An Invitation to Mathematical Physics
and its History

Jont B. Allen

Help from Steve Levinson, John D'Angelo, Michael Stone

University of Illinois at Urbana-Champaign

July 14, 2017; *Version 0.81.8*

Contents

Abstract

An understanding of physics requires knowledge of mathematics. The contrary is not true. By definition, pure mathematics contains no physics. Yet historically, mathematics has a rich history filled with physical applications. Mathematics was developed by people with intent of making things work. In my view, as an engineer, I see these creators of early mathematics, as budding engineers. This book is an attempt to tell this story, of the development of mathematical physics, as viewed by an engineer.

The book is broken down into three topics, called streams, presented as five chapters: 1) Introduction, 2) Number systems, 3) Algebra Equations, 4) Scalar Calculus, and 5) Vector Calculus. The material is delivered as 40 "Lectures" spread out over a semester of 15 weeks, three lectures per week, with a 3 lecture time-out for administrative duties. Problems are provided for each week's assignment. These problems are written out in LaTeX, with built in solutions, that may be expressed by un-commenting one line. Once the home-works are turned in, each student is given the solution. The students rated these Assignments as the most important part of the course. There is a built in interplay between these assignments and the lectures. On many occasions I solved the homework in class, as motivation for coming to class.

There were four exams, one at the end of each of the three sections, and a final. One of the exams was in class, three others and the final were evening exams. Each of the exams and assignments are LaTeX files, with solutions encoded with a one line software switch. When the exam was returned by the student, the full solution was provided, while the exam was fresh in the students mind, providing a *teaching moment*. The Exams are largely based on the Assignments. It is my philosophy that, in principle, the students could see the exam in advance of taking it. In a sense they do, since each exam was based on the homework.

Author's Personal Statement

An expert is someone who has made all possible mistakes in a small field. I don't know if I would be called and expert, but I certainly have made my share of mistakes. I openly state that "I love making mistakes, because I learn so much from them." One might call that the "expert's corollary."

This book has been written out of both my love for the topic of mathematical physics, and a frustration for wanting to share many key concepts, and even new ideas on these basic concepts. Over the years I have developed a certain *physical sense* of math, along with a related mathematical sense of physics. While doing my research,[1] I have come across what I feel are certain conceptual holes that need filling, and sense many deep relationships between math and physics, that remain unidentified. While what we presently teach is not wrong, it is missing these relationships. What is lacking is an intuition for how math "works." A good scientest "listens" to their data. In the same way we need to start listening to the language of mathematics. We need to let mathematics guide us toward our engineering goals.

It is my suspicion that over the centuries many others have had similar insights, and like me, have been unable to convey this slight misdirection. I hope these views can be useful to open young minds.

As summarized in Fig. 1, this marriage of math, physics and engineering will help us make progress in understanding the physical world. We must turn to mathematics and physics when trying to understand the universe. I have arrived in my views following a lifelong attempt to understand human

---

[1] https://auditorymodels.org/index.php/Main/Publications
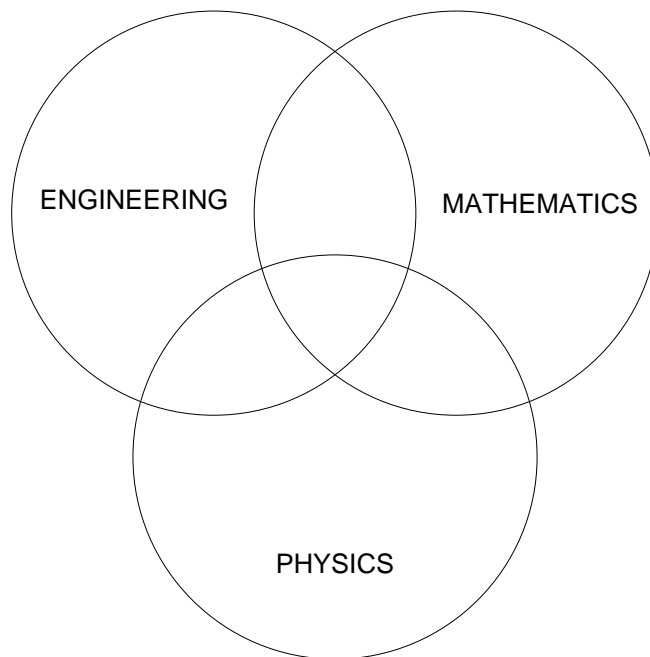
Figure 1: *There is a natural symbiotic relationship between Physics, Mathematics and Engineering, as depicted by this Venn diagram. Physics explores the boundaries. Mathematics provides the method and rigor. engineering transforms the method into technology. While these three disciplines work well together, there is poor communication due to a different vocabulary. For example, Matematics rarely, if ever, uses a system of units, whereas Physics and Engineering depend critically on them. Mathematics strives to abstract the ideas into proofs. Physics rarely, if ever, uses a proof. When they attempt abstract, they usually get into some difficulty. Engineers blunder ahead, ignoring most of these defects.*

communication, i.e., the perception and decoding of human speech sounds. This research arose from my 32 years at Bell Labs in the Acoustics Research Department. There such lifelong pursuits were not only possible, they were openly encouraged. The idea was that if you were successful at something, take it as far as you can. But on the other side, don't do something well, that's not worth doing.
5   People got fired for the latter. I should have left for University after a mere 20 years,[2] but the job was just too cushy.

In this text it is my goal to clarify conceptual errors while telling the story of physics and mathematics. My views have been inspired by classic works, as documented in the bibliography. This book was inspired by my reading of Stillwell (2002), through Chapter 21 (Fig. 2) (p. 10). Somewhere in
10   Chapter 22 I switched to the third edition (Stillwell, 2010), at which point I realized I had much more to master. It became clear that by teaching this material to first year engineers I could absorb the advanced material at a reasonable pace. This book soon followed.

Back Cover Summary

This is foremost a math book, but not the typical math book. *First*, this book is for the engineering
15   minded, for those who need to understand math to do engineering, to learn how things work. In that sense it is more about physics and engineering. Math skills are essential for to making progress in building things, be it pyramids or computers, as clearly shown by the many great civilizations of the Chinese, Egyptians, Arabs (people of Mesopotamia), and Greeks and Romans.

Second, this is a book about the math that developed to explain physics, to allow people to engineer
20   complex things. To sail around the world one needs to know how to navigate. This requires a model of the planets and stars. You can only know where you are on earth once you understand where earth is relative to the heavens. The answer to such a deep questions will depend on who you ask. The utility and accuracy of that answer depends critically on the depth of understanding of how the worlds

---

[2]I started around December 1970, fresh out of Graduate school, and retired in December 2002.

and heavens work. Who is qualified to answer such question? It is best answered by those who study mathematics applied to the physical world.

The English astronomer Edmond Halley (1656–1742) asked Newton (1643–1727) for the equation that describes the orbit of the planets. Halley was obviously interested in comets. Newton immediately answered "an ellipse." It is said that Halley was stunned by the response (Stillwell, 2010, p. 176), as this was what had been experimentally observed by Kepler (c1619), and thus he knew Newton must have some deeper insight.

When Halley asked Newton to explain how he knew, Newton responded "I calculated it." But when challenged to show the calculation, Newton was unable to reproduce it. This open challenge eventually led to Newton's grand treatise, *Philosophiae Naturalis Principia Mathematica* (July 5, 1687). It had a humble beginning, as a letter to Halley, explaining how to calculate the orbits of the planets. To do this Newton needed mathematics, a tool he had mastered. It is widely accepted that Isaac Newton and Gottfried Leibniz invented calculus. But the early record shows that perhaps Bhāskara II (1114–1185 AD) had mastered this art well before Newton.[3]

Third, the main goal of this book is to teach motivated engineers mathematics, in a way that it can be understood, mastered and remembered. How can this impossible goal be achieved? The answered is to fill in the gaps with "who did what, and when." An historical story may be easily mastered.

To be an expert in a field, one must know its history. This includes who the people were, what they did, and the credibility of their story. Do you believe the Pope or Galileo on the roles of the sun and the earth? The observables provided by science are clearly on Galileo's side. Who were those first engineers? They are names we all know: Archimedes, Pythagoras, Leonardo da Vinci, Galileo, Newton, ... All of these individuals had mastered mathematics. This book teaches the tools taught to every engineer. Do not memorize complex formulas, rather make the equations "obvious" by learning the simplicity of the underlying concepts.

Credits

---

[3] `http://www-history.mcs.st-and.ac.uk/Projects/Pearce/Chapters/Ch8_5.html`

Mathematics and its History (Stillwell, 2002)

Contents

Figure 2:   Table of contents of Stillwell (2002)

Preface

It is widely acknowledged that interdisciplinary science is the backbone of modern scientific investigation. This is embodied in the STEM (Science, Technology, Engineering, and Mathematics) programs. Contemporary research is about connecting different areas of knowledge, thus it requires an understanding of cross-disciplines. However, while STEM is being taught, interdisciplinary science is not, due to its inherent complexity and breadth. Furthermore there are few people to teach it. Mathematics, Engineering and Physics (MEP) are at the core of such studies.

STEM vs. MEP

Mathematics is about rigor. Mathematicians specifically attend to the definitions of increasingly general concepts. Thus mathematics advances slowly, as these complex definitions must be collectively agreed upon. Since it embraces rigor, mathematics shuns controversy, the opposite of uncertainty. Physics explores the fringes of uncertainty. Physicists love controversy. Engineering addresses the advancement the technology. Engineers, much like mathematicians, are uncomfortable with uncertainty, but are trained to deal with it.

To create such an interdisciplinary STEM program, a unified MEP curriculum is needed. In my view this unification could (should) take place based on a core mathematical training, from a historical perspective, starting with Euclid or before (i.e., Chinese mathematics), up to modern information theory and logic. As a bare minimum, the *fundamental theorems of mathematics* (arithmetic, algebra, calculus, vector calculus, etc.) need to be appreciated by every MEP student. The core of this curriculum is outlined in Table 1.1 (p. 21).

If first year students are taught a common MEP methodology and vocabulary, presented in terms of the history of mathematics, they will be equipped to

1. Exercise interdisciplinary science (STEM)

2. Communicate with other MEP trained (STEM) students and professors.

The goal is a *comprehensive understanding of the fundamental concepts of mathematics*, defined as those required for engineering. We assume that students with this deep understanding will end up being in the top 0.1% of Engineering. Time will tell if and when this assumption is proved to be correct.

The key tool is methodology. The traditional approach is a five to eight course sequence: Calc I, II, III, Linear Algebra IV, DiffEq V, Real analysis VI, Complex analysis VII and Number theory VIII, over a time frame of three or more years (six semesters). This was the way I learned math. Following this regime I felt I had not fully mastered the material, so I started over. I now consider myself to be self-taught. We need a more effective teaching method. I am not suggesting we replace the standard 6 semester math curriculum, rather I am suggesting replacing Calc I, II with a mathematical physics course, based on the historical thread, designed for those students who have demonstrated advanced ability. One needs more than a high school education to succeed in college engineering courses.

By teaching mathematics in the context of history, the student can fully appreciate the underlying principles. Including the mathematical history provides a uniform terminology for understanding the fundamentals of mathematics. The present teaching method, using abstract proofs, with no (or few)

figures or physical principles, by design, removes intuition and the motivation that was available to the creators of these early theories. This present six semester approach does not function for many students, leaving them with poor, or even no intuition.

Chapter 1

Introduction

Much of early mathematics centered around the love of art and music, due to our sensations of light and sound. Exploring our physiological senses required a scientific understanding of vision and hearing, as first explored by Newton (1687) and Helmholtz (1863a)[1] (Stillwell, 2010, p. 261). Our sense of color and musical pitch are determined by the frequencies of light and sound. The Chinese and Pythagoreans are well known for their early contributions to music theory.

Pythagoras strongly believed that "all is integer," meaning that every number, and every mathematical and physical concept, could be explained by integral relationships. It is likely that this belief was based on Chinese mathematics from thousands of years earlier. It is also known that his ideas about the importance of integers were based on what was known about music theory. For example it was known that the relationships between the musical notes (pitches) obey natural integral relationships.

Modern applications of number theory include:

- Public-private key encryption: which require computationally intensive prime factoring of large integers

- IEEE Floating point[2]

As acknowledged by Stillwell (2010, p. 16), the Pythagorean view is relevant today

> *With the digital computer, digital audio, and digital video coding everything, at least approximately into sequences of whole numbers, we are closer than ever to a world in which "all is number."*

Mersenne (1588-1647) contributed to our understanding of the relationship between the wavelength and the length of musical instruments. These results were extended by Galileo's father, and then by Galileo himself (1564-1642). Many of these musical contributions resulted in new mathematics, such as the discovery of the wave equation by Newton (c1687), followed by its one-dimensional general solution by d'Alembert (c1747).

By the time of Newton there was a basic understanding that sound and light traveled at very different speeds (thus why not the velocities of different falling weights?).

> Ole Rõmer first demonstrated in 1676 that light travels at a finite speed (as opposed to instantaneously) by studying the apparent motion of Jupiter's moon Io. In 1865, James Clerk Maxwell proposed that light was an electromagnetic wave, and therefore traveled at the speed c appearing in his theory of electromagnetism.[3]

In 1589 Galileo famously conceptualized an experiment where he suggested dropping two different weights from the Leaning Tower of Pisa, and showed that they must take the same time to hit the

---

[1] https://en.wikipedia.org/wiki/Acoustics
[2] https://en.wikipedia.org/wiki/IEEE_floating_point\#Formats
[3] https://en.wikipedia.org/wiki/Speed_of_light

ground. Conceptually this is an important experiment, driven by a mathematical argument in which he considered the two weights to be connected by an elastic cord. This resulted in the concept of *conservation of energy*, one of the cornerstones of physical theory, since that time.
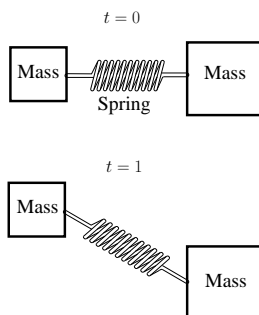
Figure 1.1: *Depiction of the argument of Galileo (unpublished book of 1638) as to why weights of different masses (weight) must fall with identical velocity. By joining them with an elastic cord they become one. Thus if the velocity were proportional to the mass, the joined masses would fall even faster. This results in a logical fallacy. This may have been the first time that the principle of* conservation of energy *was clearly stated. It seems likely that Galileo was attracted to this model of two masses connected by a spring because he was interested in planetary motion, which consist of masses (sun, earth, moon) dominated by gravity, which represents the the spring, providing the mutual attraction between the masses. There are actually three masses involved in his experiment, the earth pulling the two masses down, which in Galileo's mind would represents the sun, the larger of the two masses (the earth) and the smaller mass (the moon).*

While Newton may be best known for his studies on light, he was the first to predict the speed of sound. However his theory was in error[4] of $\sqrt{c_p/c_v} = \sqrt{1.4} = 1.183$. This famous error would not be resolved for over two hundred years, awaiting the formulation of thermodynamics and the equi-partition theorem, by Maxwell and Boltzmann, and others. What was needed was the concept of constant-heat, or *adiabatic process*. For audio frequencies (0.02-20 [kHz]), the small temperature gradients cannot diffuse the distance of a wavelength in one cycle (Pierce, 1981; Boyer and Merzbach, 2011), "trapping" the heat energy in the wave. There were several other physical enigmas, such as the observation that sound disappears in a vacuum and that a vacuum cannot draw water up a column by more than 34 feet.

There are other outstanding examples where physiology impacted mathematics. Leonardo da Vinci is well known for his studies of the human body. Helmholtz's theories of music and the perception of sound are excellent examples of under-appreciated fundamental mathematical contributions (Helmholtz, 1863a). Lord Kelvin (aka William Thompson),[5] was one of the first true engineer-scientists, equally acknowledged as a mathematical physicist, well known for his interdisciplinary research, and knighted by Queen Victoria in 1866. Lord Kelvin coined the term *thermodynamics*, a science more fully developed by Maxwell (the same Maxwell of electrodynamics). Thus the interdisciplinary nature of science has played many key roles in the development of thermodynamics.[6] Lord Rayleigh's book on the theory of sound (Rayleigh, 1896) is a classic text, read even today by those who study acoustics.

It seems that we have somewhat retracted from this venerable interdisciplinary view of science, by splitting the disciplines into into smaller parts, whenever we perceived a natural educational boundary. Reforging these natural connections into the curriculum, is essential for the proper training of STEM students (scientists, mathematicians and engineers).[7]

WEEK 1

---

[4]The square root of the ratio of the *specific heat capacity* at constant pressure c_p to that at constant volume c_v.

[5]Lord Kelvin was one of half a dozen interdisciplinary mathematical physicists, all working about the same time, that made a fundamental change in our scientific understanding. Others include Helmholtz, Stokes, Green, Heaviside, Rayleigh and Maxwell.

[6]Thermodynamics is another example of a course that needs reworking along historical lines. (**?**).

[7]Perhaps its time to put the MEP Humpty Dumpty back together.

1.1 Early Science and Mathematics

The first 5,000 years is not well document, but the basic record is clear, as outlined in Fig. 1.2. Thanks to Euclid and later Diophantus (c250 CE), we have some vague understanding Chinese thought. For example, *Euclid's formula* (Eq. 1.6, p. 38; Sec. 2.2.1, Fig. 2.3, p. 123) provides a method for computing Pythagorean triplets, a formula known long before Euclid (Stillwell, 2010, pp. 4-9). 5

Chinese Bells and stringed musical instruments were exquisitely developed in their tonal quality, as documented by ancient physical artifacts (Fletcher and Rossing, 2008). In fact this development was so rich that one must question why the Chinese failed to initiate the industrial revolution. Specifically, why did Europe eventually dominate with its innovation when it was the Chinese who did the extensive early invention? 10

According to Lin (1995) this is known as the *Needham question*:

> "Why did modern science, the mathematization of hypotheses about Nature, with all its implications for advanced technology, take its meteoric rise only in the West at the time of Galileo[, but] had not developed in Chinese civilization or Indian civilization?"

Needham cites the many developments in China:[8] 15

> "Gunpowder, the magnetic compass, and paper and printing, which Francis Bacon considered as the three most important inventions facilitating the West's transformation from the Dark Ages to the modern world, were invented in China." (Lin, 1995)

> "Needham's works attribute significant weight to the impact of Confucianism and Taoism on the pace of Chinese scientific discovery, and emphasizes what it describes as the 'diffusionist' 20 approach of Chinese science as opposed to a perceived independent inventiveness in the western world. Needham held that the notion that the Chinese script had inhibited scientific thought was 'grossly overrated' " (Grosswiler, 2004).

Lin was focused on military applications, missing the importance of non-military contributions. A large fraction of mathematics was developed to better understand the solar system, acoustics, musical 25 instruments and the theory of sound and light. Eventually the universe became a popular topic, as it still is today.

1.1.1 Lec 1 The Pythagorean theorem

While early Asian mathematics is not fully documented, it clearly defined the course for at least several thousand years. The first recorded mathematics was that of the Chinese (5000-1200 BCE) and the Egyptians (3,300 BCE). Some of the best early record were left by the people of Mesopotamia (Iraq, 1800 BCE). Thanks to Euclid's Elements (c323 BCE) we have an historical record, tracing the progress in geometry, as defined by the Pythagorean theorem *for any right triangle*

$$c^2 = a^2 + b^2, \tag{1.1}$$

having sides of lengths (a, b, c) that are either positive real numbers, or more interesting, integers, such that $c > [a, b]$ and $a + b > c$. Integer solutions were likely found by trial and error rather than by 30 Euclid's formula.

If a, b, c are lengths, then a^2, b^2, c^2 are each the area of a square. Equation 1.1 says that the square area a^2 plus the square area b^2 equals the square area c^2. Today a simple way to prove this is to compute the magnitude of the complex number $c = a + bj$, which forces the right angle

$$|c|^2 = (a + bj)(a - bj) = a^2 + b^2.$$

---

[8]`https://en.wikipedia.org/wiki/Joseph_Needham\#cite_note-11`

However, complex arithmetic was not an option for the Greek mathematicians, since complex numbers and algebra had yet to be invented.

Almost 700 years after Euclid's *Elements*, the Library of Alexandria was destroyed by fire (391 CE), taking with it much of the accumulated Greek knowledge. Thus one of the best technical records may be Euclid's Elements, along with some sparse mathematics due to Archimedes (c300 BCE) on geometrical series, computing the volume of a sphere, and the area of the parabola, and elementary hydrostatics. Additionally, a copy of a book by Diophantus *Arithmetic* was discovered by Bombelli (c1572) in the Vatican library (Stillwell, 2010, p. 51). This book became an inspirational resource for Fermat.

Chronological history pre 16th century

200th *BCE Chinese (Primes; quadratic equation; Euclidean algorithm (GCD))*

180th *BCE Babylonia (Mesopotamia/Iraq) (quadratic equation)*

6th *BCE Pythagoras (Thales) and the Pythagorean "tribe"*

4th *BCE Euclid (quadratic equation) 300BCE; Archimedes*

3th *CE Diophantus c250CE;*

4th *CE Alexandria Library destroyed 391CE;*

6th *CE Brahmagupta (negative numbers; quadratic equation) 598-670CE*

10th *CE al-Khwārizmī (algebra) 830CE; Hasan Ibn al-Haytham (Alhazen) 965-1040CE*

12th *CE Bhaskara (calculus) 1114-1183*

15th *Leonardo & Copernicus 1473-1543*

16th *Tartaglia (cubic eqs); Bombelli 1526-1572; Galileo Galilei 1564-1642*

Time Line

| 1500BCE | | |0CE | |500 | |1000 | |1400|1650 |

Chinese   Pythagoreans   Brahmagupta   Leonardo
Babylonia   Euclid   Diophantus   Bhaskara Bombelli
                   Archimedes   al–Khawarizmi Copernicus
                   Christ |830
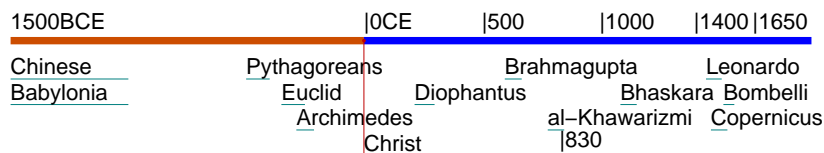
Figure 1.2:   *Mathematical time-line between 1500* BCE *and 1650* CE. *[fig:TimeLineBCE]*

1.1.2 Pythagorean Triplets

Well before Pythagoras, the Babylonians had tables of *Pythagorean triplets* (PTs), integers $[a, b, c]$ that obey Eq. 1.1. An example is $[3, 4, 5]$. A stone tablet (Plimpton-322) dating back to 1800 BCE (Fig. 1.9, p. 39) was found with integers for $[a, c]$. Given such sets of two numbers, which determined a third positive integer b such that $b = \sqrt{c^2 - a^2}$, this table is more than convincing that the Babylonians were well aware of PTs, but less convincing that they had access to Euclid's formula (Eq. 1.6 p. 38).

It seems likely that Euclid's Elements was largely the source of the fruitful 6th century era due to the Greek mathematician Diophantus (Fig. 1.2), who developed the concept of *discrete mathematics*, now known as *Diophantine analysis*.

The work of Diophantus was followed by a rich mathematical era, with the discovery of 1) early calculus (Brahmagupta, 7th CE), 2) algebra (al-Khwārizmī, 9th CE), and 3) complex arithmetic (Bombelli, 15th CE). This period overlapped with the European middle (aka, dark) ages. Presumably European intellectuals did not stop thinking during these many centuries, but what happened in Europe is presently unclear given the available records.[9]

---

[9]It might be interesting to search the archives of the monasteries, where the records were kept, to figure out what happened during this strange time.

1.1.3   What is mathematics?

Mathematics is a language, not so different from other languages. Today's mathematics is a written language with an emphasis on symbols and glyphs, biased toward Greek letters. The etymology of these symbols would be interesting to study. Each symbol is dynamically assigned a meaning, appropriate for the problem being described. These symbols are then assembled to make sentences. It is similar to Chinese in that the spoken and written version are different across dialects. In fact, like Chinese, the sentences may be read out loud in the language (dialect) of your choice, while the mathematical sentence (like Chinese) is universal.

Math (i.e., the syntax) is a language: It seems strange when people complain that they "can't learn math,"[10] but they claim to be good at languages. Math is a language, with the symbols taken from various languages, with a bias toward Greek, due to the popularity of Euclid's Elements. Learning languages is an advanced social skill. However the social outcomes are very different between learning a romance language and math. Learning a new language is fun because it opens doors to other cultures.

Math is different due to the rigor of the rules of the language, along with the way it is taught (e.g., not as a language). A third difference between math and the romance languages is that math evolved from physics, with important technical applications. This was the concept behind the Pythagorean school, a band of followers called the *Pythagoreans.*

A further problem is that pre-high-school students confuse arithmetic with math. The two topics are very different, and students need to understand this. One does not need to be good at arithmetic to be good at math (but it doesn't hurt).

There are many rules that must be mastered. These rules are defined by algebra. For example the sentence $a = b$ means that the number a has the same value as the number b. The sentence is spoken as "a equals b." The numbers are nouns and the equal sign says they are equivalent, playing the role of a verb, or action symbol. Following the rules of algebra, this sentence may be rewritten as $a - b = 0$. Here the symbols for minus and equal indicate two types of actions.

Sentences can become arbitrarily complex, such as the definition of the integral of a function, or a differential equation. But in each case, the mathematical sentence is written down, may be read out loud, has a well defined meaning, and may be manipulated into equivalent forms following the rules of algebra and calculus. This language of mathematics is powerful, with deep consequences, known as proofs.

The writer of an equation should always translate (explicitly summarize the meaning of the expression), so the reader will not miss the main point. This is simply a matter of clear writing.

Language may be thought of as mathematics (turning this idea on its head). To properly write correct English it is necessary to understand the construction of the sentence. It is important to identify the subject, verb, object, and various types of modifying phrases. If you wish to read about this, look up the distinction between the words *that* and *which*, which make a nice example of this concept. Most of us work directly with what we think "sounds right," but if you're learning English as a second language, it is very helpful to understand the rules, which are arguably easier to master than the foreign phones (i.e., speech sounds).

1.1.4   Early physics as mathematics

The role of mathematics is to summarize algorithms (i.e., sets of rules). As outlined in Fig. 1.2, it was clear to Pythagoras (and many others before him), that there was an important relationship between mathematics and the physical world. Pythagoras may have been one of the first to capitalize on this relationship, using science and mathematics to design and make things.[11] This was the beginnings of technology as we know it, coming from the relationship between physics and math, impacting

---

[10]"It looks like Greek to me."

[11]It seems likely that the Chinese and Egyptians also did this, but that is more difficult to document.
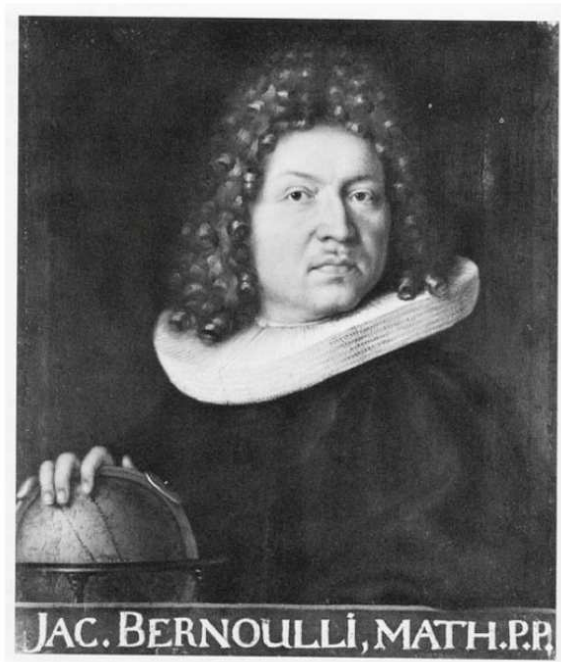
Figure 13.10: Portrait of Jakob Bernoulli by Nicholas Bernoulli

Figure 13.11: Johann Bernoulli

Figure 10.4: Leonhard Euler

Figure 1.3: *Above: Jakob (1655-1705) and Johann (1667-1748) Bernoulli; Below: Leonhard Euler (1707) and Jean le Rond d'Alembert (1717-1783). Euler was blind in his right eye, hence the left portrait view. The figure numbers are from Stillwell (2010).*

map making, tools, implements of war (the wheel, gunpowder), art (music), sound, water transport, sanitation, secure communication, food, . . . , etc.

Why is Eq. 1.1 called a *theorem*, and what exactly needs to be proved? We do not need to prove that (a, b, c) obey this relationship, since this is a condition that is observed. We do not need to prove that a^2 is the area of a square, as this is the definition of the area of a square. What needs to be proved is that this relation only holds if the angle between the two shorter sides is 90°, when the sum of the areas a^2 and b^2 equals area c^2. It follows that lengths a and b are less than length c, which is less than $a + b$ (i.e., $0 < a \le b < c < a + b$).

To appreciate the significance of this development it is helpful to trace the record back to before the time of the Greeks. The Pythagorean theorem (Eq. 1.1) did not begin with Euclid or Pythagoras. Rather Euclid and Pythagoras appreciated the importance of these ideas and documented them.

In the end the Pythagoreans were destroyed by fear. This may be the danger of mixing technology and politics:

> "Whether the complete rule of number (integers) is wise remains to be seen. It is said that when the Pythagoreans tried to extend their influence into politics they met with popular resistance. Pythagoras fled, but he was murdered in nearby Mesopotamia in 497 BCE."

–Stillwell (2010, p. 16)

Chronological history 17-18th centuries
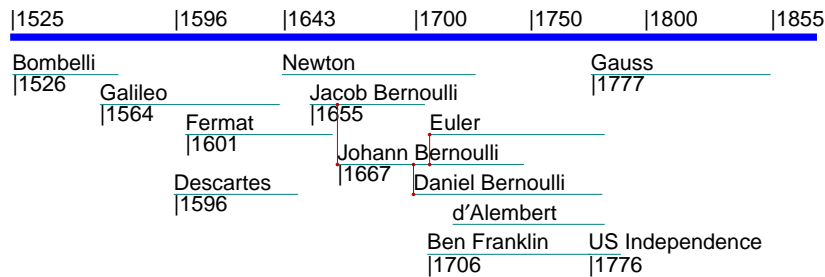
Figure 1.4: *Exploded time-line of the two centuries between the 16th and 18th CE. Given the time line, it seems likely that Bombelli's discovery of Diophantus' book on "Arithmetic" in the Vatican library, triggered many of the ideas presented by Descartes and Fermat, followed by others (i.e., Newton). Bombelli's discovery might be considered as a magic moment in mathematics. The vertical red lines connect mentor-student relationships. Figure 1.2 (p. 16) gives the time-line from 1500BCE to 1650CE. Figure 1.11 (p. 45) give the 17-20 CE (Newton–Einstein) view from 1640–1950, and Fig. 1.17 (p. 75) outlines the full (Bombelli–Einstein, 1525-1925) period.* [fig:TimeLine16_18]

1.1.5 The birth of modern mathematics

Modern mathematics (what we know today) was born in the 15-16th century, in the hands of Leonardo da Vinci, Bombelli, Galileo, Descartes, Fermat, and many others (Stillwell, 2010). Many of these early master were, like the Pythagoreans, secretive to the extreme about how they solved problems. This soon changed due to Mersenne, Descartes and Newton, causing mathematics to blossom.

The amazing Bernoulli family The first individual that seems to have openly recognized the importance of mathematics, to actually teach it, was Jacob Bernoulli (Fig. 1.3). Jacob worked on what is now view as the standard package of analytic "circular" (i.e., periodic) functions: $\sin(x)$, $\cos(x)$, $\exp(x\jmath)$, $\log(x)$.[12] Eventually the full details were developed (for real variables) by Euler (Sections 1.3.1 p. 49 and 3.1.1, p. 130).

---

[12]The log and tan functions are related by $\tan^{-1}(z) = -\frac{1\jmath}{2} \ln(\frac{1\jmath-z}{1\jmath+z})$.

From Fig. 1.17 we see that Jacob was contemporary to Mersenne, Descartes, Fermat, Huygens, Newton, and Euler. Thus it seems likely that he was strongly influenced by Newton, who in turn was influenced by Descartes,[13] Viète and Wallis (Stillwell, 2010, p. 175). With the closure of Cambridge University due to the plague of 1665, Newton returned home to Woolsthorpe-by-Colsterworth (95 [mi] north of London), to worked by himself, for over a year.

Jacob Bernoulli, like all successful mathematicians of the day, was largely self taught. Yet Jacob was in a new category of mathematicians, because he was an effective teacher. Jacob taught his sibling Johann, who then taught his sibling Daniel. But most importantly, Johann taught Leonhard Euler (Figs. 1.3, 1.17), the most prolific (thus influential) of all mathematicians. This resulted in an explosion of new ideas and understanding. It is most significant that all four mathematicians published their methods and findings. Much later, Jacob studied with students of Descartes[14] (Stillwell, 2010, p. 268-9).

Euler went far beyond all the Bernoulli family, Jacob, Johann and Daniel, (Stillwell, 2010, p. 315). A special strength of Euler was the degree to which he published. First he would master a topic, and then he would publish. His papers continued to appear long after his death (Calinger, 2015).

Another individual of that time of special note, who also published extensively, was d'Alembert (Figs. 1.3, 1.17). Some of the most important tools were first proposed by d'Alembert. Unfortunately, and perhaps somewhat unfairly, his rigor was criticized by Euler, and later by Gauss (Stillwell, 2010).

Once the tools were being openly published, mathematics grew exponentially.[15] It was one of the most creative times in mathematics. Figure 1.17 shows the list of the many famous names, and their relative time-line. To aid in understand the time line, note that Leonhard Euler was a contemporary of Benjamin Franklin, and James Clerk Maxwell of Abraham Lincoln.[16]

1.1.6   Three Streams from the Pythagorean theorem

From the outset of his presentation, Stillwell (2010, p. 1) defines "three great *streams* of mathematical thought: *Numbers, Geometry and Infinity,*" that flow from the Pythagorean theorem, as summarized in Table 1.1. The Pythagorean theorem is the spring from which flow the three streams of all mathematics. This is a useful concept, based on reasoning not as obvious as one might think. Many factors are in play here. One of these was the strongly held opinion of Pythagoras that all mathematics should be based on integers. The rest are tied up in the long, necessarily complex history of mathematics, as best summarized by the fundamental theorems (Table 1.2, p. 29), which are each discussed in detail in a relevant chapter.

Stillwell's concept of three streams following from the Pythagorean theorem is the organizing principle behind the this book, organized by chapter:

1. *Introduction* (Chapter 1) A detailed overview of the fundamentals and the three streams are presented in Sections

   1.2 Number systems

   1.3 Algebraic equations

   1.4 Scalar Calculus)

   1.5 Vector Calculus)

   Chapter 1 is intended to be self contained. Chapter 2-5 are dedicated to advanced topics. If your a novice, stick to Chapter 1.

   As of Version Version 0.81.8, Chapters 2-5 are not finished.

---

[13] https://en.wikipedia.org/wiki/Early_life_of_Isaac_Newton

[14] It seems clear that Descartes was also a teacher.

[15] There are at least three useful exponential scales: Factors of 2, factors of $e \approx 2.7$, and factors of 10. The decibel uses factors of 2 (6 [dB]) and factors of 10 (20 [dB]). Information theory uses powers of 2 (1 [bit]), 4 (2 [bits]). Circuit theory uses in factors of $e \approx 3.7$ (11.4 [dB]).

[16] Lincoln traveled through Mahomet IL (where I live) on his way to the Urbana Court house.

2. *Number Systems* (Chapter 2: Stream 1) Fundamentals of number systems, starting with prime numbers, through complex numbers, vectors and matrices.

3. *Algebraic Equations* (Chapter 3: Stream 2) Algebra and its development, as we know it today. The theory of real and complex equations and functions of real and complex variables. Complex impedance $Z(s)$ of complex frequency $s = \sigma + \omega\jmath$ is covered with some care, given its importance ₅ for engineering mathematics.

4. *Scalar Calculus* (Chapter 4: Stream 3a) Ordinary differential equations. Integral theorems. Acoustics.

5. *Vector Calculus*: (Chapter 5: Stream 3b) Vector Partial differential equations. Gradient, divergence and curl differential operators. Stokes, and Green's theorems. Maxwell's equations. ₁₀

Table 1.1: *Three streams followed from Pythagorean theorem: Number Systems (Stream 1), Geometry (Stream 2) and Infinity (Stream 3).*

- *The Pythagorean Theorem is the mathematical spring which bore the three streams.*
- *≈Several centuries per stream:*

  1) **Numbers:**

  $6^{th}BCE$ ℕ *counting numbers,* ℚ *(Rationals),* ℙ *Primes*
  $5^{th}BCE$ ℤ *Common Integers,* 𝕀 *Irrationals*
  $7^{th}CE$ *zero* ∈ ℤ

  2) **Geometry:** *(e.g., lines, circles, spheres, toroids, . . .)*

  $17^{th}CE$ *Composition of polynomials (Descartes, Fermat)*
  Euclid's Geometry + algebra ⇒ Analytic Geometry
  $18^{th}CE$ *Fundamental Theorem of Algebra*

  3) **Infinity:** *(∞ → Sets)*

  $17\text{-}18^{th}CE$ *Taylor series, Functions, Calculus (Newton)*
  $19^{th}CE$ ℝ *Real,* ℂ *Complex 1851*
  $20^{th}CE$ *Set theory*

1.2 Stream 1: Number Systems

Number theory (discrete, i.e., integer mathematics) was a starting point for many key ideas. For example, in Euclid's geometrical constructions the Pythagorean theorem for $\{a, b, c\} \in \mathbb{R}$ was accepted as true, but the emphasis in the early analysis was on integer constructions (i.e., $\{a, b, c\} \in \mathbb{N}$), such as ₁₅ Euclid's formula for Pythagorean triplets (Eq. 1.6, Fig. 1.8, p. 38). As we shall see, the Pythagorean theorem is a rich source of mathematical constructions, such as composition of polynomials, and solutions of Pell's equation by eigenvector and recursive analysis methods. Recursive difference equation solutions predate calculus, at least going back to the Chinese (c2000 BCE). These are early (pre-limit) forms of differential equations, best analyzed using an eigenvector expansion, a powerful geometrical ₂₀ concept from linear algebra, of an expansion in terms of an orthogonal set of normalized (unit-length) vectors (Appendix C, p. 175),

The first use of zero and ∞: It is hard to imagine that one would not appreciate the concept of zero and negative numbers when using an abacus. If five beads are moved up, and one is moved down, then four are left. Then if four more are move down, that leaves zero. Taking away (subtracting) is the ₂₅

opposite of giving (adding). Thus subtraction, to obtain zero beads, is no different than subtraction from zero, giving *negative* four beads. On an abacus, subtraction is obviously the inverse of addition. To assume the Romans or the Chinese did not understand negative numbers, seems a bit absurd.

However, understanding the concept of zero and negative numbers is not the same as having a
5 symbolic notation. The Roman number system has no such symbols. The first recorded use of a symbol for zero is said to be by Brahmagupta in 628 CE.[17] It does not take much imagination to go from counting numbers $\mathbb{N}$ to the set of all integers $\mathbb{Z}$, including zero, but apparently it takes 600 years to functionally develop a terminology, that represents these ideas. Likely this is more about politics and government rule, than mathematics. Defining the *rules* of subtraction required the creation of
10 algebra (c830 CE, Fig. 1.2).

The concept that caused much more difficulty was ∞. Until Riemann's thesis in 1851 it was not clear if ∞ was a number, many numbers, or even definable.

1.2.1   Lec 2: The Taxonomy of Numbers: $\mathbb{P}, \mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{I}, \mathbb{R}, \mathbb{C}$

Once symbols for zero and negative numbers were accepted, progress was made. In a similar manner, to
15 fully understand numbers, a transparent notation is required. First one must differentiate between the different classes (genus) of numbers, providing a notation that defines each of these classes, along with their relationships. It is logical to start with the most basic *counting numbers*, which we indicate with the double-bold symbol $\mathbb{N}$. All the double-bold symbols and their genus are summarized in Appendix A.

20 **Counting numbers $\mathbb{N}$:**   These are known as the "natural numbers" $\{1, 2, 3, \cdots\} \in \mathbb{N}$, denoted by the double-bold symbol $\mathbb{N}$. For increased clarity we shall refer to the natural numbers as *counting numbers*, to clarify that *natural* means *integer*. The mathematical sentence $2 \in \mathbb{N}$ is read as *2 is a member of the set of counting numbers*. The word *set* means the *sharing of a specific property*.

Primes $\mathbb{P}$:   A prime number $\mathbb{P}$ is an integer that may not be factored by other than 1 and itself.
25 Since $1 = 1 \cdot 1$, $1 \notin \mathbb{P}$, as it is seen to violate this basic definition of a prime. Prime numbers $\mathbb{P}$ are a *subset* of the counting numbers ($\mathbb{P} \subset \mathbb{N}$). We shall use the convenient notation π_n for the prime numbers, indexed by $n \in \mathbb{N}$. The first 12 primes ($n = 1, \ldots, 12$) are $\pi_n = 2, 3, 5, 7, 11, 13, 17, 19, 23,$ $29, 31, 37$. Since, $4 = 2^2$ and $6 = 2 \cdot 3$ may be factored, $\{4, 6\} \notin \mathbb{P}$ (read as: *4 and 6 are not in the set of primes*). Given this definition, multiples of a prime, i.e., $n\pi_k \equiv [2, 3, 4, 5, \ldots] \times \pi_k$ of any prime π_k,
30 cannot be prime. It follows that all primes except 2 must be odd and every integer N is unique in its factorization.

