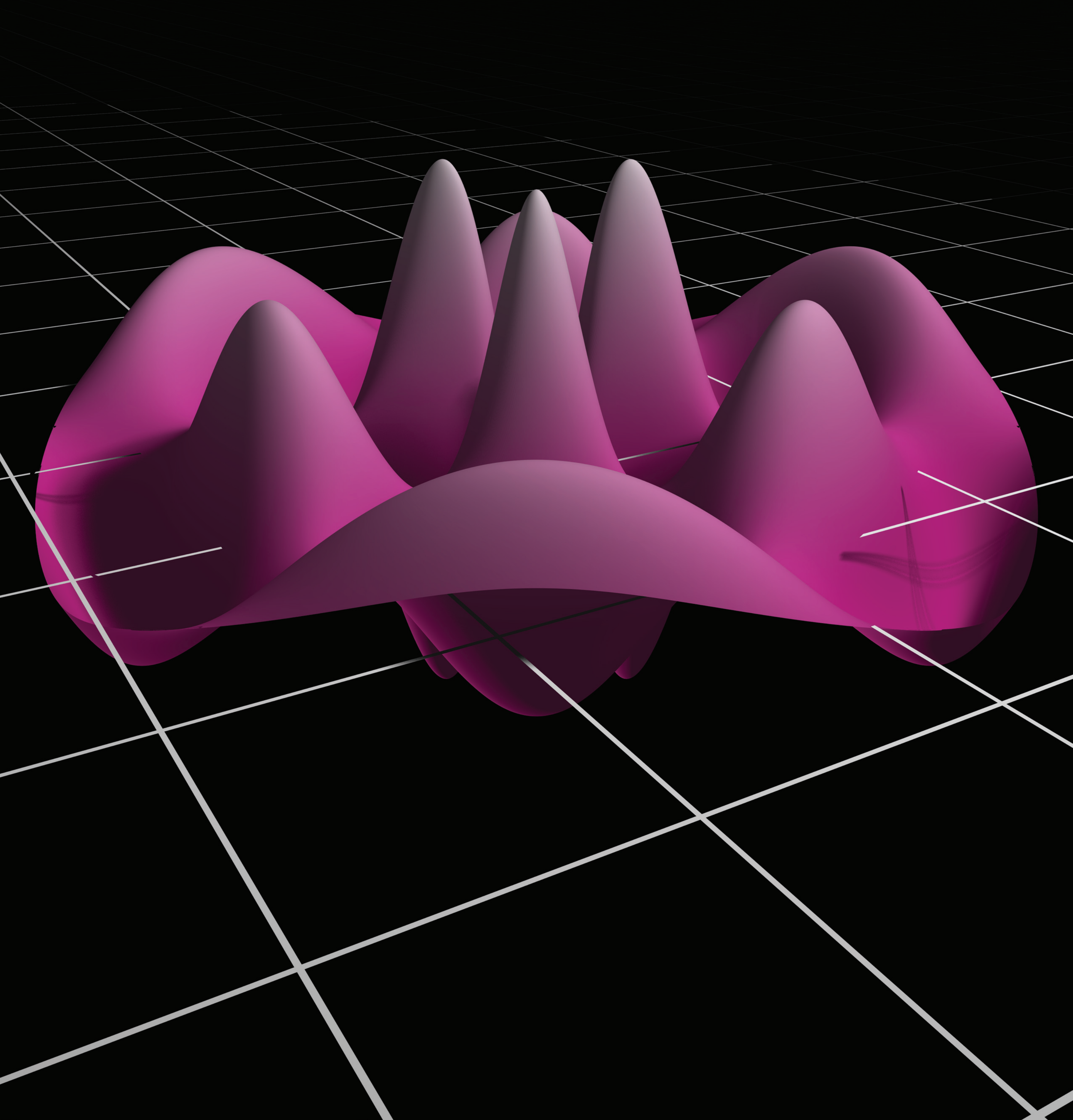
THE WATERLOO MATHEMATICS REVIEW

VOLUME I, ISSUE 2 SPRING 2011



Instructions for authors: All submissions should include the name, institutional affiliation (if applicable), and postal and electronic mail addresses of each author. General inquiries may be addressed to the editorial board at `editor@mathreview.uwaterloo.ca`.

Articles: *The Waterloo Mathematics Review* invites both original research and quality expository articles written for an advanced undergraduate audience on topics related to mathematics at the undergraduate level including: computer science, mathematical physics, statistics, mathematical finance, and actuarial science.

Articles may be submitted electronically or in hard copy. Electronic submissions should be in Adobe PDF format, accompanied by the relevant source files (figures, TeX files, etc.), and sent as an attachment to `submissions@mathreview.uwaterloo.ca`. Hard copies may be sent to

The Waterloo Mathematics Review
Faculty of Mathematics
University of Waterloo
200 University Ave. West
Waterloo, Ontario, N2L 3G1

Submissions should include an abstract and bibliography (preferably in BibTeX format). Figures must be of a production quality, and submitted in a vector format (e.g. encapsulated PostScript). The *Review* can provide assistance in vectorizing hand-drawn figures.

Mathematical Art: The *Review* invites graphics of a mathematical nature. Submissions of graphics should include both black-and-white and colour versions in production quality, along with a description of the mathematics being depicted and a few words on how the image was produced. Source code submission is not required when images are produced using a custom program, however, should an artist request it the *Review* will publish the source code on our website, to accompany the electronic version of the issue.

Join The Review: Are you passionate about mathematics? The *Review* is currently looking for new members for the review board. If you are an interested undergraduate at the University of Waterloo please contact the editorial board at `editor@mathreview.uwaterloo.ca`. If reviewing papers does not interest you but you would like to be involved (or you are a student at another institution and would like to join the effort) the *Review* is also looking for a new general manager, a webmaster, and local distribution partners at other institutions. Contact the editorial board at `editor@mathreview.uwaterloo.ca` for more information.

The Cover: The cover is an artistic rendering of the $\nu_{3,2}$ eigenfunction of the Laplacian on the disk in $\mathbb{R}^2$. Eigenfunctions of the Laplacian are discussed in Mihai Nica's article in this issue. The cover was designed by Edgar A. Bering IV using the POV-Ray ray tracer with assistance from Mihai Nica.

Archive: Past, present, and future issues of the *Review* are available in whole or in part on our website:

`mathreview.uwaterloo.ca`

Machine-readable citations and bibliography data from articles are also available.

Subscriptions: The *Review* currently does not offer subscriptions. Instead it is distributed free of charge at institutions across Canada. If you would like to see the *Review* at your school, please contact the editors.

ISSN 1927-1417 (Print) ISSN 1927-1425 (Online)

The Waterloo Mathematics Review
Volume I, Issue 2

Spring 2011

Contents

Remarks

From the Editors

Dear Reader,

Thank you for taking a look at the second issue of *The Waterloo Mathematics Review*. We are greatly motivated by the positive feedback the first issue generated, and hope that this one will do the same. The second issue has brought considerable growth to the *Review's* organizational structure; we have more reviewers and an expanded general management staff. This issue also marks an important landmark for the *Review*: the first article submitted from outside the University of Waterloo. As the *Review* grows we become able to cover more areas of mathematics. While there is little overlap between the content of this issue and the last, there are still several subfields that are underrepresented. In particular, while mathematical physicists have been very active contributors, other applied mathematicians have been mostly silent. We strongly encourage students in mathematical finance, mathematical economics, mathematical biology, optimization in practice and theory, and computer science to submit articles to the *Review*.

In addition to looking for a greater diversity in submissions, we would like to see a greater diversity in readership. Several readers have commented that they only read one or two articles in the first issue. One reader commented that she did this because to her mathematics is a tribe of disciplines, connected by a common method and tool set. We disagree with this point of view, to us mathematics is a single work, diverse in its appearance and expression but fundamentally interconnected by more than just method. We submit as evidence the many places in mathematics where two seemingly unrelated theories are revealed to be facets of a common deeper theory. Finding and appreciating these interactions can only be accomplished with a broad study of mathematics, so we encourage you to read (or at least attempt to read) every article. We recognize that this may be a difficult task. As an example, suppose an author wishes to discuss algebraic geometry. To read and understand such an article article, you would need knowledge of ring theory and commutative algebra at the very least. Even if the article were relatively self contained it would not be practical to build up all of the necessary prerequisites. Striking a balance between accessibility and length of an article is a difficult task from the perspective of both an editor and a writer, and we will appreciate any future feedback regarding this balance.

We would again like to sincerely thank Dr. Frank Zorzitto for his continued guidance and support. Additionally, we would like to thank Dean Ian Goulden and the Faculty of Mathematics, the Mathematics Society, and *mathNEWS* for their financial and logistic support in the production of the second issue. We are also grateful for the actions of the 2011 CUMC Organizers, both for distributing the Review at the conference and for organizing an excellent conference; we wish the 2012 organizers similar success. Finally we once again thank our general manager Richard Zsolt for his work. This will be the last issue produced with him as general manager, and going into the fall the position is still vacant. We greatly appreciate his support through the early days of the project, and wish him all the best as he starts law school at Western this fall.

Yours truly,
Edgar A. Bering IV
Eeshan Wagh
Frank Ban
Editors-in-Chief
editor@mathreview.uwaterloo.ca

From the Dean

Congratulations to the Editors-in-Chief of this second issue of *The Waterloo Mathematics Review*! It takes a lot of talent and dedication to tackle a project like this and be successful.

Talent and dedication are common traits among the students in Waterloo's Faculty of Mathematics. As Dean of the Faculty, I'm proud that students have developed this *Review* as a showcase. Readers will enjoy the selection of original articles and, I hope, be inspired to continue digging deeper into the challenging problems that define our discipline.

With best wishes for continued success,
Ian Goulden
Dean, Faculty of Mathematics

From the CMS Student Committee

The Canadian Mathematical Society's Student Committee (Studc) has been gaining momentum over the past couple of years. Our ongoing projects, such as the successful CUMC 2011 held in Laval and CUMC 2012 to be held in Kelowna at UBC Okanagan, are becoming more popular than ever. Also, we have introduced a number of new events at the semi-annual CMS meetings, and the upcoming 2011 CMS Winter Meeting, hosted by Ryerson in Toronto, promises to be an exciting conference filled with student-related activities. In addition to the always-popular student social, we are hosting a panel discussion on using mathematics to succeed in industry and also a professional CV writing workshop.

Feel free to check out our new and improved newsletter, "Notes from the Margin", and stay tuned for our newly redesigned website to be launched this Fall. For more information about the Studc, as well as information on our CUMC scholarship and our support for local conferences, go to our website: http://www.cms.math.ca/students.

The CMS Student Committee

A Combinatorial Approach to Finding Dirichlet Generating Function Identities

Aleksandar Vlasev
Simon Fraser University
azv@sfu.ca

Abstract: This paper explores an integer partitions-based method for obtaining Dirichlet generating function identities. In the process we shall generalize a previous result, obtain previously unknown formulae for the Möbius and Liouville Dirichlet generating functions, and obtain a formula on unit fractions.

1 Introduction

The positive integers can be expressed as sums of positive integers in many ways. For example: $6 = 3 + 3 = 1 + 5 = 2 + 2 + 2$, where 6 is also a partition of 6. The problems concerning such a representation are additive in nature. If order matters, the sum is called a composition. If the order of the summands does not matter, the sum is called a partition. Pak [Pak09] gives a history of partition identities and an overview of various constructions that are used to obtain those identities.

One could also think about the multiplicative properties of the integers. Each positive integer can be represented as a product of positive integers. Each such representation is a multiplicative partition. For example, $24 = 2 \times 12 = 4 \times 6$, where 24 is also its own multiplicative partition. The set of numbers in such a representation is called a factorization. If the order of the summands matters, the factorization is called ordered. When order is immaterial, the factorization is called unordered. Therefore unordered factorizations are the multiplicative equivalent of integer partitions. Sometimes we shall call them multiplicative partitions as well.

Knopfmacher and Mays [KM05, KM03] give results about factorizations. They use algebraic manipulation and bijections to obtain factorization identities. The goal of this paper is to show that symbolic methods, like the ones used by Pak, can be applied with some modifications to obtaining multiplicative partition identities. We will essentially follow both papers and generalize some of the ideas presented. The following two identities appear in the process

$$\frac{\zeta(2s)}{\zeta(s)} = 1 - \sum_{k=1}^{\infty} \frac{p_k^{-s}}{(1+2^{-s})(1+3^{-s})(1+5^{-s})\dots(1+p_k^{-s})}$$
$$\frac{1}{\zeta(s)} = 1 - \sum_{k=1}^{\infty} p_k^{-s}(1-2^{-s})(1-3^{-s})(1-5^{-s})\dots(1-p_{k-1}^{-s}),$$

where p_k is the k-th prime. These are Dirichlet generating functions for the Möbius and Liouville functions from number theory. The identities are a beautiful complement to the following two identities, obtained by Knopfmacher and Mays

$$\zeta(s) = 1 + \sum_{k=1}^{\infty} \frac{p_k^{-s}}{(1-2^{-s})(1-3^{-s})(1-5^{-s})\dots(1-p_k^{-s})}$$
$$\frac{\zeta(s)}{\zeta(2s)} = 1 + \sum_{k=1}^{\infty}(1+2^{-s})(1+3^{-s})(1+5^{-s})\dots(1+p_{k-1}^{-s}).$$

In Section 2 we will introduce the notation that will be used throughout the rest of this paper.

We assume familiarity with Dirichlet generating functions (DGF's). The section will conclude with two basic examples that are important in the study of factorizations. In 2003, Knopfmacher and Mays [KM05] derived several identities for factorization generating functions. For example, the Dirichlet generating function for unordered factorizations and unordered factorizations with distinct divisors are

$$F(s) = \prod_{n=2}^{\infty} \frac{1}{1-n^{-s}}, \quad F_d(s) = \prod_{n=2}^{\infty} (1+n^{-s}).$$

Also, they derived the following DGF for the number of unordered factorizations with largest divisor k

$$F(s) = \frac{k^{-s}}{(1-2^{-s})(1-3^{-s})\dots(1-k^{-s})}.$$

In Section 3 we will use methods of Knopfmacher and Mays [KM05] to treat the general case of an arbitrary environment. In the end of the section we shall derive some new identities not mentioned in said paper. One of those identities will be the DGF for the number of unordered factorizations with smallest divisor k. Based on this identity, we will derive some new identities for two number theoretic DGF's (the Möbius and Liouville functions).

In Section 4 we shall use the symbolic method to obtain a new factorization identity. It is based on the idea of the Durfee square that is used for partitions. In the process we will also obtain an identity on unit fractions.

We would like to thank the reviewers of *The Waterloo Mathematics Review* for catching many errors and making many useful suggestions for improving the exposition of this paper.

2 Unordered Factorizations

We will model our notation after the notation used by Pak [Pak09].

Definition 2.1. Let an *unordered factorization* μ be an integer sequence

$$(\mu_1, \mu_2, \dots, \mu_L), \quad \text{where} \quad \mu_1 \geq \mu_2 \geq \dots \geq \mu_L.$$

The μ_i are the *parts* of the factorization. The *size* of the factorization is

$$|\mu| = \prod_n \mu_n.$$

The *length* is the number of distinct entries in μ (denoted by $L(\mu)$). If $|\mu| = n$, μ is an *unordered factorization* of n (denoted by $\mu \vdash n$). Let $a(\mu)$ and $s(\mu)$ be the largest and smallest parts of μ. The *multiplicity* m_d of an integer d is the number of times it is present in μ. We can also use the notation $\mu = (1^{m_1}, 2^{m_2}, \dots)$.

For example, let $n = 1567641600 = 2^{12}3^7 5^2 7^1$. One unrestricted unordered factorization is

$$\mu = (2^4, 3^3, 4^4, 5^2, 7, 9^2).$$

We can also use the fundamental theorem of arithmetic and decompose n in primes. One such factorization is

$$\nu = (2^{12}, 3^7, 5^2, 7^1),$$

Next

$$a(\mu) = 9, \quad s(\mu) = 2$$
$$a(\nu) = 7, \quad s(\nu) = 2.$$

When considering the additive problem for integer partitions, one is quickly confronted by Ferrers diagrams. For example, they are introduced in Pak's paper under the name Young diagrams. They are not to be confused with Young tableaux. A Ferrers diagram is a collection of squares on $\mathbb{Z}^2$ that represent an integer partition. There are different conventions but the important part is that one axis encodes size and the other encodes multiplicities. We model the definition after the one for Young diagrams used by Pak [Pak09]. Multiplicatively, such a diagram does not work and we need to modify it.

Definition 2.2. A *multiplicative partition diagram* $[\mu]$ of an unordered factorization $\mu \vdash n$ is a collection of 1×1 squares (i, j) on a square Cartesian grid. Let $1 \leq i \leq L(\mu)$ and $1 \leq j \leq m_i$. The parts of the partition are on the i-axis and their multiplicities are on the j-axis. The *complete* diagram $[\mu]_o$ is the multiplicative partition diagram where we have an empty space for every 0 power of a prime.

Note that these diagrams are similar to composition diagrams but have an important difference. In a composition diagram, the horizontal axis encodes the size of the element itself. That is, 5 squares indicate the integer 5. In a multiplicative partition diagram, the horizontal axis encodes multiplicities. That is, 5 squares for a divisor d indicate that d occurs 5 times in the particular factorization of the integer whose diagram we are looking at. Also, the first row of a diagram of an integer n encodes the multiplicity of the smallest positive integer d dividing n, $d \geq 2$. For μ and ν, the diagrams are

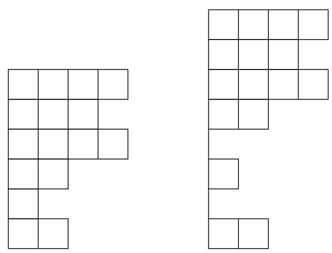


Figure 2.1: $[\mu]$ and $[\mu]_o$ for $\mu = (2^4, 3^3, 4^4, 5^2, 7, 9^2)$

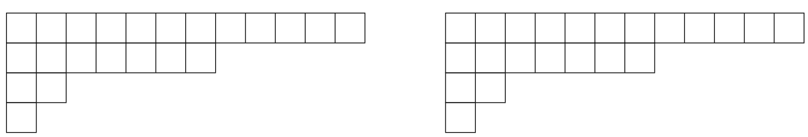


Figure 2.2: $[\nu]$ and $[\nu]_o$ for $\nu = (2^{12}, 3^7, 5^2, 7^1)$. The latter does not have empty spaces because we restrict to prime divisors.

Next, we would like to focus on certain subsets of the positive integers, i.e. the prime numbers, the residues of a certain modulo, etc. Also we would like to focus on certain allowed multiplicities of parts in the multiplicative partition, i.e. distinct divisors, multiplicities from a certain residue class, other finite sets, etc. Here we define a unifying structure for all the different sets of divisors and allowed multiplicities.

Definition 2.3. Let $\mathcal{D}$ be the *set of allowed divisors.* For $d \in \mathcal{D}$, let $\mathcal{M}_d$ be the *set of allowed multiplicities for d.* The *environment* we are working in is the pair $(\mathcal{D}, \mathcal{M})$.

For $\mathcal{M}_d$ and $\mathcal{D}$ we can have the positive integers ≥ 2, the prime numbers, denoted by $\mathbb{P}$ and the integers equivalent to b modulo a. Also, $\mathcal{M}$ is just the collection of sets of allowed multiplicities for the divisors in $\mathcal{D}$. We shall use generating functions to extract information on multiplicative problems. When considering additive partitions, the most often used generating function is the ordinary generating function (OGF for short). For multiplicative problems the DGF is the natural tool.

The most famous DGF is the Riemann Zeta function (the DGF for the sequence $(1, 1, \ldots)$), given by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{k=1}^{\infty} \frac{1}{1 - p_k^{-s}}.$$

The second equality is the Euler product representation for the zeta function. Both sides of the equation converge for s with real part greater than 1. In this paper, we will treat the DGF's formally, so their convergence properties will not be covered. This could be the topic of future work.

Let's start with a motivating example that is covered by Knopfmacher and Mays [KM05]. Let the allowed divisors be the positive integers except 1. For each divisor, let the set of multiplicities be the set of non-negative integers. That is $\mathcal{D} = \mathbb{N}\backslash\{1\}$ and $\mathcal{M}_n = \mathcal{M} = \mathbb{N} \cup \{0\}$ for each divisor n. Next, let $f(n)$ be the number of multiplicative partitions of n with divisors from $\mathcal{D}$. Note that $f(1)$ is thus not defined. For convenience, we define it to be $f(1) = 1$. Our first goal is to find the DGF for the sequence $(f(n))$. Formula (3) [KM05] gives us the following DGF

$$F(s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s} = \prod_{n=2}^{\infty} \frac{1}{1 - n^{-s}}.$$

Using the geometric series, we can expand the right hand side (RHS) as

$$\prod_{t=2}^{\infty} \frac{1}{1 - t^{-s}} = \prod_{t=2}^{\infty} \sum_{k=0}^{\infty} t^{-ks} = \prod_{t=2}^{\infty} \sum_{k=0}^{\infty} \left(t^k\right)^{-s} = \sum_{n=1}^{\infty} c_n n^{-s},$$

where the coefficients c_n count the number of ways we have arrived at n^{-s} through various multiplications. This is exactly the number of unordered factorizations of n. Note that the product runs through all the elements in the set $\mathcal{D}$. Also, the n-th term in the product can be expanded as a geometric sum. Each of those sums runs over the set of allowed multiplicities $\mathcal{M}_n$. Next we consider the following product

$$F_d(s) = \prod_{k=2}^{\infty} (1 + k^{-s}).$$

We can expand the product as in the previous example. If c_n is the coefficient of n^{-s} in the expansion of $F_d(s)$, it again counts unordered factorizations. However, the situation is different this time. For each k of the product, we can choose either k^{-s} or 1 in the expansion. This corresponds to k either being an element of $\mu \vdash n$ or not. However it cannot be an element more than once. Hence, the multiplicity of k is restricted to 0 and 1. This is reflected in the k-th term in the product. From the preceding discussion, $F_d(s)$ is the DGF for the number of unordered factorizations with distinct parts. Also, the product forces us to define $f(1) = 1$ for this problem as well. We could define $f(1) = 1$ but this will put restrictions on what we can do. It is more useful to allow it to be defined by the problem.

