paper and is inclined to learn more about related topics.

Bibliography

[1] Cox, D.: Primes of the Form $x^2 + ny^2$. 2nd edn. Wiley-Interscience (1997)

[2] Marcus, D.: Number Fields. 2nd edn. Springer International Publishing (2018)

[3] Artin, Michael.: Algebra. 2nd edn. Pearson (2010)

[4] Janusz, Gerald . 2nd edn. American Mathematical Society (2005)

[5] Milne, James. Class Field Theory. Retrieved from http://www.jmilne.org/math/CourseNotes/CFT.pdf. Version 4.02. 2013.

[6] Triantafillou, Nicholas. The Chebotarev Density Theorem. Retreived from https://math.mit.edu/ ngtriant/notes/chebotarev.pdf. (2015)

[7] Bhandarkar, Shaunak. From Hilbert Class Field Theorey to Complex Multiplication. Retreived from https://drive.google.com/file/d/1SE1GEhtP-UWMUOTcWtJOPLuPOkYjiRw3/view. (2018)

[8] Lenstra, Henrik. and Stevenhagen, Peter.: Chebotarëv and his Density Theorem. Retreived from http://websites.math.leidenuniv.nl/algebra/chebotarev.pdf(2002)

[9] Conrad, Brian.: Dirichlet Density for Global Fields. Retreived from http://math.stanford.edu/ conrad/676Page/handouts/dirdensity.pdf. (2004)

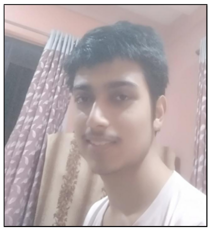
Ball State Undergraduate Mathematics Exchange
`https:// digitalresearch.bsu.edu/mathexchange`
Vol. 15, No. 1 (Fall 2021)
Pages 28 – 39

Measuring Abundance with Abundancy Index

*Kalpok Guha, Sourangshu Ghosh**

Kalpok Guha (kalpok.guha@gmail.com) is currently a Master's student at the Department of Mathematics, RWTH Aachen University, Germany. He completed his undergraduate degree in Mathematics from Presidency University, Kolkata, India. He aspires to pursue higher studies and research in the field of Analytic Number Theory. He is also interested in Abstract Algebra and Combinatorics.

Sourangshu Ghosh (sourangshug123@gmail.com) is an undergraduate student of the Department of Civil Engineering at the Indian Institute of Technology Kharagpur. He is pursuing Mathematics as a minor degree and is interested in Structural Reliability and Discrete Mathematics. He enjoys playing the violin in the Indian Classical Music Style.



Abstract

A positive integer n is called perfect if $\sigma(n) = 2n$, where $\sigma(n)$ denote the sum of divisors of n. In this paper we study the ratio $\frac{\sigma(n)}{n}$. We define the Abundancy Index $I : \mathbb{N} \to \mathbb{Q}$ with $I(n) = \frac{\sigma(n)}{n}$. Then we study different properties of Abundancy Index and discuss the set of Abundancy Indices. Using this function we define a new class of numbers known as superabundant numbers. Finally we study superabundant numbers and their connection with the Riemann Hypothesis.

1 Introduction

Definition 1. A positive integer n is called perfect if $\sigma(n) = 2n$, where $\sigma(n)$ denote the sum of positive divisors of n.

The first few perfect numbers are $6, 28, 496, 8128, ...$ (OEIS A000396), This is a well studied topic in number theory. Euclid studied properties and nature of perfect numbers in 300 BC. He proved that if $2^p - 1$ is a prime, then $2^{p-1}(2^p - 1)$ is an even perfect

***Corresponding author**: sourangshug123@gmail.com number(Elements, Prop. IX.36). Later mathematicians have spent years to study the properties of perfect numbers. But still many questions about perfect numbers remain unsolved. Two famous conjectures related to perfect numbers are

1. There exist infinitely many perfect numbers. Euler[13] proved that a number is an even perfect number if and only if it can be written as $2^{p-1}(2^p-1)$ and 2^p-1 is also a prime number. Primes number of the form 2^p-1 are known as Mersenne primes. Therefore this conjecture is equivalent to the conjecture that there exist infinitely many Mersenne primes. Some good references on this topic are [15], [9], [45].

2. There do not exist any odd perfect numbers. Computation of Lower Bounds for the smallest perfect numbers have been done by many mathematicians. Kanold (1957)[28] found the bound 10^{20}, Tuckerman (1973) [46] found the bound 10^{36}, Hagis (1973) [19]found the bound 10^{50}, Brent and Cohen (1989) [5] found the bound 10^{160}, Brent et al. (1991) [6] found the bound 10^{300}. The best bound till today is 10^{1500} by Ochem and Rao (2012)[33]. The odd perfect numbers if they exist must be of the form $p^{4\lambda+1}Q^2$, where p is a prime of the form $4n+1$ as proven by Euler[8][49].Touchard[44] and Holdener[23] proved that the odd perfect numbers if they exist must be of the form $12k+1$ or $36k+1$. Stuyvaert[11] proved that the odd perfect numbers if they exist must be must be a sum of two squares. Greathouse and Weisstein[17] alternatively write that any odd perfect number must be of the form

   $$N = p^{\alpha}{q_1}^{2\beta_1}...{q_r}^{2\beta_r}$$

   where all the primes are odd. Also $p \equiv \alpha \equiv 1(\bmod 4)$. Steuerwald[43] and Yamada[51] proved that all the β_is cannot be 1. Odd perfect numbers have a large number of distinct prime factors. The odd perfect number if one exists must have at least 6 distinct prime factors, as proved by Gradshtein[4]. This was extended to 8 by Haggis[20]. If there are 8 the number must be divisible by 15, as proved by Voight [47]. Norton[32] proved that odd perfect numbers must have at least 15 and 27 distinct prime factors if the number is not divisible by 3 or 5 and 3, 5, or 7 respectively. Nielsen[31] extended the bound by showing that odd perfect numbers should have at least 9 distinct prime factors and if it is not divisible by 3 it should have at least 12 distinct prime factors. Hare[22] shown that any odd perfect number must have at least 75 prime factors. The method used by Hare involves factorization of several large numbers[49][22].The best lower bound is by Ochem and Rao (2012)[33], who prove that any odd perfect number must have at least 101 prime factors. Odd perfect numbers must have the largest prime factor very large. The first such lower bound was proved by Haggis[21], who proved every odd perfect number has a Prime Factor which exceeds 10^6. Iannucci[25][26], Jenkins[27], Goto and Ohno[16] proved that the largest three factors must be at least 100000007, 10007, and 101[49].

Two other related concepts are abundant numbers and deficient numbers. A positive integer n is called an abundant number if $\sigma(n) > 2n$. A positive integer n is called a deficient number if $\sigma(n) < 2n$. To study these interesting properties of these beautiful numbers we define **Abundancy Index**. That was defined by Laatsch[29]. For a positive integer n, the Abundancy Index $I(n)$ is defined as $I(n) = \frac{\sigma(n)}{n}$. More generally Abundancy Index can be considered as a measure of perfection of an integer. We can easily observe a positive integer is perfect when $I(n) = 2$ and n is abundant or deficient when $I(n) > 2$ or $I(n) < 2$ respectively. Positive integers with integer valued Abundancy indices are called **multiperfect numbers**. In this article we study different properties about Abundancy Index and to try generalize the Abundancy Index of any positive integer n.

2 Properties

Theorem 2. *The Abundancy Index function $I(n)$ is a multiplicative function.*

Proof. Let m,n be any two co-prime positive integers. Using the multiplicativity of σ function as proved in Theorem 6.3 of [8],

$$I(mn) = \frac{\sigma(mn)}{mn} = \frac{\sigma(m)\sigma(n)}{mn} = \frac{\sigma(m)}{m}\frac{\sigma(n)}{n} = I(m)I(n).$$

□

Theorem 3. (Laatsch[29])*: $I(kn) \geq I(n)$ for all $k \in \mathbb{N}$. The equality condition holds iff $k = 1$.*

Corollary 4. *Every proper multiple of a perfect number is abundant and every proper divisor of a perfect number is deficient.*

Corollary 5. *There are infinitely many abundant numbers.*

It is easy to see that there are infinitely many deficient numbers. Indeed, all prime numbers are deficient, as $\sigma(p) = p+1 < 2p$. We observe for future reference that

$$I(n) = \frac{\sigma(n)}{n} = \frac{1}{n}\sum_{d|n} d = \frac{1}{n}\sum_{d|n}\frac{n}{d} = \sum_{d|n}\frac{1}{d} \tag{1}$$

Theorem 6. (Laatsch[29])*: The $I(n)$ is function is unbounded.*

Proof. We discuss two proofs of this theorem. The first proof goes like this.

Let m be any real number. We know the series $\sum_{i=1}^{\infty}\frac{1}{i}$ is divergent. Hence for a given m there exist a natural number N such that $\sum_{i=1}^{N}\frac{1}{i} > m$. Let us take $n_0 = \text{lcm}(1,2,\cdots,N)$. Using (1) we get $I(n_0) = \sum_{d|n_0}\frac{1}{d} \geq \sum_{i=1}^{N}\frac{1}{i}$. Thus for any real $m \exists n_0 \in \mathbb{N} \ni I(n_0) > m$. Therefore $I(n)$ is not bounded above.

The second proof goes like this.

Let $n_0 = 2\cdot 3\cdots p_k = \prod_{i=1}^{k} p_i$ i.e the product of first k primes. Therefore using Theorem 2.1 we have

$$I(n_0) = \prod_{i=1}^{k}(1+\frac{1}{p_i}) > \sum_{i=1}^{k}\frac{1}{p_i}.$$

Now the series $\sum_{prime}\frac{1}{p}$ is divergent, as proven by Euler[14]. Hence we can say $I(n)$ is not bounded above. □

Theorem 7. *For any $r \in \mathbb{R}$ there are infinitely many n such that $I(n) > r$.*

Proof. By Theorem 2.3 we see for any $r \in \mathbb{R} \exists n_0 \in \mathbb{N}$ such that $I(n_0) > r$. By using Theorem 2.2 we get $I(kn_0) \geq I(n_0)$ for any positive integer k. Therefore $I(kn_0) > r$ for all $k \in \mathbb{N}$. As there are infinitely many choices for k, there are infinitely many n such that $I(n) > r$. □

Theorem 8. *If $n = \prod_{i=1}^{k} p_i^{\alpha_i}$ where the p_i are distinct primes, then $\prod_{i=1}^{k} \frac{p_i+1}{p_i} \leq I(n) \leq \prod_{i=1}^{k} \frac{p_i}{p_i-1}$*

Proof. Consider p to be a prime and α any positive integer. Now as proven earlier in (1), we have

$$I(p^\alpha) = \sum_{d|p^\alpha} \frac{1}{d} = 1 + \frac{1}{p} + \frac{1}{p^2} + \cdots + \frac{1}{p^\alpha}$$

By using the inequality

$$1 + \frac{1}{p} \leq 1 + \frac{1}{p} + \frac{1}{p^2} + \cdots + \frac{1}{p^\alpha} \leq \sum_{i=1}^{\infty} \frac{1}{p^i}$$

We get

$$\frac{p+1}{p} \leq I(p^\alpha) \leq \frac{p}{p-1} \tag{2}$$

Now since I is multiplicative function(Theorem 2.1)

$$I(n) = I(\prod_{i=1}^{k} p_i^{\alpha_i}) = \prod_{i=1}^{k} I(p_i^{\alpha_i}) \tag{3}$$

Using the inequality (2) we get

$$\prod_{i=1}^{k} \frac{p_i+1}{p_i} \leq \prod_{i=1}^{k} I(p_i^{\alpha_i}) \leq \prod_{i=1}^{k} \frac{p_i}{p_i-1}$$

Using the identity mentioned in (3)

$$\prod_{i=1}^{k} \frac{p_i+1}{p_i} \leq I(n) \leq \prod_{i=1}^{k} \frac{p_i}{p_i-1}$$

So we get our desired result. □

3 Set of Abundancy Indices

As we study the function $I : \mathbb{N} \to \mathbb{Q}$, many questions arise. For example, is every rational $q \geq 1$ the Abundancy Index of some integer? Many mathematicians have tried to study the set of Abundancy Indices, Laatsch [29] shown the set $D = \{I(n) : n \geq 2\}$ is dense in $(1,\infty)$. Later Weiner[48] showed there exists rationals which are not the Abundancy Index of any integer. In 2007 Stanton and Holdener[42] defined Abundancy Outlaw. An Abundancy Outlaw is a rational greater than 1 that is not an Abundancy Index of integer, in other words it is not in the image map of the map I.

Theorem 9. (Laatsch[29])*: $D = \{I(n) : n \geq 2\}$ is dense in $(1,\infty)$.*

A rational number $q > 1$ is said to be an Abundancy Outlaw if $I(n) = q$ has no solution in $\mathbb{N}$.

Theorem 10. (Weiner[48])*: If k is relatively prime to m and $m < k < \sigma(m)$, then $\frac{k}{m}$ is an Abundancy Outlaw. Hence if r/s is an Abundancy Index with* $\gcd(r,s) = 1$*, then $r \geq \sigma(s)$.*

Example of such outlaws given by Holdener and Stanton [42] are

$$5/4, 7/6, 9/8, 10/9, 11/6, 11/8, 11/9, 11/10, 13/8, 13/10, 13/12, 15/14, 16/15, \ldots$$

The previous theorem was also proven by Anderson[3]. The theorem implies that $\frac{k+1}{k}$ is an Abundancy Index if and only if k is prime; also $\frac{k+2}{k}$ is an Abundancy Outlaw whenever k is an odd composite number. This is a very important result shown by Weiner, which concludes that there are rationals in $(1,\infty)$ which are the not Abundancy Index of any integer. This can be proven using **Theorem 3.2**.

Theorem 11. (Wein[48])*: The set of Abundacy outlaws is dense in $(1,\infty)$.*

In the next three theorems we are giving few general forms of Abundancy Outlaw, which were studied by Holdener and Stanton [42]. These are some particular cases of proven results by Holdener [24]. For the original general results someone may look into the original paper of Holdener [24]. **Theorem 3.4** is really just the special case of **Theorem 3.5** with $p = 2$.

Theorem 12. *For all primes $p > 3$,*

$$\frac{\sigma(2p)+1}{2p}$$

is an Abundancy outlaw. If $p = 2$ or $p = 3$ then $\frac{\sigma(2p)+1}{2p}$ is an Abundancy index.

For $p = 2$ or $p = 3$, it is easy to see that $\frac{\sigma(2p)+1}{2p}$ is an Abundancy index since $I(6) = \frac{\sigma(4)+1}{4}$ and $I(18) = \frac{\sigma(6)+1}{6}$. Let us assume $p > 3$.By substituting $\sigma(2p) = 3+3p$ we can get an explicit expression. Note that $\frac{\sigma(2p)+1}{2p} = \frac{3p+4}{2p}$ is in lowest terms. Therefore if $I(N) = \frac{\sigma(2p)+1}{2p}$, then $2p|N$. Now since $p > 3$, we have $I(4p) > (\sigma(2p)+1)/2p$, so $4 \not| N$. Hence we have, $\sigma(2)|\sigma(N)$. Also note that since $\sigma(2p)+1$ is not divisible by $\sigma(2) = 3$, 3 divides N. Therefore we can write

$$I(N) > I(6p) > 2 > (\sigma(2p)+1)/2p$$

We hence arrive at a contradiction. Hence $(\sigma(2p)+1)/2p$ is an Abundancy outlaw. Example of such outlaws given by Holdener and Stanton [42] are

$$\frac{19}{10},\frac{25}{14},\frac{37}{22},\frac{43}{26},\frac{55}{34},\frac{61}{38},\frac{73}{46},\frac{91}{58},\frac{97}{62},\frac{115}{74},\frac{127}{82},\frac{133}{86},\frac{145}{94},\frac{163}{106},\frac{181}{118},\frac{187}{122}\cdots$$

Theorem 13. *For primes p,q with $q>3$, $q>p$ and* $\gcd(p,q+2)=\gcd(q,p+2)=1$*,*

$$\frac{\sigma(pq)+1}{pq}$$

is an Abundancy outlaw.