Coprimes are numbers with no common factors. For example, $4 = 2 \cdot 2$ and $6 = 2 \cdot 3$ are *not* coprime, as they have 2 as a common factor, whereas $21 = 3 \cdot 7$ and $10 = 2 \cdot 5$ *are*. By definition all distinct primes are coprime. The notation $m \perp n$ indicates that m, n are coprime. The ratio of two coprimes
35 is reduced, as it has no factors to cancel. The ratio of two numbers that are not coprime, may always be reduced to a smaller ratio, once the common factors are canceled. This is called the *reduced form*, or *irreducible fraction*. When doing numerical work, it is always beneficial to work with the reduced form. This is especially important when working with the ratio of polynomials, to remove common roots, typically by long division or deconvolution.
40 The *fundamental theorem of arithmetic* states that all integers may be uniquely expressed as a product of primes. The *Prime Number Theorem* estimates the mean density of primes over $\mathbb{N}$.

Integers $\mathbb{Z}$:   These include positive and negative counting numbers and zero. Notionally we might indicate this using *set notation* as $\mathbb{Z} : \{-\mathbb{N}, 0, \mathbb{N}\}$. Read this as *The integers are in the set composed of the negative of the natural numbers $-\mathbb{N}$, zero, and $\mathbb{N}$*. Note that $\mathbb{N} \subset \mathbb{Z}$.

---

[17]The fall of the Roman Empire has been established as Sept. 4, 476 CE.

Rational numbers $\mathbb{Q}$: These are defined as numbers formed from the ratio of two integers. Since the integers $\mathbb{Z}$ include 1, it follows that integers are a subset of rational numbers ($\mathbb{Z} \subset \mathbb{Q}$). For example, the rational number $3/1 \in \mathbb{Z}$). The main utility of rational numbers is that that they can efficiently approximate any number on the real line, to any precision. For example the rational approximation $\pi \approx 22/7$, has a relative error of $\approx 0.04\%$.

Fractional number $\mathbb{F}$: The fractionals $\mathbb{F}$ are defined as the ratio of two coprimes. If $a \perp b$, then $c = a/b \in \mathbb{F}$. Given this definition, $\mathbb{F} \perp \mathbb{Z}$, $\mathbb{F} \subset \mathbb{Q} = \mathbb{Z} \cup \mathbb{F}$. Because of the powerful approximating power of rational numbers, the fractional set $\mathbb{F}$ represent the most important (and the largest) portion of the rational numbers, dwarfing the size of the integers, another good reason for defining the two distinct subsets. For example $\pi \approx 22/7 = 11 \cdot 2/7 = 3 + 1/7$ with $22 \perp 7$, and $9/6 = 3/2 = 1 + 1/2$ with $3 \perp 2$.

Irrational numbers $\mathbb{I}$: Every real number that is not rational is *irrational* ($\mathbb{Q} \perp \mathbb{I}$). Irrational numbers include π, e and the square roots of the primes ($\in \sqrt{\mathbb{P}}$). These are decimal numbers that never repeat, thus requiring infinite precision in their representation. Such numbers can not, and never will be represented on any computer, as they would require an infinite number of bits (precision).

The rationals $\mathbb{Q}$ and irrationals $\mathbb{I}$ split the reals ($\mathbb{R} : \mathbb{Q} \cup \mathbb{I}$, $\mathbb{Q} \perp \mathbb{I}$), thus each is a subset of the reals ($\mathbb{Q} \subset \mathbb{R}$, $\mathbb{I} \subset \mathbb{R}$). This is analogous to the integers $\mathbb{Z}$ and fractionals $\mathbb{F}$, which split the rationals ($\mathbb{Q} : \mathbb{Z} \cup \mathbb{F}$, $\mathbb{Z} \perp \mathbb{F}$) (thus each is a subset of the rationals ($\mathbb{Z} \subset \mathbb{Q}$, $\mathbb{F} \subset \mathbb{Q}$)).

Irrational numbers ($\mathbb{I}$) were famously problematic for the Pythagoreans, who incorrectly theorized that all numbers were rational. Like ∞, irrational numbers require a new and difficult concept before they may even be defined: It was essential to understand the factorization of counting numbers into primes (i.e., the fundamental theorem of arithmetic) before the concept of irrationals could be sorted out.

As discussed in Sect. Lec. 1.2.5 (p. 36), fractionals can approximate any irrational number with arbitrary accuracy. Integers are also important, but for a very different reason. All numerical computing today is done with $\mathbb{Q} = \mathbb{F} \cup \mathbb{Z}$. Indexing uses integers $\mathbb{Z}$, while the rest of computing (flow dynamics, differential equations, etc.) is done with fractionals $\mathbb{F}$ (i.e., IEEE-754). Computer scientists are trained on these topics, and computer engineers need to be at least conversant with them.

Real numbers $\mathbb{R}$: Reals are the union of rational and irrational numbers, namely $\mathbb{R} : \{\mathbb{I}, \mathbb{Q}\}$ ($\mathbb{R} = \mathbb{Z} \cup \mathbb{F} \cup \mathbb{I}$). The lengths in Euclidean geometry are reals. Many people assume that *IEEE 754 floating point numbers* (c1985) are real (i.e., $\in \mathbb{R}$). In fact they are rational ($\mathbb{Q} : \{\mathbb{F} \cup \mathbb{Z}\}$) approximations to real numbers, designed to have a very large dynamic range. There can be no machine realization of irrational numbers, since such a number would require infinite precision (∞ bits). The hallmark of fractional numbers ($\mathbb{F}$) is their power in making highly accurate approximations of any real number.

Using Euclid's compass and ruler methods, one can make line length proportionally shorter or longer, or (approximately) the same. A line may be made be twice as long, an angle bisected. However, the concept of an integer length in Euclid's geometry was not defined.[18] Nor can one construct an imaginary or complex line as all lines are assumed to be real.

Real numbers were first fully accepted only after set theory was developed by Cantor (1874) (Stillwell, 2010, pp. 461, 525, ...). This seems amazing, given how widely accepted real numbers are today. In some sense they were accepted by the Greeks, as lengths of real lines.

The cardinality of $\mathbb{R}$: The real line may be divided into a hierarchical groups, sorted by size: $|\mathbb{R}| > |\mathbb{I}| \ggg |\mathbb{Q}| > |\mathbb{F}| \gg |\mathbb{Z}| > |\mathbb{N}| > |\mathbb{P}|$. The reals may be split into irrationals and rationals, which may be further split into the integers and the fractionals. Thus, all is *not* integer. If a triangle has two integer sides, then the hypotenuse can be irrational ($\sqrt{2} = \sqrt{1^2 + 1^2}$). This leads us to a fundamental

---

[18] As best I know.

question: "Are there integer solutions to Eq. 1.1?" We need not look further than the simple example $\{3, 4, 5\}$. In fact the example generalizes: the formula for generating an infinite number of integer solutions is called *Euclid's Formula* (Section 1.2.5, p. 36).

5   The more important point is that the cardinality of the irrationals is much larger than any set other than the reals (i.e., complex numbers). Thus when we use computers to model physical systems, we are constantly needing to compute with irrational numbers. But this is impossible since every irrational numbers would require an infinite number of bits to represent it. Thus we must compute with rational approximations to the irrationals. This means we need to use the fractionals. In the end, we must work with the IEEE 754 floating point numbers,[19] which are rationals.

10 **Complex numbers** $\mathbb{C}$:   Complex numbers are best defined as *ordered pairs of real numbers*.[20] The word "complex," as used here, does *not* mean that the number are complicated or difficult. The word "imaginary" does not mean the numbers disappear. Complex numbers are quite special in engineering mathematics, as roots of polynomials. The most obvious example is the quadratic formula for the roots of a 2^d degree polynomial, having either rational, real or complex coefficients.

15   Today the common way to write a complex number is using the notation $z = a + bj \in \mathbb{C}$, where $a, b \in \mathbb{R}$. Here $1j = \sqrt{-1}$. We also define $1i = -1j$ to account for the two possible signs of the square root. Accordingly $1j^2 = 1i^2 = -1$.

Multiplication of complex numbers follows the basic rules of real algebra, for example, the rules of multiplying two polynomials. Multiplication of two first degree polynomials gives

$$(a + bx)(c + dx) = ac + (ad + bc)x + bdx^2$$

If we substitute $1j$ for x, and use the definition $1j^2 = -1$, we obtain the product of the two complex numbers

$$(a + bj)(c + dj) = ac - bd + (ad + bc)j.$$

Thus multiplication of complex numbers obey the accepted rules of algebra.

Polar representation: A alternative for complex multiplication is to work with polar coordinates. The polar form of complex number $z = a + bj$ is written in terms of its magnitude $\rho = \sqrt{a^2 + b^2}$ and angle $\theta = \angle z = \tan^{-1}(z) = \arctan z$, as $z = \rho e^{\theta j}$. From the definition of the complex natural log function

$$\ln \rho e^{\theta j} = \ln \rho + \theta j,$$

which is useful in engineering calculations.[21] Note that the angle $\angle \theta = \Im \log z$. Clearly this is an
20 improvement over $\angle \theta = \arctan(z)$.

Matrix representation: An alternative way to represent complex numbers is in terms of 2x2 matrices. This relationship is defined by the mapping from a complex number to a 2x2 matrix

$$a + bj \leftrightarrow \begin{bmatrix} a & -b \\ b & a \end{bmatrix}, \quad 2 \leftrightarrow \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad -j \leftrightarrow \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad e^{\theta j} \leftrightarrow \begin{bmatrix} \sin(\theta) & -\cos(\theta) \\ \cos(\theta) & \sin(\theta) \end{bmatrix} \tag{1.2}$$

By taking the inverse of the 2x2 matrix one can define the ratio of one complex number by another, Until you try out this representation, it may not seem obvious, or even that it could work.

---

[19]IEEE 754: `http://www.h-schmidt.net/FloatConverter/IEEE754.html`.

[20]A polynomial $a + bx$ and a 2-vector $[a, b]^T = \begin{bmatrix} a \\ b \end{bmatrix}$ are also examples of ordered pairs.

[21]Chapter 2 discusses the definition of the phase, i.e., how is it computed (i.e., $\arctan(e^{\theta j})$, `arctan2(x,y)`), and the importance of the unwrapped phase, as in the example (Table 1.5, p. 71) $\delta(t - \tau) \leftrightarrow e^{-\tau j}$.

Exercise: Verify that

$$\frac{a + bȷ}{c + dȷ} = \frac{ab + bd + (bc - ad)ȷ}{c^2 + d^2} \longleftrightarrow \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c & -d \\ d & c \end{bmatrix}^{-1} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c & d \\ -d & c \end{bmatrix} \frac{1}{c^2 + d^2}.$$

This representation proves that $1ȷ$ is not necessary when defining a complex number. What $1ȷ$ can do is to conceptually simplify the algebra. It is worth your time to become familiar with the matrix representation, to clarify any possible confusions you might have about multiplication and division of complex numbers. This matrix representation can save you time, heartache and messy algebra. Once you have learned how to multiply two matrices, it's a lot simpler than doing the complex algebra. In many cases we will leave the results of our analysis in matrix form, to avoid the algebra altogether.[22] Thus both representations are important. More on this topic may be found in Chapter 2.

Real versus complex numbers: As discussed in Appendix A (p. 165), all numbers may be viewed as complex. Namely every real number is complex if we take the imaginary part to be zero (Boas, 1987). For example, $2 \in \mathbb{P} \subset \mathbb{C}$. Likewise every purely imaginary number (e.g., $0 + 1ȷ$) is complex with zero real part. It follows that $2ȷ = 0 + \pi_2 ȷ \in \mathbb{P}$. Integers are a subset of reals, which are a subset of complex numbers[23] *Gaussian integers* are complex integers ($\mathbb{Z} \subset \mathbb{R} \subset \mathbb{C}$).[24] From the above discussion it should be clear that each of these different classes of number are nested in a hierarchy, with the following embedding

$$\pi_k \in \mathbb{P} \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{F} \subset \mathbb{R} \subset \mathbb{C}.$$

The roots of polynomials x_k are complex ($x_k \in \mathbb{C}$), independent of the genus of the coefficients (e.g., real integer coefficients give rise to complex roots). Each genus plays an important role in algebra, with prime numbers at the bottom (root of the tree) and complex numbers at the top. We shall explore this further in Chaps. 2 and 3.

Finally, note that complex numbers $\mathbb{C}$ do not have "order," meaning one complex number cannot be larger or smaller than another. It makes no sense to say that $ȷ > 1$ or $ȷ = 1$ (Boas, 1987). The real and imaginary parts and the magnitude and phase, have order. If time t were complex, there could be no yesterday and tomorrow.[25]

Applications of integers

The most relevant question at this point is "Why are integers important?" First we count with them, so we can keep track of "how much." But there is much more to numbers than counting: We use integers for any application where absolute accuracy is essential, such as banking transactions (making change), the precise computing of dates (Stillwell, 2010, p. 70) and location ("I'll meet you at 34 and Vine at noon on Jan. 1, 2034."), building roads or buildings out of bricks (objects built from a unit size).

To navigate we need to know how to predict the tides, the location of the moon and sun, etc. Integers are important because they are precise: Once a month there is a full moon, easily recognizable. The next day its slightly less than full. If one could represent our position as integers in time and space, we would know exactly where we are at all times.

The Pythagoreans claimed that all was integer. From a practical point of view, it seems they were right. Today all computers compute floating point numbers as fractionals. However, in theory they were wrong. The difference is a matter of precision.

---

[22]Sometimes we let the computer do the final algebra, numerically, as 2x2 matrix multiplications.

[23]The plural *complexs* (a double /s/) seems an unacceptable word in English.

[24]It follows that integers are a subset of Gaussian integers (the imaginary or real part of the Gaussian integer may be zero).

[25]One can define $\xi = x + 1j\,ct$ to be complex ($x, t \in \mathbb{R}$, with x in meters [m] and t is in seconds [s], with the speed of light c [m/s].

History of complex numbers: It is notable how long it took for complex numbers to be accepted (1851), relative to when they were first introduced by Bombelli (16th century CE). One might have thought that the solution of the quadratic, known to the Chinese, would have settled this question. It seems that complex integers (aka, *Gaussian integers*) were accepted before non-integral complex numbers. Apparently real numbers ($\mathbb{R}$) were not accepted (i.e., proved to exist, thus mathematically defined) until even later. It took the development of *set theory* in the late 19th century to sort out a proper definition of the real number, due to the existence of irrational numbers. It makes one wonder "How many angels can stand on the point of a pin?"

Numerical Representations of $\mathbb{I}, \mathbb{R}, \mathbb{C}$:   When doing numerical work, one must consider how we may compute within the family of reals (i.e., irrationals). There can be no irrational number representation on a computers. IEEE floating point numbers, which are the international standard of computation, are actually rational approximations. The mantissa and the exponent are each integers, having sign and magnitude. The size of each integer depends on the precision of the number being represented. An IEEE floating-point number is rational because it has a binary (integer) mantissa, multiplied by 2 raised to the power of a binary (integer) exponent. For example $\pi \approx a2^b$ with $a, b \in \mathbb{Z}$. In summary, IEEE floating-point numbers are not, and cannot, be irrational, since to do that would require an infinite number of bits.

True floating point numbers contain irrational numbers, which must be approximate by rational numbers. This leads to the concept of *fractional representation*, which requires the definition of the *mantissa*, *base* and *exponent*, where both the mantissa and the exponent are signed. Numerical results must not depend on the base. One could dramatically improve the resolution of the numerical representation by the use of the *Fundamental Theorem of Arithmetic* (Section 1.2.2, page 30). For example one could factor the exponent into its primes and then represent the number as $a2^b3^c5^d7^e$, etc. Such a representation would improve the resolution of the representation. But even so, the irrational numbers would be approximate. For example, base ten[26] is natural using this representation since $10^n = 2^n5^n$. Thus

$$\pi \cdot 10^5 \approx 314\,159.27\ldots = 3 \cdot 2^55^5 + 1 \cdot 2^45^4 + 4 \cdot 2^35^3 + \cdots + 9 \cdot 2^05^0 + \cancel{2} \cdot 2^{\cancel{-1}}5^{-1} \cdots.$$

If we approximate π by $22/7$, then according to the Matlab/Octave DEC2BIN() routine, the binary representation is

$$\pi \cdot 2^{17} \approx 131072_{10} \cdot 22/7 = 110, 0100, 1001, 0010, 0101_2,$$

where 1 and 0 are multipliers of powers of 2, which are then added together as follows

$$2^{18} + 2^{17} + 2^{14} + 2^{11} + 2^8 + 2^5 + 2^2 + 2^0.$$

In base 16 (i.e, hexadecimal) $2^{17} \cdot 22/7 = 2^{18} \cdot 8_{16}/7_{16}$.

Computers keep track of the decimal point using the exponent, which in this case is the factor 2^{17} = 131072_{10}. The concept of the decimal point is replaced by an integer, having the desired precision, and a scale factor of any base (radix). This scale factor may be thought of as moving the decimal point to the right (larger number) or left (smaller number). The mantissa "fine-tunes" the value about a scale factor (the exponent).

Here is $x = 2^{17} \times 22/7$ using IEEE 754 double precision, as computed by an IEEE-754 floating point converter:[27] $x = 411, 940.5625_{10} = 2^{54} \times 1198372 = 0, 10010, 00, 110010, 010010, 010010, 010010_2 = 0x48c92492_{16}$. The commas in the binary (0,1) string are to help visualize the quasi-periodic nature of the bit-stream. The mantissa is $4, 793, 490_{10}$ and the exponent is 2^{18}. The numbers are stored in a 32 bit format, with 1 bit for sign, 8 bits for the exponent and 23 bits for the mantissa. Perhaps a more instructive number is $x = 4, 793, 490.0 = 0, 100, 1010, 100, 100, 100, 100, 100, 100, 100, 100_2 = 0x4a924924_{16}$, which has a repeating binary bit pattern of $((100))_3$, broken by the scale factor

---

[26]Base 10 is the natural world-wide standard simply because we have 10 fingers.

[27]http://www.h-schmidt.net/FloatConverter/IEEE754.html

$0x4a$. Another with even higher symmetry is $x = 6.344, 131, 191, 146, 9 \times 10^{-17} = 0x24, 924, 924_{16} =$ $00, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100_2$. In this example the repeating pattern is clear in the Hex representation as a repeating $((942))_3$, as represented by the double brackets, with the subscript indicating the period, in this case, 3 digits. As before, the commas are to help with readability, and have no other meaning.

There are other important types of representations. As pairs of reals, complex numbers have similar approximate representations. An important representations of complex numbers is $e^z = \cosh(z) + j \sinh(z)$ with $z \in \mathbb{C}$, which includes the famous formula of Euler ($\theta \in \mathbb{R}$)

$$e^{j\theta} = \cos\theta + j\sin\theta \leftrightarrow \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Some of these concepts can be generalized to include vectors, matrices and polynomials.

Integers and the Pythagoreans   The integer is the corner stone of the Pythagorean doctrine, so much so that it caused a fracture within the Pythagoreans when it was discovered that not all numbers are rational. The famous example is the isosceles triangle $1, 1, \sqrt{2}$, which lead to the next triangle $[1, 2, \sqrt{3}]$, etc. This is known as the Spiral of Theodorus: the short side is 1 and the long side is recursively incremented by one, using a simple compass-ruler construction.

There are right-triangles with integral lengths, the best known being $[3, 4, 5]$. Such triplets of integers $[a, b, c]$ that satisfy the Pythagorean formula (Eq. 1.1) are denoted *Pythagorean triplets*, which may be verified using *Euclid's formula* (Eq. 1.6, p. 123).

To form triangles with perfect $90°$ angles, the lengths need to satisfy Eq. 1.1. Such triangles are useful in constructing buildings or roads made from of bricks having a uniform size.

Public-private key Security: An important application of prime numbers is public-private key (RSA) encryption, essential for internet security applications (e.g., online banking). To send secure messages the security (i.e., utility) of the internet is dependent on key encryption. Most people assume this is done by a personal login and passwords. Passwords are simply not secure, for many reasons. A proper method depends on factoring integers formed from products of primes having thousands of bits.[28] The security is based on the relative ease in multiplying large primes, but the virtual impossibility of factoring them.

When a computation is easy in one direction, but its inverse is impossible, it is called a *trap-door function.* We shall explore the reasons for this in Chapter 2. If everyone were to switch from passwords to public key encryption, the internet would be much more secure.

Prob → HW

Puzzles: A third application of integers are imaginative problems that use integers. An example is the classic Chinese *Four stone problem:* "Find the weight of four stones that can be used with a scale to weigh any object (e.g., salt, gold) between 0, 1, 2, ..., 40 [gm]." As with the other problems, the answer is not as interesting as the method, since the problem may be easily recast into a related one. This type of problem can be found in airline magazines as entertain on a long flight. The solution to this problem is best cast as a linear algebra problem, with integer solutions. Again, once you know the trick, it is "easy."[29]

---

[28]It would seem that public key encryption could work by having two numbers with a common prime, and then by using Euclidean algorithm, that GCD could be worked out. One of the integers could be the public key and the second could be the private key. Given the difficulty of factoring the numbers into their primes, and ease of finding the GCD using Euclidean algorithm, a practical scheme may be possible. One problem with this idea is that one may store all the known primes.

[29]When ever someone tells you something is "easy," you should immediately appreciate that it is very hard, but there is a concept, that once you learn, the difficulty evaporates.

1.2.2   Lec 3: The role of physics in mathematics

Bells, chimes and Eigenmodes: Integers naturally arose in art, music and science. Examples include the relations between musical notes, the natural eigenmodes (tones) of strings and other musical instruments. These relations were so common and well studied, it appeared that to understand the physical world (aka, the Universe), one needed to understand integers. This was a seductive view, but not actually correct. As will be discussed in Sections 1.3.1 (p. 43) and 3.1.1 (p. 129), it is best to view the relationship between acoustics, music and mathematics as historical, since these topics played such an important role in the development of mathematics. Also interesting is the role that integers play in quantum mechanics, also based on eigenmodes, but in this case, those of atoms. Eigenmodes follow from solutions of the wave equation, which has natural delay due to d'Alembert's solution, along with reflecting boundary conditions (Eq. 1.117, p. 102), resulting in roots of the characteristic equation.

Engineers are so accustomed to working with real (or complex) numbers, the distinction between real (i.e., irrational) and fractional numbers are rarely acknowledged. Integers on the other hand arise in many contexts. One cannot master programming computers without understanding integer, hexadecimal, octal, and binary representations, since all numbers in a computer are represented in numerical computations in terms of rationals ($\mathbb{Q} = \mathbb{Z} \cup \mathbb{F}$).

As discussed in Section 1.2.1 (p. 22), the primary reason integers are so important is their absolute precision. Every integer $n \in \mathbb{Z}$ is unique,[30] and has the *indexing property*, which is essential for making lists that are ordered, so that one can quickly look things up. The alphabet also has this property (e.g., a book's index). Other than for hexadecimal numbers, which for notional reasons use the alphabet, letters (base 26), are equivalent to integers (base 10).

Because of the integer's absolute precision, the digital computer overtook the analog computer, once it was practical to make logic circuits that were fast. The first digital computer was thought to be the *Eniac* at the University of Pennsylvania, but it turned out that the code-breaking effort in Bletchley Park, England, under the guidance of Alan Turing, created the first digital computer (The Colossus) to break the WWII German "Enigma" code. Due to the high secrecy of this war effort, the credit was only acknowledged in the 1970s when the project was finally declassified.

There is zero possibility of analog computing displacing digital computing, due to the importance of precision (and speed). But even with binary representation, there is a non-zero probability of error, for example on a hard drive, due to physical noise. To deal with this, error correcting codes have been developed, to reduce the error by several orders of magnitude. Today this is a science, and billions of dollars are invested to increase the density of bits per area, to increasingly larger factors. A few years ago the terabyte drive was unheard of; today it is the standard. In a few years petabyte drives will certainly become available. It is hard to comprehend how these will be used by individuals, but they are essential for on-line (cloud) computing.

The role of mathematics in physics

Modern mathematics is build on a hierarchical construct of *fundamental theorems*, as summarized in Table 1.2. The importance of such theorems cannot be overemphasized. Gauss' and Stokes' Laws play a major role in understanding and manipulating Maxwell's equations. Every engineering student needs to fully appreciate the significance of these key theorems. If necessary, memorize them. But that will not do over the long run, as each and every theorem must be fully understood. Fortunately most students already know several of these theorems, but perhaps not by name. In such cases, it is a matter of mastering the vocabulary.

The theorems are naturally organized and may be thought of in terms of Stillwell's three streams. For Stream 1 there is the *Fundamental Theorem of Arithmetic* and the *Prime Number Theorem*. For Stream 2 there is the *Fundamental Theorem of Algebra* and Bézout's theorem, while for Stream 3 there are a host of theorems on calculus, ordered by their dimensionality. Some of these theorems verge

---

[30]Check out the history of $1729 = 1^3 + 12^3 = 9^3 + 10^3$.

Table 1.2:   *The fundamental theorems of mathematics*

---

1. Fundamental theorems of:

    (a) **Number systems: Stream 1**
        - arithmetic
        - prime number

    (b) **Geometry: Stream 2**
        - algebra
        - Bézout

    (c) **Calculus: Stream 3**[a]
        - Leibniz $\mathbb{R}^1$
        - complex $\mathbb{C} \subset \mathbb{R}^2$
        - vectors $\mathbb{R}^3, \mathbb{R}^n, \mathbb{R}^\infty$
            - Gauss' Law (Divergence theorem)
            - Stokes' Law (Curl theorem, or Green's theorem)
            - Vector calculus (Helmholtz's theorem)

2. Other key concepts:

    - Complex analytic functions (complex roots are finally accepted!)
        - Complex Taylor series (complex analytic functions)
        - Region of convergence (RoC) of complex analytic series
        - Laplace transform, and its inverse
        - Causal time $\implies$ complex frequency s
        - Cauchy Integral Theorem
        - Residue integration (i.e., Green's Thm in $\mathbb{R}^2$)
    - Riemann mapping theorem (Gray, 1994; Walsh, 1973)
    - Complex Impedance (Ohm's Law) (Kennelly, 1893)

---

[a]Flanders, Harley (June–July 1973). "Differentiation under the integral sign." *American Mathematical Monthly* 80 (6): 615-627. doi:10.2307/2319163. JSTOR 2319163.

on trivial (e.g., the Fundamental Theorem of Arithmetic). Others are more challenging, such as the *Fundamental Theorem of Vector Calculus* and *Green's theorem.*

Complexity should not be confused with importance. Each of these theorems is, as stated, fundamental. Taken as a whole, they are a powerful way of summarizing mathematics.

Stream 1: Prime Number theorems:

There are two fundamental theorems about primes,

1. *The fundamental theorem of arithmetic*: This states that every integer $n \in \mathbb{Z}$ may be uniquely factored into prime numbers. This raises the question of the meaning of *factor* (split into a product). The *product* of two integers $m, n \in \mathbb{Z}$ is $mn = \sum_m n = \sum_n m$.

2. *The Prime Number Theorem:* One would like to know how many primes there are. That is easy: $|\mathbb{P}| = \infty$ (The *cardinality*, or size of the set of primes, is infinite). The proper way of asking this questions is *What is the average density of primes, in the limit as $n \to \infty$?* This question was answered, for all practical purposes, by Gauss, who in his pastime, computed the first three million primes by hand. He discovered that, to a good approximation, the primes are equally likely, on a log scale. This is nicely summarized by the jingle, attributed to the mathematician Pafnuty Chebyshev (Stillwell, 2010, p. 585)

   Chebyshev said, and I say it again: *There is always a prime between n and 2n.*

When the ratio (interval) of two frequencies (pitch) is 2, the relationship is called an *octave.* Thus we might say, with a slight stretch of terminology, there is at least one prime per octave. Thus one might wonder about the maximum number of primes per octave. In modern music the octave is further divided into 12 intervals called *semitones* (factors), equal to the $\sqrt[12]{2}$. The product of 12 semitones is an octave. Thus one might ask how many primes there are per semitone? In the end, it is a question of the density of primes on a log (i.e., ratio) scale.

Stream 2: Fundamental theorem of algebra

This theorem states that every polynomial has at least one root. When that root is removed then the degree of the polynomial is reduced by 1. Thus when applied recursively, a polynomial of degree N has N roots.

Besides the fundamental theorem of algebra, a second important theorem is Bézout's theorem, which is a generalization of the fundamental theorem of algebra. It says[31] that the *composition* of two polynomials has degree equal to the product of the degrees of each polynomial. For example, if $P_3(x) = x^3$ and $P_5(x) = x^5$, then $P_3(P_5)(x) = (x^5)^3 = x^{15}$. It further states that when counting the N roots of a polynomial of degree N, one must include the imaginary roots, double roots and roots at infinity, some of which may difficult to identify.

One must wonder what happens when the degree is fractional, or worse, irrational? Are these cases covered by Bézout?

Stream 3: Fundamental theorems of calculus

In Sections 1.5.4 and 1.5.5 we will deal with each of the theorems for Stream 3, where we consider the several fundamental theorems of integration, starting with Leibniz's formula for integration on the real line ($\mathbb{R}$), then progressing to complex integration in the complex plane ($\mathbb{C}$) (*Cauchy's theorem*), which is required for computing the inverse Laplace transform. Gauss' and Stokes' Laws for $\mathbb{R}^2$ require closed and open surfaces, respectively. One cannot manipulate Maxwell's equations, fluid flow, or acoustics

---

[31]Statements of the theorem speak of *intersections* and constructions of curves, rather than compositions. I find this somewhat confusing. For example, how does intersection differ from elimination, or construction from composition (Stillwell, 2010, p. 119)?

without understanding these theorems. Any problem that deals with the wave equation in more than one dimension, requires an understanding of these theorems, thus are the basis of the derivation of the Kirchhoff voltage and current laws. The ∇ symbol is pronounced as "del" (preferred) or "nabla."

Finally we define the four basic vector operations based on the ∇ "operator:" the *gradient* $\nabla()$, *divergence* $\nabla\cdot()$ *curl* $\nabla\times()$ and the *Laplacian* $\nabla\cdot\nabla() = \nabla^2()$. The first three operations are defined in terms of integral operations on a surface in 1, 2 or 3 dimensions, by then taking the limit as that surface, and the volume contained within, goes to zero. These three differential operators are necessary if you wish to understand Maxwell's Equations, the crown jewel of mathematical physics. Hence mathematics plays a key role in physics, as does physics in math.

Other key concepts

Besides the widely recognized fundamental theorems for the three streams, there are a number of equally important theorems that have not yet been labeled as "fundamental."[32]

The widely recognized *Cauchy Integral Theorem* is an excellent example, since it is a stepping stone to Green's theorem and the *fundamental theorem of complex calculus*. In Section 1.5.3 (p. 100) we clarify the contributions of each of these special theorems.

Once these fundamental theorems of integration (Stream 3) have been mastered, the student is ready for the *complex frequency domain*, which takes us back to Stream 2 and the *complex frequency plane* ($s = \sigma + \omega\jmath \in \mathbb{C}$). While the Fourier and Laplace transforms are taught in mathematics courses, typically few physical connections are made, accordingly the concept of *complex frequency* is rarely mentioned. The *complex frequency domain* and *causality* are fundamentally related (Sects. 1.4.6–1.4.8, p. 85–86), and critical for the analysis of signals and systems, and especially the concept of impedance (Sect. 1.4.3, p. 78).

Without the concept of time and frequency, one cannot develop an intuition for the Fourier and Laplace transform relationships, especially within the context of engineering and mathematical physics.

WEEK 2

1.2.3 Lec 4: Prime numbers

If someone came up to you and asked for a theory of counting numbers, I suspect you would look them in the eye with a blank stare, and start counting. It sounds like either a stupid question or a bad joke. Yet integers are rich topic, so the question is not even slightly dumb. It is somewhat amazing that even birds and bees can count. While I doubt birds and bees can recognize primes, cicadas and other insects only crawl out of the ground in multiples of prime years, (e.g., 13 or 17 year cycles). If you have ever witnessed such an event (I have), you will never forget it. Somehow they know. Finally, there is an analytic function, first introduced by Euler, based on his analysis of the sieve, now known as the *Riemann zeta function* $\zeta(s)$, which is complex analytic, with its poles at the logs of the prime numbers. The exact relationship between the primes and the poles will be discussed in Sections 1.4.12 (p. 91) and **??** (p. **??**). The properties of this function are truly amazing, even fun.[33] It follows that primes are a subset of the counting numbers (Section 1.2.1), that the theory of numbers (and primes) is an important topic of study. Many of the questions, and some of the answers, go back to at least the time of the Chinese (Stillwell, 2010).

---

[32]It is not clear what it takes to reach this more official sounding category.

[33]The Riemann zeta function is known as the million dollar equation as there is a cash reward for a proof of the Riemann Hypothesis.

The importance of prime numbers

Likely the first insight into the counting numbers started with the *sieve*, shown in Fig. 1.5. A sieve answers the question "How can one list the prime numbers?" The answer comes from looking for irregular patterns in the counting numbers, by playing the counting numbers against themselves.

5   A prime is that subset of positive integers $\mathbb{P} \in \mathbb{N}$ that cannot be factored. The number 1 is not a prime, for some non-obvious reasons, but there is no pattern in it since it is always a (useless) factor of every counting number.

Sieves

A recursive sieve method for finding primes was first devised by the Greek Eratosthenes,[34] and sum-
10   marized in Fig. 1.5.

1. Write $N - 1$ counting number from 2 to N (<u>List</u>)

2. Define loop index $k = 1$ and a <u>multiplier</u> $n \in \mathbb{N}$ denoted $n := \{2, \cdots, N\}$.

3. The next number on the list is prime $\pi_k \in \mathbb{P}$

4. Remove (Cross out) all multiples $n \cdot \pi_n$ of π_k

15   5. $k = k + 1$: return to step 3.

Starting from the first prime $\pi_1 = 2$, one successively strikes out all the multiples of that prime. For example, starting from $\pi_1 = 2$ one strikes out $2 \cdot 2, 2 \cdot 3, 2 \cdot 4, 2 \cdot 5, \cdots, 2 \cdot (N/2)$. By definition the multiples are products of the target prime (2 in our example) and every another integer ($n \geq 2$). In this way all the even numbers are removed in this first iteration. The next remaining integer (3 in our
20   example) is identified as the next (second) prime π_2. Then all the $(N - 2)/2$ multiples of $\pi_2 = 3$ are removed. The next remaining number is $\pi_3 = 5$, so all multiples of $\pi_3 = 5$ are removed (i.e., ~~10~~, ~~15~~, ~~25~~ etc., $\cdots$). This process is repeated until all the numbers of the list have been either canceled, or identified as prime.

As the word sieve implies, this sifting process takes a heavy toll on the integers, rapidly pruning
25   the non-primes. In four iterations of the sieve algorithm, all the primes below $N = 50$ are identified in **red**. The final set of primes are displayed in the caption of Fig. 1.5.

Once a prime greater than $N/2$ has been identified, we may stop, since twice that prime is greater than N, the maximum number under consideration. Once you have reached $\sqrt{N}$ all the primes have been identified (this follows from the fact that the next prime π_n is multiplied by an integer $n = 1, \dots N$).
30   Once this number $n\pi_n > N$ the list has been exhausted, which must be $n < \sqrt{N}$.

When using a computer, memory efficiency and speed are the main considerations. There are various schemes for making the sieve more efficient. For example the recursion $n\pi_k = (n-1)\pi_k + \pi_k$ will speed up the process, by replacing the multiply with an addition of π_k.

Two fundamental theorems of primes: Early theories of numbers revealed two fundamental theorems (there are many more than two), as discussed in Section 1.2.2 and 2.1.1 (p. 117). The first of these is the *fundamental theorem of arithmetic*, which says that every integer greater than 1 may be uniquely factored into a product of primes

$$n = \prod_{k=1}^{K} \pi_k^{\beta_k}, \tag{1.3}$$

where $k = 1, \dots, K$ indexes the integer's K *prime factors* $\pi_k \in \mathbb{P}$. Typically prime factors appear
35   more than once, for example $25 = 5^2$. To make the notation compact we define the *multiplicity* β_k

---

[34]https://en.wikipedia.org/wiki/Sieve_of_Eratosthenes\#Euler.27s_Sieve

1. Write N integers from 2 to $N-1$. Let $k=1$. The first element $\pi_1 = 2$ is a prime. Cross out $n \cdot \pi_n$: (e.g., $n \cdot \pi_1 = 4, 8, 16, 32, \cdots$).

$$
\begin{array}{cccccccccc}
\mathbf{2} & 3 & \not{4} & 5 & \not{6} & 7 & \not{8} & 9 & \not{10} \\
11 & \not{12} & 13 & \not{14} & 15 & \not{16} & 17 & \not{18} & 19 & \not{20} \\
21 & \not{22} & 23 & \not{24} & 25 & \not{26} & 27 & \not{28} & 29 & \not{30} \\
31 & \not{32} & 33 & \not{34} & 35 & \not{36} & 37 & \not{38} & 39 & \not{40} \\
41 & \not{42} & 43 & \not{44} & 45 & \not{46} & 47 & \not{48} & 49 & \not{50}
\end{array}
$$

2. Let $k=2$, $\pi_2 = 3$. Cross out $n\pi_k$ $(6, 9, 12, 15, \ldots)$:

$$
\begin{array}{cccccccccc}
\mathbf{2} & \mathbf{3} & \not{4} & 5 & \not{6} & 7 & \not{8} & \not{9} & \not{10} \\
11 & \not{12} & 13 & \not{14} & \not{15} & \not{16} & 17 & \not{18} & 19 & \not{20} \\
\not{21} & \not{22} & 23 & \not{24} & 25 & \not{26} & \not{27} & \not{28} & 29 & \not{30} \\
31 & \not{32} & \not{33} & \not{34} & 35 & \not{36} & 37 & \not{38} & \not{39} & \not{40} \\
41 & \not{42} & 43 & \not{44} & \not{45} & \not{46} & 47 & \not{48} & 49 & \not{50}
\end{array}
$$

3. Let $k=3, \pi_3 = 5$. cross out $n\pi_3$. (Cross out 25, 35).

$$
\begin{array}{cccccccccc}
\mathbf{2} & \mathbf{3} & \not{4} & \mathbf{5} & \not{6} & \mathbf{7} & \not{8} & \not{9} & \not{10} \\
\mathbf{11} & \not{12} & \mathbf{13} & \not{14} & \not{15} & \not{16} & \mathbf{17} & \not{18} & \mathbf{19} & \not{20} \\
\not{21} & \not{22} & \mathbf{23} & \not{24} & \not{25} & \not{26} & \not{27} & \not{28} & \mathbf{29} & \not{30} \\
\mathbf{31} & \not{32} & \not{33} & \not{34} & \not{35} & \not{36} & \mathbf{37} & \not{38} & \not{39} & \not{40} \\
\mathbf{41} & \not{42} & \mathbf{43} & \not{44} & \not{45} & \not{46} & \mathbf{47} & \not{48} & \not{49} & \not{50}
\end{array}
$$

4. Finally let $k=4, \pi_4 = 7$. Remove $n\pi_4$: (Cross out 49). Thus there are 15 primes less than $N = 50$: $\pi_k = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47\}$.

Figure 1.5: *Sieve of Eratosthenes for the case of $N = 50$.*

of each prime factor π_k. For example $2312 = 2^3 \cdot 17^2 = \pi_1^3 \, \pi_7^2$ (i.e., $\pi_1 = 2, \beta_1 = 3; \pi_7 = 17, \beta_7 = 2$) and $2313 = 3^2 \cdot 257 = \pi_3^2 \, \pi_{55}$ (i.e., $\pi_2 = 3, \beta_3 = 2; \pi_{55} = 257, \beta_{55} = 1$). Our demonstration of this is empirical, using the Matlab/Octave `factor(N)` routine, which factors N.[35]

What seems amazing is the unique nature of this theorem. Each counting number is *uniquely* represented as a product of primes. There cannot be two integers with the same factorization. Once you multiply the factors out, the result is unique (N). Note that it's easy to multiply integers (e.g., primes), but expensive to factor them. And factoring the product of three primes is significantly more difficult than factoring two.

Factoring is much more expensive that division. This is not due to the difference in the cost of division over multiplication. The division of one number by another is only slightly more expensive than multiplying two numbers, maybe by a factor of 2.[36] Dividing the product of two primes, given one, is trivial, slightly more expensive that multiplying. Factoring the product of two primes is nearly impossible, as one needs to know what to divide by. This is an explosion in probability, measured by *Shannon Entropy*. Factoring means dividing by some integer and obtaining another integer *with remainder zero*. Thus one could factor factor a product of primes $N = \pi_k\pi_l$ by doing M divisions, where M is the number of primes less than N. This assumes the list of primes less than N are known.