Naturally, if we want to restrict ourselves to a finite set of allowed divisors, all we need to do is truncate the product and consider terms only from $\mathcal{D}$. Similarly, we can do the same on each term of the product.

3 General approach

We will now generalize the above discussion. Also, we will use the multiplicative factorization diagrams to our advantage. We begin by writing down another definition

Definition 3.1. Let $(\mathcal{D}, \mathcal{M})$ be an environment. Let C be a constraint. Then n is *reachable* if there is at least one $\mu \vdash n$ from the environment under the constraints C. Let $f(n, C)$ count the number of unordered factorizations of n. Let $f(n) = f(n, \emptyset)$ (unconstrained). Let $\mathcal{M}_d(s)$ be the DGF for the constraints for the divisor d. Then we define

$$F(s;C) = \sum_{n=1}^{\infty} \frac{f(n,C)}{n^s}, \quad M_d(s) = \sum_{m \in \mathcal{M}_d} \frac{1}{d^{ms}}.$$

We also write $F(s;\emptyset) = F(s)$.

The kinds of constraints we will focus on are simple—divisors and/or multiplicities will be restricted in size or number. For example, we can set the largest admissible divisor to some integer or we can set the maximum number of occurrences of a divisor. In the above definition, one could ask why are we taking the sum over all n when all we need is $d \in \mathcal{D}$. The reason is that product identities for $F(s;C)$ may give terms n^{-s}, unreachable in our environment. Therefore, the coefficients $f(n,C)$ for such terms must be defined through the product. For example, we set $f(1) = 1$ in the determination of the DGF for all multiplicative partitions. In that example, the only element unreachable from our environment is $n = 1$. This is true in general—either $f(1) = 0$ or we have to define it as $f(1) = 1$. There is no other case of unreachable elements n, for which we have to define $f(n)$. This is not covered by Knopfmacher and Mays in 2003 [KM05]. Next, we have our result about a general environment. We were unable to find a previous proof of this result in the literature

Proposition 3.1. Let $(\mathcal{D}, \mathcal{M})$ be an environment. Then the generating function for $f(n)$ is

$$F(s) = \prod_{d \in \mathcal{D}} M_d(s)$$

The only unreachable element that we may need to define $f(n)$ for is $n = 1$. We have to set $f(1) = 1$ if and only if 0 is in $\mathcal{M}_d(s)$ for every $d \in \mathcal{D}$.

Proof. As in the discussion above, we need to formally expand the product on the RHS. We use the definition for $\mathcal{M}_d(s)$

$$\prod_{d \in \mathcal{D}} M_d(s) = \prod_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}_d} d^{-ms} = \sum_{n=1}^{\infty} \frac{c_n}{n^s}.$$

The coefficient c_n counts the number of times that elements from the product multiply to n^{-s}. If $c_n = 0$, then n is not reachable, since it is not the product of elements from the environment. If $c_n \geq 1$ we have two cases. If $n \geq 2$, there must be some $d \in \mathcal{D}$, $d \geq 2$ with allowed multiplicity greater than 0, such that $d|n$. Therefore, n is reachable and $c_n = f(n)$. Since all divisors are ≥ 1, if $c_n = 1$, this means that $0 \in \mathcal{M}_d$ for every d. Otherwise, we would never reach it. So in this case we need to set $f(1) = 1$ for our environment.

Now, suppose that $f(1) = 1$ and the latter half of the statement is not true. Then there is a d such that $0 \notin \mathcal{M}_d$. Then there is no way we could get $c_1 > 0$, since every product in the expansion of the RHS would have a factor of at least d^{-s}. Then this is a contradiction and the proof is complete. □

The proposition allows us to construct the DGF for an environment $(\mathcal{D}, \mathcal{M})$ in an intuitive and efficient way. For example, if we are unrestricted or if we use only even or distinct divisors we get the DGF's

$$F(s) = \prod_{n=1}^{\infty} \frac{1}{1-n^{-s}}, \quad F_e(s) = \prod_{n=1}^{\infty} \frac{1}{1-(2n)^{-s}}, \quad F_d(s) = \prod_{n=2}^{\infty} (1+n^{-s}).$$

If we only want to use a finite subset of divisors, we just truncate the products. Also, we can readily construct hybrids. The only thing left to do is to check whether 0 is in the allowed multiplicities for each $d \in \mathcal{D}$. This is the case for all the identities so far. This scheme is very similar to the scheme for partitions. The main difference is that in partitions we use OGF's and in the multiplicative case we use DGF's. For example, if we are unrestricted or if we use only even or distinct summands, we get the OGF's

$$P(s) = \prod_{n=1}^{\infty} \frac{1}{1-z^n}, \quad P_e(s) = \prod_{n=1}^{\infty} \frac{1}{1-z^{2n}}, \quad P_d(s) = \prod_{n=1}^{\infty} (1+z^n).$$

We shall look at some simple constraints. The multiplicative partition diagram can help us visualize some transformations and help us derive identities. The first constraint we can look at is that of greatest or smallest elements. The greatest element constraint is covered for a few cases by Knopfmacher and Mays [KM05]. However, the minimum element constraint is not covered. We will do so here and we will obtain some interesting identities. If we have the constraint $C : a(\mu) \leq d$, then we are essentially just truncating $\mathcal{D}$ to elements smaller than or equal to d. If $C : s(a) \geq d$, we have a truncation from below. Similarly one can consider a constraint on the multiplicity of an element. The truncation follows there as well.

Proposition 3.2. Let $(\mathcal{D}, \mathcal{M})$ be an environment. Then

$$F(s; a(\mu) \leq u) = \prod_{d \in \mathcal{D}, d \leq u} M_d(s), \quad F(s; s(\mu) \geq u) = \prod_{d \in \mathcal{D}, d \geq u} M_d(s).$$

Proof. The proof follows from the previous discussion and Proposition 3.1. □

There is an analogous result for partitions. In Section 5 [KM05] Knopfmacher and Mays perform a largest element decomposition. That is, they find the generating function for unordered factorizations of n with largest factor k. We shall imitate their exposition from the point of view of environments and then we shall generalize it. Let $\mathcal{D} = \mathbb{N} \backslash \{1\}$ and $\mathcal{M}_d = \mathbb{N} \cup \{0\}$ for all d in $\mathcal{D}$. We use the constraint $C : a(\mu) = k$. Let n be a positive integer divisible by k. From the paper, the factorization of n "can be written as $k \times a$ where a represents a factorization of n/k into factors $\leq k$". Then authors give the generating function

$$F(s; a(\mu) = k) = \sum_{n=1}^{\infty} \frac{f(n, a(\mu) = k)}{n^s} = \frac{k^{-s}}{(1 - 2^{-s})(1 - 3^{-s}) \ldots (1 - k^{-s})},$$

where we have written it in the new notation. Looking at this function, we can see that the right-hand side is just $k^{-s} F(s; a(\mu) \leq k)$. Next the authors sum the above identity over all k and obtain the identity

$$\prod_{n=2}^{\infty} \frac{1}{1 - n^{-s}} = 1 + \sum_{k=2}^{\infty} \frac{k^{-s}}{(1 - 2^{-s})(1 - 3^{-s}) \ldots (1 - k^{-s})}.$$

In the later part of that section they reapply this method to derive more "Series—Product" identities of similar form. We shall generalize this method and extend it. To do so we will go through the same argument in an arbitrary environment. However, we have to modify the approach a little bit because when we divide by k, we cannot be sure if the remaining multiplicity of k is allowed. The following diagrams show what an arbitrary multiplicative diagram looks like with maximum and minimum element constraints. Note that the bars for u can be of any length. We start with the discussion for the maximum element decomposition and later we will just state the results for the minimum element decomposition without proof because the proofs are analogous.

Proposition 3.3. Let $(\mathcal{D}, \mathcal{M})$ be an environment. Let the largest part of μ be u and let it have multiplicity v (denote this by $C_u^v : a(\mu) = u, m_u = v$). Then

$$F(s, C_u^v) = u^{-vs} \prod_{d \in \mathcal{D}, d < u} M_d(s).$$

Furthermore,

$$F(s, a(\mu) = u) = \prod_{d < u, d \in \mathcal{D}} M_d(s) \sum_{m \in \mathcal{M}_u, m \neq 0} u^{-ms}.$$

If $0 \notin \mathcal{M}_u$, then

$$F(s, a(\mu) = u) = \prod_{d \leq u, d \in \mathcal{D}} M_d(s).$$

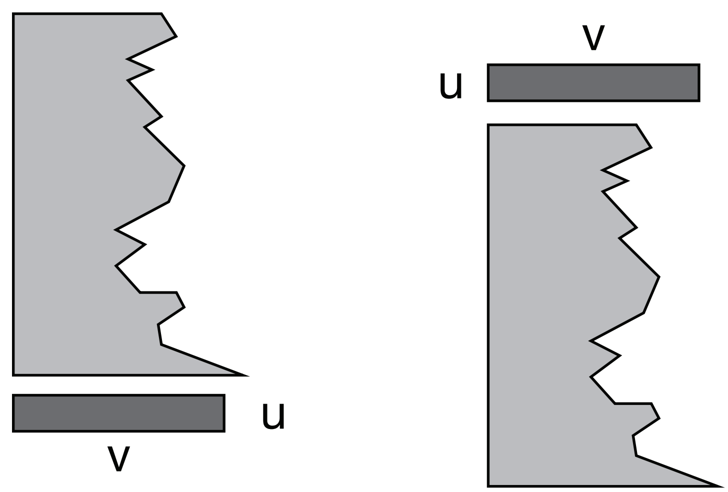


Figure 3.1: Constraints $a(\mu) = u$ and $s(\mu) = u$ with multiplicity v

If $0 \in \mathcal{M}_u$, then

$$F(s, a(\mu) = u) = (M_u(s) - 1) \prod_{d<u, d\in\mathcal{D}} M_d(s).$$

Proof. Suppose that $\mu \vdash n$ is a multiplicative partition satisfying condition C_u^v. Its largest part is u and it has multiplicity v. All multiplicative partitions of n contain v copies of u. We can map each partition of n to a partition of n/u^v by removing u^v. Furthermore this map is a bijection between the multiplicative partition diagrams of n and those of n/u^v. Thus $f(n, C_u^v) = f(n/u^v, a(\mu) < u)$ and we can perform the following summation

$$\begin{aligned} F(s, C_u^v) &= \sum_{n=1}^{\infty} \frac{f(n, C_u^v)}{n^s} = \sum_{n=1}^{\infty} \frac{f(n/u^v, a(\mu) < u)}{(\frac{n}{u^v})^s (u^v)^s} \\ &= u^{-vs} \sum_{n=1}^{\infty} \frac{f(n, a(\mu) < u)}{n^s} = u^{-vs} F(s, a(\mu) < u) \\ &= u^{-vs} \prod_{d\in\mathcal{D}, d<u} M_d(s) \end{aligned}$$

Next, we need to prove the second identity. If μ is a multiplicative partition and $u \in \mu$, then u has some multiplicity. The classes of multiplicative partitions containing μ are disjoint with respect to multiplicity. Therefore, the DGF for all μ containing u is just the sum $F(s, C_u^m)$ over all non-zero multiplicities m in $\mathcal{M}_u$ the identity follows directly. If $0 \notin \mathcal{M}_u$, then

$$\sum_{m\in\mathcal{M}_u, m\neq 0} u^{-ms} = \sum_{m\in\mathcal{M}_u} u^{-ms} = M_u(s).$$

If $0 \in \mathcal{M}_u$, then

$$\sum_{m\in\mathcal{M}_u, m\neq 0} u^{-ms} = -1 + \sum_{m\in\mathcal{M}_u} u^{-ms} = M_u(s) - 1.$$

$\square$

For the example before the proposition, we have that for d, $\mathcal{M}_d$ is the set all possible multiplicities. Therefore

$$F(s, a(\mu) = k) = \left(\frac{1}{1-k^{-s}} - 1\right) \prod_{t=2}^{k-1} \frac{1}{1-t^{-s}} = k^{-s} \prod_{t=2}^{k} \frac{1}{1-t^{-s}},$$

which is what we had before. Similarly, if our divisors are distinct, we have

$$F(s, a(\mu) = k) = (1 + k^{-s} - 1) \prod_{t=2}^{k-1} (1 + t^{-s}) = k^{-s} \prod_{t=2}^{k-1} (1 + t^{-s}),$$

giving us the product inside the summation sign in Identity 28 of Knopfmacher and Mays [KM05]. Similarly, we can perform the same analysis on unordered factorizations on primes to give us Identity 29 of said paper [KM05]

$$F(s, a(\mu) = p_k) = \frac{p_k^{-s}}{(1-2^{-s})(1-3^{-s})\dots(1-p_k^{-s})}.$$

Thus, Proposition 8 unifies and generalizes the approach of Knopfmacher and Mays. The next step is to get identities like Identities 27, 28, 30, and 31 from Knopfmacher and Mays [KM05] right away.

Theorem 3.4. Let $(\mathcal{D}, \mathcal{M})$ be an environment. Then

$$F(s) = c + \sum_{d \in D} F(s, a(\mu) = d),$$

where $c = 1$ if 1 or 0.

Proof. To prove this we need to look at both sides of the equation and compare. Let $n > 1$. Each $F(s, a(\mu) = d)$ enumerates all multiplicative partitions $\mu \vdash n$ that contain a largest factor d. Therefore, for different d, they enumerate disjoint classes. Since we are summing over all $d \in \mathcal{D}$, we are summing over all possible largest factors of μ. Hence the sum on the RHS enumerates all multiplicative partitions of n. The function on the LHS is just the DGF for all multiplicative partitions of n in the current environment. Hence the two sides formally expand to the same sums of n^{-s} for $n > 1$. For the case $n = 1$, if the LHS contains a non-zero coefficient for n^{-s}, we just need to add it to the RHS. By proposition 3.1, this coefficient can only be 1. If the coefficient of 1^{-s} is zero, then $c = 0$. $\square$

Thus Identities 27, 28, 30, and 31 from Knopfmacher and Mays [KM05] fall out of this formula right away. Here are Identities 30 and 31

$$\zeta(s) = 1 + \sum_{k=1}^{\infty} \frac{p_k^{-s}}{(1-2^{-s})(1-3^{-s})(1-5^{-s})\dots(1-p_k^{-s})}$$

$$\frac{\zeta(s)}{\zeta(2s)} = 1 + \sum_{k=1}^{\infty} (1+2^{-s})(1+3^{-s})(1+5^{-s})\dots(1+p_{k-1}^{-s}).$$

Similar analysis can be done considering the smallest element of the multiplicative partition. The following theorem sums up the new results.

Theorem 3.5. Let $(\mathcal{D}, \mathcal{M})$ be an environment. Let $S_u^v : s(\mu) = u, m_u = v$, that is we have the constraint that the smallest part of μ is u and it has multiplicity v. Then

$$F(s, S_u^v) = u^{-vs} \prod_{d \in \mathcal{D}, d > u} M_d(s).$$

Furthermore,

$$F(s, s(\mu) = u) = \prod_{d > u, d \in \mathcal{D}} M_d(s) \sum_{m \in \mathcal{M}_u, m \neq 0} u^{-ms}.$$

Also

$$F(s) = c + \sum_{d \in D} F(s, s(\mu) = d),$$

where $c = 1$ is 1 or 0.

This scheme leads to some interesting consequences. The identity for $F(s)$ from the proposition can be re-written. First,

$$\begin{aligned}
F(s, s(\mu) = d) &= \prod_{d<u, d\in\mathcal{D}} M_d(s) \sum_{m\in\mathcal{M}_u, m\neq 0} u^{-ms} \\
&= F(s)\left(\prod_{d\geq u, d\in\mathcal{D}} M_d(s)\right)^{-1} \sum_{m\in\mathcal{M}_u, m\neq 0} u^{-ms} \\
&= \frac{F(s)}{F(s, s(\mu)\geq d)} \sum_{m\in\mathcal{M}_u, m\neq 0} u^{-ms}.
\end{aligned}$$

If $c = 1$, we get

$$F(s, s(\mu) = d) = \frac{F(s)(M_d(s) - 1)}{F(s, s(\mu) \geq d)}.$$

Then the identity from the last proposition can be re-written as

$$F(s) = 1 + \sum_{d\in\mathcal{D}} F(s, s(\mu) = d) = 1 + F(s) \sum_{d\in\mathcal{D}} \frac{(M_d(s) - 1)}{F(s, s(\mu) \geq d)},$$

giving us the identity

$$\frac{1}{F(s)} = 1 - \sum_{d\in\mathcal{D}} \frac{(M_d(s) - 1)}{F(s, s(\mu) \geq d)}.$$

Next, if $c = 0$, we get

$$F(s, s(\mu) = d) = \frac{F(s)}{F(s(\mu) > d)}$$

and

$$\sum_{d\in\mathcal{D}} \frac{1}{F(s, s(\mu) > d)} = 1.$$

The propositions lead to some interesting identities

$$\begin{aligned}
\frac{1}{\zeta(s)} &= 1 - \sum_{k=1}^{\infty} \frac{p_k^{-s}}{(1+2^{-s})(1+3^{-s})(1+5^{-s})\ldots(1+p_k^{-s})} \\
\frac{\zeta(2s)}{\zeta(s)} &= 1 - \sum_{k=1}^{\infty} p_k^{-s}(1-2^{-s})(1-3^{-s})(1-5^{-s})\ldots(1-p_{k-1}^{-s}).
\end{aligned}$$

They are the DGF's for the Möbius and Liouville functions from number theory. Note the beautiful symmetry between those identities and Identities 30 and 31 from Knopfmacher and Mays [KM05].

Now that we have gone over the general approach we may continue drawing analogies between additive and multiplicative partitions. We may pick many of the constructions given by Pak and transform them into multiplicative analogs but the one for the Durfee square seems simple enough to consider without difficulties, yet complex enough to illustrate the method.

4 Durfee Square equivalent

A Durfee square is the largest square in a Ferrers diagram. We can define a similar object for multiplicative partitions. The *largest square (rectangle)* in a complete multiplicative partition diagram is the largest square (rectangle) rooted at the top-left corner that fits in the diagram. In the case of rectangles, largest refers to area.

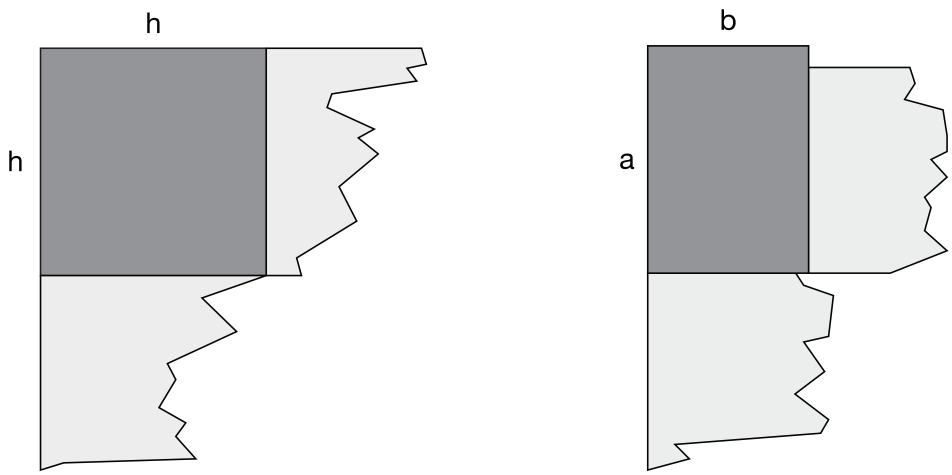


Figure 4.1: A largest square and a largest rectangle.

This can be done for a general environment but is messy. For the unrestricted environment we have two cases. Since the square is the largest, there must be some kind of obstruction on one of its sides. Let's consider two cases. Either $h+2$ has multiplicity $\leq h$ or $\geq h+1$. In the first case, the obstruction is on the lower side of the square. However, it could also be on the right. In the second case the obstruction is strictly on the right. The following diagram shows the two cases. We can look at parts A and B as environments

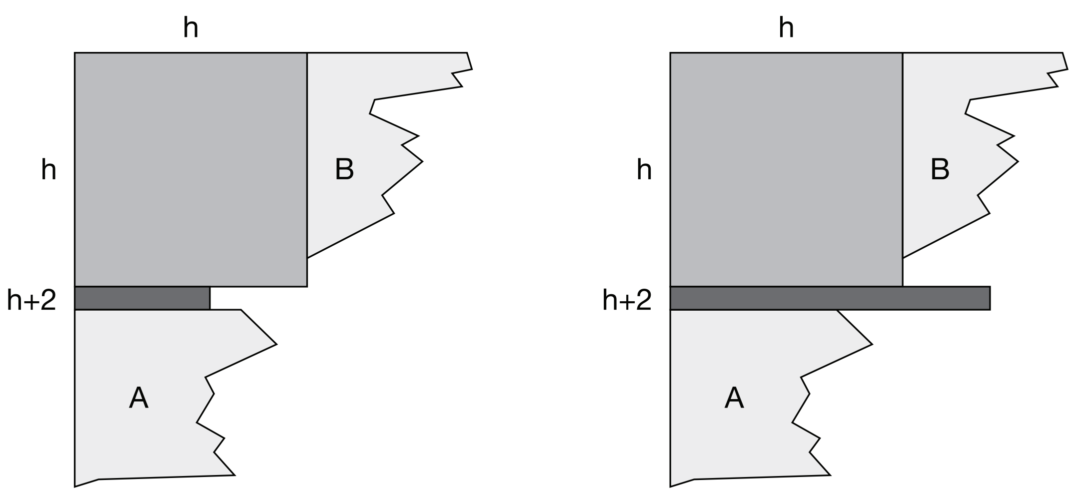


Figure 4.2: The two cases for a largest square.

with some restrictions. We decompose the problem into smaller problems for each case.

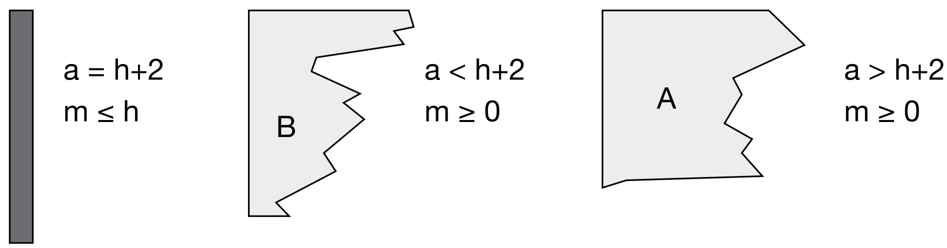


Figure 4.3: Case one.