Note that if p and $q=p+2$ are twin primes then **Theorem 3.5** does not hold true. We get

$$\frac{\sigma(p(p+2))+1}{p(p+2)}=\frac{\sigma(p)+1}{p}=\frac{p+2}{p}$$

Abundancy index satisfying $I(x)=\frac{p+2}{p}$ has been studied by Ryan[39]. It is still not known whether any such example exist. The existence of such a solution is important since if $\frac{5}{3}=\frac{3+2}{3}$ is an Abundancy index then there must exist an odd perfect number. We state a state an important result of Weiner which proves this.

Theorem 14. (Weiner[48])*: If there is a positive integer n with $I(n)=5/3$, then $5n$ is an odd perfect number.*

This theorem was further generalized by Ryan[40].

Theorem 15. (Ryan[40])*: If there exist positive integers m and n such that m is odd, $2m-1$ is prime, $2m-1$ does not divide n, and $I(n)=(2m-1)/m$, then $n(2m-1)$ is an odd perfect number.*

He further showed that if m is even but not a power of 2 then $(2m-1)/m$ is an Abundancy Outlaw. Both of these results are further generalized by Holdener[24].

Theorem 16. (Holdener[24])*: There is an odd perfect number if and only if there are positive integers p,n, and k such that p is prime and does not divide n and also satisfies $p\equiv k\equiv 1(\bmod 4)$, and*

$$I(n)=\frac{2p^k(p-1)}{p^{k+1}-1}$$

A similar example can be made about **Theorem 3.5** as we have done earlier for **Theorem 3.4**. For this we assume that the two odd primes p,q, satisfying $q\equiv 1(\bmod p)$. Then $p\not| q+2$ and $q\not| p+2$ .Now by Dirichlet's theorem on arithmetic progressions of primes, we know that there are infinitely many such pairs of odd primes p,q. Example of such outlaws given by Holdener and Stanton [42] are
For $p=5$

$$\frac{73}{55},\frac{193}{155},\frac{253}{205},\frac{373}{305},\frac{433}{355},\frac{613}{505},\frac{793}{655},\frac{913}{755},\frac{1093}{905},\frac{1153}{955},\frac{1273}{1055},\frac{1513}{1255},\frac{1633}{1355}$$

$$\frac{1693}{1405},\frac{1873}{1555},\frac{1993}{1655},\frac{2413}{2005},\frac{2533}{2105},\cdots$$

For $p = 7$

$$\frac{241}{203}, \frac{353}{301}, \frac{577}{497}, \frac{913}{791}, \frac{1025}{889}, \frac{1585}{1379}, \frac{1697}{1477}, \frac{1921}{1673}, \frac{2257}{1967}, \frac{2705}{2359}, \frac{3041}{2653}, \frac{3377}{2947}, \frac{3601}{3143}$$

$$\frac{3713}{3241}, \frac{3937}{3437}, \frac{4385}{3829}, \frac{4945}{4319}..$$

For $p = 11$

$$\frac{289}{253}, \frac{817}{737}, \frac{1081}{979}, \frac{2401}{2189}, \frac{3985}{3641}, \frac{4249}{3883}, \frac{4777}{4367}, \frac{5041}{4609}, \frac{5569}{5093}, \frac{7417}{6787}, \frac{7945}{7271}, \frac{8209}{7513}, \frac{8737}{7997}$$

$$\frac{10321}{9449}, \frac{10585}{9691}, \frac{11377}{10417}, \ldots$$

Theorem 17. *If N is an even perfect number, then $\frac{\sigma(2N)+1}{2N}$ is an abundancy outlaw.*

4 Superabundant Numbers

A positive integer n is called superabundant if $I(m) < I(n)\ \forall m < n$.

The first few superabundant numbers are 1, 2, 4, 6, 12, 24, 36, 48, 60, 120, 180. Ramanujan [34][35][36] in 1915 first introduced the idea of superabundant numbers. In 30 pages of Ramanujan's paper "Highly Composite Numbers" Ramanujan defined generalized highly composite numbers, which is a generalized case of superabundant numbers. Ramanujan's work remained unpublished till 1997 when it was published in Ramanujan Journal. The idea of Superabundant numbers were also independently defined by Alaoglu and Erdős [2] in 1944, who are unknown to the unpublished work done by Ramanujan earlier in 1915. They proved that if n is superabundant, then there exist a k and $a_1, a_2, ..., a_k$ satisfying $a_1 \geq a_2 \geq \cdots \geq a_k \geq 1$ such that

$$n = \prod_{i=1}^{k} (p_i)^{a_i}$$

where p_i is the i-th prime number, and

Theorem 18. *There are infinitely many superabundant numbers.*

Proof. Let us assume there are finitely many superabundant numbers and n is the largest superabundant number. So $I(m) < I(n)$ for all $m < n$. Now let us consider the integer $2n$. By Theorem 2 we know $I(2n) > I(n)$. So $I(m) < I(2n)$. But $2n$ cannot be a superabundant number. So $\exists n_0 \ni I(n_0) > I(2n)$ and $n < n_0 < 2n$. Let us consider the least n_0. We know

$$I(n_0) > I(2n) > I(n) > I(m) \text{ for all } m < n$$

n_0 cannot be a superabundant number. Hence there exist a real number n_1 such that $I(n_0) > I(n_1)$ and $n < n_1 < n_0$. It is easy to see $I(n_1) > I(2n)$ and $n < n_1 < 2n$. But we had assumed n_0 to be least such integer. Hence we get a contradiction. □

So we can conclude there are infinitely many superabundant numbers. This type of numbers can be further generalized as colossally abundant numbers.

A number n is colossally abundant if and only if there is an $\varepsilon > 0$ such that for all $k > 1$,

$$\frac{\sigma(n)}{n^{1+\varepsilon}} \geq \frac{\sigma(k)}{k^{1+\varepsilon}}$$

Therefore all colossally abundant numbers are also superabundant numbers, but all superabundant numbers may not be a colossally abundant number. For every $\varepsilon > 0$ the function $\frac{\sigma(n)}{n^{1+\varepsilon}}$ has a maximum and that these maxima will increase as ε tends to zero. Thus there are infinitely many colossally abundant numbers [30].

Now we draw a connection between superabundant numbers and well known Riemann Hypothesis[37], which is considered as one of the most important unsolved problems in Mathematics. The Riemann Hypothesis conjectures that the Riemann zeta function defined as

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \ldots$$

valid when the real part of s exceeds 1 has non-trivial zeros only at the complex numbers with real part $\frac{1}{2}$. This conjecture is of significant interest to number theorists since this result has direct consequences in the distribution of prime numbers.

In 1984 Robin [38] proved a surprising result. He showed an equivalence between Riemann Hypothesis and a bound to the Abundancy Index.

Theorem 19. (Robin[38])*: For $n \geq 3$ we have $I(n) < e^{\gamma} \log\log n + \frac{0.6483}{\log\log n}$.*

Theorem 20. (Robin[38])*: The Riemann Hypothesis is true if and only if $I(n) < e^{\gamma} \log\log n$ for all $n \geq 5041$.*

Note: Here γ denotes Euler's Gamma Constant(also known as Euler–Mascheroni constant). It is the limiting difference between the the natural logarithm and harmonic series.

$$\gamma = \lim_{x \to \infty} (-\ln x + \sum_{k=1}^{x} \frac{1}{k})$$

The value of Euler's Gamma Constant is approximately 0.57721[41]. **Theorem 4.3**(Robin's Inequality) is the most striking result here, it gives an alternative approach to prove or disprove Riemann's Hypothesis, one of the greatest problems in Number Theory.

This result by Robin's inequality is supported by many other findings. Gronwall [18] found that

$$\limsup_{n \to \infty} \frac{I(n)}{e^{\gamma} \log\log n} = 1$$

Wojtowicz[50] further showed that the values of $f(n) = \frac{I(n)}{e^{\gamma} \log\log n}$ are close to 0 on a set of asymptotic density 1. An alternate version of Robin's inequality equivalent to

Riemann Hypothesis was found by Lagarias[30], who showed the equivalence of the Riemann hypothesis to an sequence of elementary inequalities involving the harmonic numbers H_n, the sum of the reciprocals of the integers from 1 to n:

$$\sigma(n) \leq e^{H_n} \log H_n + H_n \text{ for all } n \geq 1$$

Another alternate version of Robin's inequality is by Choie et.al [10] who have shown that the RH holds true if and only if every natural number divisible by a fifth power greater than 1 satisfies Robin's inequality. Briggs[7] describe a computational study of the successive maxima of $I(n)$. They found that the maxima of this function occur at superabundant and colossally abundant numbers and studied the density of these numbers. He then compared this with the known maximal order of $f(n)$ and found out a condition equivalent to the Riemann Hypothesis using these data.

Theorem 21. (Akbary[1])*: If there is any counterexample to Robin's inequality then the least such counterexample is a superabundant number.*

Let $S(x)$ be the number of superabundant numbers not exceeding x.

From the proof of **Theorem 4.1**, we get the inequality $S(x) \geq \log x$, since the spacing grows at most exponentially. This gives $\log x$ as the lower bound to the counting function $S(x)$. Note that **Theorem 4.4** helps us find a counterexample of the Robin's inequality by limiting our attention to only superabundant numbers. Unfortunately there is no algorithm find superabundant numbers except finding it using **Definition 4.1**. Some results in the distribution of the superabundant numbers are therefore very helpful. We now state two results in that regard.

Theorem 22. (Alaoglu[2])*:* $S(x) > c\frac{\log x \log\log x}{(\log\log\log x)^2}$

Erdős and Nicholas [12] proved a more stronger inequality.

Theorem 23. (Nicholas[12])*:* $S(x) > (\log x)^{1+\delta}$ $(x > x_0)$ *for every* $\delta < 5/48$.

So we finally see that abundancy index and superabundant numbers have a very close connection with Riemann Hypothesis. One may try to prove or disprove Riemann Hypothesis with the help of **Theorem 4.3**. To disprove Riemann's Hypothesis it enough to get a counterexample to Robin's inequality. One might try to find it computationally and **Theorem 4.4** will definitely make his or her job easier.

Bibliography

[1] Akbary, A., and Friggstad, Z. (2009). Superabundant Numbers and the Riemann Hypothesis. The American Mathematical Monthly, 116(3), 273-275.

[2] Alaoglu, L., and Erdos, P. (1944). On Highly Composite and Similar Numbers. Transactions of the American Mathematical Society, 56(3), 448-469.

[3] Anderson C.W., The solution of $\sum(n) = \sigma(n)/n = a/b, \Phi(n) = \phi(n)/n = a/b$ and some related considerations, unpublished manuscript, 1974

[4] Ball, W. W. R. and Coxeter, H. S. M. Mathematical Recreations and Essays, 13th ed. New York: Dover, 1987.

[5] Brent, R. P. and Cohen, G. L. "A New Bound for Odd Perfect Numbers." Math. Comput. 53, 431-437 and S7-S24, 1989.

[6] Brent, R. P.; Cohen, G. L.; te Riele, H. J. J. "Improved Techniques for Lower Bounds for Odd Perfect Numbers." Math. Comput. 57, 857-868, 1991.

[7] Briggs,K. Abundant numbers and the Riemann hypothesis. Experiment. Math. 15 (2006),251–256.

[8] Burton, David M. Elementary Number Theory, 7th Edition, McGraw-Hill, 2011

[9] Caldwell, Chris K., "A proof that all even perfect numbers are a power of two times a Mersenne prime", Prime Pages, retrieved 2014-12-02.

[10] Choie Y., Lichiardopol N., Moree P., and Sol ́e P., On Robin's criterion for the Riemann hypothesis, J. Th ́eor. Nombres Bordeaux 19 (2007), 357–372.

[11] Dickson, L. E. History of the Theory of Numbers, Vol. 1: Divisibility and Primality. New York: Dover, pp. 3-33, 2005.

[12] Erdős, P. and Nicolas, J.-L. R ́epartition des nombres superabondants, Bull. Soc. Math. France 103 (1975) 65–90.

[13] Euler, L. (1849), "De numeris amicibilibus" [On amicable numbers], Commentationes arithmeticae (in Latin), 2, pp. 627–636.

[14] Euler L. (1737). "Variae observationes circa series infinitas" [Various observations concerning infinite series]. Commentarii Academiae Scientiarum Petropolitanae. 9: 160–188.

[15] Gerstein, Larry (2012), Introduction to Mathematical Structures and Proofs, Undergraduate Texts in Mathematics, Springer, Theorem 6.94, p. 339, ISBN 9781461442653.

[16] Goto, T. and Ohno, Y. "Odd Perfect Numbers Have a Prime Factor Exceeding 10^8" Preprint, Mar. 2006. https://www.ma.noda.tus.ac.jp/u/tg/perfect.html.

[17] Greathouse C. and Weisstein E. W., "Odd Perfect Number." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/OddPerfectNumber.html

[18] Gronwall T.H., Some asymptotic expressions in the theory of numbers, Trans. Amer. Math. Soc. 14 (1913), 113–122.

[19] Hagis, P. Jr. "A Lower Bound for the Set of Odd Perfect Numbers." Math. Comput. 27, 951-953, 1973.

[20] Hagis, P. Jr. "An Outline of a Proof that Every Odd Perfect Number has at Least Eight Prime Factors." Math. Comput. 34, 1027-1032, 1980.

[21] Hagis, P. Jr.; and Cohen, G. L. "Every Odd Perfect Number Has a Prime Factor Which Exceeds 10^6." Math. Comput. 67, 1323-1330, 1998.

[22] Hare, K. "New Techniques for Bounds on the Total Number of Prime Factors of an Odd Perfect Number." Math. Comput. 74, 1003-1008, 2005.

[23] Holdener, J.A. "A Theorem of Touchard and the Form of Odd Perfect Numbers." Amer. Math. Monthly 109, 661-663, 2002.

[24] Holdener J.A., "Conditions equivalent to the existence of oddperfect numbers", Math. Mag.79(2006), 389–391

[25] Iannucci, D. E. "The Second Largest Prime Divisor of an Odd Perfect Number Exceeds Ten Thousand." Math. Comput. 68, 1749-1760, 1999.

[26] Iannucci, D. E. "The Third Largest Prime Divisor of an Odd Perfect Number Exceeds One Hundred." Math. Comput. 69, 867-879, 2000.

[27] Jenkins, P. M. "Odd Perfect Numbers Have a Prime Factor Exceeding 10^7." Math. Comput. 72, 1549-1554, 2003.

[28] Kanold, H.-J. "Über mehrfach vollkommene Zahlen. II." J. reine angew. Math. 197, 82-96, 1957.

[29] Laatsch, R. (1986). Measuring the Abundancy of Integers. Mathematics Magazine, 59(2), 84-92.

[30] Lagarias J.C. , An elementary problem equivalent to the Riemann hypothesis, American Mathematical Monthly 109 (2002), pp. 534–543.

[31] Nielsen, P. P. "Odd Perfect Numbers Have at Least Nine Distinct Prime Factors." 22 Feb 2006. https://arxiv.org/abs/math.NT/0602485.

[32] Norton, K. K. "Remarks on the Number of Factors of an Odd Perfect Number." Acta Arith. 6, 365-374, 1960.

[33] Ochem, P. and Rao, M. "Odd Perfect Numbers Are Greater than $10^(15000)$." Math. Comput. 81, 1869-1877, 2012.

[34] Ramanujan S., Highly composite numbers, Proc. Lond. Math. Soc. 14 (1915), 347–407.

[35] Ramanujan S., Collected Papers, Chelsea, 1962.

[36] Ramanujan S. (annotated by J.-L. Nicolas and G. Robin), Highly composite numbers, Ramanujan J. 1 (1997), 119–153.

[37] Riemann, B. (1859), "Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse", Monatsberichte der Berliner Akademie. In Gesammelte Werke, Teubner, Leipzig (1892)

[38] Robin, G . Grandes valeurs de la fonction somme de diviseurs et hypoth'ese de Riemann,J. Math. Pure Appl. (9) 63 (1984) 187–213.