But the utility has to do with the density of primes (the prime number theorem, i.e., Gauss'

---

[35]If you wish to be a Mathematician, you need to learn how to prove theorems. If you're an Engineer, you are happy that someone else has already proved them, so that you can use the result.

[36]https://streamcomputing.eu/blog/2012-07-16/how-expensive-is-an-operation-on-a-cpu/

hypothesis). If we were simply looking up a few numbers from a short list of primes, it would be easy, but the density of primes among the integers, is logarithmic (>1 per octave, Section 2.1.1, p. 118).

This take us to the *Prime Number Theorem* (PNT). The security problem is the reason why these two theorems are so important: 1) Every integer has a unique representation as a product of primes, and 2) the number of primes is very dense (their are a very large number of them, the density is proportional to the log of their number). Security reduces to the "needle in the haystack problem," the cost of a search. A more formal way to measure the density is known as *Shannon entropy*, couched in terms of the expected value of the log-probability of events "What is the probability of finding a prime between N and $2N$?"[37]

1.2.4   Lec 5: Greatest common divisor (Euclidean algorithm)

The *Euclidean algorithm* is a method to find the greatest common divisor (GCD) k between two integers n, m, denoted $k = gcd(n, m)$, where $n, m, k \in \mathbb{N}$. For example $15 = gcd(30, 105)$ since when factored $(30, 105) = (2 \cdot 3 \cdot 5, 7 \cdot 3 \cdot 5) = 3 \cdot 5 \cdot (2, 7) = 15 \cdot (2, 7)$. The Euclidean algorithm was known to the Chinese (i.e., not discovered by Euclid) (Stillwell, 2010, p. 41).

Why is the GCD important?   Computing the GCD is simple, whereas a full factoring is extremely expensive. The GCD is important, precisely because of the fundamental difficulty of factoring large integers into their primes. This utility surfaces when the two numbers are composed of very large primes. When two integers have no common factors they are said to be *coprime*, thus their GCD is 1. The ratio of two integers which are coprime is automatically in *reduced form* (they have no common factors).

For example $4/2 \in \mathbb{Q}$ is not reduced since 2=gcd(4,2). Canceling out the common factor 2, gives the reduced form $2/1=2 \in \mathbb{N}$. Thus if we wish to form the ratio of two integers, first compute the gcd, and remove it from the two numbers, to form the ratio. This assures the rational number is in its reduced form. If the GCD were 10^3 digits it is obvious that the common factor must be removed to greatly simplify further computation.

An example: Take the two integers $[873, 582]$. In factored form these are $[\pi_{25} \cdot 3^2, \pi_{25} \cdot 3 \cdot 2]$. Given the factors, we see that the largest common factor is $\pi_{25} \cdot 3 = 291$ ($\pi_{25} = 97$). When we take the ratio of the two numbers this common factor cancels

$$\frac{873}{582} = \frac{\pi_{25} \cdot \cancel{3} \cdot 3}{\pi_{25} \cdot \cancel{3} \cdot 2} = \frac{3}{2} = 1.5.$$

Of course if we divide 582 into 873 this we will numerically obtain the answer $1.5 \in \mathbb{F}$.

If the common factor is large (π_{25} in this example), a floating point number in $\mathbb{F}$ is returned, since all floating point numbers are in $\mathbb{F}$. But due to rounding errors, it may not be $3/2$. As an example, in Matlab/Octave `rat(3/2)=2+1/(-2)`, due to IEEE floating point rounding. One would expect `rat(3/2)=1+1/2`. To obtain the precise answer in $\mathbb{F}$, we need to remove the the GCD. Removing large common factors, without actually factoring the two numbers, has obvious practical utility. However this breaks down due to numerical issues. For example `rat(582,97)=` 582 (the wrong answer) yet `factor(582)=` 2*3*97 and `factor(873)` = 3*3*97 (the correct answers). If we remove factors 2 or 3, we get the correct answer: `rat(582/3,97)=` 194=2*97, `rat(582/2,97)=` 291 = 3*97. This is the nature of rounding error in the IEEE processor.

Euclidean algorithm: The algorithm is best explained by a trivial example: Let the two numbers be $6, 9$. At each step the smaller number (6) is subtracted from the larger (9) and the difference (the remainder) and the smaller numbers are saved. This process continues until the two resulting numbers are equal, at which point the GCD equals that final number. If we were to take one more step, the final numbers would be the gcd and zero. For our example step 1 gives 9-6=3, leaving 6 and

---

[37]When I understand this better, I'll do a better job of explaining it.

Greatest Common Divisor: k=gcd(m,n)

- *Examples ($m, n, k \in \mathbb{Z}$):*
  - *gcd(13*5,11*5) = 5 (The common 5 is the gcd)*
  - *gcd(13*10,11*10) = 10 (The gcd(130,110) = 10 = 2*5, is not prime)*
  - *gcd(1234,1024) = 2 (1234=2*617, 1024=2^{10})*
  - *gcd($\pi_k \pi_m$, $\pi_k \pi_n$) = π_k*
  - *k=gcd(m,n) is the part that cancels in the fraction $m/n \in F$*
  - *$m/\gcd(m,n) \in \mathbb{Z}$*

- *Coprimes ($m \perp n$) are numbers with no common factors: i.e., gcd(m,n)=1*
  - *The gcd of two primes is always 1: gcd(13,11) = 1, gcd(π_m, π_n)=1 ($m \neq n$)*
  - *$m = 7 \cdot 13, n = 5 \cdot 19 \Rightarrow (7 \cdot 13) \perp (5 \cdot 19)$*
  - *If $m \perp n$ then $\gcd(m,n) = 1$*
  - *If $\gcd(m,n) = 1$ then $m \perp n$*

- *The GCD may be extended to polynomials: e.g., $gcd(ax^2 + bx + c, \alpha x^2 + \beta x + \gamma)$*
  - *gcd($(x-3)(x-4)$, $(x-3)(x-5)$)= $(x-3)$*
  - *gcd($x^2 - 7x + 12$, $3(x^2 - 8x + 15)$)= $3(x-3)$*
  - *gcd($x^2 - 7x + 12$, $(3x^2 - 24x + 45)$)=$3(x-3)$*
  - *gcd($(x - 2\pi)(x-4)$, $(x - 2\pi)(x-5)$)=$(x - 2\pi)$ (Needs long division)*

Figure 1.6: *The Euclidean algorithm for finding the GCD of two numbers is one of the oldest algorithms in mathematics, and is highly relevant today. It is both powerful and simple. It was used by the Chinese during the Han dynasty (Stillwell, 2010, p. 70) for reducing fractions. It may be used to find pairs of integers that are coprime (their gcd must be 1), and it may be used to identify factors of polynomials by long division. It has an important sister algorithm called the continued fraction algorithm (CFA), that is so similar in concept that Gauss referred to the Euclidean algorithm as the"continued fraction algorithm" (Stillwell, 2010, p. 48).*

3. Step 2 gives 6-3=3 and 3. Since the two numbers are the same, the GCD=3. If we take one more difference we obtain (3,0). We can easily verify this result since this example is easily factored (e.g., $3 \cdot 3, 3 \cdot 2) = 3(3, 2)$. It may be numerically verified using the Matlab/Octave GCD command gcd(6,9), which returns 3.

Matrix method: The GCD may be written a matrix recursion given the starting vector $(m_0, n_0)^T$. The recursion is then

$$\begin{bmatrix} m_{k+1} \\ n_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} m_k \\ n_k \end{bmatrix}$$

This recursion continues until $m_{k+1} < n_{k+1}$, at which point m and n must be swapped. Because the operation depends on the output, this is a nonlinear recursion (Postulate P1 (Linear/nonlinear) of Section 3.5.1, p. 145).

The direct method is inefficient because in recursively subtract n many times until the resulting m is less than n, as shown in Fig. 2.1 (p. 121). It also must test for $m < n$ at each iteration, and then swap m and n once that condition is met. This recursion is repeated until $m_{k+1} = 0$. At that stage the GCD is n_{k+1}. Figure 2.1, along with the above matrix relation, give the best insight into the Euclidean Algorithm, but at the cost of low efficiency.

Below is a Matlab/Octave code to find k=gcd(m,n) based on the strict definition of the EA as described by Stillwell (2010):

```
function k = gcd(m,n)
 while m ~=0
  A=m; B=n;
  m=max(A,B); n=min(A,B); %m>n
  m=m-n;
```

```
end
```

This program keeps looping until $m = 0$. An much more efficient method is described in Section 2.1.3, p. 120, using the `floor()` function, which we call *division with rounding.*

1.2.5   Lec 6: Continued fraction algorithm

5   In its simplest form, the *Continued fraction algorithm* (CFA) starts from a single real decimal number $x \in \mathbb{R}$, and recursively expands it as a fraction $x \in \mathbb{F}$. Thus the CFA It is used in finding rational approximations to any real number. For example, $\pi \approx 22/7$. The GCD (the Euclidean algorithm) on the other hand operates on a pair of integers $m, n \in \mathbb{N}$ and finds their greatest common divisor $k \in \mathbb{N}$. Thus $m/k, n/k \in \mathbb{N}$, reducing the ratio to its irreducible form, since $m/k \perp n/k$.

10   Despite this seeming large difference in the two algorithms, apparently the CFA is closely related to the Euclidean algorithm (the GCD), so closely in fact, that Gauss referred to the Euclidean algorithm as the Continued fraction algorithm (Stillwell, 2010, P. 48). At first glance it is not clear why Gauss would call the CFA the Euclidean algorithm. One must assume that Gauss had some deeper insight into this relationship. If so, that insight would be valuable to understand. [38]

15   In the following we refine the description of the CFA and give examples that go beyond the simple cases of expanding numbers. The CFA of any positive number, say $x_o \in \mathbb{R}^+$, is defined as follows:

1. Start with $n = 0$ and a positive input target $x_o \in \mathbb{R}^+$.

2. Define $a_n = \texttt{round}(x_n)$, which rounds to the nearest integer.

3. Define $r_n = x_n - a_n$, thus $-0.5 \leq r_n \leq 0.5$. If $r_n = 0$, the recursion terminates.

20   4. Define $x_{n+1} \equiv 1/r_n$ and return to step 2, with $n = n + 1$.

The recursion may continue to any desired accuracy, since convergence is guaranteed.

An example: Let $x_o \equiv \pi \approx 3.14159\ldots$. Thus $a_o = 3$, $r_o = 0.14159$, $x_1 = 7.065 \approx 1/r_o$, and $a_1 = 7$. If we were to stop here we would have

$$\widehat{\pi}_2 \approx 3 + \frac{1}{7 + 0.0625\ldots} \approx 3 + \frac{1}{7} = \frac{22}{7}. \tag{1.4}$$

This approximation of $\pi \approx 22/7$ has a relative error of 0.04%

$$\frac{22/7 - \pi}{\pi} = 4 \times 10^{-4}.$$

For the next approximation we continue by reciprocating the remainder $1/0.0625 \approx 15.9966$ which rounds to 16 giving a negative remainder of $\approx -1/300$, resulting in the second approximation

$$\widehat{\pi}_3 \approx 3 + 1/(7 + 1/16) = 3 + 16/(7 \cdot 16 + 1) = 3 + 16/113 = 355/113.$$

Note that if we had truncated 15.9966 to 15, the remainder would have been much larger, but positive, resulting in a much less accurate rational approximation. It follows that there can be a dramatic difference, depending on the rounding scheme, which must be specified.

Notation: Writing out all the fractions can become tedious. For example, expanding $e = 2.7183$ using the Matlab/Octave command `rat(exp(1))` gives the approximation

$$\hat{e}_6 = 3 + 1/(-4 + 1/(2 + 1/(5 + 1/(-2 + 1/(-7))))) - O\left(1.75 \times 10^{-6}\right).$$

25   A compact notation for this these coefficients of the CFA is $\hat{\pi}_6 = [3., -4, 2, 5, -2, -7]$.[39] Note that the leading integer part may be noted by an optional decimal point.[40] If the process is carried further, the values of $a_n \in \mathbb{N}$ give increasingly more accurate rational approximations.

---

[38]The resolution of this interrelationship is still unresolved.

[39]$\pi - 3 = [-4, 2, 5, \cdots]$.

[40]Unfortunately Matlab/Octave does not support the bracket notation.

Rational approximation examples

$$\hat{\pi}_2 = \frac{22}{7} = [3., 7] \qquad\qquad\qquad \approx \hat{\pi}_2 + O(1.3 \times 10^{-3})$$

$$\hat{\pi}_3 = \frac{355}{113} = [3., 7, 16] \qquad\qquad \approx \hat{\pi}_3 + O(2.7 \times 10^{-7})$$

$$\hat{\pi}_4 = \frac{104348}{33215} = [3., 7, 16, -249] \qquad \approx \hat{\pi}_4 + O(3.3 \times 10^{-10})$$

Figure 1.7: The expansion of π to various orders, using the CFA, along with the order of the error of each rational approximation. For example $\hat{\pi}_2 = 22/7$ has an absolute error ($|22/7 - \pi|$) of about 0.13%.

Rounding schemes: In Matlab/Octave there are four different rounding function: `round()`, `fix()`, `floor()`, `ceil()`. If the rounding-down (`floor`) is used $\hat{\pi}_{12} = [3., 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1]$ whereas true rounding to the nearest integer (`round`) gives $\hat{\pi}_8 = [3., 7, 16, -294, 3, -4, 5, -15]$. Thus `round` introduces negative coefficients each time a number rounds up to the nearest integer.

Exercise: Based on several examples, which rounding scheme is the most accurate? Explain why.   5
Sol: Rounding will give a smaller remainder at each iteration, resulting in a smaller net error.

When the CFA is applied and the expansion terminates ($r_n = 0$), the target is rational. When the expansion does not terminate (which is not always easy to determine), the number is irrational. Thus the CFA has important theoretical applications regarding irrational numbers. You may try this yourself using Matlab's `rats(pi)` command.   10

One of the useful things about the procedure, besides its being so simple, are its generalizations, one of which will be discussed in Section 2.1.3 (p. 120).

Symmetry: A *continued fraction expansion* can have a high degree of symmetry. For example, the CFA of

$$R_1 \equiv \frac{1 + \sqrt{5}}{2} = 1 + \cfrac{1}{1 + \cfrac{1}{1+\cdots}} = 1.618033988749895\cdots, \qquad (1.5)$$

Here a_n in the CFA is *always* 1 ($R_1 \equiv [1., 1, 1, \cdots]$), thus the sequence will not terminate, *proving* that $\sqrt{5} \in \mathbb{I}$. A related example is $R_2 \equiv$ `rat(1+sqrt(2))`, which gives $R_2 = [2., 2, 2, 2, \cdots]$.

When expanding a target irrational number ($x_o \in \mathbb{I}$), and the CFA is truncated, the resulting rational fraction approximates the irrational target. For the example above, if we truncate at three coefficients ([1; 1, 1]) we obtain

$$1 + \cfrac{1}{1 + \cfrac{1}{1+0}} = 1 + 1/2 = 3/2 = 1.5 = \frac{1 + \sqrt{5}}{2} + 0.118\cdots.$$

Truncation after six steps gives

$$[1., 1, 1, 1, 1, 1, 1] = 13/8 \approx 1.6250 = \frac{1 + \sqrt{5}}{2} + .0070\cdots.$$

Because all the coefficients are 1, this example converges very slowly, When the coefficients are large   15
(i.e., remainder small), the convergence will be faster. The expansion of π is an example of faster convergence.

In summary: Every rational number $m/n \in \mathbb{F}$, with $m > n > 1$, may be uniquely expanded as a continued fraction, with coefficients a_k determined using the CFA. When the target number is irrational $(x_o \in \mathbb{Q})$, the CFA does not terminate, thus each step produces a more accurate rational approximation, converging in the limit as $n \to \infty$.

Thus the CFA expansion is an algorithm that can, in theory, determine when the target is rational, but with an important caveat: one must determine if the expansion terminates. This may not be obvious. The fraction $1/3 = 0.33333\cdots$ is an example of such a target, where the CFA terminates yet the fraction repeats. It must be that

$$1/3 = 3 \times 10^{-1} + 3 \times 10^{-2} + 3 \times 10^{-3} + \cdots.$$

Here 3*3=9. As a second example[41]

$$1/7 = 0.142857142857142857142857\cdots = 142857 \times 10^{-6} + 142857 \times 10^{-12} + \cdots$$

Note that $142857 * 7 = 999999$. This also works for $1/11 = 0.090909\cdots$ and $11 \times 0.090909 = 999999$. We might conclude that when the sequence of digits repeats, the sequence is predictable, thus it must be rational.

WEEK 3

1.2.6   Lec 7: Pythagorean triplets (Euclid's formula)

Euclid's formula is a method for finding three integer lengths $[a, b, c] \in \mathbb{N}$, that satisfy Eq. 1.1. It is important to ask "Which set are the lengths [a,b,c] drawn from?" There is a huge difference, both practical and theoretical, if they are from the real numbers $\mathbb{R}$, or the counting numbers $\mathbb{N}$. Given $p > q \in \mathbb{N}$, the three lengths $[a, b, c] \in \mathbb{N}$ of Eq. 1.1 are given by

$$a = p^2 - q^2, \qquad b = 2pq, \qquad c = p^2 + q^2. \tag{1.6}$$

This result may be directly verified, since

$$[p^2 + q^2]^2 = [p^2 - q^2]^2 + [2pq]^2$$

or

$$p^4 + q^4 + 2p^2q^2 = p^4 + q^4 - 2p^2q^2 + 4p^2q^2.$$

Thus, Eq. 1.6 is easily proven, once given. Deriving Euclid's formula is obviously more difficult.
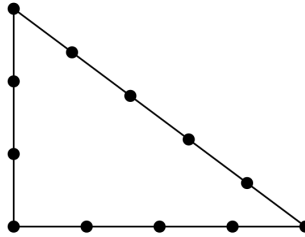
Figure 1.8: *Beads on a string form perfect right triangles when number of beads on each side satisfy Eq. 1.1. For $p = 2, q = 1$ the sides are $[3, 4, 5]$.*

---

[41]Taking the Fourier transform of the target number, represented as a sequence, could help to identify an underlying periodic component. The number $1/7 \leftrightarrow [[1, 4, 2, 8, 5, 7]]_6$ has a 50 [dB] notch at 0.8π [rad] due to its 6 digit periodicity, carried to 15 digits (Matlab/Octave maximum precision), Hamming windowed, and zero padded to 1024 samples.

A well known example is the right triangle depicted in Fig. 1.8, defined by the integers $[3, 4, 5] \in N$, having angles $[0.54, 0.65, \pi/2]$ [rad], which satisfies Eq. 1.1 (p. 15). As quantified by Euclid's formula (Eq. 1.6), there are an infinite number of *Pythagorean triplets* (PTs). Furthermore the seemingly simple triangle, having angles of $[30, 60, 90] \in N$ [deg] (i.e., $[\pi/6, \pi/3, \pi/2] \in \mathbb{I}$ [rad]), has one irrational ($\mathbb{I}$) length ($[1, \sqrt{3}, 2]$).

a	c
119	169
3367	4825
4601	6649
12709	18541
65	97
319	481
2291	3541
799	1249
481	769
4961	8161
45	75
1679	2929
161	289
1771	3229
56	106

Figure 1.3: Pairs in Plimpton 322

Figure 1.9: *"Plimpton-322" is a stone tablet from 1800 [BCE], displaying a and c values of the Pythagorean triplets $[a, b, c]$, with the property $b = \sqrt{c^2 - a^2} \in \mathbb{N}$. Several of the c values are primes, but not the a values. The stone is item 322 (item 3 from 1922), from the collection of George A. Plimpton. –Stillwell (2010)*

The technique for proving Euclid's formula for PTs $[a, b, c] \in \mathbb{Q}$, derived in Fig. 2.3 (p. 123) of Section 2.1.4, is much more interesting than the PTs themselves.

The set from which the lengths $[a, b, c]$ are drawn was not missed by the Indians, Chinese, Egyptians, Mesopotamians, Greeks, etc. Any equation whose solution is based on integers is called a *Diophantine equation*, named after the Greek mathematician Diophantus of Alexandria (c250 CE).

A stone tablet having the numbers engraved on it, as shown in Table 1.9, was discovered in Mesopotamia, from the 19th century [BCE], and cataloged in 1922 by George Plimpton.[42] These numbers are a and c pairs from PTs [a,b,c]. Given this discovery, it is clear that the Pythagoreans were following those, long before them. Recently a second similar stone, dating between 350 and 50 [BCE] has been reported, that indicates early calculus on the orbit of Jupiter's moons.[43]

1.2.7 Lec 8: Pell's Equation

Pell's equation

$$x^2 - Ny^2 = 1, \tag{1.7}$$

with non-square $N \in \mathbb{N}$ specified and $a, b \in \mathbb{N}$ unknown, is related to the Euclidean algorithm (Stillwell, 2010, 48). For example, with $N = 2$, one solution is $a = 17, b = 12$ ($17^2 - 2 \cdot 12^2 = 1$). This equation has a long history (Stillwell, 2010).

---

[42]http://www.nytimes.com/2010/11/27/arts/design/27tablets.html;https://en.wikipedia.org/wiki/Plimpton_322

[43]http://www.nytimes.com/2016/01/29/science/babylonians-clay-tablets-geometry-astronomy-jupiter.html

A 2x2 matrix recursion algorithmic was used by the Pythagoreans to investigate the $\sqrt{2}$

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_{n-1} \\ y_{n-1} \end{bmatrix}. \tag{1.8}$$

Starting with $[x_o, y_o]^T = [1, 0]^T$, results in solutions of Pell's equations (Stillwell, 2010, p. 44). Their approach was likely motivated by the Euclidean algorithm (GCD, p. 34), since $y_n/x_n \to \sqrt{2}$ (Stillwell, 2010, p. 37, 55). Note that this is a composition method, of 2x2 matrices, since the output of one matrix multiply is the input to the next.

5   **Asian solutions:**   The first solution of Pell's equation was published by Brahmagupta (c628), who independently discovered the equation (Stillwell, 2010, p. 46). Brahmagupta's novel solution introduced a different *composition method* (Stillwell, 2010, p. 69), and like the Greek result, these solutions were incomplete.

Then in 1150CE, Bhâskara II obtained solutions using Eq. 1.8 (Stillwell, 2010, p.69). This is the
10   solution method we shall explore here, as summarized in Fig. 1.10.

The best way to see how this recursion results in solutions to Pell's equation, is by example. Initializing the recursion with the trivial solution $[x_o, y_o]^T = [1, 0]^T$, gives

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad\qquad 1^2 - 2 \cdot 1^2 = -1$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad\qquad 3^2 - 2 \cdot 2^2 = 1$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} \qquad\qquad (7)^2 - 2 \cdot (5)^2 = -1$$

$$\begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} 17 \\ 12 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 5 \end{bmatrix} \qquad\qquad 17^2 - 2 \cdot 12^2 = 1$$

$$\begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \begin{bmatrix} 41 \\ 29 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 17 \\ 12 \end{bmatrix} \qquad\qquad (41)^2 - 2 \cdot (29)^2 = -1$$

Thus the recursion results in a modified version of Pell's equation

$$x_n^2 - 2y_n^2 = (-1)^n,$$

where only even values of n are solutions. This sign change had no effect on the Pythagorean goal, since they only cared about $y_n/x_n \to \sqrt{2}$.

Solution to Pell's equation: By multiplying the matrix by $1\jmath$, all the solutions to Pell's equation are determined. This solution is shown in Fig. 1.10 (p. 41) for the case of $N = 2$, and again for Eq. D.1
15   (p. 179), for $N = 3$. The math is straightforward and is easily verified using Matlab/Octave. From Fig. 1.10 we can see that every output this slightly modified matrix recursion gives solutions to Pell's equation (Eq. 1.7). The $1\jmath$ factor corrects the alternation in sign, so every iteration yields a solution. For $n = 0$ (the initial solution) $[x_o, y_o]$ is [1,0], $[x_1, y_1] = \jmath[1, 1]$, and $[x_2, y_2] = -[3, 2]$. These are easily computed by this recursion, and easily checked on a hand calculator (or using Matlab/Octave).
20   At each iteration, the ratio x_n/y_n approaches $\sqrt{2}$ with increasing accuracy, coupling it to the CFA. The value of $41/29 \approx \sqrt{2}$, with a relative error of $<0.03\%$. The solution for $N = 3$ is discussed in Eq. D.1 (p. 179).

- Case of $N = 2$ & $[x_o, y_o]^T = [1, 0]^T$
  Note: $x_n^2 - 2y_n^2 = 1, \qquad x_n/y_n \underset{\infty}{\longrightarrow} \sqrt{2}$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \jmath \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \jmath \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad\qquad \jmath^2 - 2 \cdot \jmath^2 = 1$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \jmath^2 \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \jmath \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \jmath \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad\qquad 3^2 - 2 \cdot 2^2 = 1$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \jmath^3 \begin{bmatrix} 7 \\ 5 \end{bmatrix} = \jmath \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \jmath^2 \begin{bmatrix} 3 \\ 2 \end{bmatrix} \qquad\qquad (7\jmath)^2 - 2 \cdot (5\jmath)^2 = 1$$

$$\begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} 17 \\ 12 \end{bmatrix} = \jmath \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \jmath^3 \begin{bmatrix} 7 \\ 5 \end{bmatrix} \qquad\qquad 17^2 - 2 \cdot 12^2 = 1$$

$$\begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \jmath \begin{bmatrix} 41 \\ 29 \end{bmatrix} = \jmath \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 17 \\ 12 \end{bmatrix} \qquad\qquad (41\jmath)^2 - 2 \cdot (29\jmath)^2 = 1$$

Figure 1.10: *This summarizes the solution of Pell's equation for $N = 2$ using a slightly modified matrix recursion. Note that $x_n/y_n \to \sqrt{2}$ as $n \to \infty$, which was what the Pythagoreans were pursuing.*

Relations to digital signal processing: Today we recognize Eq. 1.8 as a *difference equation,* which is a pre-limit (pre Stream 3) form of differential equation. The Greek 2x2 form is an early precursor to 17th and 18th century developments in linear algebra. Thus the Greek's recursive solution for the $\sqrt{2}$ and Bhâskara's (1030 CE) solution of Pell's equation, is an early precursor to discrete-time processing, as well as to calculus.

There are similarities between Pell's Equation and the Pythagorean theorem. As we shall see in Chapter 2, Pell's equation is related to the geometry of a hyperbola, just as the Pythagorean equation is related to the geometry of a circle. One might wonder if there is a Euclidean formula for the solutions of Pell's Equations, since these are all conic sections with closely related geometry.

WEEK 4

1.2.8 Lec 9: Fibonacci sequence

Another classic problem, formulated by the Chinese, was the Fibonacci sequence, generated by the relation

$$f_{n+1} = f_n + f_{n-1}. \tag{1.9}$$

Here the next number f_{n+1} is the sum of the previous two. If we start from $[0, 1]$, this difference equation leads to the Fibonacci sequence $f_n = [0, 1, 1, 2, 3, 5, 8, 13, \ldots]$. Alternatively, if we define $y_{n+1} = x_n$, then Eq. 1.9 may be represented as the recursion of a 2x2 matrix equation

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix}. \tag{1.10}$$

The correspondence is easily verified. Starting with $[x_n, y_n]^T = [0, 1]^T$ we obtain for the first few steps

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \cdots$$

From the above $x_n = [0, 1, 1, 2, 3, 5, \ldots]$ is the Fibonacci sequence since the next x_n is the sum of the previous two and the next y_n is x_n.

The mean-Fibonacci sequence: Suppose that the Fibonacci sequence recursion is replace by the mean of the last two values, namely let

$$f_{n+1} = \frac{f_n + f_{n-1}}{2}. \tag{1.11}$$

This seems like a small change. But how does the solution differ? Two answer this question it is helpful to look at the corresponding 2x2 matrix.

Exercise: Find the 2x2 matrix corresponding to Eq. 1.11. **Sol:** In this case we find

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix}. \tag{1.12}$$

5   The eigenvalues of this matrix are the roots of the binomial equation $-\lambda(1 - \lambda/2) - 1/2 = 0$ which has eigenvalues $[1, -1/2]$. Thus $[x_n, y_n] = [1, 1]^T$ and $[x_n, y_n] = [1, 1]^T$ are each solutions

Exercise: Starting from $[x_n, y_n]^T = [1, 0]^T$ compute the first 5 values of $[x_n, y_n]^T$.

Exercise: Show that the solution to Eq. 1.11 is bounded, unlike that of the Fibonacci sequence, which diverges. Explain what is going on. **Sol:** By studying the eigen values of Eq. 1.12 one finds that
10   the steady state solution approaches 1. Namely $f_n \to 1 = (f_{n-1} + f_{n-2})/2$ is the solution, as $n \to \infty$. Namely the average of the last two values must approach 1 for large n. **Sol:** Use the formula for the generalized diagonalization of a matrix to find the general solution of the man-Fibonacci sequence.

Summary: The GCD (Euclidean algorithm), Pell's equation and the Fibonacci sequence may all be written as compositions of 2x2 matrices. This is an important and common thread of these early mathematical findings. Thus Pell's equation and the Fibonacci sequence are special cases of 2x2 matrix composition

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix}.$$

It will turn out that this form plays a special role in mathematics and engineering. The first several thousands of years of mathematical trial and error research, set the stage for major breakthroughs, but
15   it took a long time to fully appreciate these several contributions.

1.2.9   Lec 10: Exam I (In class)

1.3   Algebraic Equations: Stream 2

The era from 1600-1850 (Fig. 1.17, p. 75) produced a stream of fundamental theorems. A few of the individuals who played a notable role in this development, in chronological (birth) order, include
20   Galileo, Mersenne, Newton, d'Alembert, Fermat, Huygens, Descartes and Helmholtz. These individuals were some of the first to develop the basic ideas, in various forms, that were then later reworked into the proofs, that today we recognize as *The fundamental theorems of mathematics.*

1.3.1 Lec 11 Algebra and geometry as physics

Following Stillwell's history of mathematics, Stream 2 is geometry, which led to the merging of Euclid's geometrical methods and the 9th century development of algebra by al-Khwarizmi (830 CE). This integration of ideas lead Descartes and Fermat to develop *analytic geometry*. While not entirely a unique and novel idea, it was late in coming, given what was known at that time. 5

The mathematics up to the time of the Greeks, documented and formalized by Euclid, served students of mathematics for more than two thousand years. Algebra and geometry were, at first, independent lines of thought. When merged, the focus returned to the Pythagorean theorem, generalized as analytic conic sections rather than as geometry in Euclid's Elements. With the introduction of Algebra, numbers, rather than lines, could be used to represent a geometrical length. Thus the appreciation for 10 geometry grew, given the addition of the rigorous analysis using numbers.

Physics inspires algebraic mathematics: The Chinese used music, art, and navigation to drive mathematics. Unfortunately much of our knowledge has been handed down either as artifacts, such as musical bells and tools, or mathematical relationships documented, but not created by, scholars such as Euclid, Archimedes, Diophantus, and perhaps Brahmagupta. With the invention of algebra by al-Khwārizmī (830CE) this paradigm did not shift, rather it expanded. During the 16th and 17th century, it had becoming clear that differential equations (DEs), such as the wave equation, can characterize a law of nature, at a single point in space and time. This principle was not obvious. A desire to understand motions of objects and planets precipitated many new discoveries. This period is organized in Fig. 1.4. Galileo investigated gravity and invented the telescope. For example the law of gravity (first formulated by Galileo to explain the dropping to two objects of different masses) must obey conservation of energy. Kepler investigated the motion of the planets. While Kepler was the first to appreciate that the planets were described by ellipses, it seems he under-appreciate the significance of this finding, and continued with his epicycle model of the planets. Newton (c1687) went on to show that there must be a gravitational potential between to masses (m_1, m_2) of the form

$$\phi(r) \propto \frac{m_1 m_2}{r}, \tag{1.13}$$

where $r = |x_1 - x_2|$ is the Euclidean distance between the two point masses at locations x_1 and x_2. Note that this a power series, but with exponent of -1, which is a *pole* with an RoC of zero. Using algebra and his calculus, Newton formalized the equation of gravity, forces and motion (Newton's three laws) and showed that Kepler's discovery of planetary elliptical motion naturally follows from these 15 laws. With the discovery of Uranus (1781) "Kepler's theory was ruined." (Stillwell, 2010, p. 23).

Once Newton proposed the basic laws of gravity, he proceed to calculate, for the first time, the speed of sound. This required some form of the *wave equation*

$$\frac{\partial^2}{\partial x^2} p(x, t) = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} p(x, t), \tag{1.14}$$

a key equation in mathematical physics. The speed of sound is $c = 343$ [m/s], which is a function of the density $\rho_o = 1.12$ [kg/m^3] and the dynamic stiffness ηP_o of air.[44]

If we substitute for the pressure

$$p(x, t) = P_o(s) e^{\jmath 2\pi (ft \pm kx)}, \tag{1.15}$$

where t is time and x is position, we find that $k = 2\pi/\lambda = 2\pi f/c$, because $k = 2\pi/\lambda$.

This classic relation $\lambda f = c$ is deceivingly simple, thus confusing, because $k = 2\pi/\lambda$ becomes complex (has both real and imaginary parts) in dispersive media (e.g., acoustic waves in tubes when

---

[44]$c = \sqrt{\eta P_o/\rho_o}$, $\eta = c_p/c_v = 1.4$ is the ratio of two thermodynamic constants, and $P_o = 10^5$ [Pa] is the barometric pressure of air.

losses are considered (Kirchhoff, 1868a); electron waves in silicon crystals (Brillouin, 1953). In these more general cases, $k(f) = 2\pi f/c$ must be replaced with a complex analytic function $\kappa(s)$ of s, i.e.,

$$p(x,t) = P_o(s)e^{st}e^{\pm\kappa(s)x}, \tag{1.16}$$

such that the wave number $\kappa(s)$ is a complex analytic function of the Laplace frequency s. This is because electron "waves" in a dispersive medium (e.g., a semi-conductor, such as silicon), are "filtered" in their magnitude and phase. This filter acts as a transmission line, or a complex "wave-filter," forcing the wavelength to be a complex function of frequency. This view is elegantly explained by Brillouin (1953, Chap. 1), in an historical context. Acoustics provides a rich source of examples (Morse, 1948; Beranek, 1954; Beranek and Mellow, 2012).

⁵ Newton's *Principia* was finally published in 1687, and the general solution to Newton's wave equation [i.e., $p(x,t) = G(t \pm x/c)$], where G is *any* function, was first published 60 years later by d'Alembert (c1747), which showed that for sounds of a single frequency, the wavelength λ and frequency f were related by

$$f\lambda = c. \tag{1.17}$$

Today d'Alembert's analytic wave solution can be written as Eq. 1.15 with a real wave number $k = 1/\lambda$ [m^{-1}]. This formulation led to the frequency domain concept of Fourier analysis, based on the *linearity* (i.e., superposition) property of the wave equation (Postulate P2: Lec. 1.3.12, p. 72).

¹⁰ While Newton's value for the speed of sound in air c_o was incorrect by the thermodynamic constant $\sqrt{\eta}$, a problem that would take more than two hundred years to resolve, his success was important because it quantified the physics behind the speed of sound, and demonstrated that momentum mv, not mass m, was transported by the wave. His concept was correct, and his formulation using algebra and calculus represented a milestone in science. If the speed of sound in-vacuo is zero. In air, assuming

¹⁵ no visco-elastic losses, it is constant (i.e., $c_o = \sqrt{\eta\rho_o/\mathcal{P}_o}$). When including losses the wave number becomes a complex function of frequency, leading to Eq. 1.16. In periodic structures, again the wave number becomes complex due to diffraction, as commonly observed in optics (e.g., diffraction gratings) and acoustics. Thus Eq. 1.17 only holds for the most simple cases, but in general it must be considered as a complex analytic function of s, as $\kappa(s)$ in Eq. 1.16.

²⁰ The corresponding discovery for the formula for the speed of light was made 174 years after Principia, by Maxwell (c1861). Maxwell's formulation also required great ingenuity, as it was necessary to hypothesize an experimentally unmeasured term in his equations, to get the mathematics to correctly predict the speed of light.

The first Algebra: Prior to the invention of algebra, people worked out problems as sentences using an obtuse description of the problem (Stillwell, 2010, p. 93). Algebra solved this problem, resulting in a compact language of mathematics, where numbers are represented as abstract symbols (e.g., x and α). The problem to be solved could be formulated in terms of sums of powers of smaller terms, the most common being powers of some independent variable (i.e., time or frequency). Today we call such an expression a *polynomial* of *degree n*

$$P_n(z) \equiv z^n + a_{n-1}z^{n-1} + \cdots + a_o z^0 = \sum_{k=0}^{n} a_k z^k = \prod_{k=0}^{n}(z - z_k). \tag{1.18}$$

Here we have set $a_n = 1$. The coefficient a_n cannot be zero, or the polynomial would not be of degree

²⁵ n. A solution is to force $a_n = 1$, since this simplifies the expression, and does not change the roots.

The key question is "What values of $z = z_k$ result in $P_n(z_k) = 0$." In other words, what are the roots z_k of the polynomial? The answer to this question consumed thousands of years, with intense efforts by many aspiring mathematicians. In the earliest attempts, it was a competition to evaluate mathematical acumen. Results were held as a secret to the death bed. It would be fair to view this

³⁰ effort as an obsession. Today the roots of any polynomial may be found by numerical methods, to

Chronological history post 17th century

17th *Newton 1642-1727, Bernoulli, Johann 1667-1748*

18th *Bernoulli, Daniel, Cauchy 1789-1857, Euler 1748-83, d'Alembert 1717-83, Gauss 1777-1855*

19th *Kirchhoff 1824-87, Helmholtz 1821-1894, Riemann 1826-1866, Maxwell 1831-1879, Rayleigh 1842-1919, Heaviside 1850-1925, Poincare 1854-1912,*

20th *Sommerfeld 1686-1951, Einstein 1879-1955, Brillouin 1889-1969 . . .*
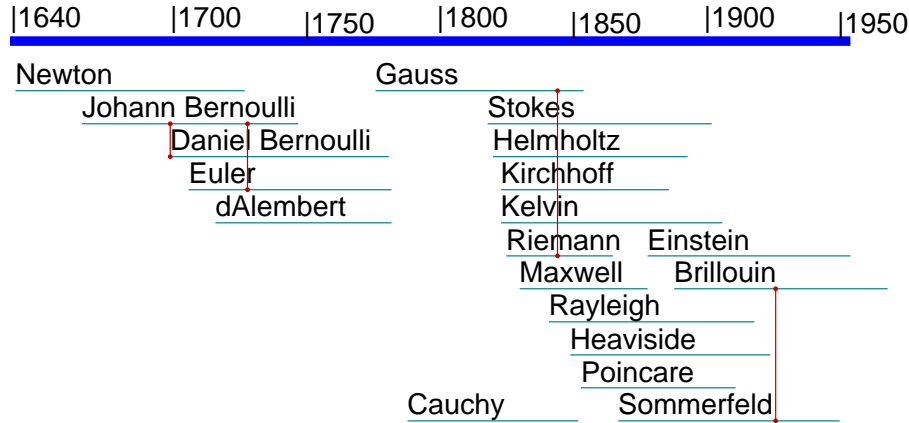
Time Line

Figure 1.11: *Time-line of the three centuries from the 18th to 20th CE. This was one of the most productive of all times, perhaps starting with the deep work of von Helmholtz, educated an experienced as a military surgeon, who mastered classical music, acoustics, physiology, vision, hearing (Helmholtz, 1863b), and, most important of all, mathematics. Kirchhoff frequently expanded on Helmholtz's contributions. It is reported that Lord Rayleigh learned German so he could read Helmholtz's great works. The history during this time is complex. For example, Lord Kelvin wrote a letter to Stokes, suggesting that Stokes try to prove what is today known as "Stokes theorem." As a result, Stokes posted a reward (the Smith Prize), searching for a prove of "Lord Kelvin's theorem," which was finally proved by Hankel (1839-73) (`https://en.wikipedia.org/wiki/Hermann_Hankel`). Many new concepts were being proved and appreciated over this productive period. In 1863-65, Maxwell published his famous equations, followed by a reformatting in modern vector notation by Heaviside, Gibbs and Hertz. The vertical red lines connect mentor-student relationships. This figure should put to rest the idea that ones best work is done in the early years. Many of these scientists were fully productive to the end of old age. Those that were not, died early, due to poor health or accidents. Figure 1.4 (p. 19) (Newton–Gauss) gives a closer look at the 15-18 CE, and Fig. 1.17 (p.75) (Bombelli–Einstein) for the full view from 16-20 CE. [fig:TimeLine19CE]*

very high accuracy. The limits to finding roots is limited by the numerical limits of the representation, namely by IEEE-754 (p. 26). There are also a number of important theorems.

Of particular interest is composing a circle with a line, when the line does not touch the circle, and finding the roots. There was no solution to this problem using geometry. This question is addressed in the assignments.

5

Finding roots of polynomials The problem of factoring polynomials has a history more than a millennium in the making. While the quadratic (degree $N = 2$) was solved by the time of the Babylonians (i.e., the earliest recorded history of mathematics), the cubic solution was finally published by Cardano in 1545. The same year, Cardano's student solved the quartic ($N = 4$). In 1826 (281 years later) it was proved that the quintic ($N = 5$) could not be factored by analytic methods.