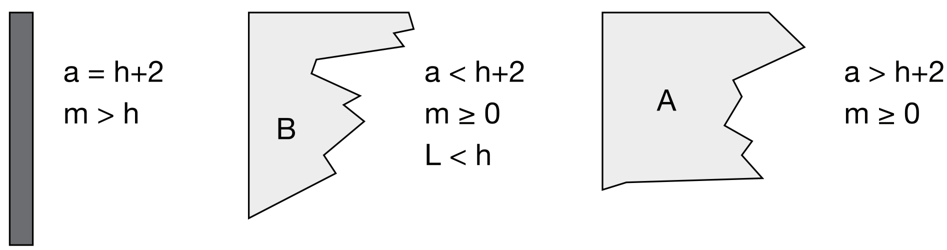


Figure 4.4: Case two.

Let's begin with the first case. The parts of μ in A and B have unrestricted multiplicity. Also A, B, the square, and part $h+2$ are independent of each other. The analysis so far gives us a bijection from the whole diagram to its smaller parts. That is we have a bijection taking the original partition μ to partitions of A, B, and part $h+2$. Since they are independent, we will multiply the generating functions together. First, the DGF for the $h+2$ part is

$$E_1(s) = \sum_{m=0}^{h} (h+2)^{-ms} = \frac{(h+2)^{-hs} - (h+2)^{s}}{1-(h+2)^{s}}.$$

The generating function for the square is

$$F_{hh}(s)^h = \prod_{n=2}^{h+1} n^{-hs} = ((h+1)!)^{-hs}.$$

The DGF's for A and B are

$$A(s) = \prod_{n=h+3}^{\infty} \frac{1}{1-n^{-s}}, \quad B_1(s) = \prod_{n=2}^{h+1} \frac{1}{1-n^{-s}}.$$

Note that

$$A(s)B_1(s) = (1-(h+2)^{-s}) \prod_{n=2}^{\infty} \frac{1}{1-n^{-s}} = (1-(h+2)^{-s})F(s).$$

Putting everything together, the generating function for this case is

$$F_1(s) = A(s)B_1(s)E_1(s)F_{hh}(s)^h.$$

Next, we work on the second case. Here A and the square have the same generating functions as in the first case. The $h+2$ part has the generating function

$$E_2(s) = \sum_{m=h+1}^{\infty} (h+2)^{-ms} = \frac{(h+2)^{-(h+1)s}}{1-(h+2)^{-s}}.$$

We need to be more careful with the DGF for B. We want to find the DGF for multiplicative partitions where the allowed divisors are from 2 to $h+1$. Also, at least one of those divisors must have multiplicity 0. If we consider the set of all allowed multiplicities, we can get the set where at least one of the divisors has multiplicity 0 by taking the whole set and subtracting the set where none of the divisors has multiplicity 0. In terms of the DGF's, this is

$$\begin{aligned} B_2(s) &= \prod_{n=2}^{h+1} \frac{1}{1-n^{-s}} - \prod_{n=2}^{h+1} \frac{n^{-s}}{1-n^{-s}} = B_1(s) - B_1(s)\prod_{n=2}^{h+1} n^{-s} \\ &= B_1(s)(1-F_{hh}(s)). \end{aligned}$$

Thus the DGF for the second case is

$$A(s)B_1(s)(1-F_{hh}(s))E_2(s)F_{hh}(s)^h.$$

Overall, since the cases are disjoint and all-including, the DGF for unordered factorizations with largest square $h \times h$ is

$$\begin{aligned} F_h(s) &= F_1(s) + F_2(s) = A(s)B_1(s)F_{hh}(s)^h(E_1(s) + (1-F_{hh}(s))E_2(s)) \\ &= (1-(h+2)^{-s})F(s)F_{hh}(s)^h(E_1(s) + (1-F_{hh}(s))E_2(s)). \end{aligned}$$

We can simplify the expression in the brackets a bit. It is

$$\begin{aligned} E_1(s) + E_2(s) - F_{hh}(s)E_2(s) &= \sum_{n=0}^{\infty}(h+2)^{-sm} - F_{hh}(s)E_2(s) \\ &= \frac{1}{1-(h+2)^{-s}} - \frac{(h+2)^{-(h+1)s}}{1-(h+2)^{-s}}((h+1)!)^{-s} \\ &= \frac{1}{1-(h+2)^{-s}} - \frac{(h+2)^{-hs}}{1-(h+2)^{-s}}((h+2)!)^{-s} \\ &= \frac{1-((h+2)!)^{-s}(h+2)^{-hs}}{1-(h+2)^{-s}}. \end{aligned}$$

Then $F_h(s)$ becomes

$$F_h(s) = \frac{F(s)}{((h+1)!)^{hs}}\left(1 - \frac{1}{((h+2)!)^s(h+2)^{hs}}\right).$$

The size h of the square can be any non-negative number. The size of the largest square defines equivalence classes among the multiplicative partitions. Hence if we sum $F_h(s)$ over all h, we will get $F(s)$ back. After cancelling $F(s)$ from both sides we have the following curious identity

Proposition 4.1. Formally, the following identity holds

$$\sum_{h=0}^{\infty} \frac{1}{((h+1)!)^{hs}}\left(1 - \frac{1}{((h+2)!)^s(h+2)^{hs}}\right) = 1.$$

Analytically, it seems that the identity holds for complex s with real part greater than 0. We have the following proposition about the partial sums of the above identity.

Proposition 4.2. If $S_N(s)$ is the partial summation of the above identity, then

$$S_N(s) = 1 - (N+2)!^{-(N+1)s}.$$

The formula is easy to prove by induction. The first few partial sums are

$$\begin{aligned} S_0(s) &= 1 - 2^{-s}, \\ S_1(s) &= 1 - 36^{-s}, \\ S_2(s) &= 1 - 13824^{-s}. \end{aligned}$$

5 Conclusion

As one can see from the survey by Pak, there are numerous constructions for additive partition identities. Yet, it seems like there are not that many constructions for multiplicative partition identities. The goal is to change this and in this paper we introduced diagrams and general theory that will be useful in this pursuit. Finally, as the Durfee square equivalent shows, this may not be straightforward but there are many other constructions that one can consider. Analyzing the multiplicative equivalents for other constructions could be the topic of future work.

References

[KM03] A. Knopfmacher and M. E. Mays, *A survey of factorization counting functions*, International Journal Number Theory **1** (2003), 563 – 581.

[KM05] ———, *Dirichlet generating functions and factorization identities*, Congressus Numerantium **173** (2005), 117 – 127.

[Pak09] Igor Pak, *Partition bijections, a survey*, Ramanujan J **12** (2009).

An Introduction to Calibration Estimators

Jennifer H. Nguyen
University of Waterloo
j6nguyen@uwaterloo.ca

Abstract: In survey sampling, the use of auxiliary information can greatly improve the precision of estimates of population total and/or means. In this paper, we explain the basic theory and use of calibration estimators proposed by Deville and Särndal, which incorporate the use of auxiliary data. Results of a simulation study conducted using real data from the 2008 Survey of Household Spending by Statistics Canada are presented, comparing the performance of two calibration estimators against the Horvitz-Thompson estimator. Limitations of calibration estimators and recent extensions made by other leading statisticians in this topic are also discussed.

1 Introduction

The technique of estimation by calibration was introduced by Deville and Särndal in 1992 [DS92]. The idea is to use auxiliary information to obtain a better estimate of a population statistic. First, consider a finite population U of size N with unit labels $1, 2, \ldots, N$. Let y_i, $i = 1, \ldots, N$ be the study variable and $\mathbf{x}_i$, $i = 1, \ldots, N$ be the k-dimensional vector of auxiliary variables associated with unit i.

Suppose we are interested in estimating the population total $t_y = \sum_{i=1}^{N} y_i$. We draw a sample $s = \{1, 2, \ldots, n\} \subset U$ using a probability sampling design P, where the first and second order inclusion probabilities are $\pi_i = Pr(i \in s)$ and $\pi_{ij} = Pr(i, j \in s)$ respectively. An estimate of t_y is the Horvitz-Thompson (HT) estimator

$$\widehat{t}_{HT} = \sum_{i \in s} d_i y_i,$$

where $d_i = 1/\pi_i$ is the sampling weight, defined as the inverse of the inclusion probability for unit i.[1] An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design P. Its variance under P is given as

$$V_p(\widehat{t}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{1.1}$$

Now let us suppose that $\{\mathbf{x}_i,\ i = 1, \ldots, N\}$ is available and $\mathbf{t}_x = \sum_{i=1}^{N} \mathbf{x}_i$, the population total for $\mathbf{x}$, is known. Ideally, we would like

$$\sum_{i \in s} d_i \mathbf{x}_i = \mathbf{t}_x,$$

but often times this is not true.

The idea behind calibration estimators is to find weights w_i, $i = 1, \ldots n$ close to d_i, based on a distance function, such that

$$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{t}_x. \tag{1.2}$$

[1]Please note that d_i and $\frac{1}{\pi_i}$ are used interchangeably in this paper.

We wish to find weights w_i similar to d_i so as to preserve the unbiased property of the HT estimator. Once w_i is found, the calibration estimator for t_y is

$$\widehat{t}_c = \sum_{i \in s} w_i y_i.$$

In Section 2, we discuss how to find w_i for a given sample s and the choice of distance function. The relationship of the calibration estimator to the generalized regression (GREG) estimator is also mentioned. In Section 3, we discuss the expectation and variance of $\widehat{t}_c$ and how to perform variance estimation. Section 4 presents the results of a simulation study to test the efficiency of $\widehat{t}_c$ against $\widehat{t}_{HT}$ using two different distance functions. In Section 5, we discuss advancements made by statisticians on the subject of calibration estimators.

2 Derivation of the Calibration Estimator

Given a sample s, we want to find w_i close to d_i based on a distance function $D(w, d)$ subject to the constraint in Equation 1.2. This is an optimization problem where we wish to minimize

$$Q(w_1, \ldots, w_n, \boldsymbol{\lambda}) = \sum_{i \in s} D(w_i, d_i) - \boldsymbol{\lambda} \left(\sum_{i \in s} w_i \mathbf{x}_i - \mathbf{t}_x \right) \tag{2.1}$$

using the method of Lagrange multipliers.

Examples of distance functions are presented in Table 2.1. We will derive the calibration weights using the Chi-squared distance $(w - d)^2/2qd$ (see Table 2.1), where q is a tuning parameter that can be manipulated to achieve the optimal minimum of Equation 2.1. Note that in practice, the choice of distance function depends on the statistician and the problem.

Table 2.1: Examples of distance functions $D(w, d)$ adapted from Deville and Särndal [DS92]

	$D(w, d)$
1. Chi-squared distance	$(w - d)^2/2qd$
2. Modified minimum entropy distance	$q^{-1}(w \log(w/d) - w - d)$
3. Hellinger distance	$2(\sqrt{w} - \sqrt{d})^2/q$
4. Minimum entropy distance	$q^{-1}(-d \log(w/d) + w - d)$
5. Modified chi-squared distance	$(w - d)^2/2qw$

Letting $D(w_i, d_i) = (w_i - d_i)^2/2q_i d_i$ in Equation 2.1 and differentiating with respect to w_i, we get

$$\frac{\partial Q}{\partial w_i} = \frac{(w_i - d_i)}{q_i d_i} - \boldsymbol{\lambda} \mathbf{x}_i. \tag{2.2}$$

Setting Equation 2.2 to zero and solving for w_i we get

$$w_i = d_i(1 + q_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\lambda}). \tag{2.3}$$

Using the constraint in Equation 1.2 we also get,

$$\boldsymbol{\lambda} = \mathbf{T}_s^{-1}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}),$$

where $\mathbf{T}_s = \sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ and $\widehat{\mathbf{t}}_{x_{HT}}$ is the HT estimator for the population total with respect to the auxiliary variable $\mathbf{x}$.

The resulting calibration estimator of t_y is then

$$
\begin{aligned}
\widehat{t}_c &= \sum_{i\in s} w_i y_i \\
&= \widehat{t}_{y_{HT}} + \sum_{i\in s} d_i q_i \mathbf{x}_i^{\mathrm{T}} \mathbf{T}_s^{-1} (\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}) y_i \\
&= \widehat{t}_{y_{HT}} + \widehat{\mathbf{B}}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}), \qquad (2.4)
\end{aligned}
$$

where $\widehat{\mathbf{B}} = \mathbf{T}_s^{-1} \sum_{i\in s} d_i q_i \mathbf{x}_i y_i$.

Written in this form, we see that $\widehat{t}_c$ is the same as the GREG estimator [CSW76]. In fact, the GREG estimator is a special case of the calibration estimator when the chosen distance function is the Chi-square distance [DS92].

It is important to note that depending on the chosen distance function $D(w,d)$, there may not exist an analytical solution to Equation 2.2 and an approximation of w_i using the Newton-Raphson or a similar method may be required. Furthermore, the solution to Equation 2.2 may yield positive and/or negative weights or extremely large weights, which may be undesirable in a survey sampling context. In terms of efficiency, Deville and Särndal showed that for medium to large samples, the choice of $D(w,d)$ does not make a large impact on the variance of $\widehat{t}_c$ [DS92]. Deville and Särndal also showed that under certain conditions, $\widehat{t}_c$ is asymptotically equivalent to $\widehat{t}_{GREG}$ for any distance function $D(w,d)$ [DS92]. Thus, the choice of distance function is unimportant for large samples, but rather depends on the computational effort of solving Equation 2.2.

3 Expectation and Variance Estimation

To find the expectation and variance of $\widehat{t}_c$, we use the linearization technique to find an approximation of $E_p(\widehat{t}_c)$ and $V_p(\widehat{t}_c)$ with respect to a probability sampling design P. Let $\mathbf{B}$ be the population-level version of $\widehat{\mathbf{B}}$. Then a linear approximation of $\widehat{t}_c$ is

$$
\widehat{t}_c \doteq \underbrace{\widehat{t}_{y_{HT}}}_{O_p(1)} + \underbrace{\mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})}_{O_p(n^{-1/2})} + \underbrace{(\widehat{\mathbf{B}} - \mathbf{B})(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})}_{O_p(n^{-1})}, \qquad (3.1)
$$

where the second term is of order $O_p(n^{-1/2})$ and the last term is of order $O_p(n^{-1})$ as shown by Deville and Särndal [DS92]. Consequently, the last term can be omitted since it is of order $O_p(n^{-1})$. Thus, we can rewrite Equation 3.1 as

$$
\widehat{t}_c \doteq \widehat{t}_{y_{HT}} + \mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}). \qquad (3.2)
$$

Using Equation 3.2, the design-based expectation of $\widehat{t}_c$ is

$$
E_p(\widehat{t}_c) \doteq E_p\left(\widehat{t}_{y_{HT}} + \mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})\right) = t_y.
$$

Thus, $\widehat{t}_c$ is an approximately designed-unbiased estimator of t_y.

Again using Equation 3.2, the designed-based asymptotic variance of $\widehat{t}_c$ is

$$
\begin{aligned}
V_p(\widehat{t}_c) &\doteq V_p\left(t_{y_{HT}} + \mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})\right) \\
&= V_p(t_{y_{HT}} - \mathbf{B}\widehat{\mathbf{t}}_{x_{HT}}) \\
&= V_p\left(\sum_{i\in s} d_i(y_i - \mathbf{B}\mathbf{x}_i)\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} (\pi_{ij} - \pi_i\pi_j)\left(d_i(y_i - \mathbf{B}\mathbf{x}_i)\right)\left(d_j(y_j - \mathbf{B}\mathbf{x}_j)\right) \text{ by Equation 1.1.}
\end{aligned}
$$

Note that since $\mathbf{B}\mathbf{t}_x$ is the true population parameter, $V(\mathbf{B}\mathbf{t}_x) = 0$. The corresponding variance estimator is given by

$$v(\widehat{t}_c) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(d_i (y_i - \widehat{\mathbf{B}} \mathbf{x}_i) \right) \left(d_j (y_j - \widehat{\mathbf{B}} \mathbf{x}_j) \right). \tag{3.3}$$

It is acceptable to use the design weights d_i in the variance estimation but Deville and Särndal suggest that the calibration weights w_i be used in Equation 3.3 as this makes the variance estimator both design-consistent and nearly model-unbiased [DS92]. Moreover, since the calibration estimator is asymptotically equivalent to the GREG estimator, it can be inferred that calibration estimators are more efficient compared to the HT estimator if there is a strong correlation between y and $\mathbf{x}$ [CSW76].

4 Simulation Study

In this section we test the performance of the calibration estimator using distance functions one and two from Table 2.1 against the HT estimator.

4.1 Background and simulation set-up

The data used is obtained from the 2008 Survey of Household Spending conducted by Statistics Canada. There are $N = 9787$ cases. The study variable, y, represents the cost of food purchased from restaurants and the auxiliary variable, x represents the household income before taxes. Figure 4.1 shows a plot of y against x, indicating some positive relationship between restaurant spending and household income, but not a linear relationship.

The statistic of interest is the mean cost of food purchased from restaurants, $\mu_y = t_y/N$, with corresponding estimator $\widehat{\mu}_y = \widehat{t}_y/N$. We treat all $N = 9787$ cases as the finite population. Thus, we know the true population means for y and x are $\mu_y = 1545.74$ and $\mu_x = 71195.52$ respectively.

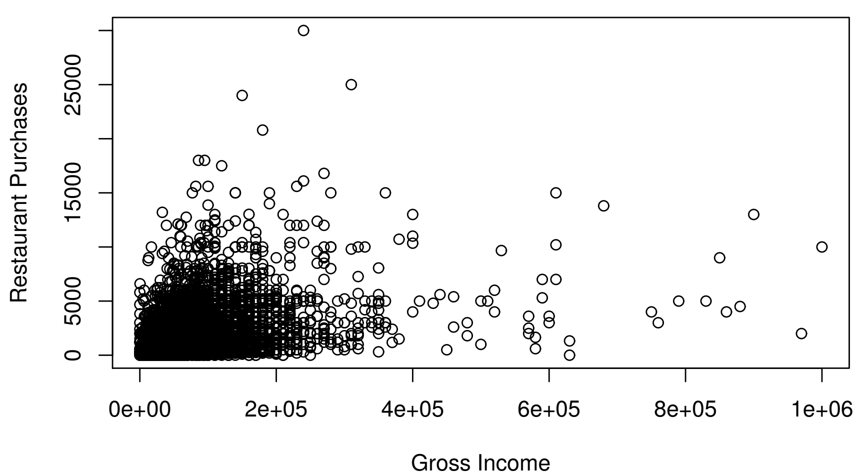


Figure 4.1: Scatter plot of restaurant spending vs. household income [Can10].

A simple regression of the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

was done to see if the model is a good fit of the data. Residual diagnostics show that the residuals meet the basic ordinary least squares assumptions. The R^2 value is only 0.149, indicating that the model explains only some of the variance. Even though the model is poor, we are still assured that $\widehat{t}_c$ is unbiased with respect to the sampling design as demonstrated in Section 3. The correlation between x and y is $\rho_{xy} = 0.387$, which is not strong, but still sufficient to imply that the calibration estimators would provide a better estimate of the total.

The simulation was conducted using the R statistical package. There were $B = 1000$ simulation runs in total. For the b-th run ($b = 1, \ldots, B$), a Bernoulli sample is drawn where each unit is selected into the sample independently with inclusion probability $\pi_i = n/N$. Here we fix $n = 100$. The corresponding HT and calibration estimators of μ_y are computed: $\widehat{\mu}_{y_{HT}}^{(b)}$, $\widehat{\mu}_{y_{c1}}^{(b)}$, and $\widehat{\mu}_{y_{c2}}^{(b)}$. For simplicity, we set the tuning parameter $q_i = 1$. The weights for $\widehat{\mu}_{y_{c1}}$ are given in Equation 2.3. For $\widehat{\mu}_{y_{c2}}$, the weights are of the form

$$w_i = d_i e^{\lambda x_i},$$

where λ is approximated using the constraint in Equation 1.2 with the Newton-Raphson method since there is no closed-form solution.

4.2 Simulation Evaluation

Since each unit is drawn independently, the variance estimators simplify to

$$\begin{aligned} v(\widehat{\mu}_{y_{HT}}) &= N^{-2} \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} y_i^2, \\ v(\widehat{\mu}_{y_{c1}}) &= N^{-2} \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} (y_i - \widehat{B} x_i)^2, \text{ and} \\ v(\widehat{\mu}_{y_{c2}}) &= N^{-2} \sum_{i \in s} (1 - \pi_i)(w_i y_i)^2. \end{aligned}$$

For each estimator of $\widehat{\mu}_y$, a 95% confidence interval $(\widehat{\mu}_L, \widehat{\mu}_U)$ is constructed, where

$$\begin{aligned} \widehat{\mu}_L &= \widehat{\mu}_y - 1.96\sqrt{v(\widehat{\mu}_y)} \text{ and} \\ \widehat{\mu}_U &= \widehat{\mu}_y + 1.96\sqrt{v(\widehat{\mu}_y)}. \end{aligned}$$

To compare the performance of each estimator, we look at four metrics: relative bias (RB), mean square error (MSE), average length of the confidence interval (AL), and the coverage probability (CP) of $\widehat{\mu}_y$. Each measure is calculated as follows:

$$\begin{aligned} RB(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} \frac{\widehat{\mu}_y^{(b)} - \mu}{\mu} \\ MSE(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} (\widehat{\mu}_y^{(b)} - \mu)^2 \\ AL(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} (\widehat{\mu}_U^{(b)} - \widehat{\mu}_L^{(b)}) \\ CP(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} I(\widehat{\mu}_L^{(b)} < \mu < \widehat{\mu}_U^{(b)}). \end{aligned}$$

4.3 Results

The results of the simulation are presented in Table 4.1. We see the relative bias for all three estimators are relatively small, but the variance for the HT estimator is significantly larger than the variances for both calibration estimators, as indicated by their respective mean squared errors. The average length of the confidence interval for calibration estimator number one is also smaller than the HT estimator, but the average length of the confidence interval for calibration estimator two is comparable to the HT estimator. The coverage probabilities for all three confidence intervals are above the 0.90 mark with the coverage probability of calibration estimator number two close to one. These results are reflected in Figure 4.2, which show greater variation in the estimates made by the HT estimator than either calibration estimators.