[39] Ryan R., Results concerning uniqueness for $\sigma(x)/x = \sigma(p^n * q^m)/(p^n * q^m)$ and related topics, Int. Math. J. 2 (2002), 497–514.

[40] Ryan R., A Simpler Dense Proof Regarding the Abundancy Index, Mathematics Magazine, Vol. 76, No. 4 (2003), 299-301.

[41] Sloane, N. J. A. (ed.). "Sequence A001620 (Decimal expansion of Euler's constant (or the Euler-Mascheroni constant), gamma)". The On-Line Encyclopedia of Integer Sequences. OEIS Foundation.

[42] Stanton, W. and Holdener, J. (2007). Abundancy "outlaws" of the form $\frac{\sigma(N)+t}{N}$. Journal of Integer Sequences [electronic only].

[43] Steuerwald, R. "Verscharfung einen notwendigen Bedingung fur die Existenz einen ungeraden vollkommenen Zahl." Sitzungsber. Bayer. Akad. Wiss., 69-72, 1937.

[44] Touchard, J. "On Prime Numbers and Perfect Numbers." Scripta Math. 19, 35-39, 1953.

[45] Travaglini, Giancarlo (2014), Number Theory, Fourier Analysis and Geometric Discrepancy, London Mathematical Society Student Texts, 81, Cambridge University Press, pp. 26–27, ISBN 9781107044036.

[46] Tuckerman, B. "Odd Perfect Numbers: A Search Procedure, and a New Lower Bound of $10^{(}36)$." Not. Amer. Math. Soc. 15, 226, 1968.

[47] Voight, J. "On the Nonexistence of Odd Perfect Numbers." MASS Selecta. Providence, RI: Amer. Math. Soc., pp. 293-300, 2003.

[48] Weiner, P. (2000). The Abundancy Ratio, a Measure of Perfection. Mathematics Magazine, 73(4), 307-310.

[49] Weisstein, Eric W. "Fermat's 4n+1 Theorem." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/Fermats4nPlus1Theorem.html

[50] Wojtowicz, Marek. (2007). Robin's inequality and the Riemann hypothesis. Proceedings of the Japan Academy, Series A, Mathematical Sciences. 83. 10.3792/pjaa.83.47.

[51] Yamada, T. "On the Divisibility of Odd Perfect Numbers by a High Power of a Prime." 16 Nov 2005. https://arxiv.org/abs/math.NT/0511410.

Ball State Undergraduate Mathematics Exchange
https:// digitalresearch.bsu.edu/mathexchange
Vol. 15, No. 1 (Fall 2021)
Pages 40 –53

Arithmetic Digit Manipulation and The Conway Base-13 Function

*Lyam K. Boylan**

Lyam K. Boylan is an undergraduate student studying theoretical mathematics and philosophy at the University of Victoria with primary interests on the implications set theory has about the universe. He enjoys finding connections between math and his other hobbies, which include creative writing, composing, chess, and video production. He values the timeless and eternally beautiful flow of patterns that connects all corners of math to each other.

Abstract

Despite positional notation being the primary way we represent numbers, it's not trivial to perform a variety of digit-manipulation with arithmetic alone. The Conway Base-13 Function is a prime example of a function who's definition is easily said in plain language, but difficult to formulate with arithmetic alone. To emphasize the difficulty, we construct a closed-form function equivalent to the Base-13 function over the integers, comprising only of arithmetic.

1 Introduction

Created by the great and late John H. Conway, the Conway Base 13 Function, $f : \mathbb{R} \to \mathbb{R}$, is a counterexample to the converse of the Intermediate Value Theorem. Despite f being discontinuous everywhere, it satisfies that for any interval (a,b), f takes all values between $f(a)$ and $f(b)$. In fact, f takes all values in $\mathbb{R}$ within every interval of non-zero length. Such a function can be defined in plain language in terms of digit-manipulation with relative ease, yet formulating f using arithmetic to perform such digit-manipulation is more difficult. Hence, the purpose of this article is to emphasize such difficulty by constructing a closed-form function equivalent to f over the integers, comprising only of arithmetic.

Imperatively, a summary of a definition will be given. Hence, let the set of digits in any base, $b \in \mathbb{Z}_{>1}$, be denoted

$$U_b = \{0, \dots, b-1\}.$$

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

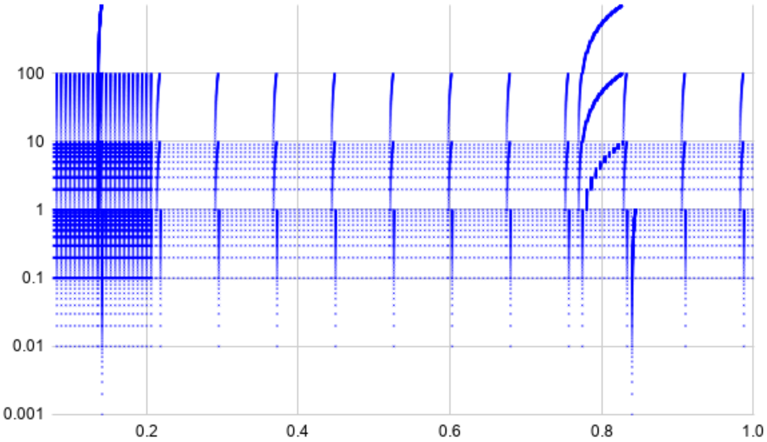


Figure 1: Log-plot of f over a subset of $\mathbb{Z}[\frac{1}{13}]$.

Thus, the set of decimal and tridecimal digits are $U_{10} = \{0,1,2,3,4,5,6,7,8,9\}$ and $U_{13} = \{0,1,2,3,4,5,6,7,8,9,A,B,C\}$, respectively. The digits A, B, and C correspond to their decimal equivalents 10, 11, and 12.

Suppose all $x \in \mathbb{R}_{\geq 0}$ have base-b expansions of the form

$$x = \ldots d_1 d_0 . d_{-1} d_{-2} \ldots_{(b)} \text{ s.t. } \sum_{k \in \mathbb{Z}} b^k d_k = x$$

where $d_k \in U_b$ are individual digits for all $k \in \mathbb{Z}$. Note that d_k corresponds to a digit to the left of the radix point only when $k \geq 0$. In reference to the position of a digit, the term *index* is used. A digit at the k^{th} index of an expansion refers to the digit k positions to the left of the units' column. Hence, a digit at index 0 is a digit in the units column. If no digit appears at the k^{th} index, the digit is assumed to be zero. Considering that some values of x and b have two expansions (such as in the cases $0.\overline{9}_{(10)} = 1_{(10)}$ or $1.2A\overline{C}_{(13)} = 1.2B_{(13)}$), we'll assume the terminating expansion is always preferred. For brevity, we'll introduce the notation $d_{j \to k(b)}$ as shorthand for $d_j d_{j-1} \ldots d_{k+1} d_{k(b)}$. Furthermore, let $d_{j \to k(b)} \subseteq x$ represent, disregarding sign and radix point, that the sequence of digits $d_{j \to k(b)}$ occurs in the base-b expansion of x. For example

$$ABC_{(13)} \subseteq -A.BC_{(13)}.$$

If $x \in \mathbb{Z}_{\geq 0}$, then $k < 0 \implies d_k = 0$. Hence, for non-negative integer values, the base-b expansion of x can simply be written $d_{m \to 0(b)}$, where $m \in \mathbb{Z}_{\geq 0}$ is the largest index such that $d_m \neq 0$ (assuming $x \neq 0$, otherwise $m = 0$).

Adapted from a definition by Greg Oman [1], the Base-13 function f is defined in plain language as follows:

For any $x \in \mathbb{R}$, $k \in \mathbb{Z}$, let d_k represent the digit at index k in the tridecimal expansion of $|x|$. A few cases are considered:

- **Case 1:** Suppose there exists a digit $A \subseteq |x|$, such that all digits to the right of such do not contain A or B, and there exists exactly one $C \subseteq |x|$ to the right of such A. Let the digits between A and C be denoted $d_{j_A-1\to j_C+1}$, where j_A and j_C are the respective indices of such A and C. Let the digits after C be denoted $d_{j_C-1\to-\infty}$. Let $f(x) = +d_{j_A-1\to j_C+1}.d_{j_C-1\to-\infty(10)}$.
- **Case 2:** Suppose there exists a digit $B \subseteq |x|$, such that all digits to the right of such do not contain A or B, and there exists exactly one $C \subseteq |x|$ to the right of such B. Let the digits between B and C be denoted $d_{j_B-1\to j_C+1}$, where j_B and j_C are the respective indices of such B and C. Let the digits after C be denoted $d_{j_C-1\to-\infty}$. Let $f(x) = -d_{j_B-1\to j_C+1}.d_{j_C-1\to-\infty(10)}$.
- **Otherwise:** $f(x) = 0$ if x is not of either form.

Here, $d_{j_C-1\to-\infty}$ is shorthand for $\lim_{k\to-\infty} d_{j_C-1\to k(13)}$. It's important to recognize that the final result in cases 1 and 2 are decimal expansions, despite using digits from the tridecimal expansion of $|x|$. This is possible because in either case, the result only uses digits after the right-most A or B. Hence, the proceeding digits do not contain A or B. There's expectantly exactly one proceeding C, (the only other possible tridecimal digit which isn't also a decimal digit) however, which incidentally is excluded in the result. Hence, all digits in the result are indeed decimal. Essentially, f is a recompilation of some of the decimal digits in the tridecimal expansion of $|x|$, using a specific C (if it exists) as a decimal point, and A or B as the sign. Here are a few examples that cover all cases:

$$\begin{aligned}
f(-B1A.3C1415\ldots_{(13)}) &= \pi \\
f(137_{(13)}) &= 0 \\
f(0.B17C11_{(13)}) &= -17.11_{(10)} \\
f(0.\overline{A1C1}_{(13)}) &= 0 \\
f(0.A1\overline{C1}_{(13)}) &= 0 \\
f(0.A999C\overline{9}_{(13)}) &= 1000_{(10)}
\end{aligned}$$

It may be easy to see why f passes through all values of $\mathbb{R}$ within every non-zero-length interval. Regardless, proofs of its properties are not the purpose of this paper. Since the digit manipulation in f is not trivial, the ability to define f using only standard mathematical operations is not immediately clear.

Theorem 1. *There exists a closed-form $g : \mathbb{Z} \to \mathbb{R}$ such that $g \subseteq f$, where f is the Conway Base-13 Function.*

Understandably, such a prospect would benefit from quantifying its cases. The condition of the existence of a digit A or $B \subseteq |x|$, such that all digits to the right of such do not contain A or B can be quantified as $\exists j_A \big[\ d_{j_A} = A \wedge\ \nexists k < j_A(d_k \in \{A,B\})\big]$ or $\exists j_B \big[\ d_{j_B} = B \wedge\ \nexists k < j_B(d_k \in \{A,B\})\big]$ respectively. With the added condition that there exists exactly one $C \subseteq |x|$ to the right of such A or B, the cases become

$$\text{case 1} \iff \exists j_A \Big[\ d_{j_A} = A \wedge\ \nexists k < j_A(d_k \in \{A,B\}) \wedge \exists! j_C < j_A(d_{j_C} = C)\Big]$$

$$\text{case 2} \iff \exists j_B \Big[\ d_{j_B} = B \wedge\ \nexists k < j_B(d_k \in \{A,B\}) \wedge \exists! j_C < j_B(d_{j_C} = C)\Big]$$

This gives rise to an equivalent piecewise formulation:

$$f(x) = \begin{cases} +d_{j_A-1\to j_C+1}.d_{j_C-1\to-\infty(10)} & : \text{case } 1 \\ -d_{j_B-1\to j_C+1}.d_{j_C-1\to-\infty(10)} & : \text{case } 2 \\ 0 & : \text{otherwise} \end{cases}$$

2 Closed Form Expressions

As indicated and motivated by Nate Eldredge [4], a construction of f using only arithmetic functions is a possible procedure, albeit tedious and logic-heavy. It requires quite the array of functions designed to arbitrarily manipulate digits and test for logical conditions. This does not guarantee that the procedure will have a closed form over the entirety of $\mathbb{R}$, however it does give credence for a closed form over $\mathbb{Z}$.

2.1 Closed Form Operations

As there is no universal definition for closed-form expressions, we assume a conservative definition.

Definition 2. Let an operation be considered closed-form if it can be equivalently expressed in a finite number of operations, of which include addition, subtraction, multiplication, division, exponentiation, principal roots, and the principal branch of the logarithm.

This definition is restrictive so that operations that fulfill this conservative definition expectantly fulfill more liberal ones [2]. As evident in following sections, a significant number of arithmetic digit manipulation relies on the floor and ceiling operations. These can be defined through the use of their relationship to the modulo operation in floored division [3]:

$$\lfloor x \rfloor := x - (x \bmod 1),$$
$$\lceil x \rceil := x + ((-x) \bmod 1).$$

Here, mod is used as a binary operation as opposed to its use in congruence relations. It can be defined though the use of the periodic nature of the principal branch of the logarithm

$$x \bmod y := \frac{y}{2\pi i}\mathrm{Log}\left(e^{\frac{2\pi i x}{y}}\right)$$

assuming $0 \leq \frac{1}{i}\mathrm{Log}\left(e^{i\theta}\right) < 2\pi \ \forall\theta \in \mathbb{R}$. Hence, the floor, ceiling, and modulo operations will be considered closed-form. Similarly, the absolute value operation can be defined closed-form through the use of the principal square root, $|x| := \sqrt{x^2}$.

2.2 Logical-Conditional Functions

Given the natural piecewise definition of the Base-13 function, a multitude of functions that act for testing logical conditions are constructed. In particular, we construct functions that check for equality and inequality relations between two real numbers.

Definition 3. Let E, the "equivalence function", be defined as

$$E(a,b) := \lfloor (1+\varepsilon)^{-|a-b|} \rfloor \text{ such that } \varepsilon > 0,\ \forall a,b \in \mathbb{R}.$$

It is easily shown that

$$E(a,b) = \begin{cases} 1 & : a = b \\ 0 & : a \neq b \end{cases}$$

For brevity, the "negation" of the equivalence function will also be used.

Definition 4. Let N, the "non-equivalence function", be defined as

$$N(a,b) := 1 - E(a,b)\ \forall a,b \in \mathbb{R}.$$

Similarly,

$$N(a,b) = \begin{cases} 1 & : a \neq b \\ 0 & : a = b \end{cases}$$

Definition 5. Let G_E, the "greater-than or equal-to function", be defined as

$$G_E(a,b) := \left\lfloor \frac{1}{2} + \frac{1}{1+(1+\varepsilon)^{b-a}} \right\rfloor \text{ such that } \varepsilon > 0,\ \forall a,b \in \mathbb{R}.$$

Although not as trivial as the equivalence function, it can be shown that

$$G_E(a,b) = \begin{cases} 1 & : a \geq b \\ 0 & : a < b \end{cases}$$

Definition 6. Let M, the "minimum function", be defined as

$$M(a,b) := aG_E(b,a) + bG_E(a,b) - aE(a,b),\ \forall a,b \in \mathbb{R}.$$

By definition of G_E, it is clear that

$$M(a,b) = \begin{cases} a & : a \leq b \\ b & : a > b \end{cases}$$

These functions enable the ability to arithmetically test for logical conditions. With such, some digit manipulation that is naturally a more piecewise procedure, may instead be done entirely arithmetically.

3 Digit Manipulation

Singling-out digits from an expansion is the most critical ability of digit manipulation. As such, let us introduce the following closed-form functions:

Definition 7. Let $\overleftarrow{T}$, the "trailing-digit-truncation function", be defined as

$$\overleftarrow{T}_b^n(x) := \left\lfloor \frac{x}{b^n} \right\rfloor$$

$$\forall x \in \mathbb{Z}_{\geq 0}, \text{ for any base } b \in \mathbb{Z}_{>1}, \text{ and any digit-index } n \in \mathbb{Z}_{\geq 0}.$$

In essence, $\overleftarrow{T}$ removes the right-most n digits from a base-b expansion of x. More formally, it removes digits with indices less than a given index n.