10

As a concrete example we begin with trivial but important case of the quadratic

$$P_2(x) = ax^2 + bx + c. \qquad (1.19)$$

The *roots* are those values of x such that $P_2(x_k) = 0$. One of the first results (recorded by the Babylonians, c2000 BCE) was the factoring of this equation by *completing the square* (Stillwell, 2010, p. 93). One may isolate x by rewriting Eq. 1.19 as

$$\frac{1}{a}P_2(x) = (x + b/2a)^2 - (b/2a)^2 + c/a. \qquad (1.20)$$

This is easily verified by expanding the squared term and canceling $(b/2a)^2$

$$\frac{1}{a}P_2(x) = [x^2 + (b/a)x + \cancel{(b/2a)^2}] - \cancel{(b/2a)^2} + c/a.$$

Setting Eq. 1.20 to zero and solving for the two roots $x_\pm$, gives the *quadratic formula*[45]

$$x_\pm = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{1.21}$$

If $b^2 - ac > 0$, the two roots are real ($x_\pm \in \mathbb{R}$). If it is exactly zero, the roots are identical. Otherwise, they are a complex pair.

In the important and common case that the coefficients are real and positive, and b is small (i.e., $c > b/2$). Since $a \neq 0$, the polynomial may be divided by a, removing it from the equation. Thus we We may set $a = 1$ with no loss of generality. In this special, but common case, the roots are given by[46]

$$x_\pm = -b/2 \pm \jmath\sqrt{c - (b/2)^2} \approx -b/2 \pm \jmath\sqrt{c}.$$

No insight is gained by memorizing the quadratic formula (Eq. 1.21). On the other hand, an important concept is gained by learning Eq. 1.20, which can be very helpful when doing analysis. It seems obvious, that instead of memorizing Eq. 1.21, learn Eq. 1.20. Arguably, the factored form (Eq. 1.20) is easier to remember, as a method, rather than as a formula that must be memorized. Importantly, the term $b/2a$ has significance $[P_2(-b/2a) = c/a - (b/2a)^2]$, the sign of which determines if the nature of the roots (real vs. complex).

In third grade I learned the trick[47]

$$9 \cdot n = (n - 1) \cdot 10 + (10 - n). \tag{1.22}$$

With this simple rule I did not need to depend on my memory for the 9 times tables. By expanding the above, one can see why it works: $9n = n10 \cancel{-10} + \cancel{10} - n = n(10 - 1)$. Learning an algorithm is much more powerful than memorization of the 9 times tables. How one thinks about a problem can have great impact.

Working with polynomials in Matlab/Octave: In Matlab/Octave there are five functions that work together as an important set, and you need to become familiar with these:

1. `R=root(A)` Vector $A = [a_N, a_{N-1}, \ldots, a_o]$ are the coefficients of polynomial $P_n(z) = \sum_{n=0}^{N} a_n z^n$, with N is the degree of the polynomial. It is convenient to force $a_N = 1$, corresponding to dividing the polynomial by this value, when it is not 1, guaranteeing it cannot be zero, as mentioned above. Define $R = [z_1, z_2, \cdots, z_n]$ as the vector of roots, such that `polyval(A,R)=0`.

2. `y=polyval(A,x)` This evaluates the polynomial defined by vector `A` at vector values of `x`, returning vector `y(x)`.

3. `A=poly(R)` This is the inverse of `root()`, returning a vector of polynomial coefficients `A`, given a vector of roots `R`. Due to IEEE-754 scaling issues, this can give strange results, that are numerically correct, but withing the limits of IEEE-754 accuracy.

4. `[K,R]=residue(A,B)`: Given the ratio of two polynomials `A,B`, `residue(A,B)` returns vectors `K,R` such that

$$\frac{A(s)}{B(s)} = \sum_k \frac{K_k}{s - s_k}. \tag{1.23}$$

where s_k are the roots of `B`.

---

[45]By direct substitution demonstrate that Eq. 1.21 is the solution of Eq. 1.19.

[46]This is the case for mechanical and electrical circuits having small damping. Physically $b > 0$ is the damping coefficient and $\sqrt{c} > 0$ is the resonant frequency. One may then simplify the form as $x^2 + 2bx + c^2 = (x + b + \jmath c)(x + b - \jmath c)$.

[47]E.G.: $9 \cdot 7 = (7 - 1) \cdot 10 + (10 - 7) = 60 + 3$ and $9 \cdot 3 = (3 - 1) \cdot 10 + (9 - 3) = 20 + 7$. As a check, note that the two terms $(n - 1)$ and $(10 - n)$, add to 9.

5. `C=conv(A,B)`: Vector `C` contain the polynomial coefficients of the convolution of the two vector of coefficients of polynomials `A,B`. For example `[1, 2, 1]=conv([1, 1], [1, 1])`.

The use of `residue()` will be discussed in Sect. 1.3.3 (p. 54), and again, in more detail, in Sect. 1.4.10 (p. 88).

Analytic Series: When the degree of the polynomial is infinite (i.e., $n = \infty$), $P_\infty(x), x, x_o, a_n \in \mathbb{R}$ the series is called a *power series*, namely

$$P(x) = \sum_0^\infty a_n (x - x_o)^n. \tag{1.24}$$

For values of x where the power series converges, $P(x)$ is said to be an *analytic function* in the neighborhood of the *expansion point* x_o, within the *region of convergence* (RoC), also known as the *radius of convergence* (RoC), for cases where the argument x is complex.

When the coefficients are determined by derivatives of $P(x)$ evaluated at $x = x_o$, then $P(x)$ is called a *Taylor series*, where

$$a_n = \frac{1}{n!} \frac{d^n}{dx^n} P(x) \Big|_{x=x_o}. \tag{1.25}$$

This formula shows, in a general way, how to uniquely define the coefficients, and without the Taylor series formula, we would have no general way of finding a_n. The proof of the Taylor formula is transparent, simply by taking the term by term derivative of Eq. 1.24 and then evaluating the result at the expansion point.

Exercise: Verify that a_o and a_1 of Eq. 1.24 follow from Eq. 1.25.

Taylor series play a special role in mathematics, as the coefficients of the series uniquely determine a function (e.g., via its derivatives). The implications, and limitations of the series representation are very specific. *First*, if the series fails to converge (i.e., outside the RoC), it is essentially meaningless. *Second*, the analytic function must be single valued. This follows from the fact that each term in Eq. 1.24 is single valued. *Third*, analytic functions are very "smooth," since the may be differentiated an ∞ number of times, and the sum still converges. There can be no jumps or kinks in these functions.

But these properties are both their curse and their blessing, as they represent a good starting point for solving differential equations, which is exactly how they were used, by Newton and others.

Two well known analytic functions are the *geometric series*

$$\frac{1}{1-x} = 1 + x + x^2 + x^2 + \ldots = \sum_{n=0}^\infty x^n \tag{1.26}$$

and exponential series

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{3 \cdot 2}x^3 + \frac{1}{4 \cdot 3 \cdot 2}x^4 + \ldots = \sum_{n=0}^\infty \frac{1}{n!}x^n. \tag{1.27}$$

Exercise: Verify that the coefficients of the above functions are given by Eq. 1.25.

Region of convergence: Determining the RoC for a given analytic function is quite important, and may not always be obvious. In general the RoC is a circle having a radius, centered on the expansion point, out to the nearest pole. Thus when the expansion point is moved, the RoC changes, since the location of the pole is fixed.

For the geometric series (Eq. 1.26), if the expansion point is taken as $x_o = 1$, the RoC is $|x| < 1$, since $1/(1 - x)$ has a pole at $x = 1$. We may move the expansion point by a linear transformation.

For example, by replacing x with $z + 3$. Then the series becomes $1/((z + 3) - 1) = 1/(z + 2)$, so the RoC becomes 2, because in the z plane, the pole has moved to -2. A second important example is the function $1/(x^2 + 1)$, which has the same RoC as the geometric series, since it may be expressed as the partial fraction expansion

$$\frac{1}{x^2 + 1} = \frac{1}{(x + 1\jmath)(x - 1\jmath)} = \frac{1}{2\jmath} \left(\frac{1}{x - 1\jmath} - \frac{1}{x + 1\jmath} \right).$$

This form of expansion is called a *partial fraction expansion.*

Each term has an RoC of $|x| < |1\jmath| = 1$. In other words, it is the sum of two geometric series, with poles at $\pm 1\jmath$ which are not as obvious because the roots are complex, and conjugate. Once factored, it becomes clear what is going on.

5   **Exercise:**   Verify the above expression is correct, and show that the residues are $\pm 1/2\jmath$.

The exponential series converges for every finite value of $x \in \mathbb{R}$ (the RoC is the entire real line), thus the exponential is called an *entire function.*

Analytic functions: Any function that has a Taylor series expansion is called an *analytic function.* Within the RoC, the series expansion defines a single valued function. Polynomials, $1/(1 - x)$ and e^x
10   are examples of analytic functions that are real. Because analytic functions are easily manipulated term by term, they may be used to find solutions of differential equations, since the derivatives of a series are uniquely determined within the RoC, due to Eq. 1.25.

Every analytic function has a corresponding differential equation, that is determined by the coefficients a_k of the analytic power series. An example is the exponential, which has the property that it is the eigenfunction of the derivative operation

$$\frac{d}{dx} e^{ax} = a e^{ax},$$

which may be verified using Eq. 1.27. This relationship is a common definition of the exponential function, which is a very special, because it is the eigenfunction of the derivative.

The complex analytic power series (i.e, complex analytic functions) may also be integrated, term by term, since

$$\int^x f(x) dx = \sum \frac{a_k}{k + 1} x^{k+1}. \tag{1.28}$$

15   Newton took full advantage of this property of the analytic function and used the analytic series (*Taylor series*) to solve analytic problems, especially for working out integrals, allowing him to solve differential equations. To fully understand the theory of differential equations, one must master single valued analytic functions and their analytic power series.

Single- vs. multi-valued functions: Polynomials are single valued functions: for each x there is
20   a single value of $P_n(x)$. The set of x values of a function are called the *domain* and the set of $y(x)$ values are called the *codomain.*[48] The roles of the domain and codomain may be swapped, to obtain the inverse function, which is typically quite different in its properties compared to the function. For example $y(x) = x^2 + 1$ has the inverse $x = \pm\sqrt{y - 1}$, which is double valued, and complex when $y < 1$. Periodic functions, such as $y(x) = \sin(x)$ are more exotic, since $x(y) = \arcsin(x) = \sin^{-1}(x)$ has an ∞
25   number of $x(y)$ values for each y.

Exercise: Let $y(x) = \sin(x)$. Then $dy/dx = \cos(x)$. Show that $dx/dy = -1/\sqrt{1 - x^2}$. Hint: $x(y) = \cos^{-1}(y) = \arccos(y)$.

---

[48]The codomain is also called the *range, image* or the *ordinate.*

Exercise: Let $y(x) = \sin(x)$. Then $dy/dx = \cos(x\jmath)$. Show that $dx/dy = -\jmath/\sqrt{1+x^2}$.

Exercise: Find the Taylor series coefficients of $y = \sin(x)$ and $x = \sin^{-1}(y)$.

Complex analytic functions: When the argument of an analytic function $F(x)$ is complex, that is, $x \in \mathbb{R}$ is replaced by $s = \sigma + \omega\jmath \in \mathbb{C}$ (recall that $\mathbb{R} \in \mathbb{C}$)

$$F(s) = \sum_{n=0}^{\infty} c_n(s - s_o)^n, \tag{1.29}$$

with $c_n \in \mathbb{C}$, that function is said to be a *complex analytic.*

For example, when the argument of the exponential becomes complex, it is periodic on the ω axis, since

$$e^{st} = e^{(\sigma + \omega\jmath)t} = e^{\sigma t}e^{\jmath\omega t} = e^{\sigma t}\left[\cos(\omega t) + \jmath\sin(\omega t)\right].$$

Taking the real part gives

$$\Re e^{st} = e^{\sigma t}\frac{e^{\omega\jmath t} + e^{-\omega\jmath t}}{2} = e^{\sigma t}\cos(\omega t),$$

and $\Im e^{st} = e^{\sigma t}\sin(\omega t)$. Once the argument is allowed to be complex, it becomes obvious that the exponential and circular functions are fundamentally related. This exposes the family of *entire circular functions* [i.e., $e^s, \sin(s), \cos(s), \tan(s), \cosh(s), \sinh(s)$] and their inverses [$\ln(s), \arcsin(s),$ $\arccos(s), \arctan(s), \cosh^{-1}(s), \sinh^{-1}(s)$], first fully elucidated by Euler (c1750) (Stillwell, 2010, p. 315). Note that because a function, such as $\sin(\omega t)$, is periodic, its inverse must be multi-valued. What is needed is some systematic way to account for this multi-valued properties. That methodology was provided by Riemann 100 years later, in this 1851 PhD Thesis, supervised by Gauss, in the final years of Gauss' life.

Given a complex analytic function of a complex variable, one must resort to the *extended complex plane*, *Riemann sheets* and *branch cuts*, as discussed in Section 1.3.8 (p. 63). The extended complex plane is a tool that extends the domain of complex analytic to include the point at infinity. This topic is critically important in engineering mathematics, and will be discussed in length in Sections 1.3.8-1.3.11 (pp. 63-70).

Definition of the Taylor series of a complex analytic function: However there is a fundamental problem, since we cannot formally define the Taylor series for the coefficients c_k, since we have not defined $dF(s)/ds$, the derivative with respect to the complex variable $s \in \mathbb{C}$. Thus simply substituting s for x in an analytic function is leaving a major hole in our understanding of the complex analytic function.

To gain a feeling of the nature of the problem, we make take derivatives of a function with respect to various variables. For example,

$$\frac{d}{dt}e^{st} = se^{st}.$$

Also

$$e^{\omega\jmath t}\frac{d}{d\sigma}e^{\sigma t} = \sigma e^{st}.$$

and

$$e^{\sigma t}\frac{d}{d\omega\jmath}e^{\omega\jmath} = \omega\jmath e^{st}.$$

are straightforward.

It was the work of Cauchy (1814) (Fig. 1.11), who uncovered much deeper relationships within complex analytic functions (Sect. 1.3.9, p. 65) by defining differentiation and integration in the complex plane, leading to several fundamental theorems of complex calculus, including the *Fundamental*

theorem of complex integration, and Cauchy's formula. We shall explore this in and several fundamental theorems in Sect. 1.4.1 (p. 76).

There seems to be some disagreement as to the status of multi-valued functions: Are they functions, or is a function strictly single valued? If so, then we are missing out on a host of interesting possibilities, including all the inverses of nearly every complex analytic function. For example, the inverse of a complex analytic is complex analytic (e.g., e^s and $\log(s)$).

Impact on Physics: It seems likely, if not obvious, that the success of Newton was his ability to describe physics by the use of mathematics. He was inventing new mathematics at the same time as he was explaining new physics. The same might be said for Galileo. It seems likely that Newton was extending the successful techniques and results of Galileo. Galileo died on Jan 8, 1642, and Newton was born Jan 4, 1643, just short of one year later. Certainly Newton was well aware of Galileo's great success, and naturally would have been influenced by them.

The application of complex analytic functions to physics was dramatic, as may be seen in the six volumes on physics by Arnold Sommerfeld (1868-1951), and from the productivity of his many (36) students (e.g., Debye, Lenz, Ewald, Pauli, Guillemin, Bethe, Heisenberg[49] and Seebach, to name a few), notable coworkers (i.e., Leon Brillouin) and others (i.e., John Bardeen), upon whom Sommerfeld had a strong influence. Sommerfeld is known for having many students who were awarded the Nobel Prize in Physics, yet he was not (the prize is not awarded in mathematics). Sommerfeld brought mathematical physics (the merging of physical and experimental principles with mathematics) to a new level with the use of complex integration of analytic functions to solve otherwise difficult problems, thus following the lead of Newton who used real integration of Taylor series to solve differential equations (Brillouin, 1960, Ch. 3 by Sommerfeld, A.).

1.3.2   Lec 12: **Polynomial root classification by convolution**

Following the exploration of algebraic relationships by Fermat and Descartes, the first theorem was being formulated by d'Alembert. The idea behind this theorem is that every polynomial of degree N (Eq. 1.18) has at least one root. This may be written as the product of the root and a second polynomial of degree of $N-1$. By the recursive application of this concept, it is clear that every polynomial of degree N has N roots. Today this result is known as the *fundamental theorem of algebra:*

> Every polynomial equation $P(z) = 0$ has a *solution* in the complex numbers. As Descartes observed, a solution $z = a$ implies that $P(z)$ has a factor $z - a$. The quotient
>
> $$Q(z) = \frac{P(z)}{z-a} = \frac{P(z)}{a}\left[1 + \frac{z}{a} + \left(\frac{z}{a}\right)^2 + \left(\frac{z}{a}\right)^3 + \cdots\right] \tag{1.30}$$
>
> is then a polynomial of one lower degree. ... We can go on to factorize $P(z)$ into n linear factors.

> —Stillwell (2010, p. 285).

The ultimate expression of this theorem is given by Eq. 1.18 (p. 44), which indirectly states that an nth degree polynomial has n roots.

Exercise: Explore expressing Eq. 1.30 in terms of real 2x2 matrices, as described in Section 1.2.1, p. 24. **Sol:** Consider the representation of polynomials as a special case of products of 2x2 complex numbers, as discussed in Eq. 1.2, p. 24.

Today this theorem is so widely accepted we fail to appreciate it. Certainly about the time you learned the quadratic formula, you were prepared to understand the concept of polynomials having

---

[49]https://www.aip.org/history-programs/niels-bohr-library/oral-histories/4661-1

roots. The simple quadratic case may be extended a higher degree polynomial. The Matlab/Octave command `roots([1, a_2, a_1, a_o])` provides the roots $[s_1, s_2, s_3]$ of the cubic equation, defined by the coefficient vector $[1, a_2, a_1, a_o]$. The command `poly([s_1, s_2, s_3])` returns the coefficient vector. I don't know the largest degree that can be accurately factored by Matlab/Octave, but I'm sure its well over $N = 10^3$. Today, finding the roots numerically is a solved problem.   5

Factorization versus convolution: The best way to gain insight into the polynomial factorization problem is through the inverse operation, multiplication of monomials. Given the roots x_k, there is a simple algorithm for computing the coefficients a_k of $P_n(x)$ for any n, no matter how large. This method is called *convolution*. Convolution is said to be a *trap-door* since it is easy, while the inverse, factoring (*deconvolution*), is hard, and analytically intractable for degree $N \geq 5$ (Stillwell, 2010, p. 102).   10

Convolution of monomials

As outlined by Eq. 1.18, a polynomial has two descriptions, first as a series with coefficients a_n and second in terms of its roots x_r. The question is "What is the relationship between the coefficients and the roots?" The simple answer is that they are related by *convolution.*
   Let us start with the quadratic

$$(x + a)(x + b) = x^2 + (a + b)x + ab,$$

where in vector notation $[-a, -b]$ are the roots and $[1, a + b, ab]$ are the coefficients.   15
   To see how the result generalizes, we may work out the coefficients for the cubic ($N = 3$). Multiplying the following three factors gives (p. 135)

$$(x - 1)(x - 2)(x - 3) = (x^2 - 3x + 2)(x - 3) = x(x^2 - 3x + 2) - 3(x^2 - 3x + 2) = x^3 - 6x^2 + 11x - 6.$$

When the roots are $[1, 2, 3]$ the coefficients of the polynomial are $[1, -6, 11, -6]$. To verify, substitute the roots into the polynomial, and show that they give zero. For example $r_1 = 1$ is a root since $P_3(1) = 1 - 6 + 11 - 6 = 0$.
   As the degree increases, the algebra becomes more difficult. Imagine trying to work out the coefficients for $N = 100$. What is needed is an simple way of finding the coefficients from the roots.   20
Fortunately, *convolution* keeps track of the book-keeping, by formalizing the procedure.

Convolution of two vectors: To get the coefficients by convolution, write the roots as two vectors $[1, a]$ and $[1, b]$. To find the coefficients we must convolve the root vectors, indicated by $[1, a] \star [1, b]$, where $\star$ denotes convolution. Convolution is a recursive operation. The convolution of $[1, a] \star [1, b]$ is done as follows: reverse one of the two monomials, padding unused elements with zeros. Next slide one monomial against the other, forming the local *dot product* (element-wise multiply and add):

a	1	0	0		a	1	0		a	1	0		0	a	1		0	0	a	1
0	0	1	b		0	1	b		1	b	0		1	b	0		1	b	0	0 ,
$=$	0				$=$	x^2			$=$	$(a+b)x$			$=$	abx^0			$=$	0	

resulting in coefficients $[\cdots, 0, 0, 1, a + b, ab, 0, 0, \cdots]$.
   By reversing one of the polynomials, and then taking successive dot products, all the terms in the sum of the dot product correspond to the same power of x. This explains why convolution of the coefficients gives the same answer as the product of the polynomials.   25
   As seen by the above example, the position of the first monomial coefficients are reversed, and then slid across the second set of coefficients, the dot-product is computed, and the result placed in the output vector. Outside the range shown, all the elements are zero. In summary,

$$[1, -1] \star [1, -2] = [1, -1 - 2, 2] = [1, -3, 2].$$

In general

$$[a, b] \star [c, d] = [ac, bc + ad, bd],$$

Convolving a third term $[1, -3]$ with $[1, -3, 2]$ gives (p. 135)

$$[1, -3] \star [1, -3, 2] = [1, -3 - 3, 9 + 2, -6] = [1, -6, 11, -6],$$

which is identical to the cubic example, found by the algebraic method.

By convolving one monomial factor at a time, the overlap is always two elements, thus it is never necessary to compute more than two multiplies and an add for each output coefficient. This greatly simplifies the operations (i.e., they are easily done in your head). Thus the final result is more likely to be correct. Comparing this to the algebraic method, convolution has the clear advantage.

Exercise:

1. Given the coefficient vector $[\alpha, \beta, \gamma] = [1, a + b, ab]$, find the formula for the roots $[a, b]$. **Sol:** To solve this one must complete the square of the polynomial:

$$0 = x^2 + \beta x + \gamma = (x + \beta/2)^2 - (\beta/2)^2 + \gamma$$
$$(x + \beta/2)^2 = (\beta/2)^2 - \gamma$$
$$x + \beta/2 = \pm\sqrt{(\beta/2)^2 - \gamma}$$
$$x = -\beta/2 \pm \sqrt{(\beta/2)^2 - \gamma}.$$

In terms of our assumed coefficient vector $[1, a + b, ab]$, this gives (See Eq.**??**, p. 135)

$$x_{\pm} = -\frac{a + b}{2} \pm \sqrt{\left(\frac{a + b}{2}\right)^2 - ab}. \tag{1.31}$$

While it is not obvious that $x_{\pm} = [a, b]$, it must be true. One may prove this several ways. The most obvious is working backwards from what we did above. However this gives no insight (try it), other than the fact that we did the algebra correctly. Let's try something else:

$$x_{\pm} + \frac{a + b}{2} = \pm\sqrt{\left(\frac{a + b}{2}\right)^2 - ab}$$
$$= \pm\sqrt{\frac{a^2 + b^2}{4} + ab - ab}$$
$$= \pm\frac{1}{2}\sqrt{(a + b)(a - b)}$$

No end in sight here, and there is obvious reason (that I can see) why this is.

Let's try the product

$$x_+ \cdot x_- = \left(-\frac{a + b}{2} + \sqrt{\left(\frac{a + b}{2}\right)^2 - ab}\right)\left(-\frac{a + b}{2} - \sqrt{\left(\frac{a + b}{2}\right)^2 - ab}\right)$$
$$= \left(\frac{a + b}{2}\right)^2 - \left(\sqrt{\left(\frac{a + b}{2}\right)^2 - ab}\right)^2$$
$$= \left(\frac{a + b}{2}\right)^2 - \left(\frac{a + b}{2}\right)^2 + ab.$$

We might conclude that we have simultaneous solved two nonlinear equations in the roots $[a, b]$, given the coefficients of the polynomial $[1, \beta, \gamma]$:

$$\beta = (a + b)/2$$
$$\gamma = ab$$

which are the mean and square of the geometric mean $\sqrt{ab}$.

2. What are the three nonlinear equations that one would need to solve to find the roots of a cubic? **Sol:** From our fomula for the convolution of three monomials we may find the nonlinear "deconvolution" relations between the roots[50] $[-a, -b, -c]$ and the cubic's coefficients $[1, \alpha, \beta, \gamma]$

$$
\begin{aligned}
(x + a) \star (x + b) \star (x + c) &= (x + c) \star (x^2 + (a + b)x + ab) \\
&= x \cdot (x^2 + (a + b)x + ab) + c \cdot (x^2 + (a + b)x + ab) \\
&= x^3 + (a + b + c)x^2 + (ab + ac + cb)x + abc \\
&= [1, a + b + c, ab + ac + cb, abc].
\end{aligned}
$$

It follows that the nonlinear equations must be

$$\alpha = a + b + c$$
$$\beta = ab + ac + bc$$
$$\gamma = abc.$$

Clearly these are solve by the classic cubic solution. How that works would be interesting to add to this discussion. **Verify!**

The product of monomial $P_1(x)$ with a polynomial $P_N(x)$ gives $P_{N+1}(x)$: This statement is another way of stating the *fundamental theorem of algebra*. Each time we convolve a monomial with a polynomial of degree N, we obtain a polynomial of degree $N + 1$. The convolution of two monomials results in a quadratic (degree 2 polynomial). The convolution of three monomials give a cubic (degree 3). In general, the degree k, of the product of two polynomials, of degree n, m, is the sum of the degrees $(k = n + m)$. For example, if the degrees are each 5 $(n = m = 5)$, then the resulting degree is 10.

In summary, the product of two polynomials of degree m, n having m and n roots, give a polynomial of degree $m + n$ having $m + n$ roots. This is an analysis process, of merging polynomials, by coefficient convolution. Multiplying polynomials is a merging process, into a single polynomial.

Note that the degree of a polynomial is one less than the length of the vector of coefficients. The coefficient on the lead term should always be set to 1 since it cannot be zero, resulting in an illogical result. *Always* normalize the lead term of $P_n(x)$ to 1 (i.e., $a_N = 1$). This has no effect on the roots.

While you already know this theorem from high school algebra class, it is important to explicitly state the *fundamental theorem of algebra*.

Composition of polynomials: Convolution is not the only operation between two polynomials. Another is *composition*, which may be defined for two functions $f(z), g(z)$. Then the composition $c(z) = f(z) \circ g(z) = f(g(z))$. As a specific example, suppose $f(z) = 1 + z + z^2$ and $g(z) = e^{2z}$. With these definitions

$$f(z) \circ g(z) = 1 + e^{2z} + (e^{2z})^2 = 1 + e^{2z} + e^{4z}.$$

Note that $f(z) \circ g(z) \neq g(z) \circ f(z)$.

---

[50]By working with the negative roots we may avoid an unnecessary and messy alternating sign problem.

Exercise: Find $g(z) \circ f(z)$. **Sol:** $e^{2f(z)} = e^{2(1+z+z^2)} = e^2 e^{(1+z+z^2)} = e^3 e^z e^{z^2}$.

1.3.3 **Lec 13** Residue expansions

As discussed in Section 1.3.1, p. 46, there are 5 important Matlab/Octave routines that are closely re-
lated: `root()`, `polyval()`, `poly()`, `residue()`, `conv()`. The function `residue()` is more complex
that the others, and it remains to be discussed.

When lines and planes are defined, the equations are said to be *linear* in the independent variables.
In keeping with this definition of *linear*, we say that the equations are *non-linear* when the equations
have degree greater than 1 in the independent variables. The term *bilinear* has a special meaning,
in that both the domain and codomain are linearly related by lines (or planes). As an example, an
impedance is defined in frequency as the ratio of the voltage over the current

$$Z(s) = \frac{V(\omega)}{I(\omega)} = sL_o + R_o + \sum_{k=0}^{K} \frac{K_k}{s - s_k}, \tag{1.32}$$

where $Z(s)$ is the impedance and V and I are the voltage and current at radian frequency ω.

The impedance is typically specified as the ratio of two polynomials, $N(s) = [a_n, a_{n-1}, \cdots, a_o]$ and
$D(s) = [b_K, b_{K-1}, \cdots, b_o]$, as functions of complex Laplace frequency $s = \sigma + \jmath\omega$, having simple roots.
The bilinear function may be written as $D(s)V = N(s)I$. Since $D(s)$ and $N(s)$ are both polynomials
in s, this is called bilinear, which comes from a corresponding scalar differential equation, of the form

$$\sum_{k=0}^{K} b_k \frac{d^k}{dt^k} i(t) = \sum_{n=0}^{N} a_n \frac{d^n}{dt^n} v(t) \quad \leftrightarrow \quad I(\omega) \sum_{k=0}^{K} b_k s^k = V(\omega) \sum_{n=0}^{N} a_n s^n$$

Equation 1.32 follows from the Laplace transform (on right) of the differential equation (on left), by
forming the impedance $Z(s) = V/I = A(s)/B(s)$. This form of the differential equation follows from
Kirchhoff's voltage and current laws (KCL, KVL) or from Newton's laws (mechanics).

The physical properties of an impedance: Based on d'Alembert's observation that the solution
to the wave equation is the sum of forward and backward traveling waves, the impedance may be
rewritten in terms of forward and backward traveling waves

$$Z(s) = \frac{V}{I} = \frac{V^+ + V^-}{I^+ - I^-} = r_o \frac{1 + \Gamma(s)}{1 - \Gamma(s)}, \tag{1.33}$$

where $r_o = P^+/I^+$ is called the *surge impedance* of the transmission line (e.g., wire) connected to the
load impedance $Z(s)$, and $\Gamma(s) = P^-/P_+ = I^-/I^+$ is the reflection coefficient corresponding to $Z(s)$.
Like $Z(s)$, $\Gamma(s)$ is causal and complex analytic. Note that the impedance and the reflectance function
must both be complex analytic, since the are connected by the bilinear (aka, Möbius) transformation,
which assures the mutual complex analytic properties.

Due to the bilinear transformation, the physical properties of $Z(s)$ and $\Gamma(s)$ are very different.
Specifically, the real part of the load impedance must be non-negative ($\Re Z(\omega \jmath) \geq 0$), if and only
if $|\Gamma(s)| \leq 1|$. In the time domain, the impedance $z(t) \leftrightarrow Z(s)$ must have a value of r_o at $t = 0$.
Correspondingly, the time domain reflectance $\gamma(t) \leftrightarrow \Gamma(s)$ must be zero at $t = 0$.

This is the basis of conservation of energy, which may be traced back to the properties of the
reflectance $\Gamma(s)$.

Exercise: Show that if the $\Re Z(s) \geq 0$ then $|\Gamma(s)| \leq 1$. **Sol:** Their two equivalent proofs, both of
which start from the relation between $Z(s)$ and $\Gamma(s)$. Taking the real part of Eq. 1.33, which must be
≥ 0, we find

$$\Re Z(s) = \frac{r_o}{2} \left[\frac{1 + \Gamma(s)}{1 - \Gamma(s)} + \frac{1 + \Gamma^*(s)}{1 - \Gamma^*(s)} \right] = r_o \frac{1 - |\Gamma(s)|^2}{|1 + \Gamma(s)|^2} \geq 0.$$

Thus $|\Gamma| \leq 1$.

1.3.4 Lec 14: Introduction to Analytic Geometry

Analytic geometry is the natural consequence of Euclid's Geometry (which deals with conical geometry (e.g., points, lines, triangles, circles, spheres, ellipses, cones, etc., in two and three physical dimensions), merged with algebra (which deals with simultaneous equations, roots of polynomials, analytic functions, and ultimately, solutions of differential equations). The combination of Euclid's (323 BCE) geometry and al-Khwarizmi's (830 CE) algebra provides a new powerful tool, *analytic geometry.*

There are many important relationships between Euclidean geometry and 16th century algebra. Important similarities include vectors, their Pythagorean lengths $[a, b, c]$

$$c = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \tag{1.34}$$

$a = x_2 - x_1$ and $b = y_2 - y_1$, and the angles. Euclid's geometry had no concept of coordinates, thus of vectors. One of the main differences is the ability of analytic geometry is that one may compute with numbers. A detailed comparison is given in Table 1.3.

There are several new concepts that come with the development of analytic geometry:

1. Composition of functions: If $y = f(x)$ and $z = g(y)$ then the composition of functions f and g is denoted $z(x) = g \circ f(x) = g(f(x))$.

2. Elimination: Given two functions $f(x, y)$ and $g(x, y)$, elimination removes either x or y. This procedure, called *Gaussian Elimination*, was known to the Chinese. This is not always possible.

3. Intersection: While one may speak of the intersection of two lines to give a point, or two planes to give a line, the term *intersection* is also an important but very different concept in set theory. This is a special case of elimination when the functions $f(x, y), g(x, y)$ are linear in their arguments.

4. Vectors: Euclidean geometry provides the new concept of a vector, as a line with length and orientation (i.e., direction). Analytic geometry defines a vector, in any number of dimensions, as an ordered set of points $\boldsymbol{x} \in \mathbb{C}$.

5. Analytic geometry extends the ideas of geometry with the introduction of the product of two vectors, the scalar (dot) product $\boldsymbol{f} \cdot \boldsymbol{g}$ and the vector (cross) product $\boldsymbol{f} \times \boldsymbol{g}$ (Fig. 1.12).

Table 1.3: Comparison between Euclidean geometry and analytic geometry. I am uncertain about the items in the third column.

Euclidean geometry	analytic geometry	Uncertain
Proof, Vector, Length, Point, Direction, ≤ 3 dimensional, Intersection, Conic Section, Dot product, Square Roots: on real line lengths.	> 3 dimensions, Numbers, Analytic series and functions, Composition, Elimination, Fund. Thm. Algebra, $\sin(\theta)$, $\cos(\theta)$, $e^{\theta j}$, $\log(z)$, Derivative, Calculus, Polynomial & Roots: on complex algebra.	Cross product, Recursion, Iteration on complex planes and circles (Newton's method).

What algebra added to geometry was the ability to compute with complex numbers. For example, the length of a line (Eq. 1.34) was measured in Geometry with a compass: numbers played no role. Once algebra was available, the line's Euclidean length could be computed from the coordinates of the two ends, defined by the *3-vector*

$$\boldsymbol{e} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}} = [x, y, z]^T,$$

which represents a point at $(x, y, z) \in \mathbb{R}^3 \subset \mathbb{C}^3$ in three dimensions, having *direction*, from the origin $(0, 0, 0)$ to (x, y, z). An alternative *matrix notation* is $\boldsymbol{e} = [x, y, z]^T$, a column vector of three numbers. These two notations are different ways of representing exactly the same thing. View them as equivalent notations.

⁵   By defining the vector, analytic geometry allows Euclidean geometry to become quantitative, beyond the physical drawing of an object (e.g., a sphere, triangle or line). With analytic geometry we have the Euclidean concept of a vector, a line having a magnitude (length) and direction, but analytic defined in terms of physical coordinates (i.e., numbers). The difference between two vectors defines a third vector, a concept already present in Euclidean geometry. For the first time, complex numbers
¹⁰ were allowed into geometry (but rarely used until Cauchy and Riemann).

     As shown in Fig. 1.12, there are two types of products, the 1) scalar $\boldsymbol{A} \cdot \boldsymbol{B}$ and 2) vector $\boldsymbol{A} \times \boldsymbol{B}$ products.

Scalar product of two vectors: When using algebra, many concepts in geometry are made precise. There are many examples of how algebra extends Euclidean geometry, the most basic being the *scalar product* (aka dot product) between two vectors

$$\boldsymbol{x} \cdot \boldsymbol{\kappa} = (x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}) \cdot (\alpha\hat{\mathbf{i}} + \beta\hat{\mathbf{j}} + \gamma\hat{\mathbf{k}}), \quad \in \mathbb{C}$$
$$= \alpha x + \beta y + \gamma z.$$

In matrix notation the scalar product is written as

$$\boldsymbol{x} \cdot \boldsymbol{\kappa} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}^T \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} x, y, z \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \alpha x + \beta y + \gamma z.$$

The dot product takes the character of κ. For example, if $\kappa(s) \in \mathbb{C}$ is a function of complex frequency s, then the dot product is complex. If $\zeta \in \mathbb{R}$ is real, then the dot product is real.

Norm (length) of a vector: The *norm* of a vector

$$||\boldsymbol{e}|| \equiv +\sqrt{\boldsymbol{e} \cdot \boldsymbol{e}} \geq 0.$$

is *defined* as the positive square root of the scalar product of the vector with itself. This is a generalization of the length, in any number of dimensions, forcing the sign of the square-root to be non-negative. The length is a concept of Euclidean geometry, and it must always be positive and real. A complex length is not physically meaningful. A zero-length vector is a point. More generally, the Euclidean length of a line is given as the *norm* of the difference between two real vectors $\boldsymbol{e}_1, \boldsymbol{e}_2 \in \mathbb{R}$

$$||\boldsymbol{e_1} - \boldsymbol{e_2}||^2 = (\boldsymbol{e}_1 - \boldsymbol{e}_2) \cdot (\boldsymbol{e}_1 - \boldsymbol{e}_2) \tag{1.35a}$$
$$= (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \geq 0. \tag{1.35b}$$

¹⁵ **Pythagorean theorem and the Schwarz inequality:**   Regarding Fig. 1.12, suppose we compute the difference between vector $\boldsymbol{A} \in \mathbb{R}$ and $\alpha\boldsymbol{B} \in \mathbb{R}$ as $L = ||\boldsymbol{A} - \alpha\boldsymbol{B}|| \in \mathbb{R}$, where $\alpha \in \mathbb{R}$ is a scalar that modifies the length of $\boldsymbol{B}$. We seek the value of α, which we denote as α^*, that minimizes the length of L. From simple geometrical considerations, $L(\alpha)$ will be minimum when the difference vector is perpendicular to $\boldsymbol{B}$, as shown in the figure by the dashed line from the tip of $\boldsymbol{A} \perp \boldsymbol{B}$.

     To show this algebraically we write out the expression for $L(\alpha)$ and take the derivative with respect to α, and set it to zero, which gives the formula for α^*. The argument does not change, but the algebra greatly simplifies, if we normalize $\boldsymbol{A}, \boldsymbol{B}$ to be unit vectors $\boldsymbol{a} = \boldsymbol{A}/||\boldsymbol{A}||$ and $\boldsymbol{b} = \boldsymbol{B}/||\boldsymbol{B}||$, which have norm $= 1$.

$$L^2 = (\boldsymbol{a} - \alpha\boldsymbol{b}) \cdot (\boldsymbol{a} - \alpha\boldsymbol{b}) = 1 - 2\alpha\boldsymbol{a} \cdot \boldsymbol{b} + \alpha^2. \tag{1.36}$$
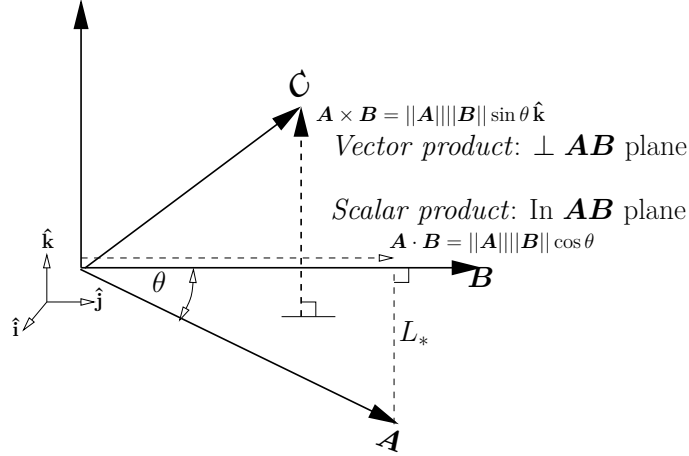
Figure 1.12: Definitions of vectors A, B, C used in the definition of $A \cdot B$, $A \times B$ and $C \cdot (A \times B)$. The cross product $A \times B$ is the area of the trapezoid formed by the two vectors, while the triple product $C \cdot (A \times B)$ defines the volume of the formed parallelepiped (i.e., prism). When all the angles are 90°, the volume becomes a *cuboid.*

Thus the length is shortest ($L = L_*$, as shown in Fig. 1.12) when

$$\frac{d}{d\alpha} L_*^2 = -2a \cdot b + 2\alpha^* = 0.$$

Solving for α^* we find $\alpha^* = a \cdot b$. Since $L_* > 0$ ($a \neq b$), Eq. 1.36 becomes

$$1 - 2|a \cdot b|^2 + |a \cdot b|^2 = 1 - |a \cdot b|^2 > 0.$$

In conclusion $\cos\theta \equiv |a \cdot b| < 1$. In terms of A, B this is $|A \cdot B| < ||A|| \, ||B|| \cos\theta$, as shown next to B in Fig. 1.12. Thus the scalar product between two vectors is their *direction cosine.* Furthermore since this forms a right triangle, the Pythagorean theorem must hold. The *triangle inequality* says that the lengths of the two sides must be greater than the hypotenuse. Note that $\Theta \in \mathbb{R} \notin \mathbb{C}$.