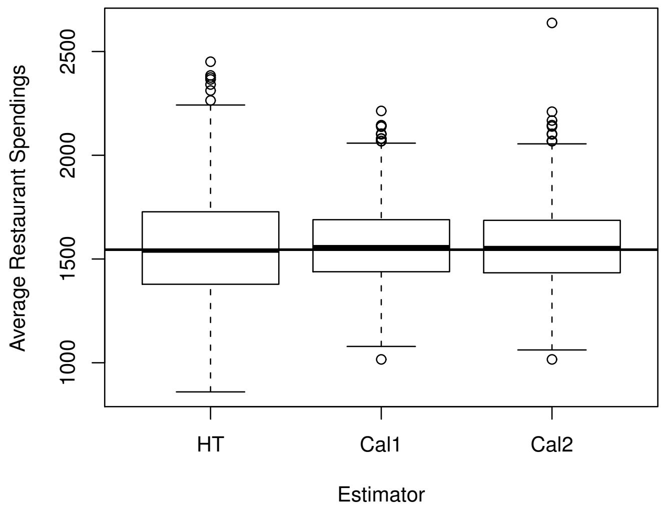


Figure 4.2: Distribution of estimates made by each estimator (the horizontal line indicates the true mean).

5 Extensions and Discussion

We have presented the concept of calibration estimators proposed by Deville and Särndal, which are simply a class of linearly weighted estimators, of which the GREG is a special member. Furthermore, it has been shown that all calibration estimators are asymptotically equivalent to the GREG [DS92].

Consequently, a limitation of the calibration estimator is that it relies on an implicit linear relationship between the study variable, y, and the auxiliary variable $\mathbf{x}$. Thus, if there exists a non-linear relationship

Table 4.1: Performance of estimators from simulation study.

Estimator	RB	MSE	AL	CP
$\widehat{\mu}_{y_{HT}}$	0.011	64669	958.5	0.938
$\widehat{\mu}_{y_{c1}}$	0.014	35615	686.8	0.920
$\widehat{\mu}_{y_{c2}}$	0.011	36269	969.9	0.988

between y and $\mathbf{x}$, the calibration estimator does not perform as well as the HT estimator, that is, if we ignore the auxiliary variable altogether [WS01].

To address this shortcoming, Wu and Sitter developed a model-assisted framework for a model-calibration technique [WS01]. The idea behind model-calibration is to rely on the predicted values, $\widehat{y}_i$, provided by a model ξ, of either linear or non-linear form. As with the original calibration method proposed by Deville and Särndal, the weights w_i are found by minimizing $D(w_i, d_i)$. Instead of using the original calibration constraint in Equation 1.2, however, the minimization is done subject to the constraints

$$\sum_{i \in s} w_i = N \qquad \text{and} \qquad \sum_{i \in s} w_i \widehat{y}_i = \sum_{i=1}^{N} \widehat{y}_i.$$

Wu and Sitter showed that the model-calibration estimator using this technique is more efficient in terms of variance reduction than the simple calibration estimator. It is guaranteed to perform better than the HT estimator, which is sometimes not the case for the original calibration estimator.

This framework paves the way for the use of a variety of models for estimation assistance and generalized the work of Briedt and Opsomer [BO00], who introduced the use of local polynomial regression for estimation. Furthermore, non-parametric models can also be used. Using this framework, Montanari and Ranalli presented a neural network model-calibrated approach [MR05].

Another limitation of the calibration estimator previously mentioned is that the weights can take on negative and/or extremely large values. Deville and Särndal recognized this issue and showed how to restrict the weights to fall within a certain range. Since then, many other methods have been developed to remedy this issue. One such method proposed by Rao and Singh uses a ridge shrinkage method to readjust the weights in an iterative fashion to meet the range restriction [RS97].

6 Acknowledgements

I would like to thank Professor Changbao Wu for introducing me to the topic of calibration estimators and for his guidance in the writing of this paper. I would also like to thank the reviewers for their constructive feedback and advice in helping to improve the clarity of this paper.

References

[BO00] F.J. Breidt and J.D. Opsomer, *Local polynomial regression estimators in survey sampling*, Annals of Statistics **28** (2000), 1026–1053.

[Can10] Statistics Canada, *Survey of household spending 2008*, 2010.

[CSW76] C.M. Cassel, C.E. Särndal, and Wretman, *Some results on generalized difference estimation and generalized estimation for finite populations*, Biometrika **63** (1976), 615–620.

[DS92] J.C. Deville and C.E. Särndal, *Calibration estimators in survey sampling*, Journal of the American Statistical Association **87** (1992), 376–382.

[MR05] G.E. Montanari and M.G. Ranalli, *Nonparametric model estimation in survey sampling*, Journal of the American Statistical Association **100** (2005), 1429–1442.

[RS97] J.N.K. Rao and A.C. Sing, *A ridge shrinkage method for range restricted weight calibration in survey sampling*, Proceedings of the Survey Research Methods Section, The American Statistical Association, 1997, pp. 57–64.

[WS01] C. Wu and R.R. Stitter, *A model-calibration approach to using complete auxiliary information from survey data*, Journal of the American Statistical Association **96** (2001), 185–193.

Eigenvalues and Eigenfunctions of the Laplacian

Mihai Nica
University of Waterloo
mcnica@uwaterloo.ca

Abstract: The problem of determining the eigenvalues and eigenvectors for linear operators acting on finite dimensional vector spaces is a problem known to every student of linear algebra. This problem has a wide range of applications and is one of the main tools for dealing with such linear operators. Some of the results concerning these eigenvalues and eigenvectors can be extended to infinite dimensional vector spaces. In this article we will consider the eigenvalue problem for the Laplace operator acting on the L^2 space of functions on a bounded domain in $\mathbb{R}^n$. We prove that the eigenfunctions form an orthonormal basis for this space of functions and that the eigenvalues of these functions grow without bound.

1 Notation

In order to avoid confusion we begin by making explicit some notation that we will frequently use.

For a bounded domain $\Omega \subset \mathbb{R}^n$ we let $L^2(\Omega)$ be the usual real Hilbert space of real valued square integrable functions Ω, with inner product $\langle u, v\rangle_2 := \int_\Omega uv\,dx$ and norm $\|u\|_2 := \left(\int_\Omega u^2\,dx\right)^{1/2}$. We will also encounter the Sobolev space, denoted $H_0^{1,2}(\Omega)$, which is a similar space of real valued function with inner product and norm given instead by

$$\langle u, v\rangle_{1,2} = \int_\Omega (\nabla u \cdot \nabla v + uv)\,dx$$
$$\|u\|_{1,2} = \left(\int_\Omega \left(|\nabla u|^2 + u^2\right)\,dx\right)^{1/2}.$$

Since this space is somewhat less common than $L^2(\Omega)$, the appendix reviews some elementary properties and theorems concerning this space which are useful in our analysis.

Our problem of interest in this article concerns the Laplace operator. This is a differential operator denoted by Δ and is given by

$$\Delta u = \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2},$$

where u is a sufficiently smooth real valued function, $u : \Omega \to \mathbb{R}$ and $x_1, x_2, \dots x_n$ are the coordinates for $\Omega \subset \mathbb{R}^n$.

2 The Eigenvalue Problem

2.1 The eigenvalue equation

We consider the eigenvalue problem for the Laplacian on a bounded domain Ω. Namely, we look for pairs (λ, u) consisting of a real number λ called an *eigenvalue* of the Laplacian and a function $u \in C^2(\Omega)$ called an *eigenfunction* so that the following condition is satisfied

$$\begin{cases} \Delta u + \lambda u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

Such eigenvalue/eigenfunction pairs have some very nice properties, some of which we will explore here. One fact of particular interest is that they form an orthonormal basis for $L^2(\Omega)$. This is an important and useful result to which we will work towards in this article.

Firstly, we will focus our attention to a weaker version of Equation 2.1. That is, we will examine a condition that is a necessary, but not sufficient, consequence of Equation 2.1. In particular, we will look for solutions u in the *Sobolev space* $H_0^{1,2}(\Omega)$ that obey the following equation for all test functions $v \in H_0^{1,2}(\Omega)$:

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \lambda \int_\Omega uv \, dx. \tag{2.2}$$

The following proposition shows that this condition is indeed weaker than Equation 2.1.

Proposition 2.1. If $u \in C^2(\Omega)$ satisfies Equation 2.1 then Equation 2.2 is satisfied too.

Proof. Suppose u is a twice differentiable function $u \in C^2(\Omega)$ that satisfies Equation 2.1. Given any $v \in H_0^{1,2}(\Omega)$, by definition of $H_0^{1,2}(\Omega)$ (see Appendix A), there is a sequence $v_k \in C_0^1(\Omega)$ so that $v_k \to v$ in the $H_0^{1,2}$ norm. We have that for any v_k

$$\begin{aligned} \int_\Omega (\Delta u + \lambda u) v_k \, dx &= 0 \\ \int_\Omega (\Delta u) v_k \, dx &= -\lambda \int_\Omega u v_k \, dx \\ -\int_\Omega \nabla u \cdot \nabla v_k \, dx &= -\lambda \int_\Omega u v_k \, dx, \end{aligned} \tag{2.3}$$

where the last swap of derivatives is justified by the divergence theorem applied to the vector field $v_k \nabla u$ and utilizing the fact that $v_k \in C_0^1(\Omega)$ is compactly supported and so v_k vanishes on the boundary $\partial\Omega$. By definition of the norm on $H_0^{1,2}(\Omega)$ we have that for any $f \in H_0^{1,2}(\Omega)$ that $\|f\|_2 \leq \|f\|_{1,2}$ and $\|\nabla f\|_2 \leq \|\nabla f\|_{1,2}$ which means that since $v_k \to v$ in $H_0^{1,2}(\Omega)$ we automatically have that $v_k \to v$ and $\nabla v_k \to \nabla v$ in $L^2(\Omega)$. In particular, $\langle u, v_k \rangle_2 \to \langle u, v \rangle_2$ and $\langle \nabla u, \nabla v_k \rangle_2 \to \langle \nabla u, \nabla v \rangle_2$. Taking the limit as $k \to \infty$ of the equality in Equation 2.3 and using these limits gives us precisely Equation 2.2 as desired. □

Remark 2.2: Even more interesting perhaps is that the converse also holds. The weak functions $u \in H_0^{1,2}(\Omega)$ that satisfy Equation 2.2 can be shown, via some regularity results, to be smooth functions in $C^\infty(\Omega)$ and will also solve the original eigenvalue problem [McO03]. The proof of these regularity results is technical and would lead us too far from the eigenvalue problem which we investigate here, so we will content ourselves to simply proving results about the eigenfunctions that solve the weak equation, Equation 2.2, in this article.

The advantage of passing from the usual eigenvalue problem, Equation 2.1, to this weak equation is that we have moved from smooth functions to the Sobolev space $H_0^{1,2}(\Omega)$. In this restricted space, we can utilize certain results that would not hold in general and will be crucial to our analysis. The main tool we gain in this space is the Rellich compactness theorem, which allows us to find convergent subsequences of bounded sequences in $H_0^{1,2}(\Omega)$. Without this powerful tool, it would be impossible to prove the results which we strive for. For this reason, we will use Equation 2.2 as our defining equation rather than Equation 2.1. From now on when we refer to "eigenfunctions" or "eigenvalues" we mean solutions in $H_0^{1,2}(\Omega)$ of Equation 2.2 (rather than solutions of Equation 2.1). We will also refer to Equation 2.2 as "*the eigenvalue equation*" to remind ourselves of its importance.

Lemma 2.1. If u_1 and u_2 are eigenfunctions with eigenvalues λ_1 and λ_2 respectively and if $\lambda_1 \neq \lambda_2$ then $\langle u_1, u_2 \rangle_2 = 0$ and moreover $\langle \nabla u_1, \nabla u_2 \rangle_2 = 0$

Proof. Since u_1 and u_2 are both eigenfunctions, they satisfy the eigenvalue equation by definition. Plugging in $v = u_2$ into the eigenvalue equation for u_1 and $v = u_1$ into the eigenvalue equation for u_2 gives

$$\int_\Omega \nabla u_1 \cdot \nabla u_2 \, dx = \lambda_1 \int_\Omega u_1 u_2 \, dx$$
$$\int_\Omega \nabla u_2 \cdot \nabla u_1 \, dx = \lambda_2 \int_\Omega u_2 u_1 \, dx.$$

Subtracting the second equations from the first gives

$$(\lambda_1 - \lambda_2) \int_\Omega u_2 u_1 \, dx = 0,$$

so the condition $\lambda_1 \neq \lambda_2$ allows us to cancel out $\lambda_1 - \lambda_2$ to conclude $\int_\Omega u_2 u_1 = \langle u_1, u_2 \rangle_2 = 0$ as desired. Finally, notice that $\langle \nabla u_1, \nabla u_2 \rangle = \int_\Omega \nabla u_1 \cdot \nabla u_2 \, dx = \lambda_1 \int_\Omega u_1 u_2 \, dx = 0$ too. ☐

2.2 Constrained optimization and the Rayleigh quotient

Consider now the functionals from $H_0^{1,2}(\Omega) \to \mathbb{R}$

$$F(u) = \int_\Omega |\nabla u|^2 \, dx = \|\nabla u\|_2^2$$
$$G(u) = \int_\Omega u^2 \, dx - 1 = \|u\|_2^2 - 1.$$

These functionals have an intimate relationship with the eigenvalue problem. The following results makes this precise.

Lemma 2.2. If $u \in H_0^{1,2}(\Omega)$ is a local extremum of the functional F subject to the condition $G(u) = 0$, then u is an eigenfunction with eigenvalue $\lambda = F(u)$.

Proof. The proof of this relies on the Lagrange multiplier theorem in the calculus of variations setting (this result is exactly analogous to the usual Lagrange multiplier theorem on $\mathbb{R}^n$ with the first variation playing the role of the gradient). The Lagrange multiplier theorem states that if F and G are C^1-functionals on a Banach space X, and if $x \in X$ is a local extremum for the functional F subject to the condition that $G(x) = 0$ then either $\delta G(x)y = 0$ for all $y \in X$ or there exists some $\lambda \in \mathbb{R}$ so that $\delta F(x)y = \lambda \delta G(x)y$ for all $y \in X$. (Here $\delta F(u)v$ denotes the first variation of the functional F at the point u and in the direction of v.)

We use this theorem with the space $H_0^{1,2}(\Omega)$ serving the role of our Banach space, and F, G as defined above playing the role of the functionals under consideration. The first variation of F and G are easily computed

$$
\begin{aligned}
\delta F(u)v &= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(F(u+\epsilon v) - F(u)\right) \\
&= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(\int_\Omega |\nabla u + \epsilon \nabla v|^2\,dx - \int_\Omega |\nabla u|^2\,dx\right) \\
&= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(\int_\Omega |\nabla u|^2 + 2\epsilon \nabla u\cdot\nabla v + \epsilon^2|\nabla v|^2 - |\nabla u|^2\,dx\right) \\
&= \lim_{\epsilon\to 0}\left(\int_\Omega 2\,\nabla u\cdot\nabla v + \epsilon|\nabla v|^2\,dx\right) \\
&= 2\int_\Omega \nabla u\cdot\nabla v\,dx = 2\left\langle \nabla u, \nabla v\right\rangle_2 .
\end{aligned}
$$

A similar calculation yields

$$
\delta G(u)v = 2\int_\Omega uv\ dx = 2\left\langle u, v\right\rangle_2 .
$$

Notice that $\delta G(u)u = 2\left\langle u,u\right\rangle_2 = 2\|u\|_2 = 2$ by the constraint $G(u) = 0$. This means that $\delta G(u)v$ is not identically zero for all $v \in H_0^{1,2}(\Omega)$. Hence, since u is given to be a local extremum of F subject to $G(u) = 0$ and $\delta G(u)$ is not identically zero, the Lagrange multiplier theorem tells us that there exists a λ so that for all $v \in H_0^{1,2}(\Omega)$ we have

$$
\begin{aligned}
\delta F(u)v &= \lambda\delta G(u)v \\
2\left\langle \nabla u, \nabla v\right\rangle_2 &= 2\lambda\left\langle u, v\right\rangle_2 .
\end{aligned}
$$

Cancelling out the constant of 2 from both sides leaves us with exactly the eigenvalue equation! Hence u is an eigenfunction of eigenvalue λ as desired. Moreover, we can calculate λ directly using the fact that the above holds for all $v \in H_0^{1,2}(\Omega)$:

$$
\begin{aligned}
F(u) &= \left\langle \nabla u, \nabla u\right\rangle_2 \\
&= \lambda\left\langle u, u\right\rangle_2 \\
&= \lambda,
\end{aligned}
$$

where we have used $\left\langle u,u\right\rangle_2 = G(u) + 1 = 1$ since $G(u) = 0$ is given. □

Theorem 2.3. There exists some $u \in H_0^{1,2}(\Omega)$ so that u is a global minimum for F subject to the constraint $G(u) = 0$.

Proof. Let us denote by $\mathcal{C}$ the constraint set we are working on, namely $\mathcal{C} = \{u \in H_0^{1,2}(\Omega) : G(u) = 0\}$. Notice that $G(u) = 0$ precisely when $\|u\|_2 = 1$ so $\mathcal{C}$ is the set of unit norm functions. Let $I = \inf\{F(u) : u \in \mathcal{C}\}$ be the infimum of F taken over this constraint set. We will prove that this infimum is actually achieved at some point $u \in \mathcal{C}$. By the definition of an infimum, we can find a sequence $\{u_j\}_{j=1}^\infty \subset \mathcal{C}$ so that $F(u_j) \le I + \frac{1}{j}$. In particular then, $\lim_{j\to\infty} F(u_j) = I$ and we also have that $F(u_j) = \|\nabla u_j\|_2^2 \le I + 1$ for all $j \in \mathbb{N}$. By the Poincaré inequality (Theorem A.1) we have then that $\|u_j\|_2 \le C\|\nabla u_j\|_2 \le C(I+1)$ for some constant C. Adding these inequalities together we see that

$$
\begin{aligned}
\|u_j\|_{1,2}^2 &= \int_\Omega |\nabla u_j|^2 + u_j^2\,dx \\
&= \|\nabla u_j\|_2^2 + \|u_j\|_2^2 \\
&\le (I+1)^2 + C^2(I+1)^2 \\
&< \infty.
\end{aligned}
$$

In particular, this shows that u_j is a *bounded* sequence in $H_0^{1,2}(\Omega)$. Calling upon the Rellich compactness theorem (Theorem A.2), we know that we can find a subsequence $\{u_{j_k}\}_{k=1}^\infty$ of $\{u_j\}_{j=1}^\infty$ that converges in the L^2 sense to some element $\overline{u} \in \overline{\{u_j\}_{j=1}^\infty} \subset L^2(\Omega)$. Moreover, since $H_0^{1,2}(\Omega)$ is a Hilbert space, every bounded sequence contains a weak-convergent subsequence that converges in the weak topology on $H_0^{1,2}(\Omega)$. (It is a fact from the theory of functional analysis that the existence of such weak-convergent subsequences in a Banach space is equivalent to that Banach space being reflexive. As Hilbert spaces are self-dual by the Riesz representation theorem, they are certainly reflexive and hence we can always find such subsequences.) Hence, we may find a subsequence of $\{u_{j_k}\}_{k=1}^\infty$ that converges in the weak topology of $H_0^{1,2}(\Omega)$ to some $\overline{u}' \in H_0^{1,2}(\Omega)$ (for notational ease, we will continue to denote this subsequence by $\{u_{j_k}\}_{k=1}^\infty$). Of course, this subsequence still converges to $\overline{u}$ in $L^2(\Omega)$. Since $u_{j_k} \to \overline{u}$ in $L^2(\Omega)$, it follows that $\overline{u} = \overline{u}'$ i.e. we have that $u_{j_k} \to \overline{u}$ in the weak topology on $H_0^{1,2}(\Omega)$. This allows us to prove the following claim.

Claim. $\|\overline{u}\|_{1,2} \leq \liminf_{k\to\infty} \|u_{j_k}\|_{1,2}$

Proof of claim. Since $u_{j_k} \to \overline{u}$ in the weak topology on $H_0^{1,2}(\Omega)$, we have

$$\begin{aligned}
\|\overline{u}\|_{1,2}^2 &= \langle \overline{u}, \overline{u} \rangle_{1,2} \\
&= \lim_{k\to\infty} \langle \overline{u}, u_{j_k} \rangle_{1,2} \\
&= \liminf_{k\to\infty} \langle \overline{u}, u_{j_k} \rangle_{1,2} \\
&\leq \liminf_{k\to\infty} \|\overline{u}\|_{1,2} \|u_{j_k}\|_{1,2} \\
&= \|\overline{u}\|_{1,2} \liminf_{k\to\infty} \|u_{j_k}\|_{1,2}.
\end{aligned}$$

Cancelling out $\|\overline{u}\|_{1,2}$ from both sides yields the desired result. □

Using the above inequality and the fact that $\|\overline{u}\|_2 = \lim_{k\to\infty} \|u_{j_k}\|_2 = 1$ since $u_{j_k} \to \overline{u}$ in $L^2(\Omega)$, we can compute

$$\begin{aligned}
F(\overline{u}) &= \int_\Omega |\nabla \overline{u}|^2 \, dx \\
&= \int_\Omega \left(|\nabla \overline{u}|^2 + \overline{u}^2 \right) \, dx - \int_\Omega \overline{u}^2 \, dx \\
&= \|u\|_{1,2}^2 - \|u\|_2^2 \\
&\leq \liminf_{k\to\infty} \|u_{j_k}\|_{1,2}^2 - \lim \|u_{j_k}\|_2^2 \\
&= \liminf_{k\to\infty} \left(\|u_{j_k}\|_{1,2}^2 - \|u_{j_k}\|_2^2 \right) \\
&= \liminf_{k\to\infty} \left(\int_\Omega \left(|\nabla u_{j_k}|^2 + {u_{j_k}}^2 \right) \, dx - \int_\Omega {u_{j_k}}^2 \, dx \right) \\
&= \liminf_{k\to\infty} \left(\int_\Omega |\nabla u_{j_k}|^2 \, dx \right) \\
&= \liminf_{k\to\infty} F(u_{j_k}) \\
&\leq \liminf_{k\to\infty} \left(I + \frac{1}{j_k} \right) \\
&\leq I,
\end{aligned}$$

but now, since $\|\overline{u}\|_2 = 1$, we have $\overline{u} \in \mathcal{C}$ so we have $F(\overline{u}) \geq I = \inf\{F(u) : u \in (C)\}$. Hence, combining the inequalities, we see that $F(\overline{u}) = I$ achieves the minimum for F restricted to $\mathcal{C}$ as desired. □

Remark 2.4: Theorem 2.3 shows that $\overline{u}$ is a global minimum of F subject to $G(u) = 0$. In particular then, it is a *local* extremum for F subject to $G(u) = 0$ so applying the result of Lemma 2.2 informs us that $\overline{u}$ is an eigenfunction with eigenvalue $\lambda = F(\overline{u})$. Since this is the smallest possible value of F subject to $G(u) = 0$, this is the smallest possible eigenvalue one could obtain. For this reason we shall call this eigenvalue λ_1 and the associated eigenfunction u_1.