Lemma 8. $x = d_{m \to 0(b)} \implies \overleftarrow{T}_b^n(x) = d_{m \to n(b)}$

Proof: Suppose $x = d_{m \to 0(b)}$. By definition of positional notation, $x = \sum_{k=0}^m b^k d_k$. Plugging this into $\overleftarrow{T}$ yields

$$\overleftarrow{T}_b^n(x) = \left\lfloor \frac{\sum_{k=0}^m b^k d_k}{b^n} \right\rfloor = \left\lfloor \sum_{k=0}^m b^{k-n} d_k \right\rfloor$$

which can be split into a whole and fractional part.

$$\begin{aligned}
&= \left\lfloor \sum_{k=n}^m b^{k-n} d_k + \sum_{k=0}^{n-1} b^{k-n} d_k \right\rfloor \\
&= \sum_{k=n}^m b^{k-n} d_k + \left\lfloor \sum_{k=0}^{n-1} b^{k-n} d_k \right\rfloor \\
&= \sum_{k=n}^m b^{k-n} d_k
\end{aligned}$$

We are left with a recompilation of the digits $d_{m \to n}$, such that d_n is now directly to the left of the radix point. In our notation, this is written $d_{m \to n(b)}$. □

For example, $\overleftarrow{T}_{10}^2(123456_{(10)}) = 1234_{(10)}$. In conjunction, the selection of an arbitrary digit at a given index is possible.

Definition 9. Let D, the "digit-selection function", be defined as

$$D_b^n(x) := \overleftarrow{T}_b^n(x) - b\overleftarrow{T}_b^{n+1}(x)$$

$$\forall x \in \mathbb{Z}_{\geq 0}, \text{ for any base } b \in \mathbb{Z}_{>1}, \text{ and any digit-index } n \in \mathbb{Z}_{\geq 0}.$$

This grants the ability to retrieve a digit at the n^{th} index of the base-b expansion of x within a closed-form manner. This ability is most critical in construction of the Base-13 Function.

Lemma 10. $x = d_{m \to 0(b)} \implies D_b^n(x) = d_n$

Proof: Suppose $x = d_{m\to 0(b)}$. Using Lemma 8, D becomes

$$\begin{aligned} D_b^n(x) &= \sum_{k=n}^{m} b^{k-n} d_k - b \sum_{k=n+1}^{m} b^{k-n-1} d_k \\ &= \sum_{k=n}^{m} b^{k-n} d_k - \sum_{k=n+1}^{m} b^{k-n} d_k \\ &= d_n + \sum_{k=n+1}^{m} b^{k-n} d_k - \sum_{k=n+1}^{m} b^{k-n} d_k \\ &= d_n \end{aligned}$$

□

For example, $D_{10}^2(123456_{(10)}) = 4_{(10)}$. Not surprisingly, the number of digits in an expansion can also be deduced arithmetically.

Definition 11. Let L, the "length function", be defined as

$$L_b(x) := \lceil \log_b(x+1) \rceil + E(x,0)$$

$$\forall x \in \mathbb{Z}_{\geq 0}, \text{ for any base } b \in \mathbb{Z}_{>1}.$$

This is variant of the usual method to count the number of digits: $\lfloor \log_b(x) \rfloor + 1$. However the latter is undefined for the case $x = 0$, whereas $L_b(0) = 1$. Otherwise both methods are equivalent over the positive integers.

Lemma 12. $x = d_{m\to 0(b)} \land x > 0 \implies L_b(x) = m+1$

Proof: Suppose $x = d_{m\to 0(b)} \land x > 0$,

$$\begin{aligned} &\implies L_b(x) = \left\lceil \log_b \left(1 + \sum_{k=0}^{m} b^k d_k\right) \right\rceil \\ &\implies \left\lceil \log_b(b^m) \right\rceil < L_b(x) \leq \left\lceil \log_b(b^{m+1}) \right\rceil \\ &\implies m < L_b(x) \leq m+1 \\ &\implies L_b(x) = m+1 \end{aligned}$$

□

For example, $L_{10}(10_{(10)}) = L_{10}(99_{(10)}) = 2$. If $x \in \mathbb{Z}_{\geq 0}$ and $d \in U_b$, then functions D, E, L can be used to count the occurrences of d in the base-b expansion of x.

Definition 13. Let O, the "digit-occurrence-counting function", be defined as

$$O_b^p(x) := \sum_{k=0}^{L_b(x)-1} E(D_b^k(x), p)$$

$$\forall x \in \mathbb{Z}_{\geq 0}, \text{ for any base } b \in \mathbb{Z}_{>1}, \text{ and any digit } p \in U_b.$$

It should be evident that as O loops through all possible indices, k, for digits in the base-b expansion of x, the summation increments by 1 iff the digit at index k is equivalent to the given digit p, which we are looking to count the occurrences of. In other words, O counts the number of occurrences of a digit p in the base-b expansion of x.

Less trivial is a method to deduce the a specific index of an occurrence of a given digit.

Definition 14. Let I, the "digit-occurrence-index function", be defined as

$$I_b^p(x) := \sum_{k=1}^{L_b(x)} E\Big(O_b^p\big(\overleftarrow{T}_b^k(x)\big), O_b^p(x)\Big)$$

$$\forall x \in \mathbb{Z}_{\geq 0},\ \text{for any base } b \in \mathbb{Z}_{>1},\ \text{and any digit } p \in U_b.$$

The purpose of I is to return the index of the right-most digit p in the base-b expansion of x. If there isn't such an index, then I returns $L_b(x)$, which is by definition a number higher than the maximum index of a nonzero digit.

Lemma 15.

$$x = d_{m \to 0(b)} \implies I_b^p(x) = \begin{cases} j & : \exists j\big[d_j = p \wedge \forall k < j(d_k \neq p)\big] \\ L_b(x) & : \text{otherwise} \end{cases}$$

Proof: Suppose $x = d_{m \to 0(b)}$. We'll look at the case where there does exist a right-most digit p in the base-b expansion of x.

Case 1: $\exists j\big[d_j = p \wedge \forall k < j(d_k \neq p)\big]$

Thus, with such a digit having index j, truncating off digits of x with indices less than k for $k \leq j$, yields a number with no occurrences of p removed. Likewise, truncating for $k > j$ yields a number with at least one less occurrence of p.

$$k \leq j \iff O_b^p\big(\overleftarrow{T}_b^k(x)\big) = O_b^p(x)$$
$$\implies E\Big(O_b^p\big(\overleftarrow{T}_b^k(x)\big), O_b^p(x)\Big) = \begin{cases} 1 & : k \leq j \\ 0 & : k > j \end{cases}$$

Thus, the summation can be split into

$$I_b^p(x) = \sum_{k=1}^{j} 1 + \sum_{k=j+1}^{L_b(x)} 0 = j$$

Resulting in the index, j.

Case 2: $\nexists j\big[d_j = p \wedge \forall k < j(d_k \neq p)\big]$

In the other case, since x is an integer of finite digits, there not being a right-most digit p implies that there are no occurrences.

$$\begin{aligned} O_b^p\big(\overleftarrow{T}_b^k(x)\big) &= O_b^p(x) = 0\ \forall k \\ \implies E\Big(O_b^p\big(\overleftarrow{T}_b^k(x)\big), O_b^p(x)\Big) &= 1 \\ \implies I_b^p(x) &= \sum_{k=1}^{L_b(x)} 1 \\ &= L_b(x). \end{aligned}$$

As such, the sum is trivially the bound, $L_b(x)$.

$$\text{Therefore } x = d_{m\to 0(b)} \implies I_b^p(x) = \begin{cases} j & : \exists j\big[d_j = p \land \forall k < j(d_k \neq p)\big] \\ L_b(x) & : \text{otherwise} \end{cases}$$

□

In parody to $\overleftarrow{T}$, we'll define a function that virtually removes the left-most n digits from a base-b expansion of x.

Definition 16. Let $\overrightarrow{T}$, the "leading-digit-truncation function", be defined as

$$\overrightarrow{T}_b^n(x) := \sum_{k=0}^{L_b(x)-n-1} b^k D_b^k(x)$$

$$\forall x \in \mathbb{Z}_{\geq 0},\ \text{for any base } b \in \mathbb{Z}_{>1},\ \text{for any digit-index } n \in \mathbb{Z}_{\geq 0}$$

Clearly, $\overrightarrow{T}$ reassembles the digits in the base-b expansion of x into their original position, save for the last n digits.

Definition 17. Let K, the "cut-to-index function", be defined as

$$K_b^p(x) := \sum_{k=0}^{I_b^p(x)} b^k D_b^k(x)$$

$$\forall x \in \mathbb{Z}_{\geq 0},\ \text{for any base } b \in \mathbb{Z}_{>1},\ \text{for any digit } p \in U_b.$$

Similar to $\overrightarrow{T}$, K reassembles the digits in the base-b expansion of x into their original position, save for the last digits with indices greater than $I_b^p(x)$. For the case where $p \nsubseteq x$, we find that $I_b^p(x) = L_b(x)$, which implies that $K_b^d(x) = x$.

4 Assembling The Conway Base-13 Function

Perhaps the most daunting of tasks to replicate in the Conway Base-13 Function is recompiling digits in an expansion from one base to another, and replacing a digit with a radix-point.

Definition 18. Let X, the "base-to-base re-radix function", be defined as

$$X^p_{b_1,b_2}(x) := \sum_{k=0}^{L_{b_1}(x)-1} N(D^k_{b_1}(x),p)D^k_{b_1}(x)b_2^{k-I^p_{b_1}(x)-G_E(I^p_{b_1}(x),k)}$$

$$\forall x \in \mathbb{Z}_{\geq 0}, \text{ for any bases } b_1, b_2 \in \mathbb{Z}_{>1}, \text{ and any digit } p \in U_{b_1}.$$

X removes a specific digit p, with index j in the base-b_1 expansion of x. This position will be virtually used as a new radix-point. Digits to the left of p (with indices $k > j$) are placed directly to left of this new radix, and digits to the right of p (with indices $k < j$) are placed directly to the right. The final result is treated as a base-b_2 expansion. For the instances where there are multiple occurrences of p, such a case is evidently disregarded in further construction of the Base-13 function.

Lemma 19.

$$x = d_{m\to 0(b_1)} \wedge \exists! j(d_j = p) \implies X^p_{b_1,b_2}(x) = d_{m\to j+1}.d_{j-1\to 0(b_2)} \ \forall b_2 \in \mathbb{Z}_{\geq b_1}$$

Proof: Suppose $x = d_{m\to 0(b_1)}$ and $\exists! j(d_j = p)$. Thus the index, j, is given by $I^p_{b_1}(x) = j$. A digit at index k is given by $D^k_{b_1}(x) = d_k$. Hence $\forall b_2 \in \mathbb{Z}_{\geq b_1}$, substituting for our positional notation,

$$N(D^k_{b_1}(x),p)D^k_{b_1}(x)b_2^{k-I^p_{b_1}(x)-G_E(I^p_{b_1}(x),k)} = \begin{cases} d_k b_2^{k-j} & \text{if } k < j \\ 0 & \text{if } k = j \\ d_k b_2^{k-j-1} & \text{if } k > j \end{cases}$$

which can be used to split the sum into

$$X^p_{b_1,b_2}(x) = \sum_{k=0}^{j-1} d_k b_2^{k-j} + \sum_{k=j+1}^{m} d_k b_2^{k-j-1}$$

We are left with two recompilations of digits from base-b_1 to base-b_2, with the digits to the left of p directly to left of the radix, and digits to the right of p to the right. In our positional notation, this is equivalent to $d_{m\to j+1}.d_{j-1\to 0(b_2)}$.

□

For example, $X^C_{13,10}(1C3_{(13)}) = 1.3_{(10)}$. Lastly, we'll introduce a method to detect whether one of two given digits are contained within a base-b expansion. This will act as the step in determining if the final expansion of Conway's Base-13 function will be positive or negative.

Definition 20. Let S, the "resulting-sign function", be defined as

$$S^{p_1,p_2}_b(x) := E\big(O^{p_1}_b(x),1\big) - E\big(O^{p_2}_b(x),1\big)$$

$$\forall x \in \mathbb{Z}_{\geq 0}, \text{ for any base } b \in \mathbb{Z}_{>1}, \text{ for any digits } p_1, p_2 \in \{0,\ldots,b-1\}.$$

Unlike the previous function, S is much simpler in description. If there exists exactly one $p_1 \subseteq x$, and not exactly one $p_2 \subseteq x$ (assuming a base-b expansion), then $S_b^{p_1,p_2}(x) = 1$. Similarly, $S_b^{p_1,p_2}(x) = -1$ if there exists exactly one $p_2 \subseteq x$, and not exactly one $p_1 \subseteq x$. Otherwise the result is zero.

Lemma 21.

$$x = d_{m\to 0(b)} \implies S_b^{p_1,p_2}(x) = \begin{cases} +1 & : \exists! j_1(d_{j_1} = p_1) \wedge \nexists! j_2(d_{j_2} = p_2) \\ -1 & : \nexists! j_1(d_{j_1} = p_1) \wedge \exists! j_2(d_{j_2} = p_2) \\ 0 & : \text{otherwise} \end{cases}$$

Proof: Suppose $x = d_{m\to 0(b)}$. With the definitions of E and O, the values of S, defined by $E\left(O_b^{p_1}(x), 1\right) - E\left(O_b^{p_2}(x), 1\right)$, in the following case-table are straightforward.

cases	$\exists! j_1(d_{j_1} = p_1)$	$\nexists! j_1(d_{j_1} = p_1)$
$\exists! j_2(d_{j_2} = p_1)$	$S_b^{p_1,p_2}(x) = 1 - 1 = 0$	$S_b^{p_1,p_2}(x) = 0 - 1 = -1$
$\nexists! j_2(d_{j_2} = p_2)$	$S_b^{p_1,p_2}(x) = 1 - 0 = 1$	$S_b^{p_1,p_2}(x) = 0 - 0 = 0$

□

With an arsenal of closed-form logical-conditional and digit manipulating functions, the Conway Base-13 Function can be constructed.

Theorem 2 There exists a closed-form $g : \mathbb{Z} \to \mathbb{R}$ such that $g \subseteq f$, where f is the Conway Base-13 Function.