This derivation is an abbreviated version of a related discussion in Section 3.2.1 (p. 132). [5]

Vector ($\times$) product of two vectors: As shown in Fig. 1.12, the vector product (aka, cross-product) $a \times b$ is the second type of product between two vectors. The vector product defines a vector, perpendicular ($\perp$) to the plane of the two vectors being multiplied. The formula for computing the cross product is

$$a \times b = (a_1\hat{\mathbf{i}} + a_2\hat{\mathbf{j}} + a_3\hat{\mathbf{k}}) \times (b_1\hat{\mathbf{i}} + b_2\hat{\mathbf{j}} + b_3\hat{\mathbf{k}}) = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}.$$

For example, if the two vectors are in $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$, then the cross-product is $\hat{\mathbf{k}}$. It is strictly in $\hat{\mathbf{k}}$ if the two vectors are perpendicular to each other (i.e., $\hat{\mathbf{k}} = \hat{\mathbf{i}} \times \hat{\mathbf{j}} = -\hat{\mathbf{j}} \times \hat{\mathbf{i}}$). The vector product of a vector with itself (or the difference between two vectors) is zero. For example $\hat{\mathbf{i}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}} \times \hat{\mathbf{k}} = 0$. Typically $a, b \in \mathbb{R}$. If they are complex, the definition must be modified to be consistent with the physics. [10]

The scalar product of a third vector c with the vector product $a \times b$, with $a, b, c \in \mathbb{R}$,

$$c \cdot (a \times b) = \begin{vmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \quad (\in \mathbb{R}),$$

is called the *triple product,* which represents the volume of a parallelepiped.

Impact of Analytic Geometry: The most obvious impact of analytic geometry, was a detailed analysis of the conic sections, using algebra, rather than drawings via a compass and ruler. An important example is the composition of the line and circle, a venerable construction, presumably going back to before Diophantus (250CE). Once algebra was invented, the composition could be done using formulas. With this analysis came complex numbers.

The first two mathematicians to appreciate this mixture of Euclid's geometry and the new algebra were Fermat and Descartes (Stillwell, 2010, p. 111-115); soon Newton contributed to this effort, by the addition of physics (calculations in acoustics, orbits of the planets, and the theory of gravity and light Stillwell (2010, p. 115-117)), significant concepts for 1687.

Given these new methods, many new solutions to problems emerged. The complex roots of polynomials continued to appear, without any obvious physical meaning. Complex numbers seem to have been viewed as more of an inconvenience than a problem. Newton's solution to this dilemma was to simply ignore the imaginary cases (Stillwell, 2010, p. 119).

1.3.5   Development of Analytic Geometry

Composition, Intersection and Gaussian elimination: The first "algebra" (al-jabr) is credited to al-Khwarizmi (830 CE). Its invention advanced the theory of polynomial equations in one variable, Taylor series, and composition versus intersections of curves. The solution of the quadratic equation had been worked out thousands of year earlier, but with algebra a general solution could be defined. The Chinese had found the way to solve several equations in several unknowns, for example, finding the values of the intersection of two circles. With the invention of algebra by al-Khwarizmi, a powerful tool became available to solve the difficult problems.

Composition, Elimination and Intersection   In algebra there are two contrasting operations on functions: *composition* and *Elimination.*

Composition: Composition is the merging of functions, by feeding one into the other. If the two functions are f, g then their composition is indicated by $f \circ g$, meaning the function $y = f(x)$ is substituted into the function $z = g(y)$, giving $z = g(f(x))$.

Composition is not limited to linear equations, even though that is where it is most frequently applied. To compose two functions, one must substitute one equation into the other. That requires solving for that substitution variable, which is not always possible in the case of nonlinear equations. However many tricks are available that may work around this restrictions. For example if one equation is in x^2 and the other in x^3 or $\sqrt{x}$, it may be possible to multiply the first by x or square the second. The point is that one of the variables must be isolated so that when it is substituted into the other equations, the variable is removed from the mix.

Examples: Let $y = f(x) =: x^2 - 2$ and $z = g(y) =: y + 1$. Then

$$g \circ f = g(f(x)) = (x^2 - 2) + 1 = x^2 - 1. \tag{1.37}$$

In general composition does not commute (i.e., $f \circ g \neq g \circ f$), as is easily demonstrated. Swapping the order of composition for our example gives

$$f \circ g = f(g(y)) = z^2 - 2 = (y + 1)^2 - 2 = y^2 + 2y - 1. \tag{1.38}$$

Intersection: Complimentary to composition is *intersection* (i.e., decomposition) (Stillwell, 2010, pp. 119,149).   For example, the intersection of two lines is defined as the point where they meet. This Not sure ref is is not to be confused with finding roots. A polynomial of degree N has N roots, but the points where two polynomials intersect has nothing to do with the roots of the polynomials. The intersection is a function (equation) of lower degree, implemented with Gaussian elimination.

Intersection of two lines Unless they are parallel, two lines meet at a point. In terms of linear algebra this may be written as 2 linear equations (left) along with the intersection point $[x_1, x_2]^T$, given by the inverse of the 2x2 set of equations (right).[51]

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad\qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad (1.39)$$

where $\Delta = ad - bc$ is called the *determinant.* Note that the determinant may be written as

$$\frac{b}{a} = \frac{d}{c},$$

which says that the ratio of the slopes of the two lines are zero. Thus if $\Delta = 0$ the two lines are parallel.

By substituting the right expression into the left, and taking the inverse we obtain the intersection point. If $\Delta = 0$ there can be no solution, in which case the two lines are parallel (they meet at infinity.) Note the structure of the inverse: 1) The diagonal values are swapped, 2) the off-diagonal values are negative and 3) the matrix is divided by the determininant.

Exercise: Show that the equation on the right is the solution of the equation on the left. **Sol:** By a direct substitution of the left equation into the right equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \frac{1}{\Delta} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, = \frac{1}{\Delta} \begin{bmatrix} ad - bc & -ab + ab \\ cd - cd & -cb + ad \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \Delta & 0 \\ 0 & \Delta \end{bmatrix},$$

which is the identity matrix.

Algebra will give the solution when geometry cannot. When the two curves fail to intersect on the real plane, the solution still exists, but is complex valued. In such cases, geometry, which only considers the real solutions, fails. For example, when the coefficients $[a, b, c, d]$ are complex, the solution exists, but the determinant can be complex. Thus algebra is much more general than geometric. Geometry fails when the solution has a complex intersection.

A system of linear $Ax = y$ equations has many interpretations, and one should not be biased by the notation. As engineers we are trained to view x as the input and y as the output, in which case then $y = Ax$ seems natural, much like the functional relation $y = f(x)$. But what does the linear relation $x = Ay$ mean, when x is the input? The obvious answer is that $y = A^{-1}x$. But when working with systems of equations, there are many uses of equations, and we need to become more flexable in the our interpretation. For example $y = A^2 x$ has a useful meaning, and in fact we saw this type of relationship when working with Pell's equation (p. 39) and the Fibonacci sequence (p. 41). As another example consider

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} a_{1x} & a_{1y} \\ a_{2x} & a_{2y} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

which is reminiscent of a three-dimentional surface $\mathbf{z} = \mathbf{f}(x, y)$. We shall find that such generalizations are much more than a curosity.

1.3.6 Lec 15 Gaussian Elimination

The method for finding the intersection of equations is based on the recursive elimination of all the variables but one. This method, known as *Gaussian elimination*, works across a broad range of cases,

---

[51]It is very important to note when writing the equation $Ax = y$ in matrix format, the unknowns are x_1, x_2 whereas in the original equations they were y, x. The starting two equations are $ay + bx = c$ and $dy + ex = f$ (where $a, b, c, d, e, f \in \mathbb{C}$ are constants), two linear equations in (x, y). In matrix format the names are changed, where ethe two equations are $ax_1 + bx_2 = y_1$ and $dx_1 + ex_2 = y_2$ with unknowns (x_1, x_2). Each pair of equations has two unknowns. In matrix equations the coordinates of the graph are (x_1, x_2) rather than the scalars (x, y). The first time you meet this scrambling of terminology it may be very confusing.

but may be defined as a systematic algorithm when the equations are linear in the variables.[52] Rarely do we even attempt to solve problems in several variables of degree greater than 1. But Gaussian eliminations may still work in such cases (Stillwell, 2010, p. 90).

In Appendix B.2 (p. 174) the inverse of a 2x2 linear system of equations is derived. Even for a 2x2 case, the general solution requires a great deal of algebra. Working out a numeric example of Gaussian elimination is more instructive. For example, suppose we wish to find the intersection of the two equations

$$x - y = 3$$
$$2x + y = 2.$$

This 2x2 system of equations is so simple that you may immediately visualize the solution: By adding the two equations, y is eliminated, leaving $3x = 5$. But doing it this way takes advantage of the specific example, and we need a method for larger systems of equations. We need a generalized (algorithmic) approach. This general approach is called *Gaussian elimination*.

Start by writing the equations in *matrix* format (note this is not in the standard form $Ax = y$)

$$\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}. \tag{1.40}$$

Next, eliminate the lower left term ($2x$) using a scaled version of the upper left term (x). Specifically, multiply the first equation by -2, add it to the second equation, replacing the second equation with the result. This gives

$$\begin{bmatrix} 1 & -1 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 2 - 3 \cdot 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}. \tag{1.41}$$

Note that the top equation did not change. Once the matrix is "upper triangular" (zero below the diagonal) you have the solution. Starting from the bottom equation, $y = -4/3$. Then the upper equation then gives $x - (-4/3) = 3$, or $x = 3 - 4/3 = 5/3$.

In principle, Gaussian elimination is easy, but if you make a calculation mistake along the way, it is very difficult to find the error. The method requires a lot of mental labor, with a high probability of making a mistake. Thus you do not want to apply this method every time. For example suppose the elements are complex numbers, or polynomials in some other variable such as frequency. Once the coefficients become more complicated, the seeming trivial problem becomes highly error prone. There is a much better way, that is easily verified, which puts all the numerics at the end, in a single step.

The above operations may be automated by finding a carefully chosen upper-diagonalization matrix U that does the same operation. For example let

$$U = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}. \tag{1.42}$$

Multiplying Eq. 1.40 by U we find

$$\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix} \tag{1.43}$$

we obtain Eq. 1.41. At this point we can either back-substitute and obtain the solution, as we did above, or find a matrix L that finishes the job, by removing elements above the diagonal.

---

[52]`https://en.wikipedia.org/wiki/System_of_linear_equations`

Exercise: Show that $det(U) = det(UA) = 1$. **Sol:** Since

$$det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc,$$

and $det(UA) = detU \cdot det(A)$, and $det(U) = 1 - 0 = 1$, the result follows naturally. Common notation is to let $det(A) = |A|$.

In Appendix B.2 the inverse of a general 2x2 matrix is summarized in terms of three steps: 1) swap the diagonal elements, 2) reverse the signs of the off-diagonal elements and 3) divide by the determinant $\Delta = ab - cd$. Specifically

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\Delta} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{1.44}$$

There are very few things that you must memorize, but the inverse of a 2x2 is one of them. It needs to be in your tool-bag of tricks, like the quadratic formula.

While it is difficult to compute the inverse matrix from scratch (Appendix B), it takes only a few seconds to verify it (steps 1 and 2)

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad - bc & -ab + ab \\ cd - cd & -bc + ad \end{bmatrix} = \begin{bmatrix} \Delta & 0 \\ 0 & \Delta \end{bmatrix}. \tag{1.45}$$

Thus dividing by the determinant gives the 2x2 identity matrix. A good strategy, when you don't trust your memory, is to write down the inverse as best you can, and then verify.

Using the 2x2 matrix inverse on our example (Eq. 1.40), we find

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{1+2} \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 5 \\ -6 + 2 \end{bmatrix} = \begin{bmatrix} 5/3 \\ -4/3 \end{bmatrix}. \tag{1.46}$$

If you use this method, you will rarely (never) make a mistake, and the solution is easily verified. Either you can check the numbers in the inverse, as was done in Eq. 1.45, or you can substitute the solution back into the original equation.

1.3.7 **Lec 16:** Transmission (ABCD) matrix composition method

In this section we shall derive the method of linear composition of systems, known by several names as the *ABCD Transmission matrix method*, or in the mathematical literature as the Möbius (bilinear) transformation. Using the method of *matrix composition*, a linear system of 2x2 matrices can represent a large and important family of networks. By the application of Ohm's law to the circuit shown in Fig. 1.13, we can model a cascade of such cells.
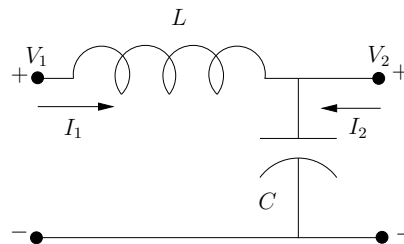

Figure 1.13: *This is a single LC cell, of an LC transmission line of Fig. 2.2 (p. 122). It may be modeled by the ABCD method, as the product of two matrices, as discussed below. The inductance L of the coil and the capacitance C of capacitor are in units of [Henry/m] and [Farad/m], thus they depend on length Δ_x [m] that the cell represents. Note the flows are always defined as into the + node.*

Matrix composition: Matrix multiplication represents a composition of 2x2 matrices, because the input to the second matrix is the output of the first (this follows from the definition of composition: $f(x) \circ g(x) = f(g(x))$). Thus the ABCD matrix is also known as the *transmission* matrix method, or occasionally the *chain matrix*. The general expression for an transmission matrix $\boldsymbol{T}(s)$ is

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}. \tag{1.47}$$

The four coefficients $\mathcal{A}(s), \mathcal{B}(s), \mathcal{C}(s), \mathcal{D}(s)$ are all complex functions of the Laplace frequency $s = \sigma + j\omega$. The derivation is repeated with more detail in Section 3.3.2 (p. 141). It is a standard convention to always define the current into the node, but since the input current (on the left) is the same as the output current on the right (I_2), the sign must be reversed to meet the convention of current into every
₅ node. When transmission matrices are cascaded, all the signs match.

We have already used 2x2 matrix composition in representing complex numbers (p. 24), and for computing the $gcd(m, n)$ of $m, n \in \mathbb{N}$ (p. 34), Pell's equation (p. 39) and the Fibonacci sequence (p. 41).

Definitions of $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$:   The definitions of the four functions of Eq. 5.8 are easily read off of the equation, as

$$\mathcal{A} = \left.\frac{V_1}{V_2}\right|_{I_2=0}, \ \ \mathcal{B} = \left.-\frac{V_1}{I_2}\right|_{V_2=0}, \ \ \mathcal{C} = \left.\frac{I_1}{V_2}\right|_{I_2=0}, \ \ \mathcal{D} = \left.-\frac{I_1}{I_2}\right|_{V_2=0}. \tag{1.48}$$

These definitions follow trivially from Eq. 5.8. These relations have general names. $\mathcal{A}, \mathcal{D}$ are called voltage and current *transfer functions*, since they are the ratio of an output over an input, whereas
₁₀ $\mathcal{B}, \mathcal{C}$ are known as the *transfer impedance* and *transfer conductance*. Each express an output (port 2) in terms of an input input at port 1.

Exercise: Explain why $\mathcal{C}$ is given as above. **Sol:** Writing out the lower equation $I_1 = \mathcal{C}V_2 - \mathcal{D}I_2$. Setting $I_2 = 0$, we obtain the equation for $\mathcal{C}$.

Thévenin parameters of a source: A very important concept in circuit theory is that of the
₁₅ *Thévenin parameters* (p. 157), the *open-circuit voltage* and the *short-circuit current*, the ratio of which define the *Thévenin impedance*. The open circuit voltage is defined as the voltage when the current is zero. In mechanics this is called the *isometric force*, the maximum force given zero velocity.

It trivially follows that the Thévenin voltage (open circuit voltage) is

$$V_{\text{Thév}} = \frac{1}{C} \equiv \left.\frac{V_2}{I_1}\right|_{I_2=0}.$$

For example, the voltage measured across a battery having no load is its Thévenin voltage.

We may find the impedance $Z_1 = \left.\frac{V_1}{I_1}\right|_{I_2=0}$ that would be seen looking into port 1, with nothing connected to port 2 ($I_2 = 0$) by taking the ratio of the upper and lower equations

$$Z_1 = \left.\frac{\mathcal{A}V_2 - \mathcal{B}I_2}{\mathcal{C}V_2 - \mathcal{D}I_2}\right|_{I_2=0} = \frac{\mathcal{A}}{\mathcal{C}} = \mathcal{A}V_{\text{Thév}}.$$

Using this relation we may find a second expression for the Thévenin voltage in terms of the a source voltage V_1 rather than in terms of the source current.

$$V_{\text{Thev}} = \frac{Z_1}{\mathcal{A}} \equiv \left.Z_1 \frac{V_2}{V_1}\right|_{I_2=0}.$$

The choice of the appropriate expression depends on the physics of the actual source.

Example of the use of the ABCD matrix composition: In Fig. 1.13 we see the network is composed of a series inductor (mass) having an impedance $Z_l = sL$, and a shunt capacitor (compliance) having an impedance $Z_c = 1/sC$, where $s \in \mathbb{C}$. As determined by Ohm's Law, each impedance is describe by a linear relation between the current and the voltage. For the inductive impedance, applying Ohm's law, we find

$$V_1 - V_2 = Z_l I_1$$

where Z_l is the complex impedance of the inductor. For the capacitive impedance, applying Ohm's law we find

$$V_2 = (I_1 + I_2)Z_c,$$

where Z_c is the complex impedance of the capacitor.

Each of these linear impedance relations may be written in matrix form. The series inductor equation gives (note $I_1 = -I_2$)

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 & Z_l \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}, \tag{1.49}$$

while the shunt capacitor equation yields (note $V_1 = V_2$)

$$\begin{bmatrix} V_2 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ Y_c & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}, \tag{1.50}$$

where $Y_c = 1/Z_c$ is called the *admittance*.

When the second matrix equation for the shunt admittance (Eq. 1.50) is substituted into the series impedance equation (Eq. 1.49), we find the ABCD matrix ($\boldsymbol{T}_1 \circ \boldsymbol{T}_2$), for the cell is simply the product of two matrices

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 & Z_l \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Y_c & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} 1 + Z_l Y_c & Z_l \\ Y_c & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}. \tag{1.51}$$

Note that the determinant of the matrix $\Delta = \mathcal{A}\mathcal{D} - \mathcal{B}\mathcal{C} = 1$.

Thus $\mathcal{A}(s) = 1 + Z_L u_c = 1 + s^2 LC$, $\mathcal{B}(s) = Z_l$, $\mathcal{C}(s) = Y_c$ and $\mathcal{D} = 1$. This equation characterizes every possible relation between the input and output voltage and current of the cell. 5

For example, the ratio of the output to input voltage with the output unloaded ($I_2 = 0$), known as the *voltage divider relation* may be found from the upper equation with $I_2 = 0$. Writing this out gives

$$\left.\frac{V_2}{V_1}\right|_{I_2=0} = \frac{1}{\mathcal{A}(s)} = \frac{1}{1 + Z_l Y_c} = \frac{Z_c}{Z_c + Z_l}.$$

To derive the formula for the current divider equation, use the lower equation and set $V_2 = 0$.

$$\left.\frac{-I_2}{I_1}\right|_{V_2=0} = \frac{1}{\mathcal{D}} = 1.$$

1.3.8 Lec 17: Riemann Sphere: 3^d extension of chord and tangent method

Once algebra was formulated c830 CE, mathematicians were able to expand beyond the limits placed on it by geometry on the real plane, and the verbose descriptions of each problem in prose (Stillwell, 2010, p. 93). The geometry of Euclid's Elements had paved the way, but after 2000 years, the addition of the language of algebra changed everything. The analytic function was a key development, heavily 10 used by both Newton and Euler. Also the investigations of Cauchy made important headway with his work on complex variables. Of special note was integration and differentiation in the complex plane of complex analytic functions, which is the topic of stream 3.

It was Riemann, working with Gauss in the final years of Gauss' life, who made the breakthrough, with the concept of the *extended complex plane*.[53] This concept was based on the composition of a line with the sphere, similar to the derivation of Euclid's formula for Pythagorean triplets (Fig. 2.3, p. 123). While the importance of the extended complex plane was unforeseen, it changed analytic mathematics forever, along with the physics it supported. It unified and thus simplified many important integrals, to the extreme. This idea is captured by the *fundamental theorem of complex integral calculus* (Table 1.7 p. 74) and **??**, p. **??**).
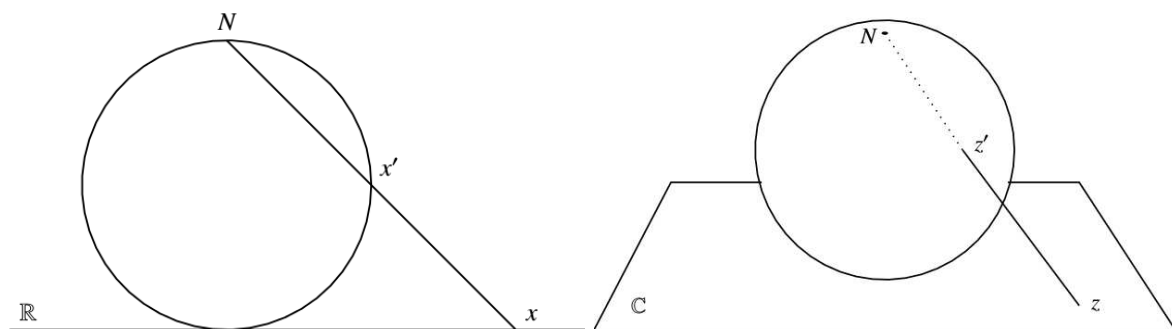
Figure 1.14: *The left panel shows how the real line may be composed with the circle. Each real x value maps to a corresponding point x' on on the unit circle. The point $x \to \infty$ maps to the north pole N. This simple idea may be extended with the composition of the complex plane with the unit sphere, thus mapping the plane onto the sphere. As with the circle, the point on the complex plane $z \to \infty$ maps onto the north pole N. This construction is important because while the plane is open (does not include $z \to \infty$), the sphere is analytic at the north pole. Thus the sphere defines the* closed extended plane. *Figure from Stillwell (2010, pp. 299-300).*

The idea is outlined in Fig. 1.14. On the left is a circle and a line. The difference between this case and the derivation of the Pythagorean Triplets is, that the line starts at the north pole, and ends on the real $x \in \mathbb{R}$ axis at point x. At point x', the line cuts through the circle. Thus the mapping from x to x' takes every point on $\mathbb{R}$ to a point on the circle. For example, the point $x = 0$ maps to the south pole (not indicated). To express x' in terms of x one must compose the line and the circle, similar to the composition used in Fig. 2.3 (p. 123). The points on the circle, indicated here by x', require a traditional polar coordinate system, having a unit radius and an angle defined between the radius and a vertical line passing through the north pole. When $x \to \infty$ the point $x' \to N$, known as the *point at infinity*. But this idea goes much further, as shown on the right half of Fig. 1.14.

Here the real tangent line is replaced by the a tangent complex plane $z \in \mathbb{C}$, and the complex puncture point $z' \in \mathbb{C}$, in this case on the complex sphere, called the *extended complex plane.* This is a natural extension of the chord/tangent method on the left, but with significant consequences. The main difference between the complex plane z and the extended complex plane, other than the coordinate system, is what happens at the north pole. The point at $|z| = \infty$ is not defined on the plane, whereas on the sphere, the point at the north pole is simply another point, like every other point on the sphere.

Open vs. closed sets: Mathematically the plane is said to be an *open set*, since the limit $z \to \infty$ is not defined, whereas on the sphere, the point z' is a member of a *closed set*, since the north pole *is* defined. The distinction between an open and closed set is important, because the closed set allows the function to be complex-analytic at the north pole, which it cannot be on the plane (since the point at infinity is not defined).

The z plane may be replaced with another plane, say the $w = F(z) \in \mathbb{C}$ plane, where w is some function F of $z \in \mathbb{C}$. For the moment we shall limit ourselves to *complex analytic functions* of z, namely $w = F(z) = u(x, y) + v(x, y)\jmath = \sum_{n=0}^{\infty} c_n z^n$.

---

[53]"Gauss did lecture to Riemann but he was only giving elementary courses and there is no evidence that at this time he recognized Riemann's genius." `http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Riemann.html` "In 1849 he [Riemann] returned to Göttingen and his Ph.D. thesis, supervised by Gauss, was submitted in 1851." `http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Riemann.html`

In summary, given a point $z = x + y\jmath$ on the open complex plane, we map it to $w = F(z) \in \mathbb{C}$, the complex $w = u + v\jmath$ plane, and from there to the closed extended complex plane $w'(z)$. The point of doing this is that it allows us to allow the function $w'(z)$ to be analytic at the north pole, meaning it can have a convergent Taylor series at the point at infinity $z \to \infty$. Since we have not yet defined $dw(z)/dz$, the concept of a complex Taylor series remains undefined.

5

Möbius bilinear transformation

In mathematics the *Möbius transformation* has special importance because it is linear in its action. In the engineering literature this transformation is known as the *bilinear transformation.* Since we are engineers we shall stick with the engineering terminology. But if you wish to read about this on the internet, be sure to also search for the mathematical term, which may be better supported.

10

When a point on the complex plane $z = x + y\jmath$ is composed with the bilinear transform ($a, b, c, d \in \mathbb{C}$), the result is $w(z) = u(x, y) + v(x, y)\jmath$

$$w = \frac{az + b}{cz + d} \tag{1.52}$$

the transformation from $z \to w$ is a cascade of four independent compositions

1. translation ($w = z + b$: $a = 1, b \in \mathbb{C}, c = 0, d = 1$),

2. scaling ($w = |a|z$: $a \in \mathbb{R}, b = 0, c = 0, d = 1$)

3. rotation ($w = \frac{a}{|a|}z$: $a \in \mathbb{C}, b = 0, c = 0, d = |a|$) and

4. inversion ($w = \frac{1}{z}$: $a = 0, b = 1, c = 1, d = 0$).

15

Each of these transformations are a special case of Eq. 1.52, with the inversion the most complicated. A wonderful video showing the effect of the bilinear (Möbius) transformation on the plane is available that I highly recommend: https://www.youtube.com/watch?v=0z1fIsUNhO4

When the extended plane (Riemann sphere) is analytic at $z = \infty$, one may take the derivatives there, defining a Taylor series with the expansion point at ∞ and then integrate through ∞. When the bilinear transformation rotates the Riemann sphere, the point at infinity is translated to a finite point on the complex plane, revealing the analytic nature at infinity. A second way to access the point at infinity is by inversion, swapping poles with zeros. Thus a zero at infinity is the same as a pole at the origin.

20

$\frac{dw}{dz}$ **not defined.**

This construction of the Riemann sphere and the M̈bious (bilinear) transformation allow us to understand the point at infinity, and treat it like any other point. If you felt that you never understood the meaning of the point at ∞ (likely), this should help.

25

1.3.9 Lec 18: Complex analytic mappings (Domain-coloring)

One of the most difficult aspects of complex functions of a complex variable is understanding what's going on. For example, $w(z) = \sin(x)$ is trivial when $z = x + y\jmath$ is real (i.e., $y = 0$), because $\sin(x)$ is real. Likewise for the case where $x = 0$, where

$$\sin(y\jmath) = \frac{e^{-y} - e^{y}}{2\jmath} = \jmath \sinh(y)$$

is purely imaginary. But the general case, $w(z) = \sin(z) \in \mathbb{C}$, is not easily visualized. Note that

$$w(z) = \sin(z\jmath) = \sin(x\jmath - y) = \jmath \sinh(z).$$

Thus when $u(z)$ and $v(z)$ are not well known, $w(z)$ are can be much more difficult to visualize.

30

Fortunately with computer software today, this problem can be solved by adding color to the chart. A Matlab/Octave script `zviz.m` has been used to make the make the charts shown here.[54] This tool is known as *Domain-coloring*. Rather than plotting $u(\sigma, \omega)$ and $v(\sigma, \omega)$ separately, domain-coloring allows us to display the entire function on one chart. Note that for this visualization we see the complex polar form of $w(s) = |w|e^{\jmath \angle w}$, rather than a four dimensional Cartesian $w(x + y\jmath) = u(x, y) + v(x, y)\jmath$ graph.

To further complicate the issue, mathematicians typically use the more abstract (i.e., non–physical) notation $w(z)$, where $w = u + v\jmath$ and $z = x + y\jmath$, whereas engineers think in terms of physical complex impedance $Z(s) = R(s) + X(s)\jmath$, having resistance $R(s)$ and reactance $X(s)$ [Ohms], as function of the complex Laplace radian frequency $s = \sigma + \omega\jmath$ [rad], as used, for example, with the Laplace transform (Sect. 1.3.11, p. 70).

In Fig. 1.15 we use both notations, with $Z(s) = s$ on the left and $w(z) = z - \sqrt{\jmath}\,\jmath$ on the right, where we show this color code as a 2x2 dimensional domain-coloring graph. Intensity (dark to light) represents the magnitude of the function, while hue (color) represents the phase, where red is $0°$, sea-green is $90°$, blue-green is $135°$, blue is $180°$, and violet is $-90°$ (or $270°$). The function $w = s$ has a dark spot (a zero) at $s = 0$, and becomes brighter away from the origin. On the right is $w(z) = z - \sqrt{\jmath}$, which shifts the zero to $z = \sqrt{\jmath}$. Thus domain–coloring gives the full picture of the complex analytic function mappings $w(x, y) = u(x, y) + v(x, y)\jmath$ in colorized polar coordinates.
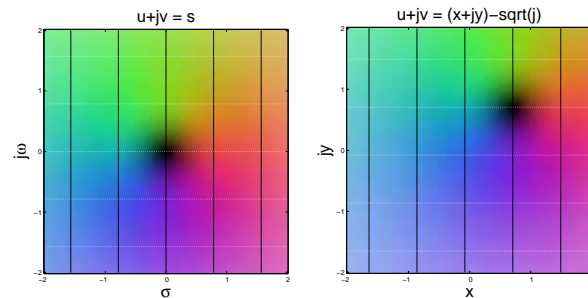

Figure 1.15:   *LEFT: Domain–colorized map showing the complex mapping from the $s = \sigma + \omega\jmath$ plane to the $w(s) = u(\sigma, \omega) + v(\sigma, \omega)\jmath$ plane. This mapping may be visualized by the use of intensity (light/dark) to indicate magnitude, and color (hue) to indicate angle (phase) of the mapping. RIGHT: This shows the $w(z) = z - \sqrt{\jmath}$ plane, shift to the right and up by $\sqrt{2}/2 = 0.707$. The white and black lines are the iso-real and iso-imaginary contours of the mapping.*

Two additional examples are given in Fig. 1.16 to help you interpret the two complex mappings $w = e^s$ (left) and its inverse $s = \ln(w)$. The exponential is relatively easy to understand because $w(s) = |e^\sigma e^{\omega\jmath}| = e^\sigma$. The red region is where $\omega \approx 0$ in which case $w \approx e^\sigma$. As σ becomes large and negative, $w \to 0$, thus the entire field becomes dark on the left. The field is becoming light on the right where $w = e^\sigma \to \infty$. If we let $\sigma = 0$ and look along the ω axis, we see that the function is changing phase, sea-green ($90°$) at the top and violet ($-90°$) at the bottom.

In the right panel note the zero for $\ln(w) = \ln |w| + \omega\jmath$ at $w = 1$. The root of the log function is $\log(w_r) = 0$ is $w_r = 1, \phi = 0$, since $\log(1) = 0$. More generally, the log of $w = |w|e^{\phi\jmath}$ is $s = \ln |w| + \phi\jmath$. Thus $s(w)$ can be zero only when the angle of w is zero.

The $\ln(w)$ function has a *branch cut* along the $\phi = 180°$ axis. As one crosses over the cut, the phase goes above $180°$, and the plane changes to the next *sheet* of the log function. The only sheet with a zero is the principle value, as shown. All others, the log function is either increasing or decreasing monotonically, and there is no zero, as seen on sheet 0 (the one showing in Fig. 1.16).

1.3.10   Lec 19: Signals: Fourier and Laplace transforms

Two basic transformations in engineering mathematics are the Fourier and the Laplace transforms, which deal with time–frequency analysis.

The Fourier transform takes a time domain signal $f(t) \in \mathbb{R}$ and transforms it to the frequency domain $\omega \in \mathbb{R}$, where it is complex ($F(\omega) \in \mathbb{C}$). For the Fourier transform, both the time $-\infty < t < \infty$ and frequency $\infty < \omega < \infty$ are strictly real. The relationship, between $f(t)$ and its transform $F(\omega)$, is indicated by the double arrow symbol

$$f(t) \leftrightarrow F(\omega).$$

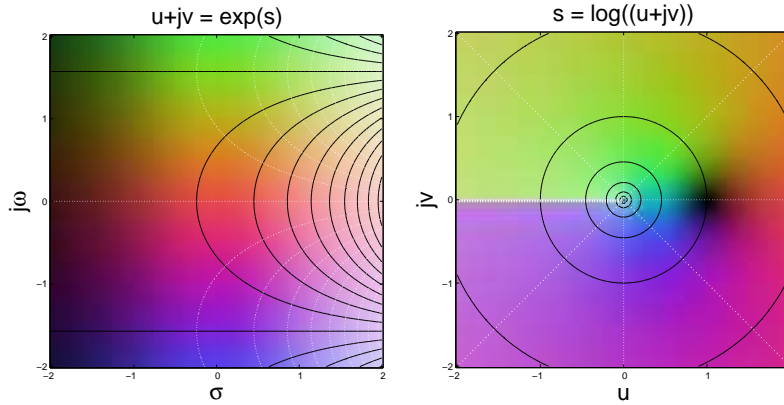[54]URL for `zviz.m`: http://jontalle.web.engr.illinois.edu/uploads/298/zviz.16.m

Figure 1.16: *This domain–color map allows one to visualize complex mappings by the use of intensity (light/dark) to indicate magnitude, and color (hue) to indicate angle (phase). The white and black lines are the iso-real and iso-imaginary contours of the mapping. LEFT: This figure shows the domain–color map for the complex mapping from the $s = \sigma + \omega_J$ plane to the $w(s) = u + v_J = e^{\sigma + \omega_J}$ plane, which goes to zero as $\sigma \to -\infty$, causing the domain–color map to become dark for $\sigma < -2$. The white and black lines are always perpendicular because e^s is complex analytic everywhere. RIGHT: This shows the principal value of the inverse function $\sigma(u,v) + \jmath\omega(u,v) = \log(u + v_J)$, which has a zero (dark) at $s = 1$, since there $\log(1) = 0$ (the imaginary part is zero). Note the branch cut from $w = 0$ to $w = -\infty_J$. On branches other than the one shown, there is are no zeros, since the phase ($\angle w = 2\pi n$) is not zero, except for the principal branch ($n = 0$).*

Since the FT obeys superposition, it is possible to define a FT of a complex time function $f(t) \in \mathbb{C}, t \in \mathbb{R}$. This is useful in certain engineering applications (i.e, Hilbert envelope, Hilbert transforms).

The *Laplace transform* takes a real signals $f(t) \in \mathbb{R}$, as a function of real time $t \in \mathbb{R}$, that are *causal*, i.e., strictly zero for negative time ($f(t) = 0$ for $t < 0$), and transforms them into complex functions $F(s) \in \mathbb{C}$ of complex frequency $s = \sigma + \omega_J$. As for the Fourier transform, there is the notation $f(t) \leftrightarrow F(s)$.

When a signal is zero for negative time $f(t < 0) = 0$, it is said to be *causal*, and the resulting transform $F(s)$ is then complex analytic over significant regions of the s plane. For a function of time to be causal, time *must* be real ($t \in \mathbb{R}$), since if it were complex, it would lose the order property (thus it could not be causal). It is helpful to emphasize the causal nature of $f(t)u(t)$ by indicating the causal nature, using the Heaviside step function $u(t)$.

Restrictions on a function (e.g., real, causal, periodic, positive real part, etc.) are called a *symmetry property*. There are many forms of symmetry (Section 1.3.12, p. 72). The concept of symmetry is very general and widely used in both mathematics and physics, where it is more generally known as *Group theory*. We shall restrict ourselves to only a few very basic symmetries (Section 3.5.1, p. 145).

Periodic signals: Besides these two basic types of time–frequency transforms, there are several variants that depend on the symmetry in time and frequency. For example, when the time signal is sampled (discrete in time), the frequency response becomes periodic, leading to the *Discrete-time Fourier transform* (DTFT). When a time response is periodic, the frequency response is sampled (discrete in frequency), leading to the *Fourier Series*. These two symmetries may be simply characterized as *periodic in time* $\Rightarrow$ discrete in frequency, and *periodic in frequency* $\Rightarrow$ discrete in time. In Section 3.4.2 (p. 144) we shall explain this concept with examples. When a function is both discrete in time and frequency, it is necessarily periodic in time and frequency, leading to the *Discrete Fourier Transform* (DFT). The DFT is typically computed with an algorithm called the *Fast Fourier Transform* (FFT), which can dramatically speed up the calculation when the data is a power of 2 in length.

A very important symmetry is when functions that are causal (in time) *and* periodic (in frequency). The best known example is the class of signals that have *z transforms*, which are causal (one-sided in time) discrete-time signals. The harmonic series (Eq. 1.26, p. 47) is the z-transform of the discrete-time step function, thus, by symmetry, analytic within the RoC in the complex frequency (z) domain.

Summary: The definitions of the FT ($\mathcal{F}$) and LT ($\mathcal{L}$) transforms are superficially similar. The key difference is that the time response of the Laplace transform is causal, leading to a complex analytic frequency responses. The frequency response of the Fourier transform is real, thus typically not analytic. These are not superficial differences. The concept of symmetry is helpful in understanding the many different types of time-frequency transforms. Two fundamental types of symmetry are causality and periodicity.

The $\mathcal{F}$ characterizes the steady-state response while the $\mathcal{L}$ characterizes both the transient and steady- state response. Given a system response $H(s) \leftrightarrow h(t)$ having input $x(t)$, then the output $y(t) = h(t) \star x(t) \leftrightarrow Y(\omega) = H(s)|_{s=\jmath\omega} X(\omega)$.

Table 1.4:   *The following table provides a short table of simple Fourier Transforms. Note $a > 0 \in \mathbb{R}$ has units [rad/s]. To flag this necessary condition, we use $|a|$ to assure this condition will be met. The other constant $T_o \in \mathbb{R}$ [s] has no restrictions, other than being real. Complex constants may not appear as the argument to a delta function, since complex numbers do not have the order property.*

$f(t) \leftrightarrow F(\omega)$	Name
$\widetilde{\delta}(t) \leftrightarrow 1(\omega) \equiv 1 \,\forall\, \omega$	Dirac
$1(t) \equiv 1 \,\forall\, t \leftrightarrow 2\pi\widetilde{\delta}(\omega)$	Dirac
$\mathrm{sgn}(t) = \dfrac{t}{|t|} \leftrightarrow \dfrac{2}{\jmath\omega}$	
$\widetilde{u}(t) = \dfrac{1(t) + \mathrm{sgn}(t)}{2} \leftrightarrow \pi\widetilde{\delta}(\omega) + \dfrac{1}{\jmath\omega} \equiv \widetilde{U}(\omega)$	step
$\widetilde{\delta}(t - T_o) \leftrightarrow e^{-\jmath\omega T_o}$	delay
$\widetilde{\delta}(t - T_o) \star f(t) \leftrightarrow F(\omega)e^{-\jmath\omega T_o}$	delay
$\widetilde{u}(t)e^{-|a|t} \leftrightarrow \dfrac{1}{\jmath\omega + |a|}$	exp
$\mathrm{rec}(t) = \dfrac{1}{T_o}\left[\widetilde{u}(t) - \widetilde{u}(t - T_o)\right] \leftrightarrow \dfrac{1}{T_o}\left(1 - e^{-\jmath\omega T_o}\right)$	pulse
$\widetilde{u}(t) \star \widetilde{u}(t) \leftrightarrow \widetilde{\delta}^2(\omega) \quad$ Not defined	NaN

FT Properties	
$\dfrac{d}{dt}v(t) \leftrightarrow \jmath\omega V(\omega)$	deriv
$f(t) \star g(t) \leftrightarrow F(\omega)G(\omega)$	conv

Definition of the Fourier transform: The forward transform takes $f(t)$ to $F(\omega)$ while the inverse transform takes $F(\omega)$ to $\widetilde{f}(t)$. The tilde symbol indicates that in general recovered inverse transform signal can be slightly different from $f(t)$. We will give examples of this below.

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-\jmath\omega t}dt \qquad\qquad \widetilde{f}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{\jmath\omega t}d\omega \qquad (1.53)$$

$$F(\omega) \leftrightarrow f(t) \qquad\qquad\qquad\qquad \widetilde{f}(t) \leftrightarrow F(\omega). \qquad\qquad (1.54)$$

Properties of Fourier Transforms:

1. Both time t and frequency ω are real.

2. For the forward transform (time to frequency), the sign of the exponential is negative.

3. The limits on the integrals in both the forward and reverse FTs are $[-\infty, \infty]$.

4. When taking the inverse FT (IFT), the normalization factor of $1/2\pi$ is required to cancel the 2π in the differential of the integral $d\omega/2\pi = df$, where f is frequency in [Hz], and ω is the radian frequency [rads].