Remark 2.5: By the definition of F, we notice that for any $u \in H_0^{1,2}(\Omega)$ and any scalar $c \in \mathbb{R}$, we have $F(cu) = c^2 F(u)$. This almost-linearity for scalars means that we can remove the condition $G(u) = 0$ from consideration in some sense by normalizing F by $\|u\|_2$. Notice that

$$\begin{aligned} \frac{F(u)}{\|u\|_2^2} &= \frac{\int_\Omega |\nabla u|^2 \, dx}{\|u\|_2^2} \\ &= \int_\Omega \left| \frac{\nabla u}{\|u\|} \right|^2 dx \\ &= F\left(\frac{u}{\|u\|} \right). \end{aligned}$$

Hence, minimizing $F(u)$ subject to $\|u\| = 1$ is the same as minimizing the quotient $\frac{\int_\Omega |\nabla u|^2 \, dx}{\int_\Omega u^2 \, dx}$ with u running in all of $H_0^{1,2}(\Omega)$. This quotient is known as the *Rayleigh quotient*. This gives us a more notationally concise way to write down our smallest eigenvalue

$$\lambda_1 = \inf_{u \in H_0^{1,2}(\Omega)} \frac{\int_\Omega |\nabla u|^2 \, dx}{\int_\Omega u^2 \, dx}.$$

2.3 The sequence of eigenvalues

To find the next eigenvalue, we can do something very similar. We first notice that the second smallest eigenvalue will have an eigenfunction that is orthogonal to u_1 by the result of Lemma 1, so we can restrict the search for this eigenfunction to the subspace $X_1 = \operatorname{span}\{u_1\}^\perp = \{u \in H_0^{1,2}(\Omega) : \langle u, u_1 \rangle_2 = 0\}$. Since this is the null space of the continuous operator $\langle \cdot, u_1 \rangle_2$, this is a closed subspace of $H_0^{1,2}(\Omega)$ and hence can be thought of as a Hilbert space in its own right. By modifying the proof of Lemma 2 slightly by using X_1 as our Banach space rather than all of $H_0^{1,2}(\Omega)$, we see that any $u \in X_1$ that is a local extrema for F subject to $G(u) = 0$ will be an eigenfunction of eigenvalue $\lambda = F(u)$. By modifying the argument of Theorem 1 slightly by changing the restriction set $\mathcal{C}$ to be $\mathcal{C} = \{u \in X_1 : G(u) = 0$, the identical argument shows that there is some $u \in \mathcal{C}$ that achieves the minimum for F on this restricted set. This will be an extremum for F on X_1 subject to the restriction $G(u) = 0$, so by modified Lemma 2 this will be an eigenfunction, call it u_2. By arguments similar to the above, we find the associated eigenvalue λ_2 is

$$\begin{aligned} \lambda_2 &= \min\{F(u) : u \in \mathcal{C} \ \subset X_1\} \\ &= \inf_{u \in X_1} \frac{\int_\Omega |\nabla u|^2 \, dx}{\int_\Omega u^2 \, dx}. \end{aligned}$$

Since $X_1 \subset H_0^{1,2}(\Omega)$, the Rayleigh quotient definition above tells us immediately that $\lambda_1 \leq \lambda_2$. Repeating this same idea inductively, we can define $X_n = \operatorname{span}\{u_1, u_2, \ldots, u_n\}^\perp = \{u \in H_0^{1,2}(\Omega) : \langle u, u_i \rangle_2 = 0 \forall i \in 1, \ldots, n\}$ and by appropriately modifying Lemma 2.2 and Theorem 2.3 we will be able to justify the fact that the nth eigenvalue can be found by

$$\lambda_n = \inf_{u \in X_n} \frac{\int_\Omega |\nabla u|^2 \, dx}{\int_\Omega u^2 \, dx}.$$

Moreover, we can always find a normalized eigenfunction u_n that achieves this lower bound. Since $H_0^{1,2}(\Omega) \supset X_1 \supset X_2 \ldots$, we can see that this generates a sequence of eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \ldots$ and eigenfunction $u_1, u_2, u_3, \ldots$ which are generated in such a way that they are all mutually orthogonal with respect to the $L^2(\Omega)$ inner product (our construction via the Rayleigh quotient restricted to X_n automatically orthogonalizes the eigenspaces of the degenerate eigenvalues). Moreover, these eigenfunctions have been normalized so that $\|u_n\|_2 = 1$ and also, by invoking the result of Lemma 2.1, we have then that $\|\nabla u_n\|_2 = \lambda_n \|u_n\|_2 = \lambda_n$. The following theorem shows that these eigenvalues tend to infinity.

Theorem 2.6. $\lim\limits_{n\to\infty} \lambda_n = \infty$

Proof. This is another result that follows with the help of the Rellich compactness theorem. Since the sequence λ_i is non-decreasing, the only way that they could not tend to infinity is if they are bounded above. Suppose by contradiction that there is some constant M so that $\lambda_n < M$ for all $n \in \mathbb{N}$. Notice then that

$$\begin{aligned}
\|\nabla u_n\|_2^2 &= \int_\Omega \nabla u_n \cdot \nabla u_n \; dx \\
&= \lambda_n \int_\Omega u_n^2 \; dx \\
&= \lambda_n \\
&\leq M,
\end{aligned}$$

where we have used the eigenvalue equation with $v = u_n$ and the fact that $\|u_n\|_2 = 1$. Notice now that the sequence of eigenfunctions is *bounded* in $H_0^{1,2}(\Omega)$ since

$$\begin{aligned}
\|u_n\|_{1,2}^2 &= \int_\Omega |\nabla u_n|^2 + u_n^2 \; dx \\
&= \|\nabla u_n\|_2^2 + \|u_n\|_2^2 \\
&\leq M + 1.
\end{aligned}$$

By the Rellich compactness theorem, we can find a convergent subsequence u_{n_k} converging to some element of $L^2(\Omega)$. This subsequence, being convergent, is an L^2-Cauchy sequence, meaning in particular that $\|u_{n_k} - u_{n_{k+1}}\|_2^2 \to 0$ as $n \to \infty$. But orthonormality of u_n prohibits this as we have

$$\begin{aligned}
\|u_{n_k} - u_{n_{k+1}}\|_2^2 &= \|u_{n_k}\|_2^2 - 2\left\langle u_{n_k}, u_{n_{k+1}} \right\rangle + \|u_{n_{k+1}}\|_2^2 \\
&= 1 - 0 + 1 \\
&> 0.
\end{aligned}$$

This contradiction shows that our original assumption that the eigenvalues are bounded above by some M is impossible. Since the eigenvalues are nondecreasing, this is enough to show $\lim_{n\to\infty} \lambda_n = \infty$, as desired. □

2.4 Orthonormal basis

Finally, we have the machinery to prove that the eigenfunctions are not only an orthonormal set in $L^2(\Omega)$, but they are are a maximal orthonormal set: an orthonormal basis for $L^2(\Omega)$.

Theorem 2.7. For any $f \in L^2(\Omega)$, we can write $f = \sum_{n=1}^{\infty} \alpha_n u_n$ where $\alpha_n = \langle f, u_n \rangle_2$, where this infinite sum converges to f in the $L^2(\Omega)$ norm.

Proof. We first prove the result for functions $f \in H_0^{1,2}(\Omega)$ so that we may freely consider the (weak) derivative of f. Since $H_0^{1,2}(\Omega)$ is dense in $L^2(\Omega)$, this result can be extended to apply to any function $f \in L^2(\Omega)$. Given any $f \in H_0^{1,2}(\Omega)$, let ρ_N be the N-th error term between f and the partial sum $\sum_{n=1}^N \alpha_n u_n$, namely $\rho_N = f - \sum_{n=1}^N \alpha_n u_n$. To show that this sum converges to f in $L^2(\Omega)$ is tantamount to showing that $\|\rho_N\|_2 \to 0$ as $N \to \infty$. Firstly notice that $\langle \nabla\rho_N, \nabla u_k\rangle_2 = 0$ for every $1 \leq k \leq N$ since

$$\begin{aligned}
\langle \nabla\rho_N, \nabla u_k\rangle_2 &= \left\langle \nabla f - \sum_{n=1}^N \alpha_n \nabla u_n, \nabla u_k \right\rangle_2 \\
&= \langle \nabla f, \nabla u_k\rangle_2 - \sum_{n=1}^N \alpha_n \langle \nabla u_n, \nabla u_k\rangle_2 \\
&= \lambda_k \langle f, u_k\rangle_2 - \sum_{n=1}^N \alpha_n \|\nabla u_n\|_2^2 \delta_{nk} \\
&= \lambda_k\alpha_k - \alpha_k \|\nabla u_k\|_2^2 \\
&= \lambda_k\alpha_k - \alpha_k\lambda_k \\
&= 0,
\end{aligned}$$

where we have used the eigenvalue equation with $v = f$ and the orthonormality of u_n. In a very similar way, we have that $\langle \rho_N, u_k\rangle_2 = 0$ for every $1 \leq k \leq N$ since

$$\begin{aligned}
\langle \rho_N, u_k\rangle_2 &= \left\langle f - \sum_{n=1}^N \alpha_n u_n, u_k \right\rangle_2 \\
&= \langle f, u_k\rangle_2 - \sum_{n=1}^N \alpha_n \langle u_n, u_k\rangle_2 \\
&= \alpha_k - \sum_{n=1}^N \alpha_n \delta_{nk} \\
&= 0.
\end{aligned}$$

Since this holds for all $1 \leq k \leq N$ we conclude that $\rho_N \in \operatorname{span}\{u_1, u_2, \ldots, u_N\}^\perp = X_N$. We hence have the following inequality which follows from the Rayleigh quotient definition of λ_{N+1}

$$\begin{aligned}
\frac{\int_\Omega |\nabla\rho_N|^2\,dx}{\int_\Omega \rho_N^2\,dx} &\geq \inf_{u\in X_N} \frac{\int_\Omega |\nabla u|^2\,dx}{\int_\Omega u^2\,dx} \\
&= \lambda_{N+1},
\end{aligned}$$

and hence:

$$\|\nabla\rho_N\|_2^2 \geq \lambda_{N+1}\|\rho_N\|_2^2.$$

This inequality is the crux of the proof, for we see that

$$\begin{aligned}
\|\nabla f\|_2^2 &= \|\nabla\rho_N + \sum_{n=1}^N \alpha_n \nabla u_n\|_2^2 \\
&= \|\nabla\rho_N\|_2^2 + \|\sum_{n=1}^N \alpha_n \nabla u_n\|_2^2 \\
&\geq \lambda_{N+1}\|\rho_N\|_2^2 + 0,
\end{aligned}$$

where we have used the fact that $\langle \nabla \rho_N, \nabla u_k \rangle_2 = 0$ for every $1 \le k \le N$ to enable the Pythagorean theorem in the second equality. Now the fact that the $\lambda_{N+1} \to \infty$ forces $\|\rho_N\|_2 \to 0$ because otherwise, the right hand side of the equation diverges as $N \to \infty$, while the left hand side is independent of N and finite as $f \in H_0^{1,2}(\Omega)$, a contradiction. Hence $\|\rho_N\|_2 \to 0$ meaning that $\displaystyle\sum_{n=1}^{\infty} \alpha_n u_n$ converges to f in the $L^2(\Omega)$ sense, as desired.

To extend this result from functions $f \in H_0^{1,2}(\Omega)$ as above to more general $f \in L^2(\Omega)$ we use the fact that $H_0^{1,2}(\Omega)$ is dense in $L^2(\Omega)$. (This is not surprising since the even more restrictive set $C_0^\infty(\Omega)$ can be shown to be dense in $L^2(\Omega)$). Given any $f \in L^2(\Omega)$, we may find some family $\{f_\epsilon\} \subset H_0^{1,2}(\Omega)$ so that $f_\epsilon \to f$ in $L^2(\Omega)$ as $\epsilon \to 0$. In particular then, by the Cauchy Shwarz inequality, we have for each $n \in \mathbb{N}$ that $\langle f - f_\epsilon, u_n \rangle_2 \to 0$ as $\epsilon \to 0$ and hence $\alpha_{n,\epsilon} = \langle f_\epsilon, u_n \rangle \to \alpha_n = \langle f, u_n \rangle$ in this limit. By careful addition and subtraction by zero, and by use of the Minkowski inequality on $L^2(\Omega)$ we have

$$\left\| f - \sum_{n=1}^{N} \alpha_n u_n \right\|_2 \le \|f - f_\epsilon\|_2 + \left\| f_\epsilon - \sum_{n=1}^{N} \alpha_{n,\epsilon} u_n \right\|_2 + \left\| \sum_{n=1}^{N} (\alpha_{n,\epsilon} - \alpha_n) u_n \right\|_2 ,$$

but now by Bessel's inequality, which holds for any orthonormal set (such as the set u_n by their construction), applied to the function $f_\epsilon - f$, we have that

$$\begin{aligned} \left\| \sum_{n=1}^{N} (\alpha_{n,\epsilon} - \alpha_n) u_n \right\|_2 &\le \left\| \sum_{n=1}^{\infty} (\alpha_{n,\epsilon} - \alpha_n) u_n \right\|_2 \\ &= \left\| \sum_{n=1}^{\infty} \langle f_\epsilon - f, u_n \rangle_2 \right\| \\ &\le \|f - f_\epsilon\|_2, \end{aligned}$$

which is then added to first inequality to get

$$\left\| f - \sum_{n=1}^{N} \alpha_n u_n \right\|_2 \le 2\|f - f_\epsilon\|_2 + \left\| f_\epsilon - \sum_{n=1}^{N} \alpha_{n,\epsilon} u_n \right\|_2 .$$

By taking ϵ small enough so that $2\|f - f_\epsilon\|_2$ becomes arbitrarily small and N large enough so that $\left\| f_\epsilon - \sum_{n=1}^{N} \alpha_{n,\epsilon} u_n \right\|_2$ is arbitrarily small, we can bound $\left\| f - \sum_{n=1}^{N} \alpha_n u_n \right\|_2$ to be arbitrarily small as well, and hence the L^2 difference between f and its N-th partial eigenfunction expansion must vanish in the limit $N \to \infty$. This shows that any $f \in L^2(\Omega)$ can be written as $f = \displaystyle\sum_{n=1}^{\infty} \alpha_n u_n$ in the L^2 sense, where $\alpha_n = \langle f, u_n \rangle_2$, meaning that the eigenfunctions do indeed form an orthonormal basis for all of $L^2(\Omega)$. □

REFERENCES

[Eva98] Lawrence C. Evans, *Partial Differential Equations*, American Mathematical Society, 1998.

[McO03] Robert C. McOwen, *Partial Differential Equations: Methods and Applications, Second Edition*, Prentice Hall, 2003.

A SOBOLEV SPACES

In this appendix we will fill in some background concerning the simplest Sobolev space, $H_0^{1,2}(\Omega)$, which is used in our investigation of the eigenvalues/eigenfunction pairs above. We also prove the Poincaré inequality, which we call on in this analysis and we very roughly motivate the ideas in the proof of the Rellich compactness theorem which is in some ways the cornerstone of many of the results about eigenvalue/eigenfunction pairs.

A.1 The Sobolev space $H_0^{1,2}(\Omega)$

The Sobolev space $H_0^{1,2}(\Omega)$ is a refinement of $L^2(\Omega)$ whose additional structure is of some use to us. One defines $H_0^{1,2}(\Omega)$ by first defining a new inner product on the the set of continuously differentiable, compactly supported functions $C_0^1(\Omega)$, namely the inner product $\langle\cdot,\cdot\rangle_{1,2}$:

$$\langle u, v\rangle_{1,2} = \int_\Omega (\nabla u \cdot \nabla v + uv)\; dx.$$

The induced norm from this inner product is

$$\|u\|_{1,2} = \sqrt{\langle u, u\rangle_{1,2}} = \left(\int_\Omega \left(|\nabla u|^2 + u^2\right)\; dx\right)^{1/2}.$$

Just as $C_0^1(\Omega)$ is not complete in the usual norm $\langle\cdot,\cdot\rangle_2$ from $L^2(\Omega)$, $C_0^1(\Omega)$ with this norm is *not complete*. However, by the definition of this norm, any sequence $\{u_k\}_{k=1}^\infty$ which is Cauchy in the $\|\cdot\|_{1,2}$ norm will be Cauchy in the $L^2(\Omega)$ norm too. This is by virtue of the fact that $\|u_k - u_j\|_2 \leq \|u_k - u_j\|_{1,2} \to 0$ since u_k is $\|\cdot\|_{1,2}$-Cauchy. (This inequality holds as the $H_0^{1,2}(\Omega)$ norm has an extra non-negative term $|\nabla u|^2$ in the integral, which gives a nonnegative contribution to this norm). Since $L^2(\Omega)$ is complete, we conclude that such a Cauchy sequence converges to some $u \in L^2(\Omega)$. By including all the limits of all the $\|\cdot\|_{1,2}$-Cauchy sequences, we get an honest Hilbert space which we denote by $H_0^{1,2}(\Omega)$, called the *Sobolev space*. In other words, the definition of this Sobolev space is

$$H_0^{1,2}(\Omega) = \overline{C_0^1(\Omega)}^{\|\cdot\|_{1,2}}.$$

This is the completion of $C_0^1(\Omega)$ with respect to the $\|\cdot\|_{1,2}$ norm. As remarked before, this completion consists of adding in some $L^2(\Omega)$ functions, and hence the resulting space is a subset of $L^2(\Omega)$.

A.2 Weak derivatives on $H_0^{1,2}(\Omega)$

Notice that by the above definition, the functions $u \in H_0^{1,2}(\Omega)$ do not necessarily have derivatives in the classical sense, but they do have *weak* derivatives defined by $\frac{\partial u}{\partial x_j} = \lim_{k\to\infty} \frac{\partial u_k}{\partial x_j}$ where u_k is any sequence in $C_0^1(\Omega)$ which converges to u in $L^2(\Omega)$. Notice that this is indeed the weak derivative since for any test function $v \in C_0^\infty(\Omega)$ we have that

$$\begin{aligned}\int_\Omega (u - u_k)\left(-\frac{\partial v}{\partial x_j}\right)\; dx &= \left\langle u - u_k, -\frac{\partial v}{\partial x_j}\right\rangle_2 \\ &\leq \|u - u_k\|_2 \left\|\frac{\partial v}{\partial x_j}\right\|_2 \\ &\to 0,\end{aligned}$$

and hence we have that

$$\begin{aligned}\int_\Omega u\left(-\frac{\partial v}{\partial x_j}\right)\; dx &= \int_\Omega \lim_{k\to\infty} u_k \left(-\frac{\partial v}{\partial x_j}\right)\; dx \\ &= \int_\Omega \left(\lim_{k\to\infty} \frac{\partial u_k}{\partial x_j}\right) v\, dx,\end{aligned}$$

where the swap of derivatives is justified by the divergence theorem because both functions are at least $C_0^1(\Omega)$ and have compact support. Since this holds for any test function v, then u is the weak solution to $\frac{\partial u}{\partial x_j} = \lim_{k\to\infty} \frac{\partial u_k}{\partial x_j}$ and this is what we mean when we say the weak derivative of u exists and is equal to this limit.

A.3 The Poincaré inequality

Theorem A.1 (Poincaré Inequality). If Ω is a bounded domain, then there is a constant C depending only on Ω so that

$$\int_\Omega u^2\,dx \le C\int_\Omega |\nabla u|^2\,dx$$

for all $u \in C_0^1(\Omega)$ and by completion for all $u \in H_0^{1,2}(\Omega)$.

Proof. For $u \in C_0^1(\Omega)$, we find an $a \in \mathbb{R}$ large enough so that the cube $Q = \{x \in \mathbb{R}^n : |x_j| < a, 1 \le j \le n\}$ contains Ω. Performing an integration by parts in the x_1-direction then gives (the non-integral terms vanish since $u = 0$ on the boundary of Q)

$$\begin{aligned}\int_\Omega u^2\,dx &= \int_\Omega 1\cdot u^2\,dx\\ &= -\int_\Omega x_1\frac{\partial u^2}{\partial x_1}\,dx\\ &= -2\int_\Omega x_1 u\frac{\partial u}{\partial x_1}\,dx\\ &= 2a\int_\Omega |u||\frac{\partial u}{\partial x_1}|\,dx.\end{aligned}$$

Using the Cauchy-Schwarz inequality for $L^2(\Omega)$ now gives

$$\begin{aligned}\int_\Omega u^2\,dx &\le 2a\int_\Omega |u||\frac{\partial u}{\partial x_1}|\,dx\\ &\le 2a\|u\|_2\|\frac{\partial u}{\partial x_1}\|_2\\ &\le 2a\|u\|_2\|\nabla u\|_2.\end{aligned}$$

Dividing through by $\|u\|_2$ gives the desired result with $C = (2a)^2$. For $u \in H_0^{1,2}(\Omega)$, we find a sequence $\{u_k\}_{k=1}^\infty \subset C_0^1(\Omega)$ converging to u in the $H_0^{1,2}(\Omega)$ norm (this is by definition of $H_0^{1,2}(\Omega)$). We have then that $\|u - u_j\|_2 \le \|u - u_j\|_{1,2} \to 0$ as $j \to \infty$ and similarly $\|\nabla u - \nabla u_j\|_2 \le \|u - u_j\|_{1,2} \to 0$. Hence, by making use of the Cauchy-Schwarz inequality, we have that $\|u_j\|_2 \to \|u\|_2$ and $\|\nabla u_j\|_2 \to \|\nabla u\|_2$ in the limit $j \to \infty$, which allows us to use the Poincaré inequality on $u_j \in C_0^1(\Omega)$ in the limit $j \to \infty$ to conclude that $\int_\Omega u^2\,dx \le C\int_\Omega |\nabla u|^2\,dx$ as desired. □

Theorem A.2 (Rellich Compactness). For a bounded domain Ω, the inclusion map $I : H_0^{1,2}(\Omega) \to L^2(\Omega)$ is a *compact* operator meaning that it takes bounded sets in $H_0^{1,2}(\Omega)$ to *totally bounded* sets (also known as precompact) in $L^2(\Omega)$. By the sequential compactness characterization of compact sets, this is equivalent to saying that for any bounded sequence $\{u_n\}_{n=1}^\infty \in H_0^{1,2}(\Omega)$, there is a subsequence $\{u_{n_k}\}_{k=1}^\infty$ that converges in the L^2 sense to some $u \in L^2(\Omega)$.