Proof. Let $f_1(x) = M\left(K_{13}^A|x|, K_{13}^B|x|\right)$. After applying f_1 to an integer x, digits directly to the left of the rightmost A or B in the tridecimal expansion of x are truncated. As the sign of the input is disregarded in the Base-13 Function, the absolute value of x is taken for each instance of x in f_1. For any $k \in \mathbb{Z}$, let d_k represent the digit at index k in the tridecimal expansion of $|x|$. Let the rightmost-index of A be written as $I_{13}^A|x| = j_A$ and the rightmost-index of B be written as $I_{13}^B|x| = j_B$. Note that by our definitions of I and K,

$$A \nsubseteq x \implies j_A = L_{13}|x| \implies K_{13}^A|x| = |x|$$

$$B \nsubseteq x \implies j_B = L_{13}|x| \implies K_{13}^B|x| = |x|$$

Since L is monotonically increasing, the inequality relation between j_A and j_B implies

$$j_A \le j_B \iff L_{13}\left(K_{13}^A|x|\right) \le L_{13}\left(K_{13}^B|x|\right) \iff K_{13}^A|x| \le K_{13}^B|x|$$

$$j_B \le j_A \iff L_{13}\left(K_{13}^B|x|\right) \le L_{13}\left(K_{13}^A|x|\right) \iff K_{13}^B|x| \le K_{13}^A|x|$$

Therefore, they determine the value of M by

$$j_A \le j_B \iff M\left(K_{13}^A|x|, K_{13}^B|x|\right) = K_{13}^A|x|$$

$$j_B \le j_A \iff M\left(K_{13}^A|x|, K_{13}^B|x|\right) = K_{13}^B|x|$$

Consequently f_1 becomes

$$f_1(x) = \begin{cases} K_{13}^{A}|x| & : j_A < j_B \\ K_{13}^{B}|x| & : j_B < j_A \\ |x| & : \text{otherwise} \end{cases}$$

And by definition of K, the function f_1 can be represented

$$f_1(x) = \begin{cases} Ad_{j_A-1\to 0(13)} & : A \subseteq x \wedge j_A < j_B \\ Bd_{j_B-1\to 0(13)} & : B \subseteq x \wedge j_B < j_A \\ |x| & : \text{otherwise} \end{cases}$$

Next, let $f_2(x) = f_1(x)E\Big(O_{13}^{C}\big(f_1(x)\big),1\Big)$. In f_2, we are checking if after such an A or B, there exists exactly one C leftover in the tridecimal expansion of $f_1(x)$. If there does not exist exactly one such C,

$$O_{13}^{C}\big(f_1(x)\big) \neq 1 \implies E\Big(O_{13}^{C}\big(f_1(x)\big),1\Big) = 0 \implies f_2(x) = 0$$

Otherwise, let the index of such be denoted $I_{13}^{C}\big(f_1(x)\big) = j_C$. Therefore

$$f_2(x) = \begin{cases} Ad_{j_A-1\to j_C+1}Cd_{j_C-1\to 0(13)} & : A \subseteq f_1(x) \wedge O_{13}^{C}(f_1(x)) = 1 \\ Bd_{j_B-1\to j_C+1}Cd_{j_C-1\to 0(13)} & : B \subseteq f_1(x) \wedge O_{13}^{C}(f_1(x)) = 1 \\ f_1(x) & : A,B \nsubseteq f_1(x) \wedge O_{13}^{C}(f_1(x)) = 1 \\ 0 & : O_{13}^{C}(f_1(x)) \neq 1 \end{cases}$$

Lastly, let $f_3(x) = S_{13}^{A,B}\Big(f_2(x)\Big)X_{13,10}^{C}\Big(\overrightarrow{T}_{13}^{1}\big(f_2(x)\big)\Big)$. This final function determines the sign of the final result, truncates off the leftover A or B, recompiles the tridecimal expansion into decimal, and essentially replaces C with a decimal point (assuming $C \subseteq f_2(x)$). By the definition of S and f_2, we find that

$$S_{13}^{A,B}\Big(f_2(x)\Big) = \begin{cases} +1 & : A \subseteq f_2(x) \\ -1 & : B \subseteq f_2(x) \\ 0 & : \text{otherwise} \end{cases}$$

and by our definition of $\overrightarrow{T}^{*}$,

$$\overrightarrow{T}_{13}^{1}\big(f_2(x)\big) = \begin{cases} d_{j_A-1\to j_C+1}Cd_{j_C-1\to 0(13)} & : A \subseteq f_1(x) \wedge O_{13}^{C}(f_1(x)) = 1 \\ d_{j_B-1\to j_C+1}Cd_{j_C-1\to 0(13)} & : B \subseteq f_1(x) \wedge O_{13}^{C}(f_1(x)) = 1 \\ d_{L_{13}|x|-2\to 0(13)} & : A,B \nsubseteq f_1(x) \wedge O_{13}^{C}(f_1(x)) = 1 \\ 0 & : O_{13}^{C}(f_1(x)) \neq 1 \end{cases}$$

Hence, by definition of X,

$$X^{C}_{13,10}\left(\overrightarrow{T}^{1}_{13}(f_2(x))\right) =$$

$$\begin{cases} d_{j_A-1\to j_C+1}.d_{j_C-1\to 0(10)} & : A \subseteq f_1(x) \wedge O^{C}_{13}(f_1(x)) = 1 \\ d_{j_B-1\to j_C+1}.d_{j_C-1\to 0(10)} & : B \subseteq f_1(x) \wedge O^{C}_{13}(f_1(x)) = 1 \\ d_{L_{13}|x|-2\to j_C+1}.d_{j_C-1\to 0(10)} & : A,B \nsubseteq f_1(x) \wedge O^{C}_{13}(f_1(x)) = 1 \\ 0 & : O^{C}_{13}(f_1(x)) \neq 1 \end{cases}$$

Therefore, the final result is of form

$$f_3(x) = \begin{cases} +d_{j_A-1\to j_C+1}.d_{j_C-1\to 0\cdot(10)} & : A \subseteq f_2(x) \wedge C \subseteq f_2(x) \\ -d_{j_B-1\to j_C+1}.d_{j_C-1\to 0\cdot(10)} & : B \subseteq f_2(x) \wedge C \subseteq f_2(x) \\ 0 & : \text{otherwise} \end{cases}$$

It should be seen that when x is an integer, the result for each case in f_3 is equivalent to f, the Base-13 function. Furthermore, as the cases from our original quantification from the plain-language definition hold equivalently, we find that $f_3 \subseteq f$, directly satisfying that f_3 is a closed-form representation of the Conway Base-13 Function over the integers.

□

5 Concluding Remarks

Due to the fractal-like symmetry of f, such that $f(13^n x) = f(x) \forall n \in \mathbb{Z}$, it's possible to extend $f_3 : \mathbb{Z} \to \mathbb{R}$ to $f_3 : \mathbb{Z}[\frac{1}{13}] \to \mathbb{R}$ by imposing that if $x = \frac{y}{13^n} \forall y \in \mathbb{Z}$, $\forall n \in \mathbb{Z}_{\geq 0}$, then $f_3(x) = f_3(y)$. This was done in the creation of Figure 1. It may be possible to extend f_3 to even larger sets of numbers whose distribution of digits are known, but a closed-form for f over the entirety of $\mathbb{R}$ is impossible, as the digit-distribution for every real number is not computable [4].

It is no doubt that the computational efficiency of these algorithms is far from optimal. A computer can perform a variety of digit manipulation tasks directly and quite efficiently, without the use of arithmetic closed-form functions such as these. The purpose, rather, was to fulfill the recreational endeavor of finding the first equation for Conway's Base-13 Function, based solely on finite arithmetic. This work was inspired by the recent passing of John H. Conway (1937→2020).

Acknowledgement Humblest gratitude to my mother, Loral Boylan, to my editor, and to Dr. Peter Dukes for suggestions on improving the readability of this paper.

Bibliography

[1] Greg Oman "The Converse of the Intermediate Value Theorem: From Conway to Cantor to Cosets and Beyond," Missouri Journal of Mathematical Sciences, Missouri J. Math. Sci. 26(2), 134-150, (November 2014)

Note that the value in the third case is irrelevant, as such a case results in 0 in f_3.

[2] Timothy Y. Chow (1999) What Is a Closed-Form Number?, The American Mathematical Monthly, 106:5, 440-448, DOI: 10.1080/00029890.1999.12005066

[3] Knuth, Donald. E. (1972). The Art of Computer Programming. Addison-Wesley.

[4] Willie Wong, et al. Is Conway's Base-13 Function Measurable? 20 July 2010, mathoverflow.net/questions/32641/is-conways-base-13-function-measurable.

Ball State Undergraduate Mathematics Exchange
`https:// digitalresearch.bsu.edu/mathexchange`
Vol. 15, No. 1 (Fall 2021)
Pages 54 – 64

Dynamic Optimization in Building Personal Emergency Fund

Aqsa Ahad, Aylara Alleyne,Worku T. Bitew , Michael De Oliveira, Courtney Schordine, Nicholas Seaton*

Worku T. Bitew is a Professor of Mathematics and Director for the Center for Applied Mathematical Sciences at State University of New York-Farmingdale. He works in applied mathematical analysis: calculus of variations and optimal control theory (optimization) and its applications in science, engineering, and natural and environmental economics. He has published research papers in the areas of analysis (calculus of variations), natural resource management (fisheries) and environmental management (pollution), and image recognition (engineering).

Michael L. De Oliveira was an undergraduate student studying Applied Mathematics with a minor in Computer Programming and Information Systems at Farmingdale State College. He graduated in December of 2019 and is currently pursuing a Master of Arts in Mathematics with a specialization in the Mathematics of Finance at Columbia University. Michael plans to pursue his interest in quantitative investment management following the completion of his education.

Courtney Schordine graduated from Farmingdale State College with a bachelor's degree in Applied Mathematics on the Financial Track with a minor in Economics. She is currently a Corporate Financial Analyst at Northwell Health and is attending Hofstra University to obtain an MBA in Finance..

Aylara Alleyne is a graduate from Farmingdale State College with a Bachelor of Science in Financial Mathematics. Currently, she serves as dedicated Homemaker/SAHM (Stay at Home Mother) of two. After maternity leave, she plans to continue advancing her education in analytics and finance .

***Corresponding author**: biteww@farmingdale.edu

Abstract

The 2018 National Financial Capability Study found that 46 percent of Americans do not have the recommended three months' worth of expenses in the case of an emergency. It is of immense importance to provide the best financial strategies towards building a solid financial foundation. In this paper, we examine how to build an emergency fund while maximizing the utility of consumption, allowing for a balance of consumer gratification and necessary future planning. This problem was approached utilizing the method of dynamic optimization. The necessary conditions for optimality were obtained and computations were performed to determine the optimal solution. The optimal savings trajectory was adjusted monthly by incorporating sensitivity factors with respect to each parameter involved in the model to get the actual monthly savings. Finally, we performed numerical simulations to create a financial plan that achieves a prescribed amount of emergency fund goal in a given planning year utilizing simulation data from an entry-level college graduate's salary, current high-yield return rates, and treasury yield-to-maturity rates.

1 Introduction

Everyday, Americans strive to achieve their "American Dream" which could be finding a better job, getting a better education, buying a property, etc. However, with the acquisition of a better life style, Americans' face greater financial responsibilities. Without an appropriate financial understanding, by the end of each paycheck, one's "American Dream" can easily turn into a nightmare. According to the 2018 National Financial Capability Study, the subject of personal finances is a source of anxiety. The study says, "more than half (53%) agree that thinking about their finances makes them anxious, and 44% feel that discussing their finances is stressful, with respondents ages 18-34 reporting the highest levels of stress (63%) and anxiety (55%)." With that being said, Bankrate, a personal finance website, conducted a survey that shows only 40% of Americans are comfortable covering $1,000 of unexpected expenses. Therefore, emergencies that might cost more, such as losing a job or getting injured, can be out of the question. After further research on the subject of personal finances, there is one common topic that is covered in almost every website and financial literacy book- the subject of building an emergency fund. An emergency fund is a saved currency that is easy to liquidate in case of emergencies. Having such a fund promises financial security. Despite the importance of having "rainy day" funds, the 2018 National Financial Capability Study found that 46% of Americans have not yet set aside funds enough to cover three months worth of expenses in case of an emergency. We feel it is of immense importance to provide a reference for those looking to get on the right track toward building a solid financial foundation.

Both savings and emergency funds bring positive changes into well-being of households. The first and most obvious benefit is interest that can be earned on money in savings accounts. Today most of the banks have low interest rates, however by investing into a high yield savings account (as we are going to present in our paper), one can maximize the interest that can be earned. Another great benefit is that there is basically no risk involved. Unlike, the stock market, which is considered a high-risk investment type, putting money in a savings account will not cause any losses. In addition, today most of the banks are insured, which ensures that one's money is safe in banks. Today's automatic deposits make it convenient to save without having to physically be present in banks. There are a lot more benefits that come with having extra money saved. Therefore, we hope that this paper will bring essential input on presenting how individuals can achieve their saving goals.

In this paper we attempt to present how Dynamic Programming and Optimal Control Theory can be applied in optimization models that deal with efficiency of savings and consumption. For simulation purposes we are assuming the year 2003 as a current time. The data that we have collected is assumed as a predicted data. The problem that we consider is a system that continuously evolves over time and we are looking for an optimal solution or trajectory for the state variable using dynamic optimization. While applications of dynamic programming and optimization are still new in economics, we believe that further work can lead to an improvement of the welfare of households. With this practical experiment we will implement mathematical techniques used in optimization. Our goal is to maximize the utility of consumption (which has an effect on overall satisfaction of a given person) while building up the savings fund. Solution to such a problem will involve structuring it into multiple stages that can be performed by using dynamic-programming approach. This approach also constitutes the states of the process. The state variable trajectory help us to evaluate future actions taken based on present decisions. Despite the rich theoretical concept behind our project, it will have a realistic connection because of the factual data, including treasury yield-to-maturity rates and treasury bond rates. We believe that this paper will bring further contributions in applications of mathematics and dynamic programming in fields of economics and finance.

2 Mathematical Model

Suppose a teacher wants to set up an emergency fund that will allow them to save a certain amount of money over a specific time period while also maximizing their utility. We seek an optimal balance between saving money while also having the most satisfaction from the money you are able to consume. Let $s(t)$ be the state of the savings account at time t and changes at the rate $\frac{ds}{dt}(t) = s'(t)$ with $s(0) = s_0$. Let $c(t)$ be the amount of money available for consumption at time t after saving. The teacher's total consumption at any moment t is

$$c(t) = E(t) + \rho(t)s(t) - s'(t),$$

where $E(t)$ is after tax earnings at time t and $\rho(t)s(t)$ is a return from the savings account. The natural log is frequently used in economics to capture the relationship between consumption and utility, $u(t) = \ln(c(t))$. It shows that the utility of each additional dollar of consumption declines as the level of consumption increases.

The teacher's objective is to determine the trajectory of the state, savings accumulation

$s(t)$, to maximize the functional

$$J = \int_0^T \ln(c(t))e^{-r(t)t}\,dt = \int_0^T \left[\ln(E(t) + \rho(t)s(t) - s'(t))\right] e^{-r(t)t}\,dt$$

subject to $s(0) = s_0 = 0$, $s(15) = 60,000$ and $s'(t) \geq 0$, where $\rho(t) =$ the return rate for the savings account at time t, $r(t)$ is the treasury yield-to-maturity rate at time t, $e^{-r(t)t}$ is the discount factor, and $\ln(c(t))e^{-r(t)t}$ is the present value of utility.

To formulate our problem in a control theory set up we summarize the state and control variables and parameters involved in our model as:

$c(t) =$ consumption at time t
$s(t) =$ the state of the savings account at time t
$r(t) =$ treasury yield-to-maturity rate at time t
$E(t) =$ after tax earnings at time t
$\rho(t) =$ return rate for the savings account at time t
$s'(t) =$ what is being deposited into the savings at time t

We assume that in each year the monthly salary to be a constant (which is also true in most professions), $E(t) = E$. We also assume that the return rate on savings and treasury yield-to-maturity rate (both depend on the market) are constant for short intervals of time by taking the average predicted values. Then we will perform sensitivity analysis of the optimal solution with respect to these parameters and adjust our solutions accordingly.

3 Mathematical Tools

Calculus of Variations (modern Optimal Control Theory) is used in mathematics to find minimums and maximums of functionals that involve functions that change over time. The origins are traced back to $1696 - 1697$ when John Bernoulli and his brother James were solving the brachistochrone problem. Later the search for the necessary conditions for an extremal to be a minimizer led to the development of the Euler-Lagrange equation. It was widely used in mathematics to solve problems of optimization, which led to fruitful outcomes in many fields such as aerospace engineering and machine learning. In early 1930 mathematicians and economic theorists such as Ramsey and Hotelling started developing optimization theories related to the field of economics (Kamien [1]). To find the optimal solution we will derive the Euler-Lagrange equation. Suppose that we have the functional

$$J(s(t)) = \int_0^T F(t, s(t), s'(t))\,dt.$$

We wish to find a function $s(t)$ that satisfies the boundary conditions $s(0) = 0$ and $s(T) = s_T$ and maximizes the functional J. Suppose that $s^*(t)$ is such a function. Then any small perturbations of $s^*(t)$ that preserves the boundary conditions will decrease the value of J since $s^*(t)$ is a maximizer.