5. The Fourier step function may be defined by the use of superposition of 1 and $\text{sgn}(t) = t/|t|$ as

$$\widetilde{u}(t) \equiv \frac{1 + \text{sgn}(t)}{2} = \begin{cases} 1 & \text{if } t > 0 \\ 1/2 & t = 0 \\ 0 & \text{if } t < 0 \end{cases}.$$

   Taking the FT of a delayed step function

$$\widetilde{u}(t - T_o) \leftrightarrow \frac{1}{2}\int_{-\infty}^{\infty} \left[1 - \text{sgn}(t - T_o)\right] e^{-j\omega t} dt = \pi\tilde{\delta}(\omega) + \frac{e^{-j\omega T_o}}{j\omega}$$

   Thus the FT of the step function has the term $\pi\delta(\omega)$ due to the 1 in the definition of the Fourier step. This term introduces a serious flaw with the FT of the step function: While it appears to be causal, it is not.

6. The convolution $\widetilde{u}(t) \star \widetilde{u}(t)$ is not defined because both $1 \star 1$ and $\widetilde{\delta}^2(\omega)$ do not exist (and cannot be defined).

7. The inverse FT has convergence problems whenever there is a discontinuity in the time response. This we indicate with a hat over the reconstructed time response. The error between the target time function and the reconstructed is zero in the root-mean sense, but not point-wise.

   Specifically, at the discontinuity point for the Fourier step function ($t = 0$), $\widetilde{u}(t) \neq u(t)$, yet $\int |\widetilde{u}(t) - u(t)|^2 dt = 0$. At the point of the discontinuity the reconstructed function displays *Gibbs ringing* (it oscillates around the step, hence does not converge at the jump).[55] The $\mathcal{L}$ does not exhibit Gibbs ringing. More about this in Section 3.4.2.

8. The FT is not always analytic in ω, as in this example of the step function. The step function cannot be expanded in a Taylor series about $\omega = 0$, because $\widetilde{\delta}(\omega)$ is not analytic in ω.

9. The Fourier δ function is denoted $\widetilde{\delta}(t)$, to differentiate it from the Laplace delta function $\delta(t)$. They differ because the step functions differ, due to the convergence problem described above.

10. One may define

$$\widetilde{u}(t) = \int_{-\infty}^{t} \widetilde{\delta}(t) dt,$$

   and define the somewhat questionable notation

$$\widetilde{\delta}(t) =: \frac{d}{dt}\widetilde{u}(t),$$

   since the Fourier step function is not analytic.

---

[55]https://en.wikipedia.org/wiki/Gibbs_phenomenon

11. The rec(t) function is defined as

$$\text{rec}(t) = \frac{\widetilde{u}(t) - \widetilde{u}(t - T_o)}{T_o} = \begin{cases} 0 & \text{if } t > 0 \\ 1/T_o & 0 < t < T_o \\ 0 & \text{if } t < 0 \end{cases}.$$

It follows that $\widetilde{\delta}(t) = \lim_{T_o \to 0}$. Like $\widetilde{\delta}(t)$, the rec(t) has unit area.

1.3.11   Lec 20: Laplace transforms

When dealing with engineering problems it is convenient to separate the *signals* we use from the *systems* that process them. We do this by treating signals, such as a music signal, differently from a system, such as a filter. In general signals may start and end at any time. The concept of causality has no mathematical meaning in signal space. Systems, on the other hand, obey very rigid rules (to assure that they remain physical). These physical restrictions are described in terms of the *Network Postulates*, which are briefly discussed in Sect. 1.3.12, and in greater detail in Sect. 3.5.1.

Definition of the Laplace transform: The forward and inverse Laplace transforms are

$$F(s) = \int_{0^-}^{\infty} f(t)e^{-st}dt \qquad\qquad f(t) = \frac{1}{2\pi\jmath} \int_{\sigma_o - \infty\jmath}^{\sigma_o + \infty\jmath} F(s)e^{st}ds \qquad (1.55)$$

$$F(s) \leftrightarrow f(t) \qquad\qquad f(t) \leftrightarrow F(s) \qquad (1.56)$$

1. Time $t \in \mathbb{R}$ and *Laplace frequency* is defined as $s = \sigma + \omega\jmath \in \mathbb{C}$.

2. When taking the forward transform ($t \to s$), the sign of the exponential is negative. This is necessary to assure that the integral converges when the integrand $f(t) \to \infty$ as $t \to \infty$. For example, if $f(t) = e^t u(t)$ (i.e., without the negative σ exponent), the integral does not converge.

3. The target time function $f(t < 0) = 0$ (i.e., it must be causal). The time limits are $0^- < t < \infty$. Thus the integral must start from slightly below $t = 0$ to integrate over a delta functions at $t = 0$. For example if $f(t) = \delta(t)$, the integral must include both sides of the impulse. If you wish to include non-causal functions such as $\delta(t + 1)$ it is necessary to extend the lower time limit. In such cases simply set the lower limit of the integral to $-\infty$, and let the integrand ($f(t)$) determine the limits.

4. The limits on the integrals of the forward transform are $t : (0^-, \infty) \in \mathbb{R}$, and the reverse LTs are $[\sigma_o - \infty\jmath, \sigma_o + \infty\jmath] \in \mathbb{C}$. These limits will be further discussed in Section 1.4.9 (p. 87).

5. When taking the inverse FT (IFT), the normalization factor of $1/2\pi\jmath$ is required to cancel the $2\pi\jmath$ in the differential ds of the integral.

6. The frequency for the LT must be is complex, and in general $F(s)$ is complex analytic for $\sigma > \sigma_o$. It follows that the real and imaginary parts of $F(s)$ are related. Given $\Re F(s)$ it is possible to find $\Im F(s)$ (Boas, 1987). More on this in Section 3.4.2 (p. 144).

7. To take the inverse Laplace transform, we must learn how to integrate in the complex s plane. This will be explained in Sections 1.4.5-1.4.9 (p. 84-87).

8. The Laplace step function is defined as

$$u(t) = \int_{-\infty}^{t} \delta(t)dt = \begin{cases} 1 & \text{if } t > 0 \\ \text{NaN} & t = 0 \\ 0 & \text{if } t < 0 \end{cases}.$$

Alternatively one could define $\delta(t) = du(t)/dt$.

Table 1.5:   *The following table provides a brief table of useful Laplace Transforms. Assume that: $f(t), \delta(t),$ $u(t), \text{rect}(t), T_o, p, e, \in \mathbb{R}$ and $F(s), G(s), s, a \in \mathbb{C}$. Note that: K_1 is the Bessel function of the second kind (i.e., J_o is of the first kind).*

$f(t) \leftrightarrow F(s)$	Name
$\delta(t) \leftrightarrow 1$	Dirac
$\delta(\lvert a \rvert t) \leftrightarrow \lvert a \rvert$	$a \neq 0;$ Time-scaled Dirac
$\delta(t - T_o) \leftrightarrow e^{-sT_o}$	delay
$\delta(t - T_o) \star f(t) \leftrightarrow F(s)e^{-sT_o}$	–
$u(t) \leftrightarrow \dfrac{1}{s}$	pure step
$e^{-at}u(t) \leftrightarrow \dfrac{1}{s+a}$	Damped step
$u(at) \leftrightarrow \dfrac{a}{s}$	$a > 0;$ dilate
$u(t - T) \leftrightarrow \dfrac{1}{s}e^{-sT}$	time-shift
$u(t) \star u(t) = tu(t) \leftrightarrow 1/s^2$	ramp
$u(t) \star u(t) \star u(t) = \dfrac{1}{2}t^2 u(t) \leftrightarrow 1/s^3$	Triple convolution
$t^p u(t) \leftrightarrow \dfrac{\Gamma(p+1)}{s^{p+1}}$	$p \in \mathbb{R} \geq 0$
$\dfrac{1}{\sqrt{\pi t}}u(t) \leftrightarrow \dfrac{1}{\sqrt{s}}$	fractional integral
$e^{\mp at}u(t) \leftrightarrow \dfrac{1}{s \pm a},\qquad a \in \mathbb{C}$	modulate/damped
$\text{rect}(t) = \dfrac{1}{T_o}\left[u(t) - u(t - T_o)\right] \leftrightarrow \dfrac{1}{T_o}\left(1 - e^{-sT_o}\right)$	pulse
$\displaystyle\sum_{n=0}^{\infty}\delta(t - nT_o) = \dfrac{1}{1 - \delta(t - T_o)} \leftrightarrow \dfrac{1}{1 - e^{-sT_o}} = \sum_{n=0}^{\infty}e^{-snT_o}$	$--$
$J_o(at)u(t) \leftrightarrow \dfrac{1}{s^2 + a^2}$	Bessel
$J_1(t)u(t)/t \leftrightarrow \sqrt{s^2 + 1} - s$	
$J_1(t)u(t)/t + 2u(t) \leftrightarrow \sqrt{s^2 + 1} + s = e^{\sinh^{-1}(s)}$	
$\delta(t) + J_1(t)u(t)/t \leftrightarrow \sqrt{s^2 + 1}$	
$I_o(t)u(t) \leftrightarrow 1/\sqrt{s^2 - 1}$	
$u(t)/\sqrt{t+1} \leftrightarrow e^s\sqrt{\dfrac{\pi}{s}}\text{erfc}(\sqrt{s})$	
$\sqrt{t}u(t) * \sqrt{1 + t}u(t) \leftrightarrow e^{s/2}K_1(s/2)/2s$	

Table 1.6: *Functional relationships between Laplace Transforms.*

$\mathcal{L}$ functional properties

$$f(at)u(at) \leftrightarrow aF(as) \quad \text{Verify} \quad a \in \mathbb{R} \neq 0 \qquad \text{scaling}$$

$$f(t)e^{-at}u(t) \leftrightarrow F(s+a) \qquad \text{damped}$$

$$f(t-T)e^{-a(t-T)}u(t-T) \leftrightarrow e^{-sT}F(s+a) \qquad \text{damped and delayed}$$

$$f(-t)u(-t) \leftrightarrow F(-s) \qquad \text{reverse time}$$

$$f(-t)e^{-at}u(-t) \leftrightarrow F(a-s) \qquad \text{time-reversed \& damped}$$

$$\frac{d}{dt}f(t) = \delta'(t) \star f(t) \leftrightarrow sF(s) \qquad \text{deriv}$$

$$f(t) \star g(t) = \int_{t=0}^{t} f(t-\tau)g(\tau)d\tau \leftrightarrow F(s)G(s) \qquad \text{conv}$$

$$u(t) \star f(t) = \int_{0^-}^{t} f(t)dt \leftrightarrow \frac{F(s)}{s} \qquad \text{conv}$$

9. It is easily shown that $u(t) \leftrightarrow 1/s$ by direct integration

$$F(s) = \int_0^\infty u(t)\,e^{-st}dt = -\left.\frac{e^{-st}}{s}\right|_o^\infty = \frac{1}{s}.$$

With the LT step $(u(t))$ there is no Gibbs ringing effect.

10. In many physical applications, the Laplace transform takes the form of a ratio of two polynomials. In such case the roots of the numerator polynomial are call the *zeros* while the roots of the denominator polynomial are called the *poles*. For example the LT of $u(t) \leftrightarrow 1/s$ has a pole at $s = 0$, which represents integration, since

$$u(t) \star f(t) = \int_{-\infty}^{r} f(\tau)d\tau \leftrightarrow \frac{F(s)}{s}.$$

11. The LT is quite different from the FT in terms of its analytic properties. For example, the step function $u(t) \leftrightarrow 1/s$ is complex analytic everywhere, except at $s = 0$. The FT of $1 \leftrightarrow 2\pi\widetilde{\delta}(\omega)$ is not analytic anywhere.

Once complex integration in the complex plane has been defined (Section 1.4.2, p. 77), we can justify the definition of the inverse LT (Eq. 1.55).

1.3.12 Lec 21: Ten network postulates

Solutions of differential equations, such as the wave equation, are conveniently described in terms of mathematical properties, which we present here in terms of 10 *network postulates*:

(P1) *causality* (non-causal/acausal): Causal systems respond when acted upon. Virtually all physical systems obey causality. An example of a causal system is an integrator, which has a response of a step function. Filters are also examples of causal systems. Signals represent acausal responses. They do not have a clear beginning or end, such as the sound of the wind or traffic noise.

(P2) *linearity* (nonlinear): Linear systems obey superposition. If two signals $x(t)$ and $y(t)$ are the inputs to a linear system, producing outputs $x'(t)$ and $y'(t)$, then if the inputs are presented together as $ax(t) + by(t)$, with weights $a, b \in \mathbb{R}$, then the output will be $ax'(t) + by'(t)$. If either a, b are zero, that signal is removed from the output.

Nonlinear system mix two inputs, thereby producing other signals not present in the input. For example, if the inputs to a nonlinear system are two sine waves, the output will contain distortion components, having frequencies not present at the input. An example of a nonlinear system is one that multiplies the two inputs. A second is a diode, which rectifies a signal, letting current flow only in one direction. Most physical systems have some degree of nonlinear response, but this is not always desired. Other systems are designed to be nonlinear, such as the diode example. ₁₀

(P3) *passive* (active): An active system has a power source, such as a battery while a passive system has no power source. While you may consider a transistor amplifier to be active, it is only so when connected to a power source.

(P4) *real* (complex) time response : Typically systems are "real in, real out." They do not naturally have complex responses (real and imaginary parts). While a Fourier transform takes real inputs ₁₅ and produces complex outputs, this is not an example of a complex time response. P4 is a characterization of the input signal, not its Fourier transform.

(P5) *time-invariant* (time varying): For a system to be time varying system its properties mus depend on when the signal starts and stops. If the output, relative to the input, is independent of the starting time, then the system is time-invariant. ₂₀

(P6) *reciprocal* (non- or anti-reciprocal): In many ways this is the most difficult propriety to understand. It is best characterized by the ABCD matrix. If $B = C$ it is said to be reciprocal. If $B = -C$ it is said to be anti-reciprocal. The loudspeaker is anti-reciprocal, which is why it is modeled by the gyrator rather than a transformer. All non-reciprocal systems are modeled by such gyrator, which swap the force and flow variables. In some ways this property is beyond the ₂₅ scope of this book.

(P7) *reversibility* (non-reversible): If the system can be flipped, between input and output, the it is said to be reversible. In terms of the ABCD parameters, if $A = D$ is is reversible.

(P8) *space-invariant* (space-variant): If a system operates independently as a function of where it physically is in space, then it is space-invariant. When the parameters that characterize the ₃₀ system depend on position, it is space-variant.

(P9) *Deterministic* (random): Predictability depends on the light cone. Given the wave equation, along with the boundary conditions, the solution may be determinisitc, or not, depending on the extent of the "light cone." Consider the wave with a wave propagating out into unchartered teritory. When the wave hits an object, the refletion will return a wave that might considered ₃₅ not be predicted. This is an examle of the boundary condition being the radiation impedance, which is not known in advance. One must do the experiment to do solve the problem, and the radiation impedance is not known in advance. To determine it, one must do the experiment. In this example, that determination could take light years. Thus is, I believe, an example of a "random" system. But the system is deterministic only if the radiation impedance is known ₄₀ accurately, in advance. To define the impedance requires that it be time invariant (P5). One cannot look beyond the light-cone, which depends on the system to be time invariant.

(P10) *quasi-static* (multi-modal): Quasi-statics follows from systems that are small compared to the wavelength. This is a very general assumption, that must be false when the frequency is raised and the wavelength becomes short. Thus this is also known as the *long-wavelength* approximation. ₄₅ It is a very powerful tool in modeling systems such as transmission lines.

Each postulate has two (or more) categories. For example for (P2) a system is either linear or non-linear and for (P1) is either causal, non-causal or acausal. P6 and P9 only apply to 2-port networks (those having an input and an output). The others can apply to both a 2– or 1–port networks (e.g., an impedance is a 1-port).

Examples.

5   Related forms of these postulates had been circulating in the literature for many years, widely accepted in the network theory literature (Van Valkenburg, 1964a,b; Ramo et al., 1965). The first six of these were formally introduced Carlin and Giordano (1964), and (P7-P9) were added by Kim et al. (2016).

1.3.13   Lec 22: Exam II (Evening Exam)

10   ## 1.4   Stream 3: Scalar (Ordinary) Differential Equations

Stream 3 is ∞, a concept which typically means unbounded (immeasurably large), but in the case of calculus, ∞ means *infinitesimal* (immeasurably small), since taking a limit requires small numbers. Taking a limit means you may never reaching the target, a concept that the Greeks called *Zeno's paradox* (Stillwell, 2010, p. 76).

15   When speaking of the class of *ordinary* (versus *vector*) differential equations, the term *scalar* is preferable, since the term "ordinary" is vague, if not a meaningless label. There are a special subset of fundamental theorems for scalar calculus, all of which are about integration, as summarized in Table 1.7, starting with Leibniz's theorem. These will be discussed below, and more extensively in Chapters 4 and 5.

Table 1.7:   *Summary of the fundamental theorems of integral calculus, each of which deals with integration. There are at least two main theorems related to scalar calculus, and three more for vector calculus.*

Name	Mapping	p.	Description
Leibniz	$\mathbb{R}^1 \to \mathbb{R}^0$	76	Area under a real curve.
Cauchy	$\mathbb{C}^1 \to \mathbb{C}^0$	76	Residue integration and complex analytic functions.
Gauss's Law	$\mathbb{R}^3 \to \mathbb{R}^2$	106	Conservation of mass and charge crossing a closed surface.
Stokes	$\mathbb{R}^3 \to \mathbb{R}^1$	107	Line integral $\equiv \boldsymbol{\nabla} \times$ flux crossing an open surface $\in \mathbb{R}^3$.
Green	$\mathbb{R}^2 \to \mathbb{R}^0$	78	Special case of Stokes for a plane
Helmholtz		108	Vector decomposed as dilatation and rotation.

20   Following the integral theorems on scalar calculus, are those on vector calculus, without which there could be no understanding of Maxwell's equations. Of these, the *fundamental theorem of complex calculus* (aka, Helmholtz decomposition), Gauss's Law and Stokes theorem, form the corner stone of modern vector field analysis. These theorems allow one to connect the differential (point) and macroscopic (integral) relationships. For example, Maxwell's equations may be written as either vector

25   differential equations, as shown by Heaviside (along with Gibbs and Hertz),[56] or in integral form. It is helpful to place these two forms side-by-side, to fully appreciate their significance. To understand the differential (microscopic) view, one must understand the integral (macroscopic) view. These are presented in Sections 1.124 (p. 105) and Fig. 1.24 (p. 107).

The beginning of modern mathematics

30   The birth of mathematics as we know it today, occurred during the 16th to 18th centuries, perhaps starting with Galileo, Descartes, Fermat, Newton, the Bernoulli family, and Euler, as outlined in Fig. 1.17 (p. 75). Galileo was formidable, due to his fame, fortune, and his "successful" stance against

---

[56]`https://en.wikipedia.org/wiki/History_of_Maxwell\%27s_equations`

Chronological history post 16th century

16th Bombelli *1526-1572*

17th *Galileo 1564-1642, Kepler 1571-1630, Newton 1642-1727 Principia 1687; Mersenne; Huygen; Pascal; Fermat, Descartes (analytic geometry); Bernoullis Jakob, Johann & son Daniel*

18th *Euler 1707-1783 Student of Johann Bernoulli; d'Alembert 1717-1783; Kirchhoff; Lagrange; Laplace; Gauss 1777-1855*

19th *Möbius, Riemann 1826-1866,*
       Cauchy 1789-1857, Helmholtz 1821-1894, Maxwell 1831-1879, Heaviside 1850-1925, Rayleigh 1842-1919

20th *Hilbert; Einstein; . . .*

Time Line
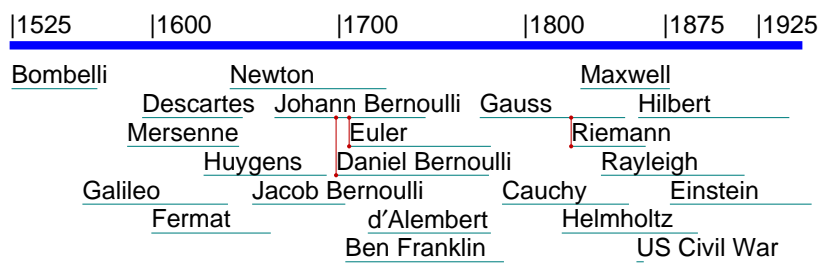
Figure 1.17: *Time-line of the four centuries from the 16th and 20th CE. Given the time line, it is likely that Bombelli's discovery of Diophantus' book on "Arithmetic" in the Vatican library, triggered many of the ideas presented by Descartes and Fermat, followed by others (i.e., Newton). Bombelli's discovery might be considered as a magic moment in mathematics. The vertical red lines connect mentor-student relationships. Figure 1.4 (p. 19) for a closer look at the 15-18c, and Fig. 1.11 (p.45) for a closer look at the 18-20c. . [fig:TimeLineCE].*

the powerful Catholic establishment. His creativity in scientific circles was certainly well known due to his many skills and accomplishments. Fermat and Descartes were at the forefront of merging algebra and geometry, but while Fermat kept meticulous notebooks, he did not publish what he knew, and tended to be secretive.

Regarding the development of calculus, much was yet to be done by Newton and Leibniz, using term by term integration of functions, based on a Taylor series representation. This was a powerful technique, but as stated, incomplete, because the series is only valid for single-valued functions, within the RoC. But more importantly, Newton (and others) failed to recognize the powerful generalization to complex analytic functions. The first breakthrough was Newton's publication of Principia (1687), and a second was Riemann (1851), with his thesis, advised by Gauss.

Following Newton's lead, the secretive and introverted behavior of the typical mathematician dramatically changed with the Bernoulli family (Fig. 1.17, p. 75). The oldest brother, Jacob, taught his much younger brother Johann, who then taught his son Daniel, and Johann's star pupil, Euler. Euler first mastered all the tools and then published, with a prolific voracity previously unknown.

Euler and the circular functions: The first major task was to understand the family of analytic circular functions, $e^x, \sin(x), \cos(x)$, and $\log(x)$, a task begun by the Bernoulli family, but mastered by Euler. Euler sought relations between these many functions, some of which may not be thought of as being related, such as the log and sin functions. The connection that may "easily" be made is through their Taylor series (Eq. 1.25, p. 47). By the manipulation of the analytic series representations, the relationship between e^x, and the $\sin(x)$ and $\cos(x)$ was precisely captured by

$$e^{x\jmath} = \cos(x) + \jmath \sin(x),$$

and its analytic inverse (Greenberg, 1988, p. 1135)

$$\ln(x) = \tan^{-1}\left(\frac{x - 1\jmath}{x + 1\jmath}\right).$$

These not only relate the exponential function to the circular functions, but also contains $\jmath = \sqrt{-1}$. While every high school student has learned Euler's relation, it is doubtful (i.e., rare) that they fully appreciate the significance of *complex analytic functions* (Eq. 1.29, p. 49). While Euler found these relationships using real power series (i.e., analytic functions), neither he, nor those who followed, seemed to fully appreciated the importance of complex analytic, and its potential role. For example, Newton famously ignored imaginary numbers, and called them imaginary in a disparaging (pejorative) way. Given Newton's prominence, that certainly must have attenuated any interest in any complex algebra, even though it had been previously quantified by Bombelli in 1525, likely based on his serendipitous finding of Diophantus' book *Arithmetic* in the Vatican library.

While Euler was fluent with $\jmath = \sqrt{-1}$, he did not consider functions to be complex analytic. That concept was first explored by Cauchy, almost a century later. The missing link to the concept of complex analytic is the definition of the derivative with respect to the complex argument (i.e., $dF(s)/ds$ with $s = \sigma + \omega\jmath$), without which the complex analytic Taylor coefficients may not be defined.

1.4.1   Lec 23: Fundamental theorems of calculus

History of the fundamental theorem of calculus: It some sense, the story of calculus begins with the *fundamental theorem of calculus* (FTC), also known generically as *Leibniz's formula*. The simplest integral is the length of a line $L = \int_0^L dx$. If we label a point on a line as $x = 0$ and wish to measure the distance to any other point x, we form the line integral between the two points. If the line is straight, this integral is simply the Euclidean length given by the difference between the two ends (Eq. 1.35, p. 56).

If $F(x)$ describes a height above the line, then the area under $F(x)$ is

$$f(x) - f(0) = \int_{x=0}^{x} F(\chi)d\chi, \tag{1.57}$$

where χ is a dummy variable of integration. Thus the area under $F(x)$ ($F(x), x \in \mathbb{R}$) only depends on the difference of the area evaluated at the end points. This makes intuitive sense, by viewing $f(x)$ as the area of a graph of $F(x)$.

This propriety of the area as an integral only depending on the end points, has important consequences in physics in terms of conservation of energy. Thus generalizations are important.

Fundamental theorems of real calculus: If $F(x)$ is analytic (Eq. 1.24, p. 47), then

$$F(x) = \frac{d}{dx}f(x), \tag{1.58}$$

which is known as the *Fundamental theorem of (real) calculus* (FTC). Thus Eq. 1.58 may be viewed as an *anti-derivative*, or *exact differential*. This is easily shown by evaluating

$$\frac{d}{dx}f(x) = \lim_{\delta \to 0}\frac{f(x + \delta) - f(x)}{\delta} = F(x),$$

starting from Eq. 1.57. If $F(x)$ is not analytic then the limit may not exist, so this seems like a necessary condition. There are many important variations on this very basic theorem (see Sect. **??**).

Exercises: Generate some exercises using Matlab's `root(), poly(), polyval(), residue(), conv(),` manipulating roots, polynomials and residues on page 46.

Case of a Complex path The *fundamental Theorem of complex calculus* (FTCC) states that for any *complex analytic function $F(s) \in \mathbb{C}$ with $s = \sigma + \omega\jmath \in \mathbb{C}$*

$$f(s) - f(s_o) = \int_{s_o}^{s} F(\zeta)d\zeta. \tag{1.59}$$

(Greenberg, 1988, p. 1197). If we compare this to Eq. 1.58, they differ in that the path of the integral is complex, meaning that the integral is over $s \in \mathbb{C}$, rather than a real integral over $x \in \mathbb{R}$. The *Fundamental theorems of complex calculus* (FTCC) states that the integral only depends on the end points, since

Verify

$$F(s) = \frac{d}{ds}f(s). \tag{1.60}$$

Comparing Eq. 1.58 (FTC) with Eq. 1.60 (FTCC), it would appear that this can only be true if $F(s) \in \mathbb{C}$ is *complex analytic,* meaning it must have a Taylor series in powers of $s \in \mathbb{C}$.

Complex analytic functions: The definition of a *complex analytic function $F(s)$ of $s \in \mathbb{C}$* is that the function may be expanded in a Taylor series (Eq. 1.29, p. 49) about an *expansion point $s_o \in \mathbb{C}$.* This definition follows the same logic as the FTC. Thus we need a definition for the coefficients $c_n \in \mathbb{C}$, which most naturally follow from Taylor's formula

$$c_n = \frac{1}{n!}\frac{d^n}{ds^n}F(s)\bigg|_{s=s_o}. \tag{1.61}$$

The requirement that $F(s)$ have a Taylor series naturally follows by taking derivatives with respect to s at s_o. *The problem is that both integration and differentiation of functions of complex Laplace frequency $s = \sigma + \omega\jmath$ have not yet been defined.*

    Thus the question is: "What does it mean to take the derivative of a function $F(s) \in \mathbb{C}$, $s = \sigma + \omega\jmath \in \mathbb{C}$, with respect to s, where s defines a plane rather than a real line?" We learned how to form the derivative on the real line. Can the same derivative concept be extended to the complex plane?

    The answer is affirmative. The question may be resolved by applying the rules of the real derivative when defining the derivative in the complex plane. However for the complex case, there is an issue, regarding direction. Given any analytic function $F(s)$, is the partial derivative with respect to σ different from the partial derivative with respect to $\omega\jmath$? For complex analytic functions, the FTCC states that the integral is independent of the path in the s plane. *Based on the chain rule, the derivative must also be independent of direction at s_o.* This directly follows from the FTCC. *If the integral of a function of a complex variable is to be independent of the path, the derivative of a function with respect to a complex variable must be independent of the direction.* This follows from Taylor's formula, Eq. 1.61 for the coefficients of the complex analytic formula. The FTC is an integral over an area. The FTCC is an integral over a complex function along a complex path. Both relate an area under an integral. But what is the meaning of "area" for the complex line integral? The *Cauchy-Riemann conditions* provide the answer.

1.4.2 Lec 24: Cauchy-Riemann conditions

For the integral of $Z(s) = R(\sigma,\omega) + X(\sigma,\omega)\jmath$ to be independent of the path, the derivative of $Z(s)$ must be independent of the direction of the derivative. As we show next, this leads to a pair of equations known as the *Cauchy-Riemann* conditions. This is an important generalization of Eq. 1.1, p. 15 which goes from real integration ($x \in \mathbb{R}$) to complex integration ($s \in \mathbb{C}$), based on lengths, thus on area.

    Taking partial derivatives of $Z(s)$, with respect to σ and $\jmath\omega$, and then equating the real and imaginary parts, gives

$$\text{CR-1: } \frac{\partial R(\sigma,\omega)}{\partial\sigma} = \jmath\frac{\partial X(\sigma,\omega)}{\partial\omega\jmath} \quad \text{and} \quad \text{CR-2: } \frac{\partial R(\sigma,\omega)}{\partial\omega\jmath} = \jmath\frac{\partial X(\sigma,\omega)}{\partial\sigma}. \tag{1.62}$$

$\jmath$ cancels in CR-1, but introduces a $\jmath^2 = -1$ in CR-2. These equations are known as the Cauchy-Riemann (CR) conditions. They may also be written in polar coordinates.

The CR conditions are necessary conditions that the integral of $Z(s)$, and thus its derivative, be independent of the path, expressed in terms of conditions on the real and imaginary parts of Z. This is a very strong condition on $Z(s)$, which follows assuming that $Z(s)$ may be approximated by a Taylor series in s

$$Z(s) = Z_o + Z_1 s + \frac{1}{2} Z_2 s^2 + \cdots, \tag{1.63}$$

where $Z_n \in \mathbb{C}$ are complex constants given by the Taylor series formula (Eq. 1.61, p. 77).

Every function that may be expressed as a *Taylor series* about a point s is said to be *complex analytic* at that point. This series, which is single valued, is said to converge within a *radius of convergence* (RoC). This highly restrictive condition has significant physical consequences. As an important example, every impedance function $Z(s)$ obeys the CR conditions over large regions of the s plane, including the entire right half plane (RHP) ($\sigma > 0$). This condition is summarized by the Brune condition $\Re Z(\sigma > 0) \geq 0$ (Eq. 1.70, Section 1.4.3). indexBrune

When this condition is generalized to volume integrals, it is called *Green's theorem*, which is a special case of both Gauss's and Stokes's Laws (Greenberg, 1988, p. 819), used heavily in the solution of boundary value problems in engineering and physics (e.g., solving problems that begin with Maxwell's equations). The last chapter of this course shall depend heavily on these concepts.

We may merge these equations into a pair of second order equations by taking a second round of partials. Specifically, eliminating the real part $R(\sigma, \omega)$ of Eq. 1.62 gives

$$\frac{\partial^2 R(\sigma, \omega)}{\partial \sigma \partial \omega} = \frac{\partial^2 X(\sigma, \omega)}{\partial^2 \omega} = -\frac{\partial^2 X(\sigma, \omega)}{\partial^2 \sigma}, \tag{1.64}$$

which may be compactly written as $\nabla^2 X(\sigma, \omega) = 0$. Eliminating the imaginary part gives

$$\frac{\partial^2 X(\sigma, \omega)}{\partial \omega \partial \sigma} = \frac{\partial^2 R(\sigma, \omega)}{\partial^2 \sigma} = -\frac{\partial^2 R(\sigma, \omega)}{\partial^2 \omega}, \tag{1.65}$$

may be written as $\nabla^2 R(\sigma, \omega) = 0$.

In summary, for a function $Z(s)$ to be analytic, the derivative dZ/ds must be independent of direction (path), which requires that the real and imaginary parts of the function obey Laplace's equation, i.e.,

$$\nabla^2 R(\sigma, \omega) = 0 \quad \text{and} \quad \nabla^2 X(\sigma, \omega) = 0. \tag{1.66}$$

The CR equations are easier to work with because they are first order, but the physical intuition is best understood by noting that the real and imaginary parts of the function each obey Laplace's equation. As you learn about solutions to Laplace's equation, this will become more clear.

As we shall see in the next few lectures, analytic functions must be very smooth in magnitude, like and area or volume, due to the above condition. They must be maximum on their boundary. For example, when you stretch a rubber sheet over a jagged form the height sheet obeys Laplace's equation. Nowhere does the height of the sheet rise above the value at the boundary. Such functions are very smooth, having no second derivatives.

Such functions define *Conservative fields*, which means that energy (like an area) is conserved. The work done in moving a mass from a to b in such a field is conserved. If you return the mass from b back to a, the energy is retrieved, and zero net work has been done.

1.4.3   Lec 25: Complex Analytic functions and Brune Impedance

It is rarely stated that the variable that we are integrating over, either x (space) or t (time), is real $x, t \in R$, since that fact is implicit, due to the physical nature of the formulation of the integral. But this intuition must be refined once complex numbers are included with $s \in \mathbb{C}$, where $s = \sigma + \omega \jmath$, as first developed by Cauchy (1789-1857), with the aid of Gauss (1777-1855).

The fact that time and space are real variables is more than an assumption, rather it is a requirement, due to the *order property*. Real numbers have order. For example, if $t = 0$ is now (the present), then $t < 0$ is the past and $t > 0$ is the future. The order property of time and space allows one to order these along a real axis. If the axis were complex, as in frequency s, the order property is lost. It follows that if we desire order, time and space must be real ($t, x \in \mathbb{R}$). Interestingly, it has was shown by d'Alembert (1747) that time and space are related by the wave speed c. To obtain a solution to the governing wave equation, that Newton first proposed for sound waves, $x, t \in \mathbb{R}^3$ may be combined as

$$\zeta = ct \pm x,$$

where c [m/s] is the phase velocity of the waves. The d'Alembert solution to the wave equation, describing waves on a string under tension, is

$$u(x, t) = F(ct - x) + G(ct + x), \tag{1.67}$$

which describes the transverse velocity (or displacement) of two independent waves $F(\xi), G(\xi)$ on the string, which represent forward and backward traveling waves. For example, starting with a string at rest, if one displaces the left end, at $x = 0$, by a step function $u(t)$, then that step displacement will propagate to the right as $u(ct - x)$, arriving at location x_o, at time ct_o. Before this time, the string will not move to the right of the wave-front, at $x = ct$, and after ct_o it will have displacement 1 at x_o. Since the wave equation obeys superposition (postulate P1, p. 72), it follows that the "plane-wave" eigen-functions of the wave equation for $\boldsymbol{x}, \boldsymbol{k} \in \mathbb{R}^3$ are given by

$$\Psi_{\pm}(\boldsymbol{x}, t) = e^{st \pm \jmath \boldsymbol{k} \cdot \boldsymbol{x}}. \tag{1.68}$$

where $s = \sigma + \omega \jmath$ and $\boldsymbol{k} = \jmath 2\pi / \boldsymbol{\lambda}$ is the *wave number*, $\lambda = |\boldsymbol{\lambda}|$ is the wavelength, and $|s/\boldsymbol{k}| = \lambda f = c$. When propagation losses are considered, we must replace $\jmath \boldsymbol{k}$ with a complex analytic wave number $\boldsymbol{\kappa}(s) = \boldsymbol{k}_r + \jmath \boldsymbol{k}$, known as the *dispersion relation*. An important example is the case of electron waves propagating in crystals (i.e., silicon). For this special case, $\boldsymbol{\kappa}(s)$ describes the crystal's dispersion relations, known as *Brillouin zones* (Brillouin, 1953).

5

Complex impedance functions

Conservation of energy (or power) is a corner stone of modern physics. It may have first been under consideration by Galileo Galilei (1564-1642). Today the question is not if it is true, the questions should be why. Specifically, what is the physics behind conservation of energy. Surprisingly, the answer is straight forward, by looking at the definition of power and impedance. The power is given by the product of two variables, sometimes call *conjugate* variables, the force and the flow. In electrical terms, these are voltage (force) ($v(t) \leftrightarrow V(\omega)$) and current (flow) ($i(t) \leftrightarrow I(\omega)$), thus the electrical power is

$$\mathcal{P}(t) = v(t)i(t) \leftrightarrow V(\omega)I(\omega). \tag{1.69}$$

One of the most important and obvious applications of complex functions of a complex variable is the impedance function.

Poor Org

Every impedance must obey conservation of energy (P3): According to Postulate P3 Section 1.3.12 (p. 72), a system is passive if it does not contain a power source. Drawing power from an impedance violates conservation of energy. This propriety is also called *positive real*, which is defined as (Brune, 1931a,b)

$$\Re Z(s \geq 0) \geq 0, \tag{1.70}$$

namely the real part of every impedance must be non-negative for $\sigma \geq 0$. When this condition holds, one cannot draw more power than stored in the impedance. A second condition requires that the

10

impedance has simple poles. If there were a pole in the region $\sigma > 0$, then the first condition would not be met. Therefore there can only be simple poles (degree of 1) in the region $\sigma \leq 0$. The region $\sigma \leq 0$ is called the *left half s plane* (LHP), and the complementary region, $\sigma > 0$, is called the *right half s plane* (RHP). The condition on the simple poles is sufficient, but not necessary, as $Z(s) = 1/\sqrt{s}$ is a physical impedance, but is not a first order pole. The impedance function $Z(s) = R(\sigma, \omega) + \jmath X(\sigma, \omega)$ has resistance R and reactance X as a function of complex frequency $s = \sigma + \jmath \omega$. The function $z(t) \leftrightarrow Z(s)$ are defined by a Laplace transform pair. From the causality postulate (P1) of Sections 1.3.12 and 3.5.1 (p. 145), $z(t < 0) = 0$.

As an example, a resistor R_o in series with an capacitor C_o has an impedance given by (Table 1.5, p. 71)

$$Z(s) = R_o + 1/sC_o \leftrightarrow R_o \delta(t) + \frac{1}{C_o} u(t) = z(t), \tag{1.71}$$

with constants $R_o, C_o \in \mathbb{R} > 0$. In mechanics an impedance composed of a dash-pot (damper) and a spring have the same form. A resonant system has an inductor, resistor and a capacitor, with an impedance given by (Table 1.6, p. 72)

$$Z(s) = \frac{sC_o}{1 + sC_o R_o + s^2 C_o M_o} \leftrightarrow C_o \frac{d}{dt} \left(c_+ e^{s^+ t} + c_- e^{s^- t} \right) = z(t) \tag{1.72}$$

is a second degree polynomial with two complex resonant frequencies $s_\pm$. When $R_o > 0$ these roots are in the left half s plane, with $z(t) \leftrightarrow Z(s)$.

Systems (networks) containing many elements, and transmission lines can be much more complicated, yet still have a simple frequency domain representation. This is the key to understanding how these physical systems work, as will be described below.

Complex analytic functions: To solve a differential equation, or integrate a function, Newton used the Taylor series to integrate one term at a time. However he only used real functions of a real variable, due to the fundamental lack of appreciation of the complex analytic series. This same method is how one finds solutions to scalar differential equations today, but using an approach that makes the solution method less obvious. Rather than working directly with the Taylor series, today we use the complex exponential, since the complex exponential is the eigenfunction of the derivative

$$\frac{d}{dt} e^{st} = s e^{st}.$$

Since e^{st} may be expressed as a Taylor series, having coefficients $c_n = 1/n!$, in some real sense the modern approach is a compact way of doing what Newton did. Thus every linear constant coefficient differential equation in time may be simply transformed into a polynomial in complex Laplace frequency s, by looking for solutions of the form $A(s)e^{st}$, transforming the differential equation into a polynomial $A(s)$ in complex frequency. For example

$$\frac{d}{dt} f(t) + a f(t) \leftrightarrow (s + a) F(s).$$

The root of $A(s_r) = s_r + a = 0$ is the eigenvalue of the differential equation. The key to understanding the solutions of differential equations, both scalar and vector, is to work in the Laplace frequency domain. The Taylor series has been replaced by e^{st}, transforming Newton's real Taylor series into the complex exponential eigenfunction. In some sense, these are the same method, since

$$e^{st} = \sum_{n=0}^{\infty} \frac{(st)^n}{n!}. \tag{1.73}$$

Taking the derivative with respect to time gives

$$\frac{d}{dt} e^{st} = s e^{st} = s \sum_{n=0}^{\infty} \frac{(st)^n}{n!}, \tag{1.74}$$

which is also complex analytic. Thus if the series for $F(s)$ is valid (i.e., it converges), then its derivative is also valid. This was a very powerful concept, exploited by Newton for real functions of a real variable, and later by Cauchy and Riemann for complex functions of a complex variable. The key here is "Where does the series fail to converge?" in which case, the entire representation fails. This is the main message behind the FTCC (Eq. 1.60).