Proof sketch. To do in full detail, the proof is rather long and technical, so we will omit most of the details and instead sketch the main themes of the proof. Given any bounded sequence $\{f_n\}_{n=1}^{\infty} \subset H_0^{1,2}(\Omega)$, the idea is to first smooth out the sequence of functions by convolving them with a so-called mollifier function η_ϵ depending on a choice of ϵ so that the resultant sequence of smoothed (also called mollified) functions $\{\eta_\epsilon * f_n\}_{n=1}^{\infty}$ is better behaved than the original sequence $\{f_n\}_{n=1}^{\infty}$ is. By choosing η_ϵ appropriately, so that η_ϵ is bounded and with bounded derivative, one can verify that the resulting sequence of smoothed functions $\{\eta_\epsilon * f_n\}_{n=1}^{\infty}$ will also be bounded and with bounded derivative. This derivative bound is enough to see that this family is equicontinuous, so one can invoke the Arzela-Ascoli theorem to see that these smoothed functions have a uniformly convergent subsequence. Using the boundedness of $\{f_n\}_{n=1}^{\infty}$ in $H_0^{1,2}(\Omega)$ allows one to argue that as $\epsilon \to 0$, these mollified functions converge uniformly back to the original sequence of functions. Since the mollified functions have convergent subsequences and since the mollified functions return to the original sequence, a little more analysis allows one to verify that the original sequence will enjoy a convergent subsequence as well. □

Remark A.3: This theorem is sometimes filed under the title "The Kondrachov compactness theorem", after V. Kondrachov who generalized Franz Rellich's result in the more general compact map $H_0^{1,p}(\Omega)$ into $L^q(\Omega)$ whenever $1 \leq q \leq np/(n-p)$.

A Systematic Construction of Almost Integers

Maysum Panju
University of Waterloo
mhpanju@uwaterloo.ca

Abstract: Motivated by the search for "almost integers", we describe the algebraic integers known as Pisot numbers, and explain how they can be used to easily find irrational values that can be arbitrarily close to whole numbers. Some properties of the set of Pisot numbers are briefly discussed, as well as some applications of these numbers to other areas of mathematics.

1 Introduction

It is a curious occurrence when an expression that is known to be a non-integer ends up having a value surprisingly close to a whole number. Some examples of this phenomenon include:

$$e^{\pi} - \pi = 19.9990999791\ldots$$
$$\left(\frac{23}{9}\right)^5 = 109.0000338701\ldots$$
$$88\ln 89 = 395.0000005364\ldots$$

These peculiar numbers are often referred to as "almost integers", and there are many known examples. Almost integers have attracted considerable interest among recreational mathematicians, who not only try to generate elegant examples, but also try to justify the unusual behaviour of these numbers. In most cases, almost integers exist merely as numerical coincidences, where the value of some expression just happens to be very close to an integer. However, sometimes there actually is a clear, mathematical reason why certain irrational numbers *should* be very close to whole numbers. In this paper, we'll look at the a set of numbers called the Pisot numbers, and how they can be used to systematically construct infinitely many examples of almost integers.

In Section 2, we will prove a result about powers of roots of polynomials, and use this as motivation to define the Pisot numbers. We will also show how Pisot numbers can generate many almost integers. In Section 3, we will explore the set S of Pisot numbers in more detail, and in Section 4, we will list some other properties and applications of the Pisot numbers. Finally, Section 5 will present some concluding remarks.

2 Pisot Numbers and Almost Integers

We begin by recalling some definitions from the study of polynomials.

Definition 2.1. A number is called an *algebraic integer* if it is the root of some polynomial of the form $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, where the coefficients a_i are all integers. If $f(x)$ is the minimal polynomial for some algebraic integer α, then the roots of $f(x)$ other than α are called the *Galois conjugates* of α.

The main result used to motivate the construction of Pisot numbers is that for any monic polynomial $f(x)$, the sum of the powers of the roots of $f(x)$ will always be exactly an integer. This is made precise in the following theorem.

Theorem 2.1. Let $f(x)$ be a monic, irreducible polynomial of degree d with (not necessarily distinct) roots $\theta_1, \ldots, \theta_d$. Then $\theta_1^n + \cdots + \theta_d^n$ is an integer for all integers $n \geq 0$.

Proof. We may write $f(x) = (x - \theta_1) \cdots (x - \theta_d)$. Taking the natural logarithm of both sides gives

$$\log f(x) = \sum_{i=1}^{d} \log (x - \theta_i).$$

Differentiating, we obtain

$$\frac{d}{dx} \log f(x) = \frac{f'(x)}{f(x)} = \frac{1}{x - \theta_1} + \cdots + \frac{1}{x - \theta_d}.$$

If we now substitute $1/x$ for x in the above equation, then we get

$$\frac{f'(1/x)}{f(1/x)} = \frac{1}{1/x - \theta_1} + \cdots + \frac{1}{1/x - \theta_d},$$

and so

$$\begin{aligned} \frac{x^{d-1} f'(1/x)}{x^d f(1/x)} &= \frac{1}{1 - x\theta_1} + \cdots + \frac{1}{1 - x\theta_d} \\ &= \sum_{i=1}^{d} \frac{1}{1 - x\theta_i}. \end{aligned}$$

By expressing the ratio $1/(1 - x\theta_i)$ as an infinite geometric series, this equation becomes

$$\begin{aligned} \frac{x^{d-1} f'(1/x)}{x^d f(1/x)} &= \sum_{i=1}^{d} \sum_{n=0}^{\infty} x^n \theta_i^n \\ &= \sum_{n=0}^{\infty} \left(\sum_{i=0}^{d} \theta_i^n \right) x^n \\ &= \sum_{n=0}^{\infty} t_n x^n, \end{aligned}$$

where we let

$$t_n = \sum_{i=0}^{d} \theta_i^n.$$

To prove the theorem, it remains to show that $t_n \in \mathbb{Z}$ for all integers $n \geq 0$.

We do this by first writing $f(x)$ in a different way. There exist integers $a_0, \ldots, a_{d-1}$ such that we may write

$$f(x) = x^d + a_{d-1} x^{d-1} + \cdots + a_1 x + a_0$$

and so

$$f'(x) = dx^{d-1} + (d-1) a_{d-1} x^{d-2} + \cdots + 2a_2 x + a_1.$$

This gives us that

$$\begin{aligned} x^{d-1} f'(1/x) &= d + (d-1) a_{d-1} x + \cdots + 2a_2 x^{d-2} + a_1 x^{d-1} \\ x^d f(1/x) &= 1 + a_{d-1} x + \cdots + a_1 x^{d-1} + a_0 x^d. \end{aligned}$$

Putting all of this together, we get

$$\begin{aligned}\frac{x^{d-1}f'(1/x)}{x^d f(1/x)} &= \sum_{n=0}^{\infty} t_n x^n \\ x^{d-1}f'(1/x) &= \left(x^d f(1/x)\right)\sum_{n=0}^{\infty} t_n x^n \\ d+(d-1)a_{d-1}x+\cdots+a_1x^{d-1} &= \left(1+a_{d-1}x+\cdots+a_0x^d\right)\sum_{n=0}^{\infty} t_n x^n.\end{aligned}$$

We now compare coefficients on both sides of this equation, to obtain the system

$$\begin{aligned}d &= t_0 \\ (d-1)a_{d-1} &= t_1 + a_{d-1}t_0 \\ (d-2)a_{d-2} &= t_2 + a_{d-1}t_1 + a_{d-2}t_2 \\ &\ \vdots\end{aligned}$$

from which we see that all of the t_i do, in fact, take integer values. Hence $\theta_1^n + \cdots + \theta_d^n \in \mathbb{Z}$ for all integers $n \geq 0$, as desired. $\square$

We now present the definition of a Pisot number. These numbers were first studied by Thue [Thu12] in 1912, and were later looked at by Hardy [Har19] in 1919. However, they only gained popularity in the wider mathematical community after Pisot's dissertation concerning them in 1938 [Pis38].[1]

Definition 2.2. A *Pisot number* is a real, algebraic integer larger than 1 whose Galois conjugates all have absolute value less than 1.

Pisot numbers can be identified by looking at the roots of their minimal polynomials on the complex plane. If $f(x)$ is the minimal polynomial for a Pisot number α, then all of the roots of $f(x)$ lie strictly within the unit disc on the complex plane, except for α, which lies outside the disc on the positive real axis.

Using Theorem 2.1, we will show that by taking high powers of Pisot numbers, we obtain values that are very close to whole numbers. First, however, we need to make more precise what we mean by "closeness to a whole number".

Definition 2.3. Given a real number x, we define the *distance (from x) to the nearest integer* to be $\|x\| = |x - n|$, where n is taken to be the closest integer to x.

For example, we have $\|7\| = 0$, $\|\pi\| = 0.14159\ldots$, and $\|2.6\| = 0.4$. Note that for any real number x, $\|x\|$ ranges between 0 and 0.5, and $\|x\| = 0$ if and only if x is an integer. Effectively, $\|x\|$ measures how close x is to being a whole number, with small values of $\|x\|$ corresponding to almost integers.

Theorem 2.2. If α is a Pisot number, then $\lim_{n\to\infty} \|\alpha^n\| = 0$.

[1] Pisot's preliminary results were independently proven in 1941 by Vijayaraghavan [Vij41], who was also interested in studying this class of numbers. Some mathematicians therefore refer to these numbers as *Pisot-Vijayaraghavan numbers*, or PV numbers, in recognition of the contributions of both mathematicians, as suggested by Salem in 1943. Unfortunately for Vijayaraghavan, however, most of the literature regarding these numbers in refers to them simply as Pisot numbers.

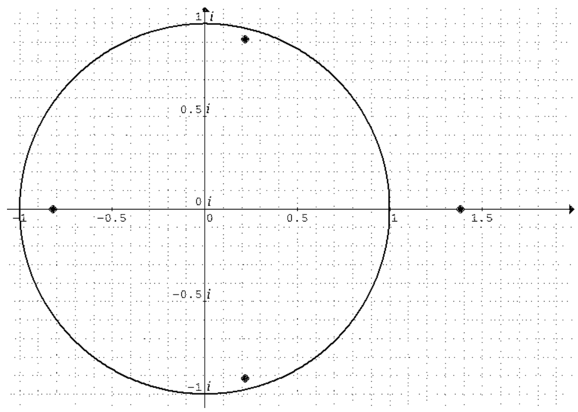


Figure 2.1: A plot of the roots of the polynomial $f(x) = x^4 - x^3 - 1$ in the complex plane, along with the unit circle. Since only one root of the polynomial, approximately 1.38, lies outside the unit disc, this root is a Pisot number.

Proof. Let α have Galois conjugates $\theta_2, \dots, \theta_d$. Since $|\theta_i| < 1$ for all $i = 2, \dots, d$, we have

$$\lim_{n\to\infty} \theta_i^n = 0,$$

for all $i = 2, \dots, d$, and in particular,

$$\lim_{n\to\infty} (\theta_2^n + \theta_3^n + \cdots + \theta_d^n) = 0.$$

From the result of Theorem 2.1, we have for any $n \geq 0$, there exists some integer b_n such that

$$\alpha^n + \theta_2^n + \cdots + \theta_d^n = b_n.$$

We now see that

$$\begin{aligned} b_n - \alpha^n &= \theta_2^n + \cdots + \theta_d^n \\ \lim_{n\to\infty} (b_n - \alpha^n) &= \lim_{n\to\infty} (\theta_2^n + \cdots + \theta_d^n) \\ &= 0 \end{aligned}$$

Thus as n grows large, α^n gets arbitrarily close to the integer b_n, and correspondingly, we have that $\|\alpha^n\|$ approaches 0, as desired. $\square$

At last, we see a systematic way of constructing almost integers. Given any Pisot number α, by taking successively higher powers of α, we obtain values that get closer and closer to whole numbers. The time is ripe for us to look at some examples of Pisot numbers, and the almost integers that they generate.

- For every integer $n \geq 2$, we have n as the root of $f(x) = x - n$, and so every positive integer larger than 1 is a Pisot number. The powers of integers, however, are exactly integers already, so these Pisot numbers are not too useful in generating almost integers.
- The golden ratio $\phi = (1 + \sqrt{5})/2 = 1.61803\dots$ is the root of $f(x) = x^2 - x - 1$, with Galois conjugate $-\phi^{-1} = -0.61803\dots$ having absolute value less than 1. Thus ϕ is a Pisot number.
- $1 + \sqrt{2} = 2.41421\dots$ is a Pisot number, with minimal polynomial $f(x) = x^2 - 2x - 1$ and $1 - \sqrt{2} = -0.41421\dots$ as its Galois conjugate.
- $1.46557\dots$ is a cubic Pisot number, with minimal polynomial $f(x) = x^3 - x^2 - 1$ and $-0.23278\dots \pm i0.79255\dots$ as its Galois conjugates.

Table 2.1: Some Pisot numbers and corresponding almost integers.

α	$1.61803\ldots = \frac{1+\sqrt{5}}{2}$	$2.41421\ldots = 1+\sqrt{2}$	$1.46557\ldots$
α^2	2.6180...	5.8284...	2.1478...
α^3	4.2360...	14.0710...	3.1478...
α^4	6.8541...	33.97056...	4.6134...
α^5	11.0901...	82.01219...	6.7613...
α^{10}	122.99186...	6725.9998513...	45.7161...
α^{15}	1364.000731...	551614.0000018128...	309.10353...
α^{20}	15126.9999338...	45239073.99999997789...	2089.96315...
α^{25}	167761.00000596...	3710155682.000000000269...	14131.01273...

Some of of the almost integers corresponding to powers of these Pisot numbers are listed in Table 2.1.

It is apparent that by taking higher powers of these Pisot numbers, we obtain irrational numbers that get closer and closer to whole numbers. As can be deduced from Theorem 2.2, the rate at which these powers approach whole numbers depends on the how large the absolute values of the Galois conjugates of the Pisot numbers are. We can see that the powers of $1+\sqrt{2}$ approach integers rapidly, since its Galois conjugate has a relatively small modulus of $0.414\ldots$. On the other hand, the Galois conjugates of $1.46557\ldots$ both have absolute value $0.82603\ldots$, which is much closer to 1, resulting in fairly slow convergence to almost integers.

One of Pisot's original and noteworthy results was that the Pisot numbers are the only algebraic numbers that can generate almost integers this way. In particular, he was able to show that if $\alpha > 1$ is an algebraic number, then the existence of some nonzero, real λ such that

$$\lim_{n\to\infty} \|\lambda\alpha^n\| = 0$$

is sufficient to conclude that α is a Pisot number [Pis38]. It is unknown if this result remains true for non-algebraic α, as no transcendental counterexamples have been found.

3 The Structure of S

In the previous section, we have shown that a clever method to construct almost integers is to find a Pisot number and evaluate it at high powers. Only one detail was missing from our discussion; we have not described a process for obtaining a non-trivial Pisot number to begin with, or mentioned anything on the distribution of Pisot numbers. It turns out that the set of Pisot numbers, commonly referred to as S, has been studied in great detail, and is very well understood. In this section, we will look at some properties of the structure of this set.

The set S is countably infinite.

We saw earlier that every natural number larger than 1 is a Pisot number, and hence the set of Pisot numbers must be infinite. On the other hand, S is a strict subset of the set of algebraic integers, which is known to be countable, so S is also a countable set.

The set S is closed.

Recall that a set is closed when it contains all of its limit points. The highly nontrivial fact that S is a closed subset of $\mathbb{R}$ is due to a proof by Salem [Sal44],[2] who clarified that the set of Pisot numbers is not dense in $\mathbb{R}$ (for if S were dense, then every point in $\mathbb{R}$ would be either in S or a limit point of S).

The smallest element of S is known.

The set of Pisot numbers is unbounded from above, since every integer larger than 1 is an element of S. However, it is bounded from below, since Pisot numbers are all strictly larger than 1. Since S is closed, it must have a least element; Siegel [Sie44] proved that this smallest Pisot number is the positive root of $x^3 - x - 1$, which is approximately 1.32472.

The set S has infinitely many limit points.

It is known that the set of limit points of S has limit points of its own. In fact, more than that can be said. Let us define the *derived sets* of S as a sequence of sets $S^{(0)}, S^{(1)}, S^{(2)}, \dots$ such that $S^{(0)} = S$, and for $n \geq 1$, $S^{(n+1)}$ is the set of limit points of $S^{(n)}$. Then it is known that $S^{(n)}$ is nonempty for any finite n. The smallest element of $S^{(2)}$ is known to be 2; that is, 2 is a limit point of limit points of S [Ber80]. It was determined by Bertin [Ber80], in fact, that $n \in S^{(2n-2)}$ for all n. A consequence of this is that there are a huge amount of Pisot numbers clustered around the real line, particularly near the integers.

The subset $S \cap [1, 2]$ is completely understood.

Amara [Ama66] has given a complete characterization of the (infinitely many) limit points of the Pisot numbers less than 2. Talmoudi [Tal78] gave the surprising result that for any of these limit points, there is some small neighbourhood around the limit point such that all Pisot numbers in the neighbourhood can be completely determined using highly structured sequences of polynomials.[3]

For example, the smallest limit point of the Pisot numbers is $\phi = (1+\sqrt{5})/2$, the root of $f(x) = x^2 - x - 1$. According to Talmoudi's classification, any Pisot number sufficiently close to ϕ must be the root of the a polynomial of the form $f(x)x^n + g(x)$, for some $n \geq 1$ and $g(x) \in \{\pm 1, \pm x, \pm(x^2 - 1)\}$. Conversely, any such polynomial will have a Pisot number as a root, for sufficiently large values of n (although the polynomial may not be irreducible in general). Furthermore, as n grows arbitrarily large, the Pisot root of this polynomial will approach the limit point ϕ. These special polynomials described by Amara and Talmoudi give rise to what are called "regular Pisot numbers"; any Pisot number not fitting one of these patterns is called "irregular". The irregular Pisot numbers are much less common than the regular Pisot numbers.

Pisot numbers can be found algorithmically.

Boyd [Boy78, Boy85, Boy84] has presented a remarkable algorithm that deterministically finds all Pisot numbers within any interval $[a, b]$ of the real line, and it is able to detect and compensate for any limit points of S that may occur there. Boyd's algorithm, which was developed over the course of three papers, is particularly useful for finding the irregular Pisot numbers, and has marked a big achievement to further the study of Pisot numbers and related areas. Due to this algorithm, along with the other facts known about S, the set of Pisot numbers is very well understood, and Pisot numbers can be obtained very easily.

[2] Salem's proof in 1944 that S is closed was a strong motivation to continue the study of Pisot numbers, and Pisot later mentioned that he called the set S in order to honour Salem for this contribution.

[3] Although Amara and Talmoudi described their highly nontrivial classification of Pisot number sequences in French, the main results have been summarized in various English papers, for example, by Hare [Har07] and Boyd [Boy96].

4 Other Applications of Pisot Numbers

It turns out that Pisot numbers are useful for more than just generating almost integers. In fact, the Pisot numbers have broad applications that arise in a variety of different areas of mathematics, making them a rich class of objects to study. Some examples are as follows.

Salem numbers.

Closely related to the set of Pisot numbers is the set T of Salem numbers.[4] A Salem number is an algebraic integer whose Galois conjugates are all less than or equal to 1 in absolute value, but with at least one root having an absolute value of exactly 1. Although the definition is very similar to that of Pisot numbers, the set of Salem numbers is much less understood than S. It is known that T is not closed; a long standing open conjecture is whether or not the set T has a least element. The currently known smallest element is the root of a degree ten polynomial found by Lehmer [Leh33], but there is no proof that a smaller one does not exist.[5] Salem numbers exhibit a close relation with the Pisot numbers in that every Pisot number is a limit point for a sequence of Salem numbers. However, whether these are the only limit points of T, like so many other questions concerning Salem numbers, remains unknown [BP90, Boy77].

Mahler measure problems.

The Mahler measure of a polynomial is the product of all of the complex roots of the polynomial that have absolute value larger than 1. The problem of finding polynomials of very small Mahler measure has interested mathematicians for a long time, and continues to be an active area of study [Mos98]. In particular, the Mahler measure of a minimal polynomial of a Pisot or Salem root α is always equal to α, so in this case the problem reduces to finding small Pisot and Salem numbers. The smallest Pisot number is known, as mentioned in Section 3; however, the smallest Salem number (if it exists) is not known, and the study of Mahler measures has therefore motivated the search for the smallest element of T.

Beta expansions.

Rényi [Rén57] introduced the notion of beta expansions as a number representation system, where numbers are written not using base 10, but base β where β may not necessarily be an integer. It was found that surprising things happen when the base of representation is not a whole number; for example, expansions are frequently not unique, and expansions of rational numbers may neither terminate nor repeat. In general, the expansions are chaotic and unpredictable; however, when the base β is chosen to be a Pisot number, then the expansions are much more well behaved. There has been a lot of study in identifying the patterns that arise in these expansions when the base is chosen to be a Pisot number [Bas02, HT08, Pan11].

Fractal tilings, quasicrystals, and more.

Pisot numbers have many applications in dynamical systems, mainly due to the nonuniform distribution of powers of Pisot numbers modulo one. The patterns in the beta expansions involving Pisot numbers can be used to generate fractal tilings of the plane [AI01]. More recently, Pisot numbers have been used to study the aperiodic tilings of quasicrystals [EF05].