Let $h(t)$ be a function that is continuous and differentiable on $[0,T]$ and $(0,T)$, respectively, such that $h(0)=h(T)=0$ and let $\varepsilon \in \mathbb{R}$. Then define

$$J(\varepsilon)=\int_0^T F\left(t,s^*(t)+\varepsilon h(t),s^{*'}(t)+\varepsilon h'(t)\right)\,dt$$

to be the resulting functional under the slight perturbations. We wish to find the total derivative of $J(\varepsilon)$ with respect to ε. Therefore we have

$$\begin{aligned}\frac{d}{d\varepsilon}(J(\varepsilon)) &= \frac{d}{d\varepsilon}\int_0^T F\left(t,s^*(t)+\varepsilon h(t),s^{*'}(t)+\varepsilon h'(t)\right)\,dt\\ &= \int_0^T \frac{d}{d\varepsilon}F\left(t,s^*(t)+\varepsilon h(t),s^{*'}(t)+\varepsilon h'(t)\right)\,dt.\end{aligned}$$

Let $x_\varepsilon^* = s^*(t)+\varepsilon h(t)$, $x_\varepsilon^{*'} = s^{*'}(t)+\varepsilon h^{'}(t)$ and $F_\varepsilon = F\left(t,s^*(t)+\varepsilon h(t),s^{*'}(t)+\varepsilon h'(t)\right)$. Then the inside derivative becomes

$$\begin{aligned}\frac{dF_\varepsilon}{d\varepsilon} &= \frac{dt}{d\varepsilon}\frac{\partial F_\varepsilon}{\partial t}+\frac{dx_\varepsilon^*}{d\varepsilon}\frac{\partial F_\varepsilon}{\partial s}+\frac{dx_\varepsilon^{*'}}{d\varepsilon}\frac{\partial F_\varepsilon}{\partial s'}\\ &= h(t)\frac{\partial F_\varepsilon}{\partial s}+h'(t)\frac{\partial F_\varepsilon}{\partial s'}.\end{aligned}$$

Therefore the integral becomes

$$\frac{d}{d\varepsilon}(J(\varepsilon)) = \int_0^T \left[h(t)\frac{\partial F_\varepsilon}{\partial s}+h'(t)\frac{\partial F_\varepsilon}{\partial s'}\right]\,dt.$$

When $\varepsilon = 0$ we have that J is at its maximum since we chose $s^*(t)$ to be the function that maximizes J. Therefore

$$\left.\frac{dF_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = \int_0^T \left[h(t)\frac{\partial F}{\partial s}+h'(t)\frac{\partial F}{\partial s'}\right]\,dt = 0.$$

Using integration by parts and the condition that $h(0)=0=h(T)$ the above equation can be rewritten as

$$\int_0^T (\frac{\partial F}{\partial s}-\frac{d}{dt}\frac{\partial F}{\partial s'})h(t)\,dt = 0.$$

For the last step in our derivation we use the following well-know fundamental theorem in Calculus of Variations.

Theorem 1. *If $f(t)$ is a continuous function and*

$$\int_0^T f(t)h(t)\,dt = 0$$

for all continuous and differentiable functions $h(t)$ over $[0,T]$ with $h(0)=0=h(T)$, then $f(t)=0$, for all $t \in [0,T]$.

By the above Theorem, we obtain the Euler-Lagrange equation

$$\frac{\partial F}{\partial s}(t,s^*(t),s^{*'}(t))-\frac{d}{dt}(\frac{\partial F}{\partial s'}(t,s^*(t),s^{*'}(t)))=0.$$

4 The Necessary Condition and Optimal Solutions

4.1 The Necessary Condition

We can write the performance function as:

$$F(t,s,s') = \ln(E+\rho s - s')e^{-rt}$$

Taking the partial derivative with respect to s, we get the equation

$$F_s = \frac{\rho e^{-rt}}{E+\rho s - s'}$$

Taking the partial derivative with respect to s', we get the equation

$$F_{s'} = \frac{-e^{-rt}}{E+\rho s - s'}$$

Then replacing $s = s(t)$ and $s' = s'(t)$ and differentiating with respect to t, we get

$$\frac{dF_{s'}}{dt} = \frac{re^{-rt}}{E+\rho s(t) - s'(t)} + \frac{e^{-rt}(\rho s'(t) - s''(t))}{(E+\rho s(t) - s'(t))^2}$$

Note that E, ρ, and r assumed to be independent of time. Please see section 2 for the details.

By taking the derivatives and substituting in the Euler-Lagrange equation that corresponds to our problem, we get the following second order differential equation:

$$(-r+\rho)(\rho s(t) + E - s'(t)) - (\rho s'(t) - s''(t)) = 0,$$

with boundary conditions $s(0) = s_0$ and $s(T) = s_T$.

4.2 Numerical Results: Optimal Monthly Savings and Consumption Plan

We collected data from the National Center for Education Statistics (NCES). The NCES data gives us the average salary of classroom teachers in public elementary and secondary schools in the United States from 1980 to 2017. We then took the average salary and divided it by twelve to find the average monthly salary. Because we were interested in safe investment options, we used data from the U.S. Treasury and the International Monetary Fund to obtain the monthly treasury bond rates which behave similarly to the high yield saving accounts like High Yield American Express account. Interest rates were obtained for 180 months for years $2003-2017$. We also obtained the corresponding monthly U.S. treasury constant maturity (yield-to-maturity) rates data, which was used as a risk-free discounting rate for our numerical simulation.

The descriptive statistics shown by Table 1 provide information about our parameters: salary, treasury yield-to-maturity rate, and treasury 1-year bond yield rate. On average, a teacher makes about $\$53,391$ per year. The minimum starting salary was recorded in

Table 1: Descriptive Statistics

Variable	Observations	Mean	SD	Min	Max
Salary	180	53391	4371.6	45757	58875
Treasury yield-to-maturity	180	0.032	0.0105	0.0150	0.0511
Treasury bond yield	180	0.0144	0.0162	0.0010	0.0522

2003 at $\$45,757$ per year and increased by approximately \$874 every year, reaching the maximum salary of $\$58,875$ per year in 2017. The average treasury yield-to-maturity rates recorded between the years 2003 to 2017 was approximately 3.20%, with a low of 1.5%, which occurred in 2016 and a high of 5.11% between the years 2006 and 2007. The average treasury bond yield rates was around 1.44%, with a minimum of 0.1% from the years 2011 and 2014 and a maximum of 5.22% in 2006.

We want to find our adjusted monthly savings that allows us to continue maximizing the overall satisfaction and meeting the $\$60,000$ target assuming that the teacher continues earning their monthly salary for 15 years. To achieve this goal, we follow the following dynamic programming or scheduling steps.

1. We solve the second order differential equation

$$(-r+\rho)(\rho s(t)+E-s'(t))-(\rho s'(t)-s''(t))=0,$$

with boundary conditions, $s(0)=0$ and $s(15)=60000$. The solution depends on t, ρ, r and E. Let $s(t,\rho,r,E)$ be the solution. Substituting the constant monthly salary for the first year, E_1 into $s_1(t,\rho,r,E)$, we get $s_1(t,\rho,r,E_1)$.
Assume that the monthly treasury bond rates and treasury yield-to-maturity rates are forcasted in advance for the first year (it can also be done quarterly or biannually). Let the first year average treasury bond and treasury yield-to-maturity rates be ρ_1 and r_1. Then the projected monthly savings at the i^{th} month of the first year is $s_1(\frac{i}{12},\rho_1,r_1,E_1)$ for $i=1..12$. These predicted values can be adjusted month-by-month using the first order Taylor's series expansion of $s_1(t,\rho,r,E_1)$:

$$s_1(t_i,\rho_i,r_i,E_1)\cong s_1(t_i,\rho_1,r_1,E_1)+\frac{\partial s_1}{\partial\rho}(t_i,\rho_1,r_1,E_1)(\rho_i-\rho_1)+ \quad (1)$$

$$\frac{\partial s_1}{\partial r}(t_i,\rho_1,r_1,E_1)(r_i-r_1) \quad (2)$$

The first term is the monthly savings predicted using the averages, the second term $\frac{\partial s_1}{\partial\rho}(t_i,\rho_1,r_1,E_1)(\rho_i-\rho_1)$ is the monthly adjustment due to the relative change in treasury bond return rates and the third term $\frac{\partial s_1}{\partial r}(t_i,\rho_1,r_1,E_1)(r_i-r_1)$ is the monthly adjustment due to the relative change in treasury yield-to-maturity rates. Therefore, the adjusted saving at i^{th} month is the difference of the savings account balance between two consecutive months:

$$s_1(\frac{i}{12},\rho_i,r_i,E_1)-s_1(\frac{i-1}{12},\rho_{i-1},r_{i-1},E_1)$$

and our i^{th} month out-of-pocket monthly savings contribution will be

$$\begin{aligned} S_1(i) = {} & s_1(\frac{i}{12},\rho_i,r_i,E_1) \\ & -r_{i-1}\left[\sum_{m=1}^{i-1}\left(s_1(\frac{m}{12},\rho_m,r_m,E_1)-s_1(\frac{m-1}{12},\rho_{m-1},r_{m-1},E_1)\right)\right] \end{aligned} \tag{3}$$

The first year total savings account balance becomes

$$B_1 = \sum_{i=1}^{i=12}\left[s_1(\frac{i}{12},\rho_i,r_i,E_1)-s_1(\frac{i-1}{12},\rho_{i-1},r_{i-1},E_1)\right],$$

by assumption $s_1(0,\rho_0,r_0,E_1)=0$.

2. After we finished the first year, month-by-month calculations and got B_1, we solve the same second order differential equation with different boundary conditions, $s(1)=B_1$ and $S(15)=60000$. Let $s_2(t,\rho,r,E)$ be the solution for this boundary value problem. Again substituting the second year constant monthly salary E_2 in $s_2(t,\rho,r,E)$, we get $s_2(t,\rho,r,E_2)$. Then we repeat step (1) using $s_2(t,\rho_2,r_2,E_2)$ for $i=13..24$, where ρ_2 and r_2 are the average values of the monthly treasury bond yield and treasury yield-to-maturity rates in year 2. We continue this process recursively for the remaining 13 years.

   We evaluated the monthly savings, monthly consumptions, and cumulative savings account balance for the years $2003-2017$ and presented the results in Figure (1), (2) and Figure (3) below.

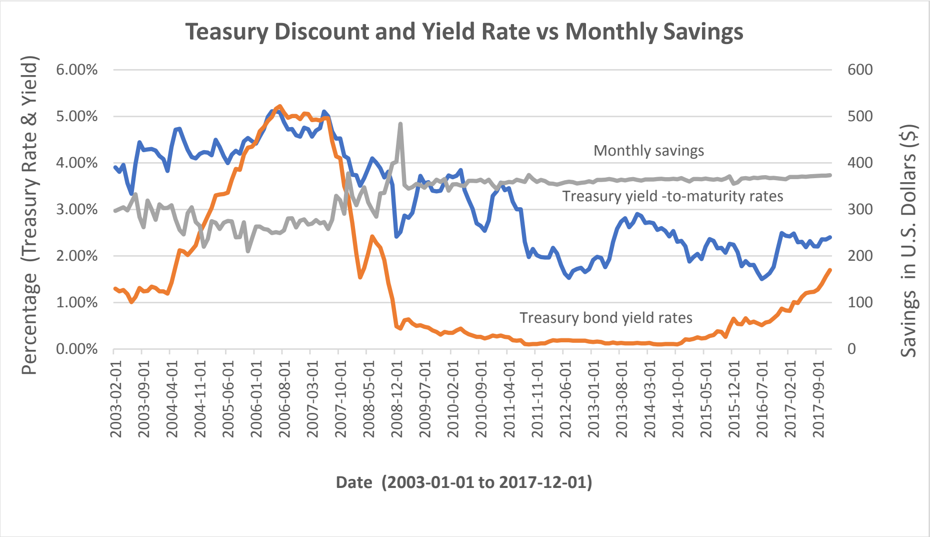


Figure 1: Monthly Adjusted Savings

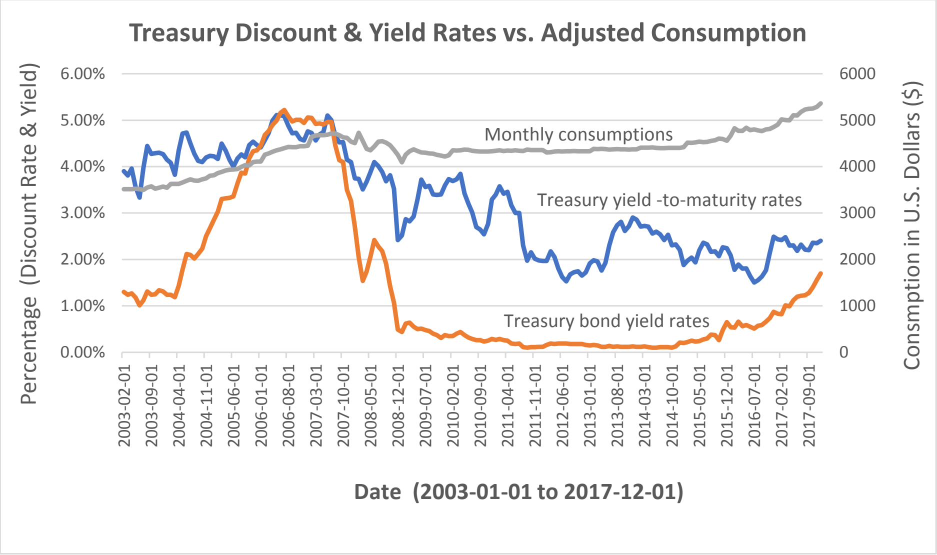


Figure 2: Monthly Adjusted Consumption

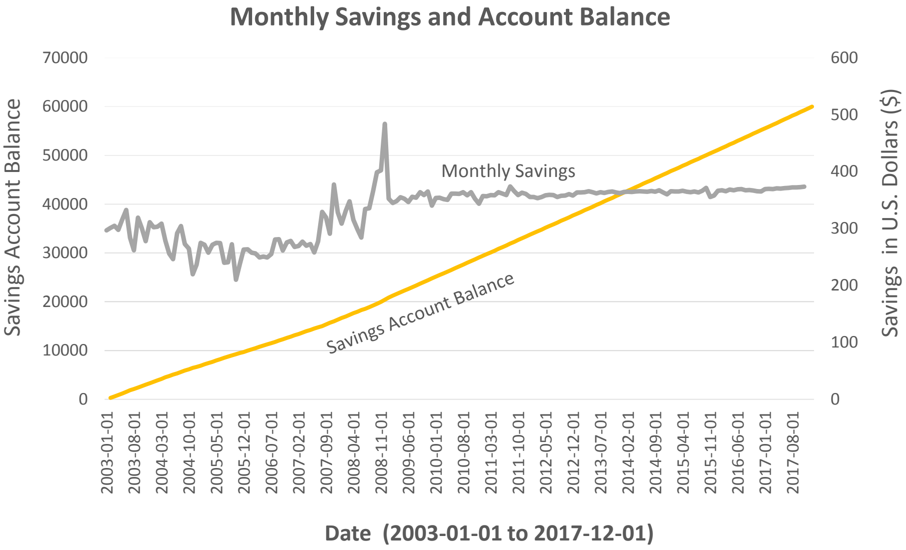


Figure 3: Monthly Savings and Savings Account Balance

Figure (1) and Figure (2) shows adjusted monthly saving and corresponding consumption that will allow the teacher to reach their savings goal of $\$60,000$ in 15 years (please see Figure (3)). From Figure (1) and Figure (2), between the years 2004 to 2007, the teacher would have been able to consume more and save less from their salary. This is because the treasury bond yield rate increased from 1.91% to 5.22%. This allowed the teacher to reach their monthly savings goal by using more of what was coming from the return rather than their salary. However, when the stock market crashed in 2008 and treasury yield rates fell, the teacher would have had to save more and consume less to make up for the low return rate they were getting on their savings account shown by the very low yield rate and keep on track with their goal. When the market began to steady around 2010, the teacher stayed on a fairly constant track with their monthly savings to make sure they reached their goal in the time frame. They were able to consume more towards the end of the fifteen years due to the return increasing again as well as the constant increase in salary per year while getting closer to and eventually reaching their goal of $60,000 in 15 years. It seems as though treasury bond yield rates have a fairly high inverse correlation with adjusted savings, but a very low correlation with adjusted consumption. This makes sense because adjusted savings is based on the fluctuations in the treasury yield-to-maturity rate and treasury bond yield while attempting to reach the $60,000 goal at the end of year 15. On the other hand, adjusted consumption (though impacted by the treasury yield-to-maturity rates and treasury bond yield rates) is more heavily dependent on changes in salary.