The FTCC (Eq. 1.60) is formally the same as the FTC (Eq. 1.58) (Leibniz formula), the key (and significant) difference being that the argument of the integrand $s \in \mathbb{C}$. Thus this integration is a line integral in the complex plane. One would naturally assume that the value of the integral depends on the path of integration.

But, according to FTCC, it does not. In fact it is indistinguishable from its much simple cousin, the fundamental theorem of calculus. And the reasoning is the same. If $F(s) = df(s)/ds$ is complex analytic (i.e., has a power series $f(s) = \sum_k c_k s^k$, with $f(s), c_k, s \in \mathbb{C}$), then it may be integrated, and *the integral does not depend on the path.* This is sort of amazing. The key is that $F(s)$ and $f(s)$ must be complex analytic, which means they are differentiable. This all follows from the Taylor series formula Eq. 1.61 (p. 77) for the coefficients of the complex analytic series. For Eq. 1.60 to hold, the derivatives must be independent of the direction, as discussed in Section 1.4.2. The concept of a complex analytic function therefore has eminent consequences.

The use of the complex Taylor series formula in the complex plane, generalizes the functions they describe, with unpredictable consequences, as nicely shown by the domain coloring diagrams presented in Section 1.3.9 (p. 65). The tool of complex integration was first exploited in physics by Sommerfeld (1952), to explain the onset transients in waves, as explained in detail in Brillouin (1960, Chap. 3). Up to 1910, when Sommerfeld first published his results using complex analytic signals and saddle point integration in the complex plane, there was a poor understanding of the implications of the causal wave-front. It would be reasonable to say that his insights changed our understanding of wave propagation, for both light and sound. Sadly this insight has not been fully appreciated, even to this day. If you question this summary, please read Brillouin (1960, Chap. 1).

The full power of the complex analytic function was first appreciated by Bernard Riemann (1826-1866), in his University of Göttingen PhD Thesis of 1851, under the tutelage of Carl Friedrich Gauss (1777-1855), and drawing heavily on the work of Cauchy.

The key definition of a complex analytic function is that it has a Taylor series representation over a region of the complex frequency plane $s = \sigma + j\omega$, that converges in a *region of convergence* (RoC) about the expansion point, with a radius determined by the nearest pole of the function. A further surprising feature of all analytic functions is that within the RoC, the inverse of that function also has a complex analytic expansion. Thus given $w(s)$, one may also determine $s(w)$ to any desired accuracy, critically depending on the RoC.

1.4.4   Lec 26: Multi-valued functions

This is a time-out for an entire lecture to answer your questions that have been building, on the many new concepts presented over the last few weeks.

The best question of the week: "What is a multi-valued function?" Second best question: "What is a line integral?," "What is the difference between the impedance of a mass and inductor?" and "How are physics and math related?"

Branch Cuts: The concept of a *branch cut* allows one to manipulate (and visualize) multi-valued functions, by breaking each region into a single valued *sheets*. The concepts of the branch cut and the sheets, along with the *extended plane*, were first devised by Riemann, working with Gauss (1777-1855), first described in his thesis of 1851. Of course it was these three mathematical constructions that provide the deep insight to complex analytic functions, supplementing the important earlier work of Cauchy (1789-1857), on the calculus of complex analytic functions.

The branch cut is a line that separates the various single valued parts of a multi-valued function. For example, in Fig. 1.18 we see the double-valued mapping of $w(z) = \pm\sqrt{z}$. Since the square root function

has two overlapping regions, corresponding to the $\pm$ of the radical, there must be two connected regions, sort of like mathematical Siamese-twins, distinct, yet the same.

15.3   Branch Points   303
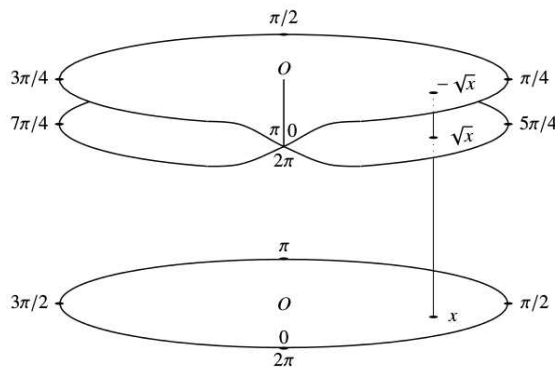
Figure 15.6: Branch point for the square root

Figure 1.18:   *Here we see the mapping for the square root function $w(z) = \pm\sqrt{z}$ which has two single-valued sheets, corresponding to the two signs of the square root. The lower sheet is $+\sqrt{z}$, and the upper sheet is $-\sqrt{z}$. The location of the branch cut may be moved by rotating the z coordinate system. For example, $w(z) = \pm\jmath\sqrt{z}$ and $w(z) = \pm\sqrt{-z}$ have a different branch cuts, as may be easily verified using the Matlab/Octave commands `j*zviz(z)` and `zviz(-z)`. A function is analytic on the branch cut, since the cut may be moved. If a Taylor series is formed on the branch cut, it will describe the function on the two different sheets. Thus the complex analytic series (i.e., the Taylor formula, Eq. 1.61) does not depend on the location of a branch cut, as it only describes the function uniquely (as a single valued function), valid in its local region of convergence. This figure has been taken from Stillwell (2010, p. 303).*

This concept of analytic inverses becomes rich when the inverse function is multi-valued. For example, if $F(s) = s^2$ then $s(F) = \pm\sqrt{F}$. Riemann dealt with such extensions with the concept of a *branch-cut* with multiple *sheets*, labeled by a sheet number. Each sheet describes an analytic function (Taylor series), that converges within some RoC, with a radius out to the nearest pole of that function. This Riemann's branch cut and sheets explicitly deal with the need to define unique single valued inverses of multi-valued functions.

Branch cuts emanate from poles that have non-integral degrees, and terminate at either another pole of the same degree, or at ∞. For example, suppose that in the neighborhood of the pole, at s_o the function is

$$f(s) = \frac{w(s)}{(s - s_o)^k},$$

where $w, s, K, k \in \mathbb{C}$. Here $w(s_o)$ is the residue, when $k = 1$, $s = \sigma + \omega\jmath$, $s_o = \sigma_o + \omega_o\jmath$ is the pole location, and k is some complex or real constant that defines the *order* of the pole (*not* the degree, as in the degree of a polynomial).

Up to this point we have assumed that the pole order is an integer $k \in \mathbb{Z}$ or fractional ($k \in \mathbb{F}$). When $k \in \mathbb{Z}$ there is no branch cut. When $k \in \mathbb{F}$ there must be a branch cut, of order k. For example, if $k = 1/2$, the pole is of order $1/2$, and there are two Riemann sheets, as shown in Fig. 1.18. An important example is the Bessel function

$$\delta(t) + \frac{1}{t}J_1(t)u(t) \leftrightarrow \sqrt{s^2 + 1},$$

which is related to the solution to the wave equation in two-dimensional cylindrical coordinates (Section 1.5.2, p. 71). Bessel functions are the solutions to guided acoustic waves in round pipes, or surface waves on the earth (seismic waves) or waves on the surface of a pond.

It is important to understand that the function is analytic on the branch cut, but not at the branch point (the pole). One is free to move the branch cut, almost at will. It does not need to be on a line, it could be cut as a spiral. The only rule is that it must start and stop at two poles of the same order, or at ∞, which must have a pole of order k.

When the pole has an irrational order ($k \in \mathbb{I}$), the branch cut has the same irrational order. Accordingly there must be an infinite number of Riemann sheets, as in the case of the log function. An example is $k = \pi$, for which

$$F(s) = \frac{1}{s^\pi} = e^{-\log(s^\pi)} = e^{-\pi \log(s)} = e^{-\pi \log(\rho)} e^{-\pi \theta \jmath},$$

where the domain is expressed in polar coordinates $s = \rho e^{\theta \jmath}$. When the irrational number is very close to 1, the branch cut could be very subtle (it could even be unnoticed), but it would have an impact on the nature of the function, and of course, on the inverse Laplace transform. In this course we shall not attempt to deal with irrational pole order.

Multivalued functions: The two basic functions we review, to answer the questions about multivalued functions and branch cuts, are $w(s) = s^2$ and $w(s) = e^s$, along with their inverse functions $w(s) = \sqrt{s}$ and $w(s) = \log(s)$. For uniformity we shall refer to the abscissa ($s = \sigma + \omega \jmath$) and the ordinate $w(s) = u + v \jmath$. The proper, but less well known terms are the *domain* and *co-domain*, today's nomenclature.

Square root function: The multi-valued nature of the square root is best understood by working with the function in polar coordinates. Let

$$s_k = r e^{\theta \jmath} e^{2\pi k \jmath},$$

where k is the sheet-index, and

$$w = \rho e^{\psi \jmath} = \sqrt{r} e^{\jmath \theta/2} e^{\jmath \pi k},$$

where $r, \rho \in \mathbb{R}$ are the magnitudes and $\theta, \psi \in \mathbb{R}$ are the angles. The domain-coloring program `zviz.m` assumes that the angles go from $-\pi < \theta < \pi$, with $\theta = 0$ being a light red and $\pm\pi$ a blue color. This angle to color map is shown in Fig. 1.15 (p. 66). The first Riemann sheet of $\sqrt{s}$ is define by $-\pi < \theta < \pi$. The second sheet picks up at $\theta = \pi$ and continues on to $\pi + 2\pi = 3\pi$. The first sheet maps the angle of w (i.e., $\phi = \angle w = \theta/2$) from $-\pi/2 < \phi < \pi/2$ ($w = \sqrt{r} e^{\jmath \theta/2}$). This corresponds to $u = \Re w(s) > 0$. The second sheet maps $\pi/2 < \psi < 3\pi/2$ (i.e., 90° to 270°), which is $\Re w = u < 0$. In summary, twice around the s plane is once around the $w(s)$ plane, because the angle is half due to the $\sqrt{s}$. This then describes the multi-valued nature of the square root function.
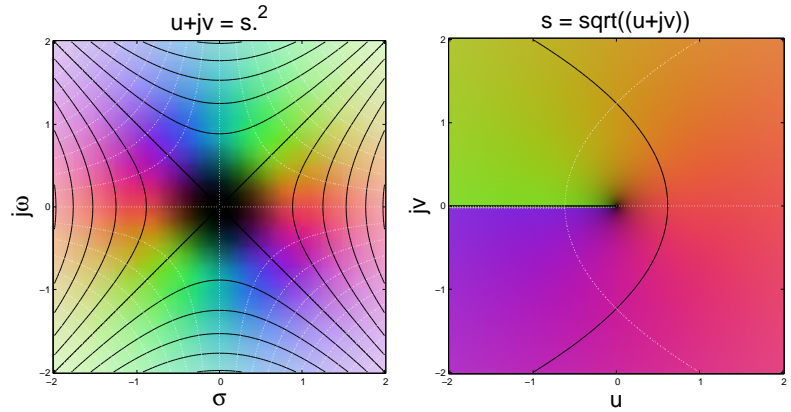
Figure 1.19: *Here the Cartesian coordinate map between $s = \sigma + \omega \jmath$ and $w = u + v \jmath$. LEFT: This shows the mapping $w(s) = s^2$. RIGHT: This shows the lower branch of the inverse $s(w) = \sqrt{w}$, as shown in Fig. 1.18 (p. 82).*

Log function: Next we discuss the multi-valued nature of the log function. In this case there are an infinite number of Riemann sheets, not well captured by Fig. 1.16 (p. 67), which only displays the principal sheet. However if we look at the formula for the log function, the nature is easily discerned. The abscissa s may be defined as multi-valued since

$$s_k = r e^{2\pi k \jmath} e^{\theta \jmath}.$$

Here we have extended the angle of s by $2\pi k$, where k is the sheet index $\in \mathbb{Z}$. Taking the log

$$\log(s) = \log(r) + (\theta + 2\pi k)\jmath.$$

When $k = 0$ we have the *principal value* sheet, which is zero when $s = 1$. For any other value of k $w(s) \neq 0$, even when $r = 1$, since the angle is not zero, except for the $k = 0$ sheet.

1.4.5   Lec 27: Three Cauchy Integral Theorems

Cauchy's theorems for integration in the complex plane

5   There are three basic definitions related to Cauchy's integral formula. They are closely related, and can greatly simplify integration in the complex plane.

1. **Cauchy's (Integral) Theorem:**

$$\oint_{\mathcal{C}} F(s)ds = 0, \tag{1.75}$$

   if and only if $F(s)$ is complex-analytic inside of the closed curve $\mathcal{C}$ (Boas, 1987, p. 45),(Stillwell, 2010, 319). The FTCC (Eq. 1.60) says that the integral only depends on the end points if $F(s)$ is complex-analytic. By closing the path (contour $\mathcal{C}$) the end points are the same, thus the integral
10   must be zero, as long as $F(s)$ is complex analytic.

2. **Cauchy's Integral Formula:**

$$\frac{1}{2\pi j} \oint_{\mathcal{C}} \frac{F(s)}{s - s_o} ds = \begin{cases} F(s_o), s_o \in \mathcal{C} \text{ (inside)} \\ 0, s_o \notin \mathcal{C} \text{ (outside)}. \end{cases} \tag{1.76}$$

   Here $F(s)$ is required to be analytic everywhere within (and on) the contour $\mathcal{C}$ (Boas, 1987, p. 51), (Stillwell, 2010, p. 220). The value $F(s_o) \in \mathbb{C}$ is called the *residue* of the pole s_o of the function $F(s) = F(s)/(s - s_o)$. Note the residue is the coefficient of a Taylor series expansion c_{-1}, if this term were included in the series. The Taylor series (Eq. **??**, p. **??**) is defined with $c_{-1} = 0$. A
15   modify series, having the $c_{-1} \neq 0$, is known as a *Laurent* series.

3. **The (Cauchy) Residue Theorem** (Greenberg, 1988, p. 1241), (Boas, 1987, p. 73)

$$\oint_{\mathcal{C}} f(s)ds = 2\pi j \sum_{k=1}^{K} c_k = \sum_{k=1}^{K} \oint \frac{f(s)}{s - s_k} ds, \tag{1.77}$$

   where the *residue* $c_k \in \mathbb{C}$, corresponding to the kth poles of $f(s)$, enclosed by the contour $\mathcal{C}$. By the use of Cauchy's integral formula, the last form of the residue theorem is equivalent to the residue form.[57]

   How to calculate the residue: The case of first order poles is special since the Brune impedance only allows simple poles and zeros, increasing the importance of this special case. The residues is defined by the coefficient of a simple pole of the function $F(s)$, which is complex analytic in the neighborhood of the pole, but not at the pole. This coefficient may be computed by removing the singularity by placing a zero at the pole frequency, followed by taking the limit as $s \to s_k$ (Boas, 1987, p. 72), namely
$$c_k = \lim_{s \to s_k} [(s - s_k)F(s)].$$

20   Consider the function $f(s) = F(s)/(s - s_k)$, where we have factored $f(s)$ to isolate the first-order pole at $s = s_k$. If $f(s)$ is analytic at s_k, then the residue of the pole at $s = s_k$ is $f(s_k)$.

---

[57]This theorem is the same as a 2D version of Stokes thm (citations).

These three theorems, all attributed to Cauchy, collectively are related to the fundamental theorems of calculus. Because the names of the three theorems are so similar, are easily confused.

In general it makes no sense to integrate *through* a pole, thus the poles (or other singularities) must not lie on $\mathcal{C}$.

The Cauchy integral theorem (1), follows trivially from the fundamental theorem of complex calculus (Eq. 1.60, p. 77), since if the integral is independent of the path, and the path returns to the starting point, the closed integral must be zero. Thus Eq. 1.75 holds when $f(s)$ is complex analytic within C.

Since the real and imaginary parts of every complex analytic function obey Laplace's equation (Eq. 1.66, p. 78), it follows that every closed integral over a Laplace field, i.e., one defined by Laplace's equation, must be zero. *In fact this is the property of a conservative system,* corresponding to many physical systems. If a closed box has fixed potentials on the walls, with any distribution what so ever, and a point charge (i.e, an electron) is placed in the box, then a force equal to $F = qE$ is required to move that charge, thus work is done. However if the point is returned to its starting location, the net work done is zero.

Work is done in charging a capacitor, and energy is stored. However when the capacitor is discharged, all of the energy is returned to the load.

Soap bubbles and rubber sheets on a wire frame, obey Laplace's equation.

These are all cases where the fields are Laplacian, thus closed line integrals must be zero. Laplacian fields are commonly observed, because they are so basic.

We have presented the impedance as the primary example of complex analytic functions. Physically, every impedance has an associated power, and every system having power, has an associated impedance. The impedance is usually defined in the frequency s domain, as the force *over* the flow (i.e., voltage over current), while the power is defined as the force *times* the flow (see Section 3.2.1, p. 131, Eq. 3.5). In summary, impedance and power are fundamentally related.

1.4.6   Lec 28: Cauchy Integral Formula & Residue Theorem

The Cauchy integral formula (Eq. 1.76) is an important extension of the Cauchy integral theorem (Eq. 1.75) in that a pole has been explicitly injected into the integrand at $s = s_o$. If the pole location is outside of the curve $\mathcal{C}$, the result of the integral is zero, in keeping with Eq. 1.75. However, when the pole is inside of $\mathcal{C}$, the integrand is no longer complex analytic, and a new result follows. By a manipulation of the contour in Eq. 1.76, the pole can be isolated with a circle around the pole, and then taking the limit, the radius goes to zero.

For the related *Cauchy residue theorem* (Eq. 1.77) the same result holds, except it is assumed that there are K simple poles in the function $F(s)$. This requires the repeated application of Eq. 1.76, K times, so it represents a minor extension of Eq. 1.76. The function $F(s)$ may be written as $f(s)/P_K(s)$, where $f(s)$ is analytic in $\mathcal{C}$ and $P_K(s)$ is a polynomial of degree K, with all of its roots $s_k \in \mathcal{C}$.

Non-integral degree poles   The key point is that this theorem applies when $n \in \mathbb{I}$, including fractionals $n \in \mathbb{F}$. The function $1/\sqrt{s}$ has a zero residue, which is strictly the amplitude of $1/s$ (Boas, 1987, p. 73). When $n \in \mathbb{F}$, the residue is, by definition, zero. When $n \in \mathbb{I}$, the residue is, by definition, zero. When $n = 1$, the residue is c_{-1}. For an intuitive discussion of Riemann sheets and branch cuts, see Boas (1987, Section 29, pp. 221-225).

This point is equally important when defining the inverse Laplace transform. When integrating over $\omega \in \mathbb{R}$, the value passes through all possible exponents, including all rational and irrational numbers. The only value of ω that has a residue, are those at the poles of the integrand.

1.4.7   Lec 29: Inverse Laplace transform (Cauchy residue theorem)

The inverse Laplace transform Eq. 1.56 transforms a function of complex frequency $F(s)$ and returns a causal function of time $f(t)$

$$f(t) \leftrightarrow F(s),$$

where $f(t) = 0$ for $t < 0$. Examples are provided in Table 1.5 (p. 71). We next discuss the details of finding the inverse transform by use of the Cauchy residue theorem, and how the causal requirement $f(t < 0) = 0$ comes about.

The integrand of the inverse transform is $F(s)e^{st}$ and the limits of integration are $-\sigma_o \mp \omega_\jmath$. To find
5   the inverse we must close the curve, at infinity, and show that the integral at $\omega_\jmath \to \infty$. There are two ways to close these limits, to the right $\sigma > 0$ (RHP), and to the left $\sigma < 0$ (LHP), but there needs to be some logical reason for this choice. That logic is the sign of t. For the integral to converge the term e^{st} must go to zero as $\omega \to \infty$. In terms of the real and imaginary parts of $s = \sigma + \omega_\jmath$, the exponential may be rewritten as $e^{\sigma t}e^{\jmath \omega t}$. Note that both t and ω go to ∞. So it is the interaction between these
10 two limits that determines how we pick the closure, RHP vs. LHP.

Case for causality $(t < 0)$**:**   Let us first consider negative time, including $t \to -\infty$. If we were to close $\mathcal{C}$ in the left half plane ($\sigma < 0$), then the product σt is positive ($\sigma < 0, t < 0$, thus $\sigma t > 0$). In this case as $\omega \to \infty$, the closure integral $|s| \to \infty$ will diverge. Thus we may not close in the LHP for negative time. If we close in the RHP $\sigma > 0$ then the product $\sigma t < 0$ and e^{st} will go to zero as $\omega \to \infty$.
15 This then justifies closing the contour, allowing for the use the Cauchy theorems.

If $F(s)$ is analytic in the RHP, the FTCC applies, and the resulting $f(t)$ must be zero, and the inverse Laplace transform must be causal. This argument holds for any $F(s)$ that is analytic in the RHP ($\sigma > 0$).

Case of unstable poles: An important but subtle point arises: If $F(s)$ has a pole in the RHP, then
20 the above argument still applies if we pick σ_o to be to the right of the RHP pole. this means that the inverse transform may still be applied to unstable poles (those in the RHP). This explains the need for the σ_o in the limits. If $F(s)$ has no RHP poles, then $\sigma_o = 0$ is adequate, and this factor may be ignored.

Case for zero time $(t = 0)$**:**   When time is zero, the integral does not, in general, converge, leaving
25 $f(t)$ undefined. This is most clear in the case of the step function $u(t) \leftrightarrow 1/s$, where the integral may not be closed, because the convergence factor $e^{st} = 1$ is lost for $t = 0$.

1.4.8   Lec 30: Inverse Laplace transform and the Cauchy Residue Theorem

Case of $t > 0$**:**   Next we investigate the convergence of the integral for positive time $t > 0$. In this case we must close the integral in the LHP ($\sigma < 0$) for convergence, so that $st < 0$ ($\sigma \le 0$ and $t > 0$).
30 When there are poles on the $\omega_\jmath = 0$ axis, $\sigma_o > 0$ assures convergence by keeping the on-axis poles inside the contour. At this point the Cauchy residue theorem (Eq. 1.77) is relevant. If we restrict ourselves to simple poles (as required for a Brune impedance), the residue theorem may be directly applied.

The most simple example is the step function, for which $F(s) = 1/s$ and thus

$$u(t) = \oint_{\text{LHP}} \frac{e^{st}}{s} \frac{ds}{2\pi\jmath} \leftrightarrow \frac{1}{s},$$

which is a direct application of Eq. 1.77. The forward transform of $u(t)$ is straight forward, as discussed in Section 1.3.11 (p. 70). This is true of most if not all of the elementary forward Laplace transforms.
35 In these cases, causality is built into the integral by the limits, so is not a result, as it must be in the inverse transform. An interesting problem is proving that $u(t)$ is not defined at $t = 0$.

The inverse Laplace transform of $F(s) = 1/(s + 1)$ has a residue of 1 at $s = -1$, thus that is the only contribution to the integral. A case that is more demanding is the Laplace transform pair

$$\frac{1}{\sqrt{t}}u(t) \leftrightarrow \sqrt{\frac{\pi}{s}} \quad \text{and} \quad J_o(t)u(t) \leftrightarrow \frac{1}{s^2 + 1} = \frac{1}{(s + \jmath)(s - \jmath)}.$$

Many of these are easily proved in the forward direction, but are much more difficult in the inverse direction due to the properties at $t = 0$, unless of course, the residue theorem (Eq. 1.77, p. 84) is invoked. The last $\mathcal{L}$-pair helps us understand the basic nature of the Bessel functions $J_0(z)$, and $H_0^{(1)}(z^2)$, with a branch cut along the negative axis (see Fig. 3.2, p. 144).

Some open questions: Without the use of the CRT (Eq. 1.77) it is difficult to see how evaluate do the inverse Laplace transform of $1/s$ directly. For example, how does one show that the above integral is zero for negative time (or that it is 1 for positive time)? The CRT neatly solves this difficult problem, by the convergence of the integral for negative and positive time. Clearly the continuity of the integral at $\omega \to \infty$ plays an important role. Perhaps the Riemann sphere plays a role in this, that has not yet been explored.

1.4.9 Lec 31: Properties of the LT (e.g., Linearity, convolution, time-shift, modulation, etc.)

As shown in the table of Laplace transforms, there are integral (i.e., integration, not integer) relationships, or properties, that are helpful to identify. The first of these is a definition not a property:[58]

$$f(t) \leftrightarrow F(s).$$

When taking the LT, the time response is given in lower case (e.g., $f(t)$) and the frequency domain transform is denoted in upper case (e.g., $F(s)$). It is required, but not always explicitly specified, that $f(t < 0) = 0$, that is, the time function must be *causal* (P1: Section 1.3.12).

Linearity: A key property so basic that it almost is forgotten, is the linearity property of the LT. These properties are summarized as P2 of Section 1.3.12, 73).

Convolution property: One of the most basic and useful properties is that the product of two LTs in frequency, results in convolution in time

$$f(t) \star g(t) = \int_0^t f(\tau)g(t - \tau)d\tau \leftrightarrow F(s)G(s),$$

where we use the $\star$ operator as a shorthand for the convolution of two time functions.

A key application of convolution is filtering, which takes on many forms. The most basic filter is the *moving average*, the moving sum of data samples, normalized by the number of samples. Such a filter has very poor performance. It also introduces a delay of half the length of the average, which may, or may not constitute a problem, depending on the application. Another important example is a low-pass filter, that removes high frequency noise, or a notch filter that removes line-noise (i.e., 60 [Hz] in the US, and its 2d and 3d harmonics, 120 and 180 [Hz]). Such noise is typically a result of poor grounding and ground loops. It is better to solve the problem at its root, than to remove it with a notch filter. Still, filters are very important in engineering.

By taking the LT of the convolution we can derive this relationship

$$\int_0^\infty [f(t) \star g(t)]e^{-st}dt = \int_{t=0}^\infty \left[\int_0^t f(\tau)g(t - \tau)d\tau\right] e^{-st}dt$$

$$= \int_0^t f(\tau) \left(\int_{t=0}^\infty g(t - \tau)e^{-st}dt\right) d\tau$$

$$= \int_0^t f(\tau) \left(e^{-s\tau} \int_{t'=0}^\infty g(t')e^{-st'}dt'\right) d\tau$$

$$= G(s) \int_0^t f(\tau)e^{-s\tau}d\tau$$

$$= G(s)F(s)$$

---

[58]Put this notional property in Appendix A.

We encountered this relationship in Section 1.3.3 (p. 54)) in the context of multiplying polynomials, which was the same as convolving their coefficients. Hopefully the parallel is obvious. In the case of polynomials, the convolution was discrete in the coefficients, and here it is continuous in time. But the relationships are the same.

Time-shift propriety: When a function is time-shifted by time T_o, the LT is modified by e^{sT_o}, leading to to the propriety

$$f(t - T_o) \leftrightarrow e^{-sT_o} F(s).$$

This is easily shown by applying the definition of the LT to a delayed time function.

Time derivative: The key to the eigen-function analysis provided by the LT, is the transformation of a time derivative on a time function, that is

$$\frac{d}{dt} f(t) \leftrightarrow sF(s).$$

Here s is the eigen value corresponding to the time derivative of e^{st}. Given the definition of the derivative of e^{st} with respect to time, this definition seems trivial. Yet that definition was not obvious to Euler. It needed to be extended to the space of complex analytic function e^{st}, which did not happen until at least Riemann (1851).

Given a differential equation of order K, the LT results in a polynomial in s, of degree K.

It follows that this LT property is the corner-stone of why the LT is so important to scalar differential equations, as it was to the early analysis of Pell's equation and the Fibonacci sequence, as presented in earlier chapters. This property was first uncovered by Euler. It is not clear if he fully appreciated its significance, but by the time of his death, it certainly would have been clear to him. Who first coined the term *eigen value* and *eigen function*? The word *eigen* is a German word meaning *of one*. It seem likely that this term became popular somewhere between the 19th and 20th century.

Initial and final value theorems: There are much more subtle relations between $f(t)$ and $F(s)$ that characterize $f(0^+)$ and $f(t \to \infty)$. While these properties can be very important in certain application, they are are beyond the scope of the present treatment. These relate to so-called *initial value theorems*. If the system under investigation has potential energy at $t = 0$, then the voltage (velocity) need not be zero for negative time. An example is a charged capacitor or a moving mass. These are important situations, but better explored in a more in depth treatment.[59]

1.4.10   Lec 32: Properties of the Brune impedance

Impedance

Impedance plays a crucial role in the solution of differential equations, and in the theory and practice of complex variables. As a theoretical concept, it is the best example of a complex analytic function, thus plays a key role in potential theory.

Theoretically, impedance plays the role of a boundary condition in the solution of differential equations, accounting for the reflectance of a wave at the boundary. It plays a critical role in conservation of energy, since if energy at the boundary (surface) must always be lost, given a passive boundary.

The first person to fully appreciate the importance of impedance at the boundary may have been Arnold Sommerfeld, who expressed this in what today is known as the *radiation impedance*, which is the impedance boundary condition as a point source wave radiates to $|R| \to \infty$.

---

[59]Discuss in Section **??**.

Brune impedance: As the concept of impedance evolved (it is a linear relation between a force and a flow), it was incorporated into more theories, such as electrical circuit theory and mechanics. The first persons to quantify the concept of impedance was Ohm, followed by Kirchhoff and Heaviside. Kennelly, not Heaviside, was the first to express the idea as a complex variable of a complex function of frequency (Kennelly, 1893). Perhaps Heaviside fully appreciated the concept, and has been given proper credit. A very interesting development was the largely forgotten but related mathematical work of Bott and Duffin (Van Valkenburg, 1964b)

There are several important theorems here that follow from *Brune's Theorem* on *positive-real functions*, defined by the relation

$$\Re Z(\sigma > 0) \geq 0, \tag{1.78}$$

where $Z(s) = R(s) + \jmath X(s) \in \mathbb{C}$ is a Brune impedance having real part $R(s) = \Re Z(s)$ and imaginary part $X(s) = \Im Z(s)$, each as functions of the Laplace frequency $s = \sigma + \omega\jmath$ (Brune, 1931a; Van Valkenburg, 1964b). Thus a Brune impedance has a positive analytic real part in the right-half plane $\sigma > 0$.

This condition has many non-obvious ramifications, including (order is significant)

1. $Z(s) \leftrightarrow z(t < 0) = 0$ is real and causal

2. $Y(s) = 1/Z(s) \leftrightarrow y(t < 0) = 0$ is real and causal

3. The phase $\angle Z(s)$ lies between $\pm\pi/2$ (i.e., $Z(s)$ is minimum–phase)

4. All the poles and zeros are first order (there are no higher order poles)

5. All poles and zeros lie in the left half plane LHP ($\sigma \leq 0$)

6. $Z(\sigma > 0)$ is complex–analytic [$Z(s)$ obeys the Cauchy-Riemann conditions in the RHP ($\sigma > 0$)]

7. All poles and zeros of $Z(s)$ on the $\omega\jmath$ or on the σ axis must alternate (Foster's Theorem Van Valkenburg (1964b))

8. All Brune impedances are quasi-static (lumped-parameter, where the size of the impedance element is much smaller than the wavelength).

9. The Brune impedance is a subset of impedance, where delay has been approximated away in the long-wavelength limit. For example, transmission lines are not in the class of Brune impedances.

A Brune impedance is defined as the ratio of the force $F(\omega)$ over the flow $U(\omega)$, and may be expressed in residue form as

$$Z(s) = sL_o + R_o + \sum_{k=1}^{K} \frac{c_k}{s - s_k} = \frac{N(s)}{D(s)}, \tag{1.79}$$

known as the *partial fraction expansion*. The term L_o represents an inductance, and R_o a resistor. Coefficients c_k are the residues, and s_k the roots of $D(s)$. It follows that

$$D(s) = \prod_{k=1}^{K}(s - s_k) \qquad \text{and} \qquad c_k = \lim_{s \to s_k}(s - s_k)Z(s) = \prod_{n'=1}^{K-1}(s - s_n),$$

where the prime on index n' means that $n = k$ is not included in the product. Every Brune impedance may be expanded in a partial fraction expansion of the form of Eq. 1.79, and every partial fraction expansion may be converted back into the ratio of two polynomials, using the Matlab/Octave commands `[C,P,K]=residue(N,D)` and `[N,D]=residue(C,P,K)`. These two representations are interchangeable.

Given a 2-port network, the input impedance and the transfer function share the same poles.[60] In fact the transfer function must be all-pole.

---

[60]Is $\mathcal{C}(s)$ of the ABCD matrix the same as the impedance denominator $D(s)$? Verify!

Exercise: Find the Laplace transform ($\mathcal{L}$) of the three impedance relations in terms of the force $F(s)$ and the velocity $V(s)$, along with the electrical equivalent impedance: Each classical law is a linear relation

1. Hooke's Law $f(t) = Kx(t)$. Sol: First $\mathcal{L}$ Hooke's Law and then write it in terms of force and velocity

    Taking the $\mathcal{L}$ gives
    $$F(s) = KX(s).$$

    Since
    $$v(t) = \frac{d}{dt}x(t) \leftrightarrow V(s) = sX(s).$$

    Thus the impedance of the spring is
    $$Z_s(s) = \frac{F(s)}{V(s)} = \frac{K}{s},$$

    which is analogous to the impedance of an electrical capacitor. If we specifying the compliance C of a spring as $C = 1/K$, the relation looks just like the electrical case.

    While Hooke's law says the force and displacement are proportional, in terms of impedance variables force and flow,
    $$f(t) = K \int^t v(t)dt.$$

2. Dash-pot resistance $f(t) = Rv(t)$. Sol: From the $\mathcal{L}$ this becomes
    $$F(s) = RV(s)$$

    and the impedance of the dash-pot is then
    $$Z_r = R,$$

    analogous to that of an electrical resistor.

3. Newton's 2d Law for a constant mass M: $f(t) = Mdv(t)/dt$. Sol: Taking the $\mathcal{L}$ gives
    $$F(s) = sMV(s)$$

    thus
    $$Z_m(s) \equiv \frac{F(s)}{V(s)} = sM,$$

    analogous to an electrical inductor.

4. Find the total impedance seen by the net force $f(t)$ applied to the mass M. Sol: Summing the forces must equal the applied force, Eq. E.1, p. 181.

5. Take the Laplace transform ($\mathcal{L}$) of Eq. E.1 (p. 181), and evaluate the total impedance $Z(s)$ of the mechanical circuit. Sol: From the properties of the $\mathcal{L}$ that $dx/dt \leftrightarrow sX(s)$, we find
    $$Ms^2X(s) + RsX(s) + KX(s) = F(s).$$

    In terms of velocity this is $(Ms + R + K/s)V(s) = F(s)$. Thus the circuit impedance is
    $$Z(s) = \frac{F}{V} = \frac{K + Rs + Ms^2}{s}.$$

6. What are $N(s)$ and $D(s)$ (e.g. Eq. 1.79, p. 89)? Sol: $D(s) = s$ and $N(s) = K + Rs + Ms^2$.

7. Assume that $M = R = K = 1$. Find the *residues* (e.g. Eq. 1.79, p. 89) of the admittance $Y(s) = 1/Z(s)$, in terms of the roots $s_\pm$. Check your answer with the Matlab/Octave command `residue`. Sol: *First* find the roots of the numerator of $Z(s)$ (the denominator of $Y(s)$):

$$s_\pm^2 + s_\pm + 1 = (s_\pm + 1/2)^2 + 3/4 = 0,$$

which is

$$s_\pm = \frac{-1 \pm \jmath\sqrt{3}}{2}.$$

Next form the *partial fraction expansion*

$$\frac{s}{1 + s + s^2} = c_o + \frac{c_+}{s - s_+} + \frac{c_-}{s - s_-} = \frac{s(c_+ + c_-) - (c_+ s_- + c_- s_+)}{1 + s + s^2}.$$

Comparing the two sides requires that $c_o = 0$. We also have two equations for the residues $c_+ + c_- = 1$ and $c_+ s_- + c_- s_+ = 0$. The best way to solve this is to set up a matrix relation and take the inverse

$$\begin{bmatrix} 1 & 1 \\ s_- & s_+ \end{bmatrix} \begin{bmatrix} c_+ \\ c_- \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{thus:} \quad \begin{bmatrix} c_+ \\ c_- \end{bmatrix} = \frac{1}{s_+ - s_-} \begin{bmatrix} s_+ & -1 \\ -s_- & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

which gives $c_\pm = \pm\frac{s_\pm}{s_+ - s_-}$ The denominator is $s_+ - s_- = \jmath\sqrt{3}$ and the numerator is $\pm 1 + \jmath\sqrt{3}$. Thus

$$c_\pm = \pm\frac{s_\pm}{s_+ - s_-} = \frac{1}{2}\left(1 \pm \frac{\jmath}{\sqrt{3}}\right).$$

As always, finding the coefficients is always the most difficult part. Using 2x2 matrix algebra can really simplify up the process, and will more likely give the correct answer.

8. By applying the CRT, find the inverse Laplace transform $(\mathcal{L}^{-1})$. Use the residue form of the expression that you derived in the previous exercise. Sol:

$$z(t) = \frac{1}{2\pi\jmath} \oint_{\mathcal{C}} Z(s)e^{st}ds.$$

were $\mathcal{C}$ is the Laplace contour which encloses the entire left-half s plane. Applying the CRT

$$z(t) = c_+ e^{s_+ t} + c_- e^{s_- t}.$$

where $s_\pm = -1/2 \pm \jmath\sqrt{3}/2$ and $c_\pm = 1/2 \pm \jmath/(2\sqrt{3})$.

1.4.11 **Lec 33: The Riemann zeta function** 5

The zeta function depends explicitly on the primes, which makes it very important (Section **??**). In 1737 Euler proposed the real-valued function $\zeta(x) \in \mathbb{R}$ and $x \in \mathbb{R}$, to prove that the number of primes is infinite (Goldstein, 1973). Euler's definition of $\zeta(x) \in \mathbb{R}$ is given by the power series

$$\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x} \qquad \text{for } x > 0 \in \mathbb{R}. \tag{1.80}$$

This series converges for $x > 0$, since $R = n^{-x} < 1, n > 1 \in \mathbb{N}$.[61]

---

[61]Sanity check: For example let $n = 2$ and $x > 0$. Then $R = 2^{-\epsilon} < 1$, where $\epsilon \equiv \lim x \to 0^+$. Taking the log gives $\log_2 R = -\epsilon \log_2 2 = -\epsilon < 0$. Since $\log R < 0$, $R < 1$.

In 1860 Riemann extended the zeta function into the complex plane, resulting in $\zeta(s)$, defined by the complex analytic power series, identical to the Euler formula, except $x \in \mathbb{R}$ has been replaced by $s \in \mathbb{C}$

$$\zeta(s) \equiv \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n^s} = \sum_{n=1}^{\infty} n^{-s} \qquad \text{for } \Re s = \sigma > 0. \tag{1.81}$$

This formula converges for $\Re s > 1$ (Goldstein, 1973). To determine the formula in other regions of the s plane one, must extend the series via analytic continuation. As it turns out, Euler's formulation provided detailed information about the structure of primes, going far beyond his original goal.

Euler product formula

As was first published by Euler in 1737, one may recursively factor out the leading prime term, resulting in Euler's product formula. Euler's procedure is an algebraic implementation of the sieve of Eratosthenes (Section 1.2.3, p. 31 and Section 1.4.11, page 91).

Multiplying $\zeta(s)$ by the factor $1/2^s$, and subtracting from $\zeta(s)$, removes all the even terms $\propto 1/(2n)^s$ (e.g., $1/2^s + 1/4^s + 1/6^s + 1/8^s + \cdots$)

$$\left(1 - \frac{1}{2^s}\right)\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s}\cdots - \left(\frac{1}{2^s} + \frac{1}{4^s} + \frac{1}{6^s} + \frac{1}{8^s} + \frac{1}{10^s} + \cdots\right), \tag{1.82}$$

which results in

$$\left(1 - \frac{1}{2^s}\right)\zeta(s) = 1 + \frac{1}{3^s} + \frac{1}{5^s} + \frac{1}{7^s} + \frac{1}{9^s} + \frac{1}{11^s} + \frac{1}{13^s} + \cdots. \tag{1.83}$$

Repeating this with a lead factor $1/3^s$ applied to Eq. 1.83 gives[62]

$$\frac{1}{3^s}\left(1 - \frac{1}{2^s}\right)\zeta(s) = \frac{1}{3^s} + \frac{1}{9^s} + \frac{1}{15^s} + \frac{1}{21^s} + \frac{1}{27^s} + \frac{1}{33^s} + \cdots. \tag{1.84}$$

Subtracting Eq. 1.84 from Eq. 1.83 cancels the RHS terms of Eq. 1.83

$$\left(1 - \frac{1}{3^s}\right)\left(1 - \frac{1}{2^s}\right)\zeta(s) = 1 + \frac{1}{5^s} + \frac{1}{7^s} + \frac{1}{11^s} + \frac{1}{13^s} + \frac{1}{17^s} + \frac{1}{19^s}\cdots.$$

Further repeating this process, with prime scale factors, (i.e., $1/5^s, 1/7^s, \cdots, 1/\pi_k^s, \cdots$), removes all the terms on the RHS but 1, results in Euler's analytic product formula ($s = x \in \mathbb{R}$) and Riemann's complex analytic product formula $s \in \mathbb{C}$

$$\zeta(s) = \prod_{\pi_k \in \mathbb{P}} \frac{1}{1 - \pi_k^{-s}} = \prod_{\pi_k \in \mathbb{P}} \zeta_k(s), \tag{1.85}$$

where π_k represents the kth prime. The above defines each prime factor

$$\zeta_k(s) = \frac{1}{1 - \pi_k^{-s}} \tag{1.86}$$

as the kth term of the complex analytic Riemann product formula, having an RoC of $|\pi_k^{-s}| < 1$, but valid for $|\pi_k^{-s}| \neq 1$. Each recursive step in this construction assures that the lead term, along with all of its multiplicative factors, are subtracted out, just as with the cancellations with the sieve of Eratosthenes. It is instructive to compare each iteration with that of the sieve (Fig. 1.5, p. 33).