[4] Although it was a nice gesture for Pisot to name his set of numbers S after Salem, this convention made notation unnecessarily awkward when Salem introduced his own related class of numbers in 1945.

[5] To date, no Salem number smaller than 1.176..., the root of the polynomial $x^{10} + x^9 - x^7 - x^6 - x^5 - x^4 - x^3 + x + 1$, is known. Despite extensive computer searches, the record still belongs to the polynomial Lehmer found using hand calculations in 1933.

5 Conclusions

The main goal of this paper was to outline a quick and easy method for producing non-integer values that were unusually close to whole numbers. In doing so, we were able to get a glimpse at the structure and properties of the set of Pisot numbers. Although these numbers are useful for generating large quantities of almost integers, we have seen that their study is rich and interesting in its own right, and that the applications of Pisot numbers in mathematics are broad.

There are many unanswered questions related to Pisot numbers, particularly ones involving the related set of Salem numbers. There is also a lot of room for extended study in the applications and properties of these numbers.

6 Acknowledgements

I would like to thank Dr. Kevin Hare for all of his guidance and instruction, and for introducing me to the study of Pisot numbers. I would also like to thank my family for their continued support.

References

[AI01] Pierre Arnoux and Shunji Ito, *Pisot substitutions and Rauzy fractals*, Bull. Belg. Math. Soc. Simon Stevin **8** (2001), no. 2, 181–207, Journées Montoises d'Informatique Théorique (Marne-la-Vallée, 2000). MR 1838930 (2002j:37018)

[Ama66] Mohamed Amara, *Ensembles fermés de nombres algébriques*, Ann. Sci. École Norm. Sup. (3) **83** (1966), 215–270 (1967). MR 0237459 (38 #5741)

[Bas02] Frédérique Bassino, *Beta-expansions for cubic Pisot numbers*, LATIN 2002: Theoretical informatics (Cancun), Lecture Notes in Comput. Sci., vol. 2286, Springer, Berlin, 2002, pp. 141–152. MR 1966122 (2003m:11175)

[Ber80] Marie-José Bertin, *Ensembles dérivés des ensembles $\Sigma_{q,h}$ et de l'ensemble S des PV-nombres*, Bull. Sci. Math. (2) **104** (1980), no. 1, 3–17. MR 560743 (81e:12002)

[Boy77] David W. Boyd, *Small Salem numbers*, Duke Math. J. **44** (1977), no. 2, 315–328. MR 0453692 (56 #11952)

[Boy78] ———, *Pisot and Salem numbers in intervals of the real line*, Math. Comp. **32** (1978), no. 144, 1244–1260. MR 0491587 (58 #10812)

[Boy84] ———, *Pisot numbers in the neighbourhood of a limit point. II*, Math. Comp. **43** (1984), no. 168, 593–602. MR 758207 (87c:11096b)

[Boy85] ———, *Pisot numbers in the neighbourhood of a limit point. I*, J. Number Theory **21** (1985), no. 1, 17–43. MR 804914 (87c:11096a)

[Boy96] ———, *On beta expansions for Pisot numbers*, Math. Comp. **65** (1996), no. 214, 841–860. MR 1325863 (96g:11090)

[BP90] David W. Boyd and Walter Parry, *Limit points of the Salem numbers*, Number theory (Banff, AB, 1988), de Gruyter, Berlin, 1990, pp. 27–35. MR 1106648 (92i:11114)

[EF05] Avi Elkharrat and Christiane Frougny, *Voronoi cells of beta-integers*, Developments in language theory, Lecture Notes in Comput. Sci., vol. 3572, Springer, Berlin, 2005, pp. 209–223. MR 2187264 (2006i:37040)

[Har19] G. Hardy, *A problem of diophantine approximation*, Journal Ind. Math. Soc. **11** (1919), 205–243.

[Har07] Kevin G. Hare, *Beta-expansions of Pisot and Salem numbers*, Computer algebra 2006, World Sci. Publ., Hackensack, NJ, 2007, pp. 67–84. MR 2427721 (2010g:11014)

[HT08] Kevin G. Hare and David Tweedle, *Beta-expansions for infinite families of Pisot and Salem numbers*, J. Number Theory **128** (2008), no. 9, 2756–2765. MR 2444222 (2009e:11143)

[Leh33] D. H. Lehmer, *Factorization of certain cyclotomic functions*, Ann. of Math. (2) **34** (1933), no. 3, 461–479. MR 1503118

[Mos98] Michael J. Mossinghoff, *Polynomials with small Mahler measure*, Math. Comp. **67** (1998), no. 224, 1697–1705, S11–S14. MR 1604391 (99a:11119)

[Pan11] Maysum Panju, *Beta expansions for regular pisot numbers*, Journal of Integer Sequences **14** (2011), no. 11.6.4.

[Pis38] Charles Pisot, *La répartition modulo 1 et les nombres algébriques*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (2) **7** (1938), no. 3-4, 205–248. MR 1556807

[Rén57] A. Rényi, *Representations for real numbers and their ergodic properties*, Acta Math. Acad. Sci. Hungar **8** (1957), 477–493. MR 0097374 (20 #3843)

[Sal44] R. Salem, *A remarkable class of algebraic integers. Proof of a conjecture of Vijayaraghavan*, Duke Math. J. **11** (1944), 103–108. MR 0010149 (5,254a)

[Sie44] Carl Ludwig Siegel, *Algebraic integers whose conjugates lie in the unit circle*, Duke Math. J. **11** (1944), 597–602. MR 0010579 (6,39b)

[Tal78] Faouzia Lazami Talmoudi, *Sur les nombres de* $S \cap [1, 2[$, C. R. Acad. Sci. Paris Sér. A-B **287** (1978), no. 10, A739–A741. MR 516773 (80a:12004)

[Thu12] A. Thue, *Über eine Eigenschaft, die keine transzendente Größe haben kann.*, Videnskapsselskapets Skrifter. I Mat.-naturv (1912) (Norwegian).

[Vij41] T. Vijayaraghavan, *On the fractional parts of the powers of a number. II*, Proc. Cambridge Philos. Soc. **37** (1941), 349–357. MR 0006217 (3,274c)

Relativistic Fluid Dynamics

Jason Olsthoorn
University of Waterloo
jolsthoo@uwaterloo.ca

Abstract: Understanding the evolution of a many bodied system is still a very important problem in modern physics. Fluid mechanics provides a mechanism to determine the macroscopic motion of the system. These equations are additionally complicated when we consider a fluid moving in a curved spacetime. The following paper discusses the derivation of the relativistic equations of motion, uses numerical methods to provide solutions to these equations and describes how the curvature of spacetime is modified by the fluid.

1 Introduction

Traditionally, a fluid is defined as a substance that does not support a shear stress. This definition is somewhat lacking, but it does present the idea that fluids "flow" and distort. Any non-rigid multi-bodied state can, under a suitable continuum hypothesis, be thus described as a fluid and will follow certain equations of motion. Here, we define a relativistic fluid as classical fluid modified by the laws of special relativity and/or curved spacetime (general relativity). The following paper attempts to provide a basic introduction to these equations of motion of a relativistic fluid.

Fluid dynamics is an approximation of the motion of a many body system. A true description of the evolution of a fluid would, in principle, need to account for the motion of each individual particle. However, this description is impractical and of no substantial worth when modelling sufficiently large systems. Therefore, provided that the desired level of accuracy is much lower than the continuum approximation, it is acceptable to consider a system as a fluid. The applications of such an approximation to relativistic fluids are varied and have been applied to the many different domains from plasma physics to astrophysics.

In this discussion, we begin with introducing the relevant equations found in Newtonian fluid mechanics. We follow this with an introduction to the necessary mathematics to describe a four dimensional curved spacetime. The stress-energy tensor of a perfect fluid is introduced and the equations of motion of a relativistic fluid are derived. We briefly mention the modification of the stress-energy tensor in the presence of viscosity. We finish off with a simple calculation of how the stress-energy of the fluid in question modifies the curvature of space-time. The reader is assumed here to have a basic understanding of relativity along with a low-level understanding of Newtonian fluid dynamics.

We note that for the remainder of this paper with will use units such that $c = G = 1$.

2 Introductory Mathematics

Classical fluids have, from a theoretical perspective, played a very important role in developing a great deal of the mathematics of vector calculus of partial differential equations (PDEs), and forms the core of our understanding of problems in multi-body physics. Extension of this classical field to the domain of relativity requires the use of the understanding of motion in a curved spacetime. An introduction to the mathematics required for this development is provided here.

2.1 Classical fluids

Determining solutions to the classical equations of motion of a fluid is still a very active area of research. If we just consider a Newtonian fluid (water and air are both good examples of this type), the strain tensor can be written as

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x^j} - \frac{\partial u_j}{\partial x^i}\right),$$

where u^j, x^j are the j^{th} components of the fluid velocity and coordinate vectors. For the remainder of this paper, we will employed the Einstein summation convection $x^j x_j = \sum_j x^j x_j$ over the range of the indices.

In this domain, there are still four basic equations to satisfy

1. *Continuity equation.* This equation is derived through the hypothesis of conservation of mass. In its typical form, it can be written down as

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{u} = 0,$$

   where $\frac{D}{Dt} = \frac{\partial}{\partial t} + u_j\frac{\partial}{\partial x^j}$ called the material derivative, and ρ the density of the fluid.

2. *Momentum equation.* The fluid must also conserve momentum. We ensure this by requiring that

$$\frac{Du_j}{Dt} = \frac{\partial T_{ij}}{\partial x_j} - \rho g_j$$
$$T_{ij} = -P\delta_{ij} + 2\mu e_{ij} + \lambda e_{mm}\delta_{ij},$$

   where T_{ij} is the stress tensor, P is the pressure , g_j is the constant gravitational acceleration, δ_{ij} is the unity matrix, with μ and λ are fluid dependent scalars. If $\mathbf{u}$ is incompressible, ($\nabla\cdot\mathbf{u} = 0$) these equations reduce to the Navier-Stokes Equations.

3. *Equation of state.* This equation defines the relation between pressure (P), temperature (T), and density (ρ). This equation can vary depending on the fluid in question. For an ideal gas it can be written

$$P = \rho RT$$

   with constant R.

4. *Temperature/Energy equation.* This final equation is needed to deal with the thermodynamic effects within the medium. If we consider the heat flux vector q_i at any given point, we need to solve for internal energy e

$$\rho\frac{De}{Dt} = -\frac{\partial q_i}{\partial x_i} - P\left(\frac{\partial u_i}{\partial x_i}\right) + \phi$$

   with density (ρ), pressure (P), velocity (u_i), and viscous dissipation (ϕ). This equation indicates that the change in energy is due to convergence of heat, volume compression and viscous dissipation.

All this gives us a system of six, non-linear coupled PDEs. These equations have been included to help guide the reader in understanding how the following equations reduce in the Newtonian limit. It is important to realize that these equations have still not been solved and currently represent one of the most challenging problems in applied mathematics. For further details, Kundu [Kun90] has a well written text on classical fluid mechanics.

2.2 Curved spacetime

In order to understand relativistic fluids, if becomes important to develop the mathematical tools to look at curves in a curved spacetime. While the reader is assumed to have a basic knowledge of differential geometry, a brief outline of the some of the mathematics is presented here.

The length of an infinitesimally small line element in 4-space can be found by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu,$$

where μ,ν run from $\{0,1,2,3\}$ or, equivalently, $\{t,x,y,z\}$ in Cartesian spacetime. Note that this line element is invariant of the chosen coordinate system, that is, it is a scalar. Here, the metric $g_{\mu\nu}$ (of form - + + +) serves the role of a weighting function, used in defining the length of a path. In a curved space the placement of the index is very important, and we use the metric $g_{\mu\nu}$ to raise and lower the indices.

$$V^\mu = g^{\mu\nu} V_\nu \qquad \text{or} \qquad V_\mu = g_{\mu\nu} V^\nu.$$

Consequently,

$$V^\mu = g^{\mu\nu} g_{\nu\rho} V^\rho \qquad \rightarrow \qquad g^{\mu\nu} g_{\nu\rho} = \delta^\mu_\rho, \tag{2.1}$$

where δ^μ_ρ is the Kronecker delta.

The infinitesimal length of a curve ds^2 divides up into three different regimes.

1. *Timelike.* If $ds^2 < 0$, the curve is called timelike. Two events are timelike separate if there exists some rest frame, in which both events occur at the same location at different times.

2. *Null.* If $ds^2 = 0$, the curve is null. There does not exist a rest frame.

3. *Spacelike.* If $ds^2 > 0$, the curve is spacelike. Two events are spacelike separated if there exists some rest frame, in which both events occur at the same time at different locations.

It turns out that all matter travels along timelike curves and light moves along null paths. As such, it is possible to define, for timelike curves, a proper time (τ) which is the time measured by an observer in a rest frame.

$$d\tau^2 = -ds^2. \tag{2.2}$$

Using this definition, it is possible to define the 4-vector velocity

$$u^\mu = \frac{dx^\mu}{d\tau}.$$

As a quick aside, the purpose of introducing this tensor calculus is to allow for a derivation of physical laws, independent of a particular coordinate system. As such, a tensor will necessarily obey certain transformation laws. We provide here the transformation relation for a vector, with higher order tensors transforming in a consistent manner.

$$\bar{V}_\mu = \frac{\partial x^\nu}{\partial \bar{x}^\mu} V_\nu \qquad \text{or} \qquad \bar{V}^\mu = \frac{\partial \bar{x}^\mu}{\partial x^\nu} V^\nu.$$

2.2.1 The covariant derivative and the material derivative

It is important to know how to find the derivative at a given point of a vector field. In a flat spacetime, the rate of change of some vector field V^ν in a particular direction x^μ can be found simply by taking the partial derivative. However, the derivative is not so easy to define in a curved spacetime. As an example, consider the vector $\mathbf{V} = V^\mu \mathbf{e}_\mu$ where $\mathbf{e}_\mu$ is some basis vector at a point. In a flat Cartesian coordinate system, the basis vectors are constant, but in a curved spacetime, they are not. We see then that

$$\partial_\mu \left(V^\nu \mathbf{e}_\nu\right) = \left(\partial_\mu V^\nu\right) \mathbf{e}_\nu \qquad \text{(Flat Cartesian)}$$
$$\partial_\mu \left(V^\nu \mathbf{e}_\nu\right) = \left(\partial_\mu V^\nu\right) \mathbf{e}_\nu + V^\nu \partial_\mu \mathbf{e}_\nu. \qquad \text{(Curved Space)}$$

From this example it has been shown that there are two main issues to resolve when defining the derivative in a curved space. First, how does one find a limit in a curved spacetime? And second, how do we ensure that the derivative transforms correctly. It can be shown that both of these requirements can be met by defining the covariant differential operator

$$\begin{aligned} \nabla_\mu V^\nu &= \partial_\mu V^\nu + \Gamma^\nu_{\mu\sigma} V^\sigma \\ \Gamma_{\mu\nu\rho} &= \frac{1}{2}\left(\partial_\rho g_{\mu\nu} + \partial_\nu g_{\rho\mu} - \partial_\mu g_{\nu\rho}\right), \end{aligned} \tag{2.3}$$

where commas denote partial derivatives. Here we see that the connection coefficient, Γ, "corrects" for the curvature of the space. Similarly, for a rank-2 tensor, we can write

$$\nabla_\rho T^{\mu\nu} = \partial_\rho T^{\mu\nu} + \Gamma^\mu_{\sigma\rho} T^{\sigma\nu} + \Gamma^\nu_{\sigma\rho} T^{\mu\sigma}.$$

Before we continue, we quickly write down a few important identities which will be important later. First, the material derivative can be written,

$$\frac{D}{D\tau} V^\nu(x_\mu(\tau)) = \frac{\partial x^\mu}{\partial \tau} \nabla_\mu V^\nu = V^\mu \nabla_\mu V^\nu.$$

Secondly, it can also be shown that for timelike curves, by Equation 2.2, that

$$u^\mu u_\mu = -1 \qquad \Rightarrow \qquad u_\mu \nabla_\nu u^\mu = 0. \tag{2.4}$$

This identity will prove invaluable when working through the details below. Finally, we note that for Riemann manifold (considered here),

$$\nabla_\mu g_{\mu\nu} = 0 \tag{2.5}$$

as a result of the definition of the connection coefficients.

2.2.2 Curvature and the Riemann tensor

We briefly present the ideas here simply for completeness and the details of the following calculations have been omitted. This section is presented merely to remind the reader of where the Einstein field equations have their origins. Anderson [AC07] has a good discussion of many these concepts.

The measure of the curvature of space is defined in terms Riemann tensor ($R^\mu_{\nu\rho\sigma}$), Ricci tensor ($R_{\mu\nu}$), and Ricci scalar (R).

$$\begin{aligned} R^\mu_{\nu\rho\sigma} &= \Gamma^\mu_{\nu\sigma,\rho} - \Gamma^\mu_{\nu\rho,\sigma} + \Gamma^\mu_{\tau\rho}\Gamma^\tau_{\nu\sigma} - \Gamma^\mu_{\tau\sigma}\Gamma^\tau_{\nu\rho} \qquad (2.6) \\ R_{\mu\nu} &= R^\rho_{\mu\rho\nu} \\ R &= R^\rho_\rho \end{aligned}$$

From the Bianchi identities

$$\nabla_\lambda R^\mu_{\nu rho\sigma} + \nabla_\rho R^\mu_{\nu\sigma\lambda} + \nabla_\sigma R^\mu_{\nu\lambda\rho} = 0,$$

it can be shown that

$$\nabla_\nu \left[R^{\mu\nu} - \frac{1}{2} R g^{\mu\nu} \right] = 0.$$

As such, we define the Einstein tensor

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}.$$

2.3 Einstein field equations

John Wheeler once said

> "Mass tells space-time how to curve, and space-time tells mass how to move."

The Einstein tensor is a measure of the curvature of spacetime. Mass is merely a form of energy and, as such, we denote the stress-energy tensor, $T_{\mu\nu}$, containing all of the information of the energy of a system. Thus, these two tensors must be in balance, which is represented in the Einstein field equations (EFE)

$$G_{\mu\nu} = \frac{8\pi G}{c^2} T_{\mu\nu}, \tag{2.7}$$

where we include the constants c, G to present the EFE in their usual form. Recall that we are using units such that $c = G = 1$.

The EFE represent a system of ten non-linear partial differential equations. The complexity of these equations explains why few analytical solutions exist.

We've seen above that

$$\nabla_\nu G^{\mu\nu} = 0$$

applying this to Equation 2.7

$$\nabla_\mu T^{\mu\nu} = 0. \tag{2.8}$$

This equation is very important in fluid dynamics, as we shall see. This equation encapsulates the idea of energy and momentum conservation.

3 Governing Equations

One of the most difficult aspects of relativistic fluid dynamics is keeping track of "what-goes-where", and what index corresponds to what physical property. In Newtonian fluids, all of the equations clearly have their own distinct physical interpretation, but when we extend these ideas to higher dimensions it is important keep track of what physics we are referring too.

It may not appear clear, however, how Equation 2.8 relates to the standard Newtonian fluid dynamics described above. The easiest way to compare these two is to first define projection operators, which will allow us to understand this equation from a more intuitive front. Anile [Ani89] has a good description of introductory relativistic fluid mechanics and the use of these projectors.

3.1 Projections

For any timelike curve p_μ, we can project this into its timelike and spacelike components. To project it into its pure timelike contribution, we contract p_μ onto u^μ. This projection captures what occurs in the rest frame of an observer as he travels along with the fluid. This is sometimes associated with "Lagrangian" coordinates.

Alternatively, sometimes it is valuable to project an equation into its purely spacelike components. We do this by defining

$$h_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu \qquad \text{or} \qquad h^\mu_\nu = \delta^\mu_\nu + u^\mu u_\nu.$$

It is left to the reader to observe that the timelike projection u^μ and the spacelike projection h^μ_ν are orthogonal.

We then find that Equation 2.8 can be decomposed into an energy conservation component

$$u_\nu \nabla_\mu T^{\mu\nu} = 0$$

and a momentum conservation component

$$h_{\rho\nu} \nabla_\mu T^{\mu\nu} = 0.$$

3.2 Stress energy tensor

Different systems will have different stress energy tensors. Often, a lot of the problems of viscosity and other effects can be neglected compared with pressure or other more dominant effects. We will consider here the "perfect fluid" stress energy tensor which is the one typically introduced when approaching the subject for the first time.

In the rest frame of the observer it can be written

$$T^{\mu\nu} = \begin{bmatrix} e & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{bmatrix}_{\mu\nu},$$

or, in a more general frame,

$$T^{\mu\nu} = (e+p)\, u^\mu u^\nu + p g^{\mu\nu}, \tag{3.1}$$

with $g^{\mu\nu}$ the metric, p pressure, e the total energy density.

Typically, we can write out that

$$e = \rho(1+\epsilon),$$

with ρ the rest frame mass energy density and ϵ internal energy density per unit mass.

The continuity equation can be written down as the following

$$\nabla_\mu (\rho u^\mu) = 0,$$

which ensures conservation of mass.

A more general conservation energy equation of this system can then be derived by finding the timelike component of Equation 2.8, projecting it onto u_ν:

$$u^\mu \nabla_\mu e = -\,(e+p)\, \nabla_\mu u^\mu, \tag{3.2}$$

where we recall that $\nabla_\mu g^{\mu\nu} = 0$, in the space we are considering here, and we have used the identities of Equation 2.4.

Similarly, we can project Equation 2.8 into its spacelike component using $h_{\alpha\mu}$:

$$(e+p)u^\mu \nabla_\mu u^\alpha = -h^{\alpha\mu}\nabla_\mu p, \tag{3.3}$$

where, again, we have used the identities found in Equation 2.4

What we have shown here are the relativistic equivalent equations to the momentum and mass conservations equations given in the Newtonian regime. In a relativistic case, it is the conservation of energy, not mass, which concerns us. However, as we return to the Newtonian domain, other sources of energy (kinetic, etc.) tend to be dominated by mass.

The equation of state and the conservation of temperature equations are not so easy to find. These need to be derived statistically using thermodynamic principles for the fluid in question.