5 Conclusion

From our paper, we were able to observe and study how Optimal Control Theory and Dynamic Programming can be applied in fields of economics and finance. To build a savings fund with minimum stress while maximizing our utility of consumption we have used mathematical tools to derive the Euler-Lagrange equation and solve the equation to determine the optimal solution. Later, we solved our problem and considered the complexity of our equations (higher-order differential equations). To complete our numerical and sensitivity analysis and to find adjusted savings and consumption, we have obtained and used the data of teacher's salary, treasury yield-to-maturity rates, and treasury bond yield rates. In the end, we have successfully reached our goal of saving $60,000 at the end of fifteen years of simulation.

What makes our project so essential is the applicability of it in real life. Even though we were assuming the past data as current data, we were able to demonstrate how advanced mathematical computations can be used to work with it. With that being said the same calculations can be used on predicted data rather than the data obtained from the past observations. We hope that our project will bring some insight about applications of Optimal Control Theory in an undergraduate program in the field of Applied Mathematics, Economics,and Finance.

Acknowledgments The authors are indebted to the anonymous reviewer and the editor of this journal for their insightful comments. We would like to acknowledge Mathematics Department-CAMS at Farmingdale State College for supporting this Applied Learning Activity.

Bibliography

[1] Morton I. Kamien, Nancy L. Schwartz. *Dynamic Optimization, The Calculus of Variation and Optimal Control in Economics and Management.* Dower Publications, Inc., Mineola, NY, 2nd Edition, 3-4, 2012.

[2] U.S. Survey Data at a Glance. FINRA Investor Education Foundation. https://www.usfinancialcapability.org/results.php. Accessed on December 4, 2019.

[3] Board of Governors of the Federal Reserve System (US), Monthly Treasury Constant Maturity Rate [DGS10], retrieved from FRED, Federal Reserve Bank of St. Louis.
https://fred.stlouisfed.org/series/DGS10. Accessed March 23, 2021.

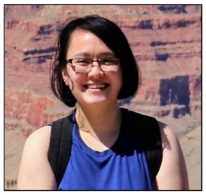
Ball State Undergraduate Mathematics Exchange
https:// digitalresearch.bsu.edu/mathexchange
Vol. 15, No. 1 (Fall 2021)
Pages 65 – 77

The Abundancy Index and Feebly Amicable Numbers

*Jamie Bishop, Abigail Bozarth, Rebekah Kuss, Benjamin Peet**

Jamie Bishop is currently double majoring in Mathematics and Secondary Education at Saint Martin's University. After graduation, she plans to teach high school math while working on a Master's in Mathematics. Her goal is to teach advanced mathematics to college students. Her favorite areas of mathematics are analytic number theory and proof theory.

Abigail Bozarth is an undergraduate student studying Mathematics at Saint Martin's University. She plans to pursue a Masters in Teaching and begin teaching math upon graduation. Abigail grew up in the small town of Kalama, WA and cites her own teachers as the inspiration for wanting to become one herself.



Rebekah Kuss is an undergraduate student in her final year at Saint Martin's University. She is currently studying Mathematics and Secondary Education. After graduation, she is planning to pursue a Ph.D. in Mathematics.

Benjamin Peet received his Ph.D. from St. Louis University in 2018 and is currently an assistant professor at St. Martin's University, WA. He has published research in 3-manifold topology, but enjoys working in all areas of mathematics, particularly alongside undergraduate students. He loves living near the water and mountains of Washington state with his wife and three young boys.

***Corresponding author**: bpeet@stmartin.edu

Abstract

This research explores the sum of divisors - $\sigma(n)$ - and the abundancy index given by the function $\frac{\sigma(n)}{n}$. We give a generalization of amicable pairs - feebly amicable pairs (also known as harmonious pairs), that is m, n such that $\frac{n}{\sigma(n)} + \frac{m}{\sigma(m)} = 1$. We first give some groundwork in introductory number theory, then the goal of the paper is to determine if all numbers are feebly amicable with at least one other number by using known results about the abundancy index. We establish that not all numbers are feebly amicable with at least one other number. We generate data using the R programming language and give some questions and conjectures.

1 Introduction

The sum of divisor function, $\sigma(n)$, for a positive integer n, is the sum of all the positive divisors including n itself. Looking at the ratio of the sum of divisor function and the number itself, $\frac{\sigma(n)}{n}$, or the abundancy index, we look into its relation to concepts such as perfect numbers, abundant numbers, deficient numbers, and amicable numbers. Through understanding these relations, we will define the concept of feebly amicable numbers also known as harmonious numbers in [12]. This is a generalization of an amicable number with weakened conditions so that the sum of divisors of the two do not need to be equal.

Formally, these are two numbers m and n such that $\frac{n}{\sigma(n)} + \frac{m}{\sigma(m)} = 1$. Examples of the first twenty feebly amicable pairs are given for illustration of this concept.

We use the R programming language to produce some abundancy indices and then a list of feebly amicable numbers. This data allows us to ask some questions about feebly amicable numbers that are unknown about amicable numbers.

Our main new results are Theorem 11 and Corollary 12 which give conditions for when a number can be feebly amicable with another and consequently amicable. The final section has some questions and conjectures that might be of interest to the reader or for future work.

2 History

The implications of this research are derived from the historical mathematical workings of famous figures, most notably Euclid, Euler, and Mersenne. Euclid further advanced our understanding of prime numbers by providing the Euclidean algorithm and showing that there are infinitely many prime numbers. Euclid also completed one of the only proofs involving perfect numbers: if $2^n - 1$ is prime, then $2^{n-1}(2^n - 1)$ is perfect. Euler continued the idea of perfect numbers by proving that every even perfect number can be expressed in Euclid's form. This research also makes use of Marin Mersenne's work on primes, and a Mersenne prime is of the form $2^n - 1$. [3]

3 Preliminary Definitions

We give some some preliminary definitions that we will rely upon throughout. We are working here with the positive integers (natural numbers). It should also be noted that it is possible to extend all this theory to the negative integers.

Firstly, a *divisor* is a number that divides into another number without a remainder.Then, a *prime* is a number that has only the divisors 1 and itself. Particular prime numbers used in this paper are Mersenne primes which are prime numbers of the form $2^p - 1$, where p is also prime.

Given a natural number n, we can define the *canonical representation* of n to be $\prod_{i=1}^{r} p_i^{a_i}$ where the p_i are the distinct prime divisors of n and a_i their multiplicities.

Two number are *coprime* (alternatively *relatively prime*) if they share only 1 as a divisor.

Then and importantly to this paper, the *sum of divisor function* $\sigma(n)$ for a positive integer n is defined as the sum of all its divisors (including n itself).

From this, we can categorize natural numbers according to:

- n is called *perfect* if $\sigma(n) = 2n$.
- n is called *abundant* if $\sigma(n) > 2n$.
- n is called *deficient* if $\sigma(n) < 2n$.

This paper generalizes the definition of amicable numbers. To be precise, *amicable numbers* are two numbers related in such a way that the sum of the proper divisors of each are equal and also equal to the sum of the numbers. That is, m,n are amicable if $m + n = \sigma(m) = \sigma(n)$.

Part of this paper is an investigation of the *abundancy index* of a number. It is defined by $\lambda(n) = \frac{\sigma(n)}{n}$ and, in some sense, measures how divisible a number is.

Multiply-perfect numbers are numbers such that their abundancy index is an integer. Note that perfect numbers by definition have abundancy index 2.

Finally, two numbers in $\mathbb{N}$ are *friendly* if they have the same abundancy index. That is, $\frac{\sigma(n)}{n} = \frac{\sigma(m)}{m}$. More generally, friendly numbers form *friendly clubs* if they all have the same abundancy index.

4 Preliminary Results

We now present some preliminary results that give some illustration of the theory and that will be used throughout the rest of the paper. Good references for these and more foundational theory are [1] and [6]. However, there are many good texts on introductory number theory.

Proposition 1. *If p,q are distinct primes then* $\sigma(p^n q^m) = \sigma(p^n)\sigma(q^m)$.

Proof. First let p and q be primes, then:

$$\sigma(p^n q^m) = \sum_{i=0}^{n}\sum_{j=0}^{m} p^i q^j = (1+p+\ldots+p^n)(1+q+\ldots+q^m) = \sigma(p^n)\sigma(q^m)$$

□

The above Proposition 1 yields the multiplicative property of the sum of divisors function. That is, if m and n are coprime, then $\sigma(mn) = \sigma(m)\sigma(n)$.

This leads to the following which is Theorem 2.24 of [1]:

Theorem 2. *If $a = \prod_{i=1}^{r} p_i^{a_i}$ where $a_i > 0$ for each i is the canonical representation of a, then*

$$\sigma(a) = \prod_{i=1}^{r} \frac{p_i^{a_i} - 1}{p_i - 1}.$$

Proof. We first establish that:

$$\sigma(p^n) = 1 + p + p^2 + p^3 + \ldots + p^n = \frac{p^{n+1} - 1}{p - 1}$$

This follows as:

$$(p-1)(1 + p + p^2 + p^3 + \ldots + p^n) = p^{n+1} - 1$$

Then by applying Proposition 1 inductively, the result follows. □

We make here a number of remarks regarding values of the abundancy index and how it relates to perfect, abundant, deficient, and friendly numbers.

The codomain of the abundancy index, $\lambda(n) = \frac{\sigma(n)}{n}$, is $\mathbb{Q} \cap (1,\infty)$. We will see later that this codomain is not in fact the range, that is, there are values in $\mathbb{Q} \cap (1,\infty)$ that are not abundnacy indices.

When $\frac{\sigma(n)}{n} = 2$, then n is perfect. When $1 < \frac{\sigma(n)}{n} < 2$, then n is deficient. And when $\frac{\sigma(n)}{n} > 2$, then n is abundant. Additionally, all perfect numbers in this case are friendly to one another. That is $\frac{\sigma(n)}{n} = \frac{\sigma(m)}{m}$; $\lambda(n) = \lambda(m)$.

We give a proof of the Euclid-Euler Theorem to illustrate the theory:

Theorem 3. *An even number is perfect if and only if it has the form $2^{p-1}(2^p - 1)$, where $2^p - 1$ is prime.*

Proof. Let $2^p - 1$ be prime. Then, by the multiplicative property, the sum of divisors of $2^{p-1}(2^p - 1)$ is equal to

$$\sigma(2^{p-1}(2^p - 1)) = \sigma(2^{p-1})\sigma(2^p - 1) = (2^p - 1)2^p = 2(2^{p-1})(2^p - 1).$$

Hence, since the sum of the divisors of $2^{p-1}(2^p - 1)$ is twice itself, $2^{p-1}(2^p - 1)$ is perfect.

For the converse, let $2^k x$ be an even perfect number, where x is odd. For $2^k x$ to be a perfect number, the sum of its divisors must be twice its value. So,

$$2(2^k x) = \sigma(2^k x) = \sigma(2^k)\sigma(x) = (2^{k+1} - 1)\sigma(x)$$

by the multiplicative property of σ.

The factor $2^{k+1} - 1$ must divide x. So $y = x/(2^{k+1} - 1)$ is a divisor of x. Now,

$$2^{k+1} y = \sigma(x) = x + y + z = 2^{k+1} y + z,$$

where z is the sum of the other divisors. Thus, for this equality to be true, there must be no other divisors, so z must be 0. Hence, y must be 1, and x must be a prime of the form $2^{k+1} - 1$. Therefore, an even number is perfect if and only if it has the form $2^{p-1}(2^p - 1)$, where $2^p - 1$ is prime. □

We also give the following proposition that we will utilize later on:

Proposition 4. *If p is prime, then $\frac{\sigma(n)}{n} = \frac{p+1}{p}$ if and only if $n = p$.*

Proof. We first show that if p is prime and $\frac{\sigma(n)}{n} = \frac{p+1}{p}$, then $n = p$. So we have $\frac{\sigma(n)}{n} = \frac{(p+1)k}{(p)k}$ for some $k = p^a L$ where p does not divide L and $n = pk$. Then suppose for contradiction that $k > 1$.

Then $\sigma(n) = (p+1)k$, but also, $\sigma(n) = \sigma(pk) = \sigma(pp^a L)$. By the multiplicative property of σ and Theorem 2, $\sigma(pp^a L) = \sigma(p^{a+1})\sigma(L) = \frac{p^{a+2}-1}{p-1}\sigma(L)$. So, $(p+1)k = \frac{p^{a+2}-1}{p-1}\sigma(L)$.

We continue the calculation with $(p+1)(p^a L) = \frac{p^{a+2}-1}{p-1}\sigma(L)$.

Hence, $(p+1)(p-1)(p^a L) = (p^{a+2} - 1)\sigma(L)$, which gives, $(p^2 - 1)(p^a L) = (p^{a+2} - 1)\sigma(L)$, and finally, $(p^{a+2} - p^a)L = (p^{a+2} - 1)\sigma(L)$.

Because $(p^{a+2} - p^a) < (p^{a+2} - 1)$ and $L < \sigma(L)$ we must have that $(p^{a+2} - p^a)L < (p^{a+2} - 1)\sigma(L)$. So, by contradiction, $n = p$.

For the converse if $n = p$, then if p is prime, as $\sigma(p) = p + 1$ we must have:

$$\frac{\sigma(n)}{n} = \frac{\sigma(p)}{p} = \frac{p+1}{p}.$$

□

5 R code

In order to explore the values of both the sum of divisors function and the abundancy index, we used some code in the R programming language [9] to generate the first 100,000 values of the sum of divisor function and abundancy indices. The following is some code that describes the algorithm:

```
sod <- function(x) %This defines the sum of divisors function%
    {s<-0
    for(i in 1:x){if(x%%i==0){s<-s+i}}
    %This loops through the values 1 through x to see which
    are factors and adds them to the sum if they are%
    return(s)}

    sigma<-c(1:100000)
    %This defines a vector of length 100,000%

for(i in 1:100000){sigma[i]<-sod(i)}
%This gives a vector of the first 100,000 values of the sum of
divisors%

abun<-c(1:100000)
%This again defines a vector of length 100,000%

for(i in 1:100000){abun[i]<-sigma[i]/i}
%This gives a vector of the first 100,000 values of the
abundancy index%
```

Figure 1 shows a histogram to reflect the data.

By our code we calculated the fraction of abundant numbers in the first $100,000$ numbers to be 0.24799. This is not within the range given by [2], but it is close. In that paper, they compute that $\alpha = \lim_{n\to\infty} \frac{A(n)}{n}$ is such that

$$0.2476171 < \alpha < 0.2476475$$

Here $A(n)$ is the number of abundant numbers less than n.

We suggest that computing larger numbers of abundancy indices would give an estimate within the proven bounds. In any case, the histogram illustrates how there are roughly three times more deficient numbers than abundant numbers.

6 Range of the Abundancy Index

From our data, $\frac{5}{4}$ does not appear as an abundancy index of any n less than $100,000$. First, we prove that $\frac{5}{4}$ does not appear for any natural number, and then, we show a generalization for finding more numbers not in the range.

Lemma 5. $\frac{5}{4} \neq \frac{\sigma(n)}{n}$ *for any natural number n.*

Proof. Suppose to the contrary. So $5k = \sigma(n)$ and $4k = n$ for some k. Thus, $n = 4k = 2^{a+2}l$ for some nonnegative integer $a \in \mathbb{N}$ and odd integer $l \in \mathbb{N}$. So then $\sigma(n) = \sigma(2^{a+2}l)$, which by the multiplicative property of σ, $\sigma(2^{a+2}l) = \sigma(2^{a+2})\sigma(l) = (2^{a+3} - 1)\sigma(l)$.

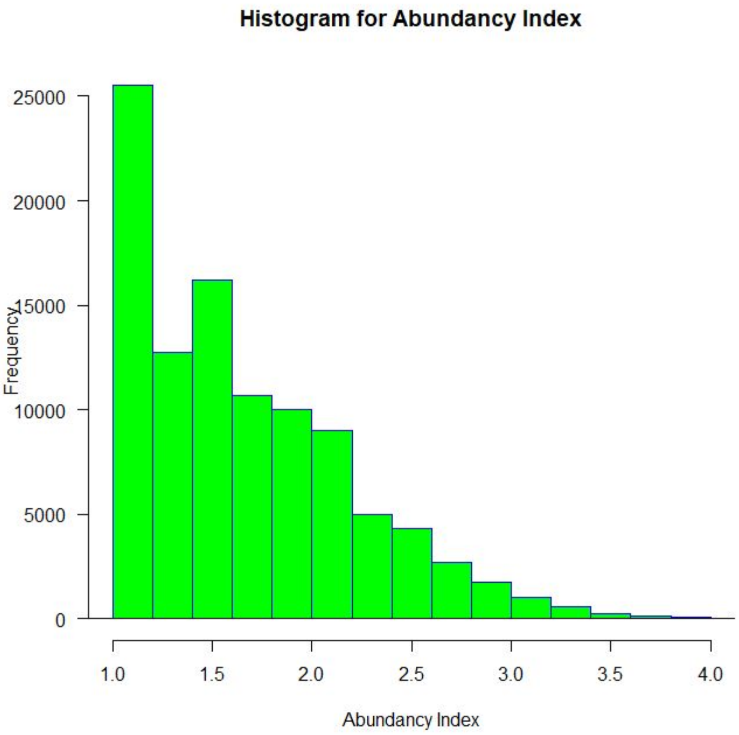


Figure 1: Histogram for Abundancy Index

This leads to:

$$\frac{\sigma(n)}{n} = \frac{(2^{a+3}-1)\sigma(l)}{2^{a+2}l} > \frac{7}{4}\frac{\sigma(l)}{l} > \frac{7}{4}.$$

Therefore, $\frac{5}{4} \neq \frac{\sigma(n)}{n}$. □

This can be generalized to Theorem 1 from [5]:

Theorem 6. *If k is coprime to m, and $m < k < \sigma(m)$, then $\frac{k}{m}$ is not the abundancy index of any integer.*

Proof. Assume $\frac{k}{m} = \frac{\sigma(n)}{n}$. Then $m\sigma(n) = kn$, so $m|kn$, hence $m|n$ because $(k,m) = 1$. But because $m|n$ implies $\frac{\sigma(m)}{m} \leq \frac{\sigma(n)}{n}$, with equality only if $m = n$, $\frac{\sigma(m)}{m} \leq \frac{\sigma(n)}{n} = \frac{k}{m}$, contradicting the assumption $k < \sigma(m)$. □

We can see that not all rational numbers greater than 1 are in the range, but we ask the question: is the range dense in rationals greater than 1? Recall that if $S \subseteq T \subseteq \mathbb{R}$, we say that S is dense in T if for any two numbers in T there exists an element of S in between them.

We refer to [4] and [5] again for the following results:

Theorem 7. *(Theorem 5 in [4]) The set*$\{\lambda(n)|n \in \mathbb{N}\}$ *is dense in* $\mathbb{Q} \cap (1,\infty)$.

We give the proofs from [5] of the following results for exposition:

Lemma 8. *Let m be a positive integer. If p is prime with $p > 2m$, then among any $2m$ consecutive integers, there is at least one integer coprime with pm.*

Proof. Let S be any set of $2m$ consecutive integers. If $p > 2m$ there is at most one multiple of p in S. But S contains at least two integers coprime with m, one of which is coprime with p and, therefore, also pm. □

Theorem 9. *(Theorem 2 in [5]) The complement of* $\{\lambda(n)|n \in \mathbb{N}\}$ *in* $\mathbb{Q} \cap (1,\infty)$ *is dense in* $\mathbb{Q} \cap (1,\infty)$.

Proof. Choose any real $x > 1$, and any $\varepsilon > 0$. We will exhibit a rational in the interval $(x-\varepsilon, x+\varepsilon)$, and that is not an abundancy ratio. By Theorem 7, choose $m > 1$ so that the abundancy index $\frac{\sigma(m)}{m}$ is in the interval $(x-\frac{\varepsilon}{2}, x+\frac{\varepsilon}{2})$. For every prime $p > 2m$, we have:

$$x-\frac{\varepsilon}{2} < \frac{\sigma(m)}{m} < \frac{\sigma(pm)}{pm} = (1+\frac{1}{p})\frac{\sigma(m)}{m} < (1+\frac{1}{p})(x+\frac{\varepsilon}{2}).$$

If we also require $p > \frac{2x+\varepsilon}{\varepsilon}$, then $(1+\frac{1}{p})(x+\frac{\varepsilon}{2}) < x+\varepsilon$, we have:

$$x-\frac{\varepsilon}{2} < \frac{\sigma(pm)}{pm} < x+\varepsilon.$$

By the Lemma 8, we know that $\sigma(pm)-k$ is coprime with pm for some k with $1 \leq k \leq 2m$. For such k, we also have:

$$\sigma(pm)-k \geq \sigma(pm)-2m \geq (p+1)(m+1)-2m > pm$$

because $p > 2m$. Therefore, by Theorem 6, $\frac{\sigma(pm)-k}{pm}$ is not an abundancy index. So, then:

$$\frac{\sigma(pm)-k}{pm} \geq \frac{\sigma(pm)-2m}{pm} = \frac{\sigma(pm)}{pm} - \frac{2}{p} > x - \frac{\varepsilon}{2} - \frac{2}{p}.$$

If $p \geq \frac{4}{\varepsilon}$, we have $x-\frac{\varepsilon}{2}-\frac{2}{p} \geq x-\varepsilon$, thus $\frac{\sigma(pm)-k}{pm} > x-\varepsilon$. All the inequalities are satisfied if $p > max\{2m, \frac{2x+\varepsilon}{\varepsilon}, \frac{4}{\varepsilon}\}$, and so:

$$x-\varepsilon < \frac{\sigma(pm)-k}{pm} < \frac{\sigma(pm)}{pm} < x+\varepsilon.$$

This gives, $\frac{\sigma(pm)-k}{pm}$ that is not an abundancy index, within ε of x. □

7 Feebly Amicable Numbers

We now proceed to generalize the definition of amicable numbers. We do so by recognizing the following result:

Proposition 10. *If two numbers m, n are amicable, then* $\frac{n}{\sigma(n)} + \frac{m}{\sigma(m)} = 1$.

Proof. Let m and n be amicable numbers. Then $\sigma(m) = \sigma(n) = m+n$ and so $\frac{m+n}{\sigma(n)} = 1$ hence $\frac{n}{\sigma(n)} + \frac{m}{\sigma(n)} = 1$.

Thus, if m and n are amicable numbers, then $\frac{m}{\sigma(m)} + \frac{n}{\sigma(n)} = 1$. □

This allows us to formulate the following definition:

Feebly amicable numbers are pairs m, n such that

$$\frac{n}{\sigma(n)} + \frac{m}{\sigma(m)} = 1$$

Alternatively, we can define in terms of the abundancy index λ:

$$\frac{1}{\lambda(n)} + \frac{1}{\lambda(m)} = 1$$

In words, two numbers are feebly amicable if the reciprocals of their abundancy indices sum to 1. Note that feebly amicable pairs are referred to as harmonious pairs in [12].

We note that pairs of perfect numbers are not amicable (no two perfect numbers have the same sum of divisors). However, they are feebly amicable, and the Venn diagram in Figure 2 illustrates the containment shown in Proposition 10.

We also note that amicable pairs have been extended to *amicable triples* which are three numbers m, n, and s such that $\sigma(m) = \sigma(n) = \sigma(s) = m+n+s$, and therefore the following is true as well: $\frac{n}{\sigma(n)} + \frac{m}{\sigma(m)} + \frac{s}{\sigma(s)} = 1$.

If amicable pairs and triples have been defined such as they have been above, then *amicable k-tuples* are numbers $n_1, \ldots, n_k$ such that

$$\sigma(n_1) = \ldots = \sigma(n_k) = n_1 + \ldots + n_k$$

Given these definitions we can also generalize to *feebly amicable triples*. These are three numbers m, n, and s such that

$$\frac{n}{\sigma(n)} + \frac{m}{\sigma(m)} + \frac{s}{\sigma(s)} = 1$$

and *feebly amicable k-tuples* as numbers $n_1, \ldots, n_k$ such that

$$\frac{n_1}{\sigma(n_1)} + \ldots + \frac{n_k}{\sigma(n_k)} = 1.$$

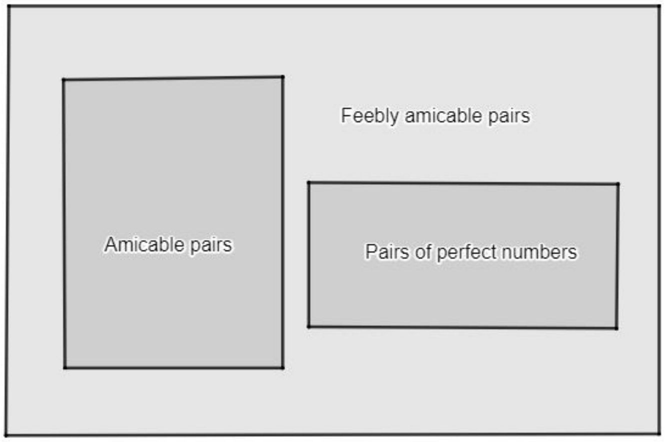


Figure 2: Venn diagram

All members of a friendly club have the same abundancy index. It is possible to talk of *feebly amicable clubs*. As if m,n are feebly amicable then m,l are feebly amicable for any l in the same friendly club as n.

8 New Results

We now ask the following question: are all integers feebly amicable with some other integer? To see that this is not true, we establish the following:

Theorem 11. *Let k and m be such that k is coprime with m and $m < k < \sigma(m)$. If some n has abundancy index $\frac{k}{k-m}$, then n is not feebly amicable with any other integer.*

Proof. Suppose that $\lambda(n) = \frac{\sigma(n)}{n} = \frac{k}{k-m}$. Since we already know $\frac{1}{\lambda(n)} + \frac{1}{\lambda(m)} = 1$, then we can substitute in and get:

$$\begin{aligned} \frac{k-m}{k} + \frac{1}{\lambda(m)} &= 1. \\ \frac{1}{\lambda(m)} &= 1 - \frac{k-m}{k} \\ &= \frac{k-(k-m)}{k} \\ &= \frac{m}{k}. \end{aligned}$$

However, there is no such abundancy index as $\lambda(m) = \frac{k}{m}$, by Theorem 6. Thus, proving the theorem. □

To see that this is not vacuous note that we have seen that $\frac{5}{4}$ is not an abundancy index, yet the abundancy index of 14182439040 is known to be $5 = \frac{5}{5-4}$ as it is a multiply perfect number of order 5. See OEIS A007539 [10].

A natural corollary of this theorem is the following:

Corollary 12. *If k is coprime with m, $m < k < \sigma(m)$, and n has abundancy index $\frac{k}{k-m}$, then n is not amicable with any other integer.*

This follows directly from the fact that amicable pairs are feebly amicable pairs. Hence, in particular 14182439040 has no amicable pair.

Proposition 4 showed that the only number with abundancy index $\frac{p+1}{p}$ was p where p was prime. The question arises, is p feebly amicable with any number? Naturally, such a number would need abundancy index $p+1$. That is, $p+1$-perfect. We state this as a result:

Proposition 13. *For p prime, p is feebly amicable with n if and only if n is $(p+1)$-perfect.*

It is unknown if there are any coprime amicable pairs [7]. Hence a natural question is whether there are any coprime feebly amicable pairs. The data reveals that there are none less than $1,000$ but that the first coprime feebly amicable pair is 1485 and 868. There are another four pairs before $5,000$.

9 Examples

We now generate the first 20 feebly amicable pairs that are not amicable or pairs of perfect numbers. To do so, we implemented the following code:

```
for(i in 1:100000)
{for(j in 1:i)
{if(1/abun[i]+1/abun[j]==1)
{print(i) print(j)}}}
```

These pairs are consistent with those listed in [10] in A253534 and A253535.

12	4
30	14
40	10
44	20
56	8
84	15
96	26
117	60
120	2
135	42
140	14
182	66
184	88
190	102
198	45
224	10
234	4
248	174
252	153
260	164

10 Questions and Conjectures

This section considers some questions and conjectures that we have encountered in the course of writing this paper.

Question 1 Are there infinitely many coprime feebly amicable pairs?

Given the regularity of coprime pairs of feebly amicable pairs, we conjecture that there are an infinite number.

In [8], Paul Erdős proved that the asymptotic density of amicable integers relative to the positive integers was 0. That is, the ratio of the number of amicable numbers less than n with n tends to zero as n tends to infinity. This gives rise to the question:

Question 2 What is the density of feebly amicable numbers relative to the positive integers?

Given the number of feebly amicable numbers in the first $5,000$ integers is 310, then 178 in the next $5,000$, and 136 in the third $5,000$, it appears that the density is decreasing and so the asymptotic density is at least less than $0.0272 = \frac{136}{15,000}$. Indeed, [12] gives an upper bound and confirms that the asymptotic density is 0.

The sum of amicable numbers conjecture [11] states that as the largest number in an amicable pair approaches infinity, the percentage of the sums of the amicable pairs divisible by ten approaches 100%. We therefore ask the question:

Question 3 As the larger number in a feebly amicable pair approaches infinity, does the percentage of the sums of the pairs divisible by ten approach 100%?

Our data tells us that in the first 5000 numbers there are 11 feebly amicable pairs that sum to a multiple of 10. There are then 8 in the next 5000 and 4 between 10000 and 15000. As a sequence of fractions of the number of feebly amicable pairs this is: $0.035, 0.045, 0.029$. This does not give us any noticeable trend and once again much higher values would be required to make a strong conjecture other than to say it does not appear that the percentage tends towards 100% as in the amicable case. Computing further is certainly possible, but it becomes much more computationally expensive to compute sum of divisors and hence abundancy indices. [13] provides an algorithm to compute in $O(n^{\frac{1}{3}})$ time, but our code was much more crude.

Bibliography

[1] Calvin T. Long. Elementary introduction to number theory. Prentice Hall, 1987.

[2] Mitsuo Kobayashi. On the density of abundant numbers. Dartmouth College, 2010.

[3] John Stillwell,J Stillwell. Mathematics and its History. Springer(3), 1989.

[4] Richard Laatsch. Measuring the abundancy of integers. Mathematics Magazine(59) no.2. Taylor & Francis,84–92, 1989.

[5] Paul A. Weiner. The abundancy ratio, a measure of perfection. Mathematics Magazine(73) no.4. Taylor & Francis,307–310, 2000.

[6] Underwood Dudley. Elementary number theory. Courier Corporation. 2012.

[7] M. García, J. M. Pedersen, H. J. J. Te Riele, Mariano García, Jan Munch Pedersen, Herman Te Riele. Amicable Pairs, a Survey.2003.

[8] Erdős, Paul. On amicable numbers. Publicationes Mathematicae Debrecen (4). 108–111, 1955.

[9] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. . 2017.

[10] OEIS Foundation Inc. (2021).The On-Line Encyclopedia of Integer Sequences. .

[11] Richard Guy. Unsolved problems in number theory. Springer Science & Business Media (1).2004.

[12] Mark Kozek, Florian Luca, Paul Pollack, Carl Pomerance. Harmonious pairs. International Journal of Number Theory(11), no.05,World Scientific, 1633–1651, 2015.

[13] Richard Sladkey. A Successive Approximation Algorithm for Computing the Divisor Summatory Function. 2012.