3.3 Relativistic Euler equations

Our goal is to write out a system of equations which can be used to solve for the flow of a fluid. At this point we have a conservation of energy equation (Equation 3.2) and a conservation of momentum equation (Equation 3.3). It is insightful to compare these equations with their classical counterparts in order to help understand what these equations mean. We will do this by expanding Equation 3.3

$$(e+p)u^\mu \nabla_\mu u_\nu = -\nabla_\nu p - u_\nu u^\mu \nabla_\mu p,$$

from which we can write out the spatial components as

$$\begin{aligned}(e+p)\frac{D\mathbf{u}}{D\tau} &= -\nabla p - \mathbf{u}\frac{Dp}{D\tau} && \text{(Momentum equation)}\\ u^\mu \nabla_\mu e &= -\left(e+p\right)\nabla_\mu u^\mu, && \text{(Continuity equation)}\end{aligned}$$

where $\frac{D}{D\tau} = u^\mu \nabla_\mu$. Now we see that if in the low velocity limit ($u_i \ll 1$), with $e \gg p$, $e \approx \rho$, and the fluid is incompressible ($\nabla \cdot \mathbf{u} = 0$) as is typical with water, we get back out typical Euler equations of Newtonian fluids. (Incompressible, viscous free Navier-Stokes equations.)

$$\begin{aligned}\rho\frac{D\mathbf{u}}{Dt} &= -\nabla p\\ \frac{D\rho}{Dt} &= 0.\end{aligned}$$

For a more detailed look at Newtonian fluids, see Kundu [Kun90].

We stop here and see that we have extended the Newtonian fluid equations into their relativistic form. Of course we are missing three very important items from this derivation. We have left out all viscosity terms, temperature evolution, and we still have not yet written down an equation of state. These are three fundamental properties which we have ignored here. The reason for this is simple; these additions are very complicated. We shall discuss these further in this article, however, they have been omitted here in order to aid the reader in understanding the current physical content.

We should also note here that we have assumed a known metric for our purposes. This is often acceptable for certain application; however, a more general relativistic treatment is required when the fluid itself causes space-time to curve.

3.4 Viscosity

Before continuing, we note that the we have omitted from the equations of motion viscosity. The addition of viscosity to a relativistic fluid will amount to an addition of non-diagonal terms to the stress energy tensor. These terms drastically increase the complexity of the equations. Viscosity plays an important role in the dispersion of energy of a system and are indispensable in the study of turbulence.

We present here the modification of the stress-energy tensor as a result of viscosity. We compare these terms to their classical counterpart. Landau [LL59] has a brief discussion on the topic (Alternatively, see the book by Wilson and Mathews [WM03]).

$$
\begin{aligned}
T_{\mu\nu} &= p g_{\mu\nu} + (e+p)u_\mu u_\nu + \tau_{\mu\nu} \\
\tau_{\mu\nu} &= -\eta\left(\left[\underbrace{\nabla_\mu u^\nu + \nabla_\nu u^\mu}\right] + u_\mu u^\alpha \nabla_\alpha u_\nu + u_\nu u^\alpha \nabla_\alpha u_\mu\right) - (\zeta - \frac{2}{3}\eta)\left[\underbrace{\nabla_\alpha u^\alpha}\right](g_{\mu\nu} + u_\mu u_\nu)
\end{aligned}
$$

Here we emphasize the relation to the Newtonian case. η and ζ are coefficients of viscosity.

4 Steady State Solution

The simplest relativistic fluid derivation is the hydrostatic problem. In the case, we can assume that the fluid is at rest and we write out that

$$
u_0 = \sqrt{-g_{00}} \qquad u_i = 0,
$$

which is simply stating that the fluid has no velocity in 3-space.

Looking back to the momentum equation (Equation 3.3) and Equation 2.3, we find that

$$
\begin{aligned}
-(e+p)\Gamma^0_{\nu 0} u_0 u^0 &= -\nabla_\nu p \\
\frac{1}{e+p}\nabla_\nu p &= -\frac{1}{2}\partial_\nu \ln\sqrt{-g_{00}},
\end{aligned} \tag{4.1}
$$

where, the metric allows $g^{00} = \frac{1}{g_{00}}$.

As Landau points out [LL59], in the weak field limit where $(e+p) \approx \rho$ and $g_{00} = -1 - 2\phi$ with

$$
\ln(1+2\phi) \approx 2\phi
$$

Equation 4.1 reduces to

$$
\begin{aligned}
\frac{1}{\rho}\nabla P &= -\nabla\phi \\
\nabla P &= \rho\mathbf{g},
\end{aligned}
$$

which is the classical condition of hydrostatics.

4.1 Numerical energy transport

We can now write out the equations that must be obeyed by the relativistic fluid. For the purposes of this paper, we will reduce down the equations under certain assumptions.

1. One-dimensional fluid flow
2. Flat Minkowski space
3. The fluid is barotropic (i.e. P = C e, where C is a constant)

4. Energy density (e) is conserved within the medium.

   As the energy density is constant we can write out the following system of equations

$$\begin{aligned}
\nabla_\mu u^\mu e &= 0 \\
(e+p)u^\mu \nabla_\mu u_\nu + u_\nu u^\mu \nabla_\mu p &= -\nabla_\nu p \\
p &= Ce,
\end{aligned}$$

which, under the barotropic fluid assumption, reduce further to

$$\nabla_\mu (u^\mu e) = 0 \tag{4.2}$$

$$u^\mu \nabla_\mu u_\nu = \frac{-C}{(1+C)e} \nabla_\nu e. \tag{4.3}$$

From the Lorentz transforms, it can be shown that

$$t = \gamma\tau \qquad \text{where} \qquad \gamma = \left(1 - v^2\right)^{\frac{-1}{2}}.$$

Under this transformation, we find that

$$u^\mu = \frac{dx^\mu}{d\tau} = \frac{dx^\mu}{dt}\frac{dt}{d\tau} = \gamma v^\mu.$$

Along with assumptions the remaining, Equations 4.2 and 4.3, become

$$\begin{aligned}
\partial_t(\gamma e) + \partial_x(\gamma e v) &= 0 \\
\gamma \partial_t(\gamma v) + \partial_x\left(\frac{(\gamma v)^2}{2}\right) &= -\frac{C}{1+C}\partial_x \ln e.
\end{aligned}$$

Using a spectral fourth order Runge-Kutta differencing scheme, we can attempt to solve these equations. The details of the method can be found in Duran's numerical methods text [Dur99]. The details have been omitted here to avoid confusion. Appendix A contains the code used to approximate the system of equations along with a brief description of the method.

In order to demonstrate the solution to this equation, we will use periodic boundary conditions. We will assume that all quantities are unit-less and we will implement initial conditions to represent a relativistic fluid with the energy density grouped into a dense region. A hyperbolic secant function was selected for the energy density function as it represents a single energy density packet centred around the origin. A relativistic velocity of $0.5c$ was selected, and a background energy of one was used to ensure a non-zero energy level throughout the domain.

$$\begin{aligned}
u(x,0) &= 0.5 \\
e(x,0) &= \operatorname{sech}\left(\frac{x}{0.5}\right) + 1
\end{aligned}$$

A time step of $\Delta t = 1e-5$ and 512 grid points were used. Figure 4.1 outputs the results of the computation for three different values of $C = \{0, \frac{1}{3}, \frac{2}{3}\}$ corresponding to non-interacting matter, relativistic matter, and cold matter respectively. For details on these values, Battaner [Bat96] has a description of the statistical mechanical derivation.

The purpose to providing this numerical solution if three-fold. First, it demonstrates the complexity of the corresponding solution. It can be clearly seen that the interaction between the velocity and the energy density is very complicated. For C=0, there is no effect of the energy density on the velocity, and

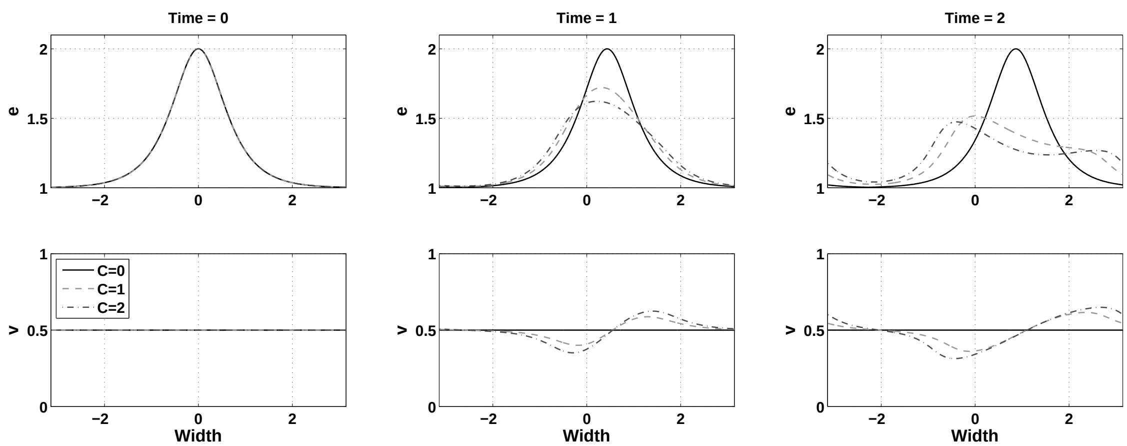


Figure 4.1: Energy transport relativistic hydrodynamics for values of P $= \{0e$ (solid line), $\frac{1}{3}e$ (dashed line), $\frac{2}{3}e$ (dash-dot line)$\}$ at time steps t=$\{0, 1, 2\}$ where both the energy density (e) and 3-velocity (v) have been output.

vice-verse. For $C > 0$, the change of one feeds back onto the other causing the solutions to distinguish themselves. Second, this serves as a basis upon which future work can be preformed. Thirdly, these numerics demonstrate the fact that, even under the simplifying assumptions of Minkowski space, the solution to the problem highly dependent upon the relativistic components of the equation. In this case, we see that the higher proportion of P to the energy density (i.e. larger values of C), the more rapid the transition from one state to another.

5 Spacetime Curvature

Up to this point we have assumed that the metric was known, that is the fluid does not substantially change the curvature of spacetime. This has many applications, however, it does leave something desired in order to get a more general theory. We return to the Einstein field equations

$$\begin{aligned} G_{\mu\nu} &= 8\pi G T_{\mu\nu} \\ R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R &= 8\pi G T_{\mu\nu}. \end{aligned}$$

It is often convenient to convert this into the form

$$R_{\mu\nu} = 8\pi G \left(T_{\mu\nu} - \frac{1}{2} g_{\mu\nu} T^{\alpha}_{\alpha} \right). \tag{5.1}$$

These ten differential equations prove very difficult to solve. We can however, show that under certain symmetries, the system reduces to a simplified form.

For the purposes of this paper, we consider an application to a star. In reference to this, we will assume that the metric should be spherically symmetric. For the present purpose, let us also assume that the metric is constant in time. In reference to this, it can be shown that the most generic metric that can be written in spherical coordinates is

$$g_{\mu\nu} = \mathrm{diag}\left[-B(r), A(r), r^2, r^2\sin^2\theta\right]_{\mu\nu}, \tag{5.2}$$

for which we can easily calculate the connection coefficients.
The components of the Ricci tensor can be found using Equation 2.6. They are

$$\begin{aligned}
R_{00} &= -\frac{1}{2A}\frac{d^2B}{dr^2} + \frac{1}{4A}\frac{dB}{dr}\left(\frac{1}{A}\frac{dA}{dr} + \frac{1}{B}\frac{dB}{dr}\right) - \frac{1}{r}\frac{1}{A}\frac{dB}{dr} \\
R_{rr} &= \frac{1}{2B}\frac{d^2B}{dr^2} - \frac{1}{4B}\frac{dB}{dr}\left(\frac{1}{A}\frac{dA}{dr} + \frac{1}{B}\frac{dB}{dr}\right) - \frac{1}{r}\frac{1}{A}\frac{dA}{dr} \\
R_{\theta\theta} &= -1 + \frac{r}{2A}\left(-\frac{1}{A}\frac{dA}{dr} + \frac{1}{B}\frac{dB}{dr}\right) + \frac{1}{A} \\
R_{\phi\phi} &= \sin^2\theta R_{\theta\theta} \\
\text{else} &= 0.
\end{aligned} \tag{5.3}$$

For a more detailed explanation, see Battaner [Bat96].

5.1 Schwarzschild metric

The previous assumptions prove reasonable when considering a star in space. In the region external to the star (provided the star is not rotating or charged) the stress energy tensor becomes null, and thus we must solve the complete set of equations

$$R_{\mu\mu} = 0.$$

The original solution to this problem was originally proposed by Schwarzschild in 1916 (the original article has recently been republished [Sch99]). Schwarzschild showed that the metric

$$g_{\mu\nu} = \mathrm{diag}\left[-\left(1 - \frac{2M}{r}\right), \left(1 - \frac{2M}{r}\right)^{-1}, r^2, r^2\sin^2\theta\right]_{\mu\nu}$$

is a solution.

5.2 Curvature Deformation

Inside the star, however, is a very different story. We will assume here that we still have spherical symmetry and the star is not rotating or charged. Many of the following details can be found in Battaner [Bat96]. We find that the basic form of the metric is the same as in (5.2), as such the Ricci tensor will in turn, have the same basic structure as Equation 5.3. However, now the field equations become

$$R_{\mu\nu} = 8\pi\left(T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T^{\alpha}_{\alpha}\right).$$

Recall that for a perfect fluid (Equation 3.1), we can write the stress energy tensor as

$$T^{\mu\nu} = (e + p)\, u^\mu u^\nu + pg^{\mu\nu}.$$

In the rest frame of the fluid we can write that

$$u_j = (-\sqrt{B}, 0, 0, 0).$$

Thus, the solution stress energy tensor becomes

$$T_{\mu\nu} = \operatorname{diag}\left[eB, pA, pr^2, pr^2\sin^2\theta\right]. \tag{5.4}$$

We can then use the metric on Equation 5.4 to find

$$T^{\alpha}_{\alpha} = 3p - e. \tag{5.5}$$

Thus, Equations 5.4 and 5.5 combined with Equation 5.1 give, (Recall that we use units such that G=1),

$$\frac{R_{\mu\nu}}{8\pi} = \operatorname{diag}\left[\frac{1}{2}(3p+e)B, -\frac{A}{2}(p-e), -\frac{r^2}{2}(p-e), -\frac{r^2\sin^2\theta}{2}(p-e)\right], \tag{5.6}$$

which, equated with Equation 5.3, provides a complete system of equations to solve for the components of the metric.

$$-\frac{1}{2A}\frac{d^2B}{dr^2} + \frac{1}{4A}\frac{dB}{dr}\left(\frac{1}{A}\frac{dA}{dr} + \frac{1}{B}\frac{dB}{dr}\right) - \frac{1}{r}\frac{1}{A}\frac{dB}{dr} = -4\pi(3p+e)B \tag{5.7}$$

$$\frac{1}{2B}\frac{d^2B}{dr^2} - \frac{1}{4B}\frac{dB}{dr}\left(\frac{1}{A}\frac{dA}{dr} + \frac{1}{B}\frac{dB}{dr}\right) - \frac{1}{r}\frac{1}{A}\frac{dA}{dr} = 4\pi A(p-e) \tag{5.8}$$

$$-1 + \frac{r}{2A}\left(-\frac{1}{A}\frac{dA}{dr} + \frac{1}{B}\frac{dB}{dr}\right) + \frac{1}{A} = 4\pi r^2(p-e) \tag{5.9}$$

$$R_{\phi\phi} = \sin^2\theta R_{\theta\theta}. \tag{5.10}$$

Combining Equations 5.7 and 5.8 gives

$$-\frac{1}{r}\left(\frac{dB}{dr} + \frac{B}{A}\frac{dA}{dr}\right) = -8\pi AB(e+p), \tag{5.11}$$

using Equation 5.9,

$$\frac{d}{dr}\left(\frac{r}{A}\right) = 1 - 8\pi r^2 e. \tag{5.12}$$

Recall that here e is the energy density of the fluid, containing mass and internal energy, so we can write out that

$$\begin{aligned} U &= \int_0^r 4\pi r^2 e\,dr \\ \frac{r}{A} &= r - 2U \\ A &= \left(1 - \frac{2U}{r}\right)^{-1}. \end{aligned}$$

Keep in mind that we have two boundary conditions on A. At $r = 0$, we want to make sure the A is finite, and for $r > R$, the radius of the star, we require A to become the Schwarzschild value, $A_R = \left(1 - \frac{2M}{R}\right)^{-1}$.

Similarly, Equations 5.11 and 5.9 provide an equation for B

$$\frac{1}{B}\frac{dB}{dr} = \frac{2A}{r^2}\left(U + 4\pi r^3 p\right). \tag{5.13}$$

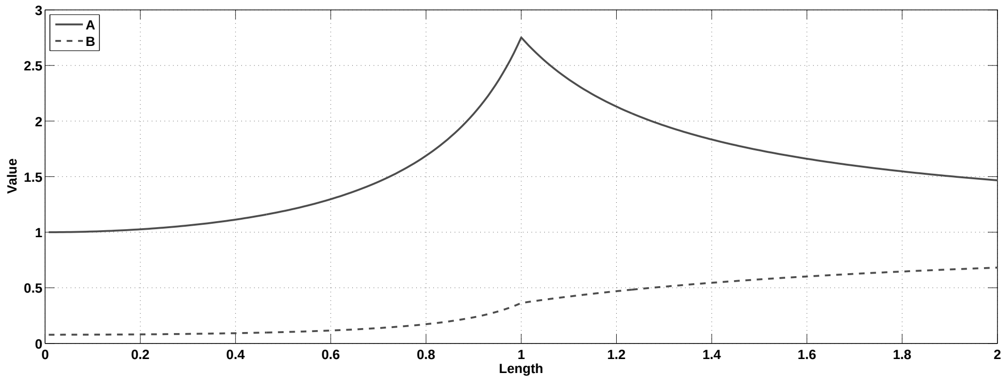


Figure 5.1: Solution to the A and B metric elements, assuming units such that the radius of the star is $R = 1$, and the mass is given as $M = 1/\pi$.

Here we can find a boundary condition such that, on the surface of the star, $p = 0$, $U = M$, thus, (where BA =1)

$$\frac{dB}{dr} = \frac{2M}{R^2}, \tag{5.14}$$

where again R is the radius of the star.

This calculation is meant to demonstrate how the stress-energy tensor modifies the curvature of space-time. In the domain without the nice symmetric properties we've introduced here, it is often necessary to solve the equations numerically. We refer the reader to further texts on the subject, such as the book written by Wilson [WM03]. These computations themselves prove very difficult and are omitted here.

5.3 Graphical solution

In order to understand the metric internal to the star, we consider the graph of A and B as a function of distance from the origin. For simplicity we assume that the density of the star is constant, and assume units such that the radius is 1 and mass is $1/\pi$. This is meant to provide a qualitative solution to the metric inside of a star. Figure 5.1 plots the resulting values of A and B under these conditions.

Notice that the solution is piecewise continuous at surface of the star ($r = 1$). Notice also that the A parameter becomes close to 1 near the centre, its asymptotic limit.

6 Conclusion

The current paper is meant to provide a brief introduction to relativistic fluids. As much as possible, this work has tried to compare the relativistic results with their Newtonian counterparts in order to provide basis for the new material. In here, the equations of motion of a perfect fluid have been written out and the static solution has been provided.

One major extension of relativistic hydrodynamics which we has not tackled here is Magnetohydrodynamics (MHD). MHD is the study of electrically charged fluids and has been applied to a wide variety of topics including stellar modelling. Golub [GP10] has a good classical approach to the topic. This is still a very active area of research.

For a more in depth discussion of relativistic hydrodynamics, the reader is referred to two well written texts on the subject. Andersson's discussion [AC07] provides a much more rigorous approach to the subject, and spends a great deal of time developing the mathematics of differential geometry. Also, Anile's book [Ani89] extends much of the above work to the case of MHD. There are many other well written texts on the subject, we simply provide the reader with two examples to serve as a basis for further research.

References

[AC07] Nils Andersson and Gregory L Comer, *Relativistic fluid dynamics: Physics for many different scales*, Living Reviews in Relativity **10** (2007), no. 1.

[Ani89] A.M. Anile, *Relativistic Fluids and Magneto-Fluids: with Applications in Astrophysics and Plasma Physics*, Cambridge University Press, Cambridge, 1989.

[Bat96] E. Battaner, *Astrophysical Fluid Dynamics*, Cambridge University Press, Cambridge, 1996.

[Dur99] Dale Durran, *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*, Springer, Berlin, 1999.

[GP10] Leon Golub and Jay M. Pasachoff, *The Solar Corona*, Cambridge University Press, Cambridge, 2010.

[Kun90] P.K. Kundu, *Fluid Mechanics*, Academic Press, San Diego, 1990.

[LL59] L. D. Landau and E.M. Lifshitz, *Fluid Mechanics*, Addison-Weasly Publishing Company, Massachusetts, 1959.

[Sch99] K. Schwarzschild, *On the gravitational field of a mass point according to einstein's theory*, ArXiv Physics e-prints (1999).

[WM03] James R. Wilson and Grant J. Mathews, *Relativistic Numerical Hydrodynamics*, Cambridge University Press, 2003.

A Numerical Hydrodynamics Methodology

This section discusses the code used to compute the solution to the relativistic fluid equations found in Section 4.1. The source code is available along side the online version of this article.

Putting the equation into the following form

$$\partial_t u = F(u),$$

this spectral method decomposes the function u_j, sampled at the grid points x_j, into its truncated Fourier series

$$u_j = \sum_{k=-\frac{(N+1)}{2}}^{\frac{N-1}{2}} a_k \exp ikx_j,$$

with N the number of grid points. Note that we have removed the $k = N/2$ wavenumber.

This technique allows us to use Matlab's built in FFT methods to compute derivatives of the corresponding function. Once the method for computing F has been established, we can then implement a

Fourth-Order Runge-Kutta method using

$$\begin{aligned}
q_1 &= \Delta t F(u^n) \\
q_2 &= \Delta t F(u^n + \frac{q_1}{2}) \\
q_3 &= \Delta t F(u^n + \frac{q_2}{2}) \\
q_4 &= \Delta t F(u^n + q_3) \\
u^{n+1} &= u^n + \frac{q_1 + 2q_2 + 2q_3 + q_4}{6},
\end{aligned}$$

where the superscripts, n, refer to the time step. The details of such a computation are included in Durran [Dur99].

ISSN 1927-1417 (Print) ISSN 1927-1425 (Online)

WATERLOO | MATHEMATICS