Exercise: Work out the RoC for $k = 2$.

---

[62]This is known as *Euler's sieve*, as distinguish from the *Eratosthenes sieve*.

Poles of $\zeta_k(s)$

Discuss RoC for each term in the product representation, and the anti-causal nature of $\zeta_k(1-s)$. Riemann's zeta function $\zeta(s) \in \mathbb{C}$ (Fig. **??**) proposed that the argument be complex (first actually explored by Chebyshev in 1850, (Bombieri, 2000)), extending $\zeta(s)$ into the complex plane, $(s \in \mathbb{C})$, making it a complex analytic function.

Given that $\zeta(s)$ is a complex analytic function, one might naturally wonder if $\zeta(s)$ has an inverse Laplace transform. There seems to be very little written on this topic (Hill, 2007). We shall explore this question further here.
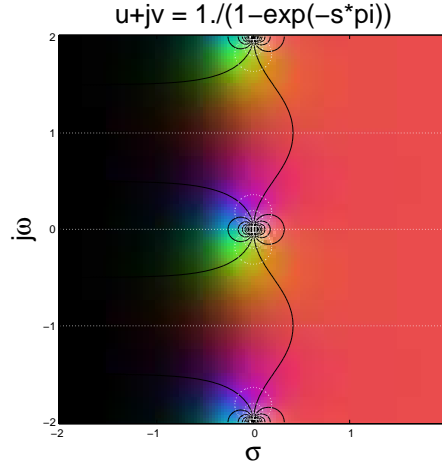
Figure 1.20: *Plot of $w(s) = \frac{1}{1-e^{-s \ln \pi_1}}$ (note $\pi_1 = 2$), which is factor $\zeta_1(s)$ (Eq. 1.85). Here $w(s)$ has poles where $e^{s_n \ln 2} = 1$, namely where $\omega_n \ln 2 = n2\pi$, as seen in the colorized–map ($s = \sigma + \omega\jmath$ is the Laplace frequency [rad]).*

One may now identify the poles of $\zeta_k(s)$ ($p \in \mathbb{N}$), which are required to determining the RoC. For example, the pth factor of Eq. 1.85, expressed as an exponential, is

$$\zeta_k(s) \equiv \frac{1}{1-\pi_k^{-s}} = \frac{1}{1-e^{-sT_k}} = \sum_{k=0}^{\infty} e^{-skT_k}, \tag{1.87}$$

where $T_k \equiv \ln \pi_k$. Thus $\zeta_p(s)$ has poles at $-s_n T_p = 2\pi n\jmath$ (when $e^{-sT_p} = 1$), thus

$$\omega_n = \frac{2\pi n}{T_k},$$

with $-\infty < n \in \mathbb{Z} < \infty$. These poles are the eigen-modes of the zeta function. A domain-colorized plot of this function is provided in Fig. 1.20. It is clear that the RoC of ζ_k is $\Re > 0$. It would be helpful to determine why $\zeta(s)$ as such a more restrictive RoC than each of its factors.

Inverse Laplace transform

The inverse Laplace transform of Eq. 1.87 is an infinite series of delays of delay T_p[63] (Table 1.5, p. 71)

$$Z_p^{eta}(t) = \delta(t))_{T_p} \equiv \sum_{k=0}^{\infty} \delta(t - kT_p) \leftrightarrow \frac{1}{1-e^{-sT_p}}. \tag{1.88}$$

Inverse transform of Product of factors

The time domain version of Eq. 1.85 may be written as the convolution of all the $Z_k^{eta}(t)$ factors

$$Z^{eta}(t) \equiv Z_2^{eta} \star Z_3^{eta}(t) \star Z_5^{eta}(t) \star Z_7^{eta}(t) \cdots \star Z_p^{eta}(t) \cdots, \tag{1.89}$$

where $\star$ represents time convolution (Table 1.5, p. 71).

---

[63]Here we use a shorthand double-parentheses notation to define the one-sided infinite sum $f(t))_T \equiv \sum_{k=0}^{\infty} f(t - kT)$.

$$Z_p^{eta}(t) = \alpha Z_p^{eta}(n - T_p) + \delta(t)$$
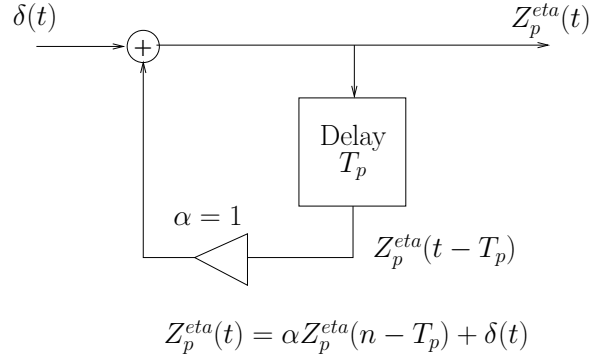
Figure 1.21:   *This feedback network is described by a time-domain difference equation with delay $T_p = \ln \pi_k$, has an all-pole transfer function, given by Eq. 1.90. Physically this delay corresponds to a delay of T_p [s].*

Physical interpretation

Such functions may be generated in the time domain as shown in Fig. 1.21, using a feedback delay of T_p [s] as described by the equation in the figure, with a unity feedback gain $\alpha = 1$.

$$Z^{eta}(t) = Z^{eta}(t - T_n) + \delta(t).$$

Taking the Laplace transform of the system equation we see that the transfer function between the state variable $q(t)$ and the input $x(t)$ is given by $Z_p^{eta}(t)$. Taking the $\mathcal{L}$, we see that $\zeta(s)$ is an all-pole function

$$\zeta_p(s) = e^{-sT_n}\zeta_p(s) + 1(t), \text{ or } \zeta_p(s) = \frac{1}{1 - e^{-sT_p}}. \tag{1.90}$$

Discussion: In terms of the physics, these transmission line equations are telling us that $\zeta(s)$ may be decomposed into an infinite cascade of transmission lines (Eq. 1.89), each having a unique delay given by $T_k = \ln \pi_k$, $\pi_k \in \mathbb{P}$, the log of the primes. The input admittance of this cascade may be interpreted as an analytic continuation of $\zeta(s)$ which defines the eigen-modes of that cascaded impedance function.

   Working in the time domain provides a key insight, as it allows us to determine the analytic continuation of the infinity of possible continuations, which are not obvious in the frequency domain. Transforming to the time domain is a form of analytic continuation of $\zeta(s)$, that depends on the assumption that $Z^{eta}(t)$ is one-sided in time (causal).

Additional relations: Some important relations provided by both Euler and Riemann (1859) are needed when studying $\zeta(s)$.

   With the goal of generalizing his result, Euler extended the definition with the functional equation

$$\zeta(s) = 2^s \pi^{s-1} \, \sin\left(\frac{\pi s}{2}\right) \, \Gamma(1 - s) \, \zeta(1 - s). \tag{1.91}$$

This seems closely related to Riemann's *time reversal symmetry* properties (Bombieri, 2000)

$$\pi^{-s/2}\Gamma\left(\frac{s}{2}\right)\zeta(s) = \pi^{-(1-s)/2}\Gamma\left(\frac{1 - s}{2}\right)\zeta(1 - s).$$

This equation is of the form $F\left(\frac{s}{2}\right)\zeta(s) = F\left(\frac{1-s}{2}\right)\zeta(1 - s)$ where $F(s) = \Gamma(s)/\pi^s$.

   As shown in Table 1.5 the $\mathcal{L}^{-1}$ of $f(-t) \leftrightarrow F(-s)$ is simply a time-reversal. This leads to a causal and anti-causal function that are symmetric about $\Re s = 1/2$ (Riemann, 1859). It seems likely this is an important insight into the Euler's functional equation.

Riemann (1859, page 2) provides an alternate integral definition of $\zeta(s)$, based on the complex contour integration[64]

$$2\sin(\pi s)\Gamma(s-1)\zeta(s) = \jmath \oint_{x=-\infty}^{\infty} \frac{(-x)^{s-1}}{e^x - 1} dx. \overset{-x \to y\jmath}{=} \oint_{y=-\infty\jmath}^{\infty\jmath} \frac{(y\jmath)^{s-1}}{e^{-y\jmath} - 1} dx\jmath.$$

Given the $\zeta_k(s)$ it seems important to look at the inverse $\mathcal{L}$ of $\zeta_k(1-s)$, to gain insight into the analytically extended $\zeta(s)$

Integral definition of the complex Gamma function $\Gamma(s)$ The definition of the complex analytic Gamma function (p. 71)

$$\Gamma(s+1) = s\Gamma(s) \equiv \int_0^\infty \xi^s e^{-\xi} d\xi,$$

which is a generalization of the real integer factorial function $n!$.

Zeros of $\zeta(s)$ We are still left with the most important question of all "Where are the zeros of $\zeta(s)$?" Equation 1.90 has no zeros, it is an all-pole system. The cascade of many such systems is also all-pole. As I see it, the issue is what is the actual formula for $\zeta(s)$?

1.4.12 Lec 34: **Exam III (Evening Exam)**

1.5 Vector Calculus (Stream 3b)

1.5.1 Lec 35: **Gradient ∇, divergence $\nabla\cdot$ Curl $\boldsymbol{\nabla}\times$, and Laplacian ∇^2**

Before we can define the vector operations $\nabla(), \boldsymbol{\nabla}\cdot(), \boldsymbol{\nabla}\times(), \nabla^2()$ we must define scalar and vector fields.

Scalar and vector fields

A field is a function of $(x, y, z) \in \mathbb{R}^3$ than is analytic, namely has a Taylor series, meaning it can take derivatives with respect to $\boldsymbol{x} = (x, y, z)$. An example of a scalar field is $T(\boldsymbol{x}) = xy$, or $\Phi(\boldsymbol{x}) = e^x \log(y)z$. In every case a scalar field is an *analytic function* of $\boldsymbol{x}$, meaning it must be a *smooth* function of $\boldsymbol{x} = (x, y, z)$.

Think of the scalar field for the case of a voltage $\Phi(x, y, z)$, or temperature $T(x, y, z)$, distributed in $\mathbb{R}^3$, say between two finite sized capacitor plates, or the temperature $T(x, y, z)$ in a room, or within a computer chip. If the furnace is on, blowing hot air into the room, it is not a smooth (analytic) function of $\boldsymbol{x}$. Likewise, if the computer chip has heat-generating transistors inside, the temperature would not be smooth, thus not analytic. Every scalar field can play the role of a *potential*.

A vector field is a ordered set of scalar fields, for example

$$\boldsymbol{A}(\boldsymbol{x}) = [\phi(\boldsymbol{x}), \psi(\boldsymbol{x}), \theta(\boldsymbol{x})]^T$$

is a vector field in $\mathbb{R}^3$, as long as each of the three function is analytic. For example $\boldsymbol{A}(\boldsymbol{x}) = [x, xy, xyz]^T$ is a legal vector field. An excellent discussion of vector potentials may be found in (Feynman, 1970, p. NN).

---

[64]Verify Riemann's use of x, which is taken to be real rather than complex. This could be more natural (i.e., modern Laplace transformation notation) if $-x \to y\jmath \to z$.

The physical meaning of a potential: The important question is: "What is the physical meaning of the scalar and vector field?" This is best answered with a few examples. If a temperature field is constant (i.e., $T(\boldsymbol{x}) = T_o$), there can be no heat flux (flow of heat), because it is the change in the temperature (potential) that causes flux. The same holds for every potential.

Note how the difference in the temperature is proportional to the heat flux. Likewise it is the voltage difference that drives an electrical current. This is simply stated by Ohm's Law

$$I = \frac{V_2 - V_1}{R},$$

5 namely the drop in voltage divided by the resistance, defines the current (electrical flux). More generally the impedance is the ratio of the voltage drop over the current. In electrical is is common to define a *ground* point all voltages are referenced to. This allows one to abstract away (hide) the difference in voltage. In fact, it is the difference across any impedance that drives the current. Other physical examples are the pressure, which drives velocity. An interesting case is the velocity of a fluid with no
10 viscosity, which may also be defined via a potential. More on this in section 1.5.5 (p. 107).

It is very helpful to look at the units of the scalar or vector potential, to help identify that it is a potential. But a proper mathematical definition seems to be that it is must be an analytic function of $\boldsymbol{x}$.

As we have learned in earlier chapters, the impedance can be complex functions of complex fre-
15 quency. A capacitor has an impedance $Z(s) = 1/sC_o$ and an inductor $Z(s) = sL_o$. For each of these examples, the voltage is a potential, which most generally is a complex analytic function of s, as well as an analytic function of $\boldsymbol{x}$. If the potential is not analytic, then the impedance would not be well defined.

Name	Input	Output	Operator	Mnemonic
Gradient	Scalar	**Vector**	$-\nabla()$	
Divergence	**Vector**	Scalar	$\nabla \cdot ()$	
Curl	**Vector**	**Vector**	$\nabla \times ()$	
Laplacian	Scalar	**Vector**	$\boldsymbol{\nabla}\cdot\nabla = \nabla^2()$	DoG
Vector Laplacian	**Vector**	**Vector**	$\nabla\boldsymbol{\nabla}\cdot = \boldsymbol{\nabla}^2()$	GoD

Table 1.8:   *The three vector operators manipulate scalar and vector fields, as indicated here. The* gradient *converts scalar fields into vector fields. The divergence eats vector fields and outputs scalar fields. Finally the curl maps vector fields into vector fields. It is helpful to have a name for second order operators (e.g., DoG, GoD: mnemonics defined in Sect. 1.5.4, p. 107).*

Vector differential operators act on scalar and vector fields

There are three key vector differential operators that are necessary for understanding partial differential equations, such as the wave equation and the diffusion equation. All of these use the differential operator ∇ operator, a multi-variate generalization of the derivative

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}.$$

20 When carefully defined, the shorthand form $\partial_{\boldsymbol{x}}$ is convenient. The official name of this operator is *nabla.* It has three basic uses: 1) the *gradient* of a scalar field, the 2) *divergence* of a vector field, and the *curl* of a vector field.

Basic differential vector operator definitions:

1. The *gradient* transforms a scalar potential field into vector field. In $\mathbb{R}^3$ the gradient of a scalar field $\nabla\phi(\boldsymbol{x})$ is defined as

$$\nabla\phi(\boldsymbol{x}) = \left(\frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}}\right)\phi(\boldsymbol{x}), = \frac{\partial\phi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\phi}{\partial y}\hat{\mathbf{j}} + \frac{\partial\phi}{\partial z}\hat{\mathbf{k}}. \tag{1.92}$$

Important examples include: the electric field vector $\boldsymbol{E}(\boldsymbol{x}) = -\nabla\phi(\boldsymbol{x})$ [Volts/m], which is the gradient of a voltage [V], and the force density $\boldsymbol{f}(\boldsymbol{x}) = -\nabla\varrho(\boldsymbol{x})$ [N/m], which is the gradient of a pressure [Pa].

2. The *divergence* of a vector field results in a scalar field. For example, the divergence of the electric field flux vector $\boldsymbol{D}(x)$ [Col/m^2] equals the scalar field charge density $\rho(\boldsymbol{x})$ [Col/m^3]

$$\boldsymbol{\nabla}\cdot\boldsymbol{D}(\boldsymbol{x}) \equiv \left(\frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}}\right) \cdot \boldsymbol{D}(\boldsymbol{x}) = \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = \rho(\boldsymbol{x}). \tag{1.93}$$

This it is analogues to the scalar product between two vectors.

When working with guided waves, narrow tubes of flux, such that the diameter is small compared with the wavelength, the divergence may be accurately approximated as (Appendix H.4, p. 195)

$$\nabla \cdot \boldsymbol{D}(\boldsymbol{x}) = \nabla_r D_r = \frac{1}{A(r)}\frac{\partial}{\partial r}A(r)D_r(r). \tag{1.94}$$

For example in spherical coordinates $A(r) = A_o r^2$, and this reduces to the radial component of the divergence of $\boldsymbol{D}(\boldsymbol{x})$ in spherical coordinates. In cylindrical coordinates $A(r) = A_o r$.

3. The *curl*

$$\boldsymbol{\nabla}\times\boldsymbol{H}(\boldsymbol{x}) \equiv \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \partial_x & \partial_y & \partial_z \\ H_x & H_y & H_z \end{vmatrix} = \boldsymbol{C}(\boldsymbol{x}) \tag{1.95}$$

transforms the vector field $\boldsymbol{H}(\boldsymbol{x})$ into the vector field $\boldsymbol{C}(\boldsymbol{x})$. The notation $|\cdot|$ indicates taking the determinant (Appendix A, p. 165).

Here the curl of the magnetic intensity $\boldsymbol{H}(\boldsymbol{x})$ [Amp/m^2] is equal to the current density $\boldsymbol{C}(x)$ [Amps/m^2]. As before, units can be helpful.

4. The *Laplacian* ($\nabla^2 \equiv \boldsymbol{\nabla}\cdot\nabla$)

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial x^2}. \tag{1.96}$$

Exercise: Construct an example which shows that the definition of the operator agrees with the type of input vector. For Example, show that the gradient must operate on a scalar, to produce a vector Sol: Given a scalar field $\Phi(\boldsymbol{x}) = xyz$ [Volts], the electric field is $\boldsymbol{E}(\boldsymbol{x}) = [\hat{\mathbf{i}}\partial_x, \hat{\mathbf{j}}\partial_y, \hat{\mathbf{k}}\partial_z]xyz = [\hat{\mathbf{i}}yz, \hat{\mathbf{j}}xz, \hat{\mathbf{k}}xy]$[V/m], and that the curl must operate on a vector, and results in a vector.

Examples of the Gradient operator ∇

When the gradient ∇ operates on a scalar field $V(\boldsymbol{x})$, the result is a vector field[65]

$$\boldsymbol{E}(\boldsymbol{x}) = [E_x(\boldsymbol{x}), E_y(\boldsymbol{x}), E_z(\boldsymbol{x})]^T = -\nabla V(\boldsymbol{x}).$$

For example, if $V(\boldsymbol{x}) = 2 + x - y^2$ then $\boldsymbol{E} = -\nabla(x + 2 - y^2) = -\hat{\mathbf{i}} + \hat{\mathbf{j}}2y$. To understand these three operations we therefore need to define the domain and range of their operation, as specified in Table 1.8. Recall that the voltage is the line integral of the electric field

$$V(a) - V(b) = \int_a^b \boldsymbol{E}(\boldsymbol{x}) \cdot d\boldsymbol{x} = -\int_a^b \nabla(V(\boldsymbol{x})) \cdot d\boldsymbol{x} = -\int_a^b \frac{dV}{dx}dx \tag{1.97}$$

---

[65]As before vectors are columns, which take up space on the page, thus we write them as rows and take the transpose to properly format them.

which is simply the fundamental theorem of calculus (Leibniz Theorem). In a charge free region, this integral is independent of the path from a to b, which is the propriety of a conservative system.

The simplest example of a scalar potential is the voltage between two very large (think ∞) conducting parallel planes, or plates (large so that we may ignore the edge effects). In this case the voltage varies linearly (the voltage is complex analytic) between the two plates. For example

$$V(x, y, z) = V_o(1 - x) \tag{1.98}$$

is a scalar potential, thus it is scalar field (i.e., potential). At $x = 0$ the voltage is V_o and at $x = 1$ the voltage is zero. Between 0 and 1 the voltage varies linearly. Thus $V(x, y, z)$ defines a *scalar field.*
5 What is the physical configuration corresponding to $V(\boldsymbol{x}) = 2 + x - y^2$ (the example above)?

If the same setup were used but the two plates were 1x1 [cm^2], with a 1 [mm] air gap, there will be a small "fringe" effect at the edges that would slightly modify the ideal fields. This effect can be made small by changing the air gap to area ratio, so that the ends do not impact the capacitor's value. If we are given a set of three scalar fields, we define a *vector field.* If the three elements of the vector are
10 potentials, then we have a *magnetic vector potential.*

Perhaps a more intuitive example is an acoustic pressure potential $\varrho(\boldsymbol{x}, t)$, which defines a force density $f(\boldsymbol{x}, t) = -\nabla \varrho(\boldsymbol{x}, t)$ (Eq. 1.99, p. 99).

1.5.2   Lec 36: Scalar Wave Equation (Acoustics)

The scalar wave equation is the wave equation for a scalar potential, such as the pressure or voltage in
15 a transmission line. The best known examples are the electrical and acoustic transmission lines, known as the *telephone equations* or simply the transmission line equations. We have previously studied these in terms of their ABCD transmission line representations, which go back to Chinese innovation, as Pell's equations.

This corresponds to the development of the basic equations of acoustics: Conservation of mass and
20 momentum, or sound in a uniform tube.

Horns provide an important generalization of the solution of the 1D wave equation, in regions where the properties (i.e., area of the tube) vary along the axis of wave propagation. Classic applications of horns is vocal tract acoustics, loudspeaker design, cochlear mechanics, and cases having wave propagation in periodic media (Brillouin, 1953). The typical experimental setup is shown in Fig. 1.22. The Laplacian may be accurately approximated for the case of wave-guides, or long tubes whose diameter is much less than a wavelength, but whose length is much greater than a wavelength. This approximation is

$$\nabla^2 \equiv \boldsymbol{\nabla} \cdot \nabla = \frac{1}{A(r)} \frac{\partial}{\partial r} A(r) \frac{\partial}{\partial r},$$

as derived in Appendix H.4, p. 195. The term on the right is an approximation that applies for thin tubes (e.g., spherical, cylindrical coordinates, etc.), with angular symmetry. This is well known in acoustics, where it is known as the *Webster Horn Equation*, following its introduction by Webster (1919).
25 The traditional formulation of the horn is widely known and is discussed in many papers and books on acoustics (Hanna and Slepian, 1924; Morse, 1948; Mawardi, 1949), and in greater detail by (Olson, 1947, p. 101) and (Pierce, 1981, p. 360). Extensive experimental testing for various types of horns (conical, exponential, parabolic) along with a review of horn theory is provided in Goldsmith and Minton (1924).
30 It is frequently stated that the Webster Horn equation (WHEN) is an approximation that only applies to low frequencies, because it is assumed that the area function $A(x)$ is the horn's cross-sectional area rather than the area of the iso-pressure wave-front (Morse, 1948; Shaw, 1970; Pierce, 1981). This "fundamental" limitation of the WHEN is discussed in the majority of the discussions Morse (1948); Shaw (1970); Pierce (1981). In Appendix H (p. 193), based on Gauss' Law, it is shown
35 that this restriction may be totally avoided, *making the Webster theory of the Horn equation exact* (but
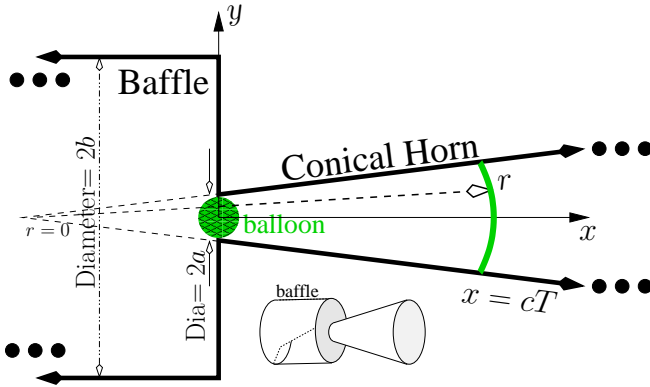
Figure 1.22: *Experimental setup: a semi-∞ long pipe (baffle) of radius b is terminated with a cap, having a hole of radius a, containing a balloon (shown in green). At time $t = 0$ the balloon is pricked and as a result, pressure pulses travel into the baffle (left) and conical horn (right). At time $t = T$ the outbound pressure pulse $p(x, t) = \delta(t - x/c)/r$ has reached position $x = Tc$, where $x = r - r_o$. The balloon is centered at $x = 0, y = 0$; r is measured from the horn's vertex at $r = 0$ (see Fig. 5.1, p. 156 from Olson (1947)).*

still subject to the quasi-statics "low-frequency" approximation). The variables are redefined as the average pressure wave-front and the corresponding volume velocity, each defined on the iso-pressure wave-front boundary (Webster, 1919; Hanna and Slepian, 1924).

The horn equation is a generalization of Sturm-Louville theory, for cases where the pressure (i.e., generalized force potential) does not form a "separable" coordinate system.

The study of wave propagation begins at least as early as Huygens (ca. 1678) and following soon after (ca. 1687) with Sir Isaac Newton's calculation of the speed of sound (Pierce, 1981). To obtain a wave, one must include two basic components, the stiffness of air, and its mass. These two equations shall be denoted *Hooke's Law* ($F = kx$) and *Newton's 2ⁿᵈ Law* ($F = ma$), respectively. In vector form these equations are *Euler's equation* (i.e., conservation of momentum density)

$$-\nabla \varrho(\mathbf{x}, t) = \rho_o \frac{\partial}{\partial t} \boldsymbol{u}(\mathbf{x}, t) \tag{1.99}$$

and the *continuity equation* (i.e., conservation of mass density)

$$-\nabla \cdot \boldsymbol{u} = \frac{1}{\eta_o P_o} \frac{\partial}{\partial t} \varrho(\boldsymbol{x}, t) \tag{1.100}$$

(Pierce, 1981, page 15).

Combining these two equations results in the scalar acoustic (pressure) wave equation

$$\nabla^2 \varrho(\boldsymbol{x}, t) = \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \varrho(\boldsymbol{x}, t).$$

General solution of the wave equation: Apparently Huygen (1690) was the first to gain insight into wave propagation. His important concept, today known as "Huygens' principle," was seriously flawed, as left unexplained why there was no backward traveling wave generated by the forward wave (Miller, 1991). In 1747 a well known French intellectual and mathematician, Jean le Rond d'Alembert, published the general form of the solution to the scalar wave equation, which is

$$\varrho(x, t) = f(x - ct) + g(x + ct) \tag{1.101}$$

where $f(\cdot)$ and $g(\cdot)$ are quite general functions of their argument. That this is the solution may be easily shown by use of the chain rule and taking partials with respect to x and t. Surprisingly, this is the solution even when the functions are not differentiable. For example $f(x - ct) = u(x - ct)$ or even $\delta(x - ct)$ are valid solutions, even though the proof of this is more difficult.

In terms of the physics, d'Alembert's general solution describes two waves, traveling at a speed c, one forward, and one reversed. Thus his solution is quite easily visualized. As previously discussed, Newton (1687) was the first to calculate the speed of sound, be it with a small error ($\sqrt{1.4}$). This was not corrected for over 200 years, following the creation of thermodynamics, and the concept of an adiabatic process.

Related partial differential equations: The wave equation would morph into several other important partial differential equations as the theory was further developed, resulting in the *diffusion equation*

$$D_o \frac{\partial T(\boldsymbol{x},t)}{\partial t} = \nabla^2 T(\boldsymbol{x},t) \tag{1.102}$$

which describes, for example, the temperature $T(\boldsymbol{x},t)$ or for the diffusion of two miscible liquids, *Poisson's equation*

$$\nabla^2 \Phi(\boldsymbol{x},t) = \rho(x,t) \tag{1.103}$$

which holds for gravitational fields, or the voltage around a charge, and *Laplace's equation*

$$\nabla^2 \Phi(\boldsymbol{x},t) = 0, \tag{1.104}$$

which describes, for example, the voltage inside a closed chamber with various voltages on the walls, or the temperature in a box with a temperature distribution on the walls.

Each of these four equations has properties that may be simply explained, and visualized, in simple geometries, and all are some variant of the Laplacian, formed from the *divergence* $\boldsymbol{\nabla}\cdot$ of the gradient ∇, i.e., $\nabla^2() = \boldsymbol{\nabla}\cdot\nabla()$, as described below. In rectangular coordinates

$$\nabla^2 T(\boldsymbol{x}) = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) T(\boldsymbol{x}). \tag{1.105}$$

We first saw the Laplacian in Section 1.4.2 (p. 77), when we studied complex analytic functions. This operator is fundamental in mathematics, starting with complex analytic functions, and then spreading into the wave equation, the diffusion equation, Poisson's equation, and Laplace's' equation. We start with the *Webster horn equation*, which is a key extension of Newton's calculation of the speed of sound.

1.5.3   Lec 37 Webster horn equation

Newton's Law Eq. 1.99 (conservation of momentum) along with the continuity equation (conservation of mass) Eq. 1.100, are modern versions (generalizations) of Newton's starting point, for calculating the speed of sound. Following the simplification of averaging the normal component of the velocity over the iso-pressure wave front (Appendix H, p. 193), two obtain the volume velocity, the two equations may be written in matrix form, in terms of the unknown pressure $\varrho(x,t) \leftrightarrow \mathcal{P}(x,\omega)$ and *volume velocity* $\nu(x,t) \leftrightarrow \mathcal{V}(x,\omega)$

$$\frac{d}{dx} \begin{bmatrix} \mathcal{P}(x,\omega) \\ \mathcal{V}(x,\omega) \end{bmatrix} = - \begin{bmatrix} 0 & \mathscr{Z}(s,x) \\ \mathscr{Y}(s,x) & 0 \end{bmatrix} \begin{bmatrix} \mathcal{P}(x,\omega) \\ \mathcal{V}(x,\omega) \end{bmatrix}, \tag{1.106}$$

as discussed on p. 193, Eqs. H.7, H.9. The Fourier-transform pair of the average pressure and volume velocity are denoted as $\varrho(x,t) \leftrightarrow \mathcal{P}(x,\omega)$ and $\nu(x,t) \leftrightarrow \mathcal{V}(x,s)$, respectively.[66] Here we use the complex *Laplace frequency* $s = \sigma + j\omega$ when referring to the *per-unit-length impedance*

$$\mathscr{Z}(s,x) \equiv s \frac{\rho_o}{A(x)} = sM(x) \tag{1.107}$$

and *per-unit-length admittance*

$$\mathscr{Y}(s,x) \equiv s \frac{A(x)}{\eta_o P_o} = sC(x), \tag{1.108}$$

to clearly indicate that these functions must be *causal*, and except at their poles, *analytic* in s.[67] Here $M(x) = \rho_o/A(x)$ is the horn's *per-unit-length mass*, $C(x) = A(x)/\eta_o P_o$ *per-unit-length compliance*, $\eta_o = c_p/c_v \approx 1.4$ (air).

---

[66]**Notation:** Lower case variables (i.e., $\varrho(x,t), \nu(x,t)$) denote time-domain functions while upper case letters (i.e., $\mathcal{P}(x,\omega), \mathcal{V}(x,\omega), Z(x,s), \mathscr{Z}, \mathcal{Z}(s)$) indicate frequency domain Fourier (ω) or Laplace (s) transforms, the latter being *analytic* in $s = \sigma + j\omega$ for $\sigma > 0$. Matrix notation provides notional transparency.

[67]A function $F(s)$ is said to be *complex analytic* in $s = \sigma + \omega j$ at point a if it may be represented by a convergent Taylor series in s the neighborhood of $s = a$. That is if $F(s) = \sum_{n=0}^{\infty} f_n(a)(s-a)^n$, where $f_n(a) = \frac{d^n F(s)}{ds^n}\Big|_{s=a}$.

Bulk acoustic parameters: The bulk acoustic parameters are the *free-field speed of sound*

$$c = \sqrt{\frac{\eta_o P_o}{\rho_o}}, \tag{1.109}$$

the *specific acoustic resistance* is

$$\rho_o c = \sqrt{\rho_o \eta_o P_o}, \tag{1.110}$$

and the *acoustic resistance* is

$$r_o(x) = \sqrt{\frac{\mathscr{Z}}{\mathscr{Y}}} = \frac{\sqrt{\rho_o \eta_o P_o}}{A(x)} = \frac{\rho_o c}{A(x)}. \tag{1.111}$$

The primitive solutions of the horn equation always depend on a complex wave *propagation function*[68] $\kappa(s)$, defined as the square root of the *product* of $\mathscr{Z}$ and $\mathscr{Y}$:

$$\kappa(s) \equiv \sqrt{\mathscr{Z}(s,x)\mathscr{Y}(s,x)} = \sqrt{\frac{s\rho_o}{A(x)} \times \frac{sA(x)}{\eta_o P_o}} = \frac{s}{c}. \tag{1.112}$$

Note that since the area $A(x)$ cancels in this expression, c is a constant. While horns are generally dispersive, meaning that the solution have a group velocity (and thus a group delay) that depends on frequency, the wave-front speed is always constant, defined by the free-field sound velocity c. These concepts seem to have been first clarified between 1907–1914 by Brillouin (1960, English version), while working with Arnold Sommerfeld (Pleshko and Plàcz, 1969).

Equation 5.4 is the differential equation in the *conjugate variables* $\mathcal{P}, \mathcal{V}$. The product of conjugate variables[69] defines the *intensity*, and the ratio, an *impedance* (Pierce, 1981, p. 37-41).

We shall discuss two alternative matrix formulations of these equations, the ABCD *transmission matrix*, use for computation, and the *impedance matrix*, used when working with experimental measurements (Pierce, 1981, Chapter 7). For each formulation *reciprocity* and *reversibility* show up as different matrix symmetries, as discussed by (Pierce, 1981, p. 195-203) and further here.

The uniform horn

All these formulae are well known for the case of the uniform, or plane-wave horn, having a constant area $A(x) = A_o$. Equation 5.4 (p. 155) then reduces to the classical 1D wave wave equation

$$\mathcal{P}_{xx} + \frac{\omega^2}{c^2}\mathcal{P} = 0. \tag{1.113}$$

Wave solutions: The solutions is the well-known d'Alembert (plane-wave) solution

$$\varrho(x,t) = \varrho^+(t - x/c) + \varrho^-(t + x/c) \;\leftrightarrow\; \alpha(s)\,e^{-\kappa x} + \beta(s)e^{-\kappa L}\,e^{\kappa x}, \tag{1.114}$$

where $a(t) \leftrightarrow \alpha(s)$ and $b(t) \leftrightarrow \beta(s)$ are $\mathcal{L}$ transform pairs, representing the forward and retrograde traveling wave amplitudes. As for the general case given above, it is convenient to define $P_o^+ = 1$ ($\alpha = 1$) and $P_L^- = 1$ ($\beta = e^{-\kappa L}$), thus the normalized primary solutions are

$$\varrho^+(x,t) = \delta(t - x/c) \leftrightarrow e^{-\kappa x} \tag{1.115}$$

and

$$\varrho^-(x,t) = \delta(t + (x - L)/c) \leftrightarrow e^{\kappa(x-L)}. \tag{1.116}$$

---

[68]The *propagation function* is commonly called the *wave-number* (Sommerfeld, 1952, p. 152) or the *propagation-constant*. However since $\kappa(s)$ is neither a number, nor constant, we have appropriately rename it. The meaning of a complex $\kappa(s)$ is addressed in (Sommerfeld, 1952, p. 154). The units of $\kappa(s)$ are reciprocal length, namely $\kappa(s) = 2\pi/\lambda(s)$.

[69]https://en.wikipedia.org/wiki/Conjugate_variables_(thermodynamics)

Radiation Admittance: Since $\partial_x \ln \mathcal{P}^{\pm} = \mp \kappa = s/c$, from Eq. 5.7 we find that $Y_{rad}^{\pm} \equiv y_0 = A_o/\rho_o c$. The signs are chosen to assure that the real admittance $y_0 > 0$ for both types of waves, where the velocity $\mathcal{V}^{\pm}$ is always *into* the port.

1^d **ABCD matrix:**   Repeating the entire argument leading to the ABCD matrix (i.e., Eq. 5.18)

$$\begin{bmatrix} \mathcal{P}(x) \\ \mathcal{V}(x) \end{bmatrix} = \begin{bmatrix} e^{-\kappa x} & e^{\kappa(x-L)} \\ y_0 e^{-\kappa x} & -y_0 e^{\kappa(x-L)} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{1.117}$$

Solving Eq. 1.117 for α and β and evaluating at $x = L$ (and fixing all the signs to be consistent with Eq. 5.8) we find the amplitudes in terms of the pressure and velocity at any point along the line

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \frac{-1}{2y_0 e^{-\kappa L}} \begin{bmatrix} -y_0\, e^{-\kappa(x-L)} & -e^{\kappa(x-L)} \\ -y_0 e^{-\kappa x} & e^{-\kappa x} \end{bmatrix}_L \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix} = \frac{1}{2} \begin{bmatrix} e^{\kappa L} & r_0 e^{\kappa L} \\ 1 & -r_0 \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}, \tag{1.118}$$

where $r_0 \equiv 1/y_0$. We have changed the sign of $-\mathcal{V}_L$ so as to be positive looking into the port at $x = L$. Since in this case $\Delta_L = -2y_0 e^{-\kappa L}$ is independent of x, evaluating it at $x = L$ has no effect. Substitution of Eq. 1.118 into Eq. 1.117 and evaluating the result at $x = 0$ gives the final ABCD matrix

$$\begin{bmatrix} \mathcal{P}_o \\ \mathcal{V}_o \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & e^{-\kappa L} \\ y_0 & -y_0 e^{-\kappa L} \end{bmatrix} \begin{bmatrix} e^{\kappa L} & \mathcal{Z} e^{\kappa L} \\ 1 & -\mathcal{Z} \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix} \tag{1.119}$$

Multiplying these out gives the final transmission matrix as

$$\begin{bmatrix} \mathcal{P}_o \\ \mathcal{V}_o \end{bmatrix} = \begin{bmatrix} \cosh(\kappa L) & r_0 \sinh(\kappa L) \\ y_0 \sinh(\kappa L) & \cosh(\kappa L) \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix},$$

with $\kappa = s/c$, $y_0 = 1/\mathcal{Z} = A_o/\rho_o c$ (Pipes, 1958). The two velocities are defined into their respective ports.

Impedance matrix: The impedance matrix (Eq. 5.20) is therefore

$$\begin{bmatrix} \mathcal{P}_o \\ \mathcal{P}_L \end{bmatrix} = \frac{\mathcal{Z}}{\sinh(\kappa L)} \begin{bmatrix} \cosh(\kappa L) & 1 \\ 1 & \cosh(\kappa L) \end{bmatrix} \begin{bmatrix} \mathcal{V}_o \\ -\mathcal{V}_L \end{bmatrix}.$$

The input admittance: Given the input admittance of the horn, it is already possible to determine if it is uniform, without further analysis. Namely if the horn is uniform and infinite in length, the input impedance at $x = 0$ is

$$Y_{in}(0, s) \equiv \frac{\mathcal{V}(0, \omega)}{\mathcal{P}(0, \omega)} = y_0,$$

since $\alpha = 1$ and $\beta = 0$. That is for an infinite uniform horn, there are no reflections.

When the horn is terminated with a fixed impedance Z_L at $x = L$, one may substitute pressure and velocity measurements into Eq. 1.118 to find α and β, and given these, one may calculate the *reflectance* at $x = L$ (see Eq. 5.19)

$$\Gamma_L(s) \equiv \left.\frac{\mathcal{P}^-}{\mathcal{P}^+}\right|_{x=L} = \frac{e^{-\kappa L}\beta}{\alpha} = \frac{\mathcal{P}(L,\omega) - \mathcal{Z}\mathcal{V}(L,\omega)}{\mathcal{P}(L,\omega) + \mathcal{Z}\mathcal{V}(L,\omega)} = \frac{Z_L - \mathcal{Z}}{Z_L + \mathcal{Z}}$$

given accurate measurements of the throat pressure $\mathcal{P}(0,\omega)$, velocity $\mathcal{V}(0,\omega)$, and the characteristic impedance of the input $\mathcal{Z} = \rho_o c/A(0)$.

The conical horn

The uniform is very basic with plane waves being reflected from the ends, leading to a simple solution of the wave equation. The next step up in difficulty is the conical horn, which is closely related to the spherical radiator. This setup is described in Fig. 1.22). Using the general expression for the divergence of the velocity, the Webster equation in pressure becomes (Eq. 1.94, p. 97, Table 5.1, p. 154)

$$\frac{1}{A(x)}\frac{\partial}{\partial x}A(x)\frac{\partial}{\partial x}\varrho(x,t) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\varrho(x,t), \tag{1.120}$$

where the area is $A(x) = A_o x^2$. Working this out results in the *spherical wave equation* (Appendix H.4, p. 195)

$$\frac{\partial^2}{\partial x^2}x\varrho(x,t) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}x\varrho(x,t) \leftrightarrow \frac{1}{c^2}\frac{\partial^2}{\partial t^2}x\mathcal{P}_c(x,\omega), \tag{1.121}$$

having d'Alembert solutions of the form

$$\mathcal{P}(x,\omega) = \mathcal{P}^+(\omega)\frac{e^{-\kappa x}}{x} + \mathcal{P}^-(\omega)\frac{e^{\kappa x}}{x}, \tag{1.122}$$

where $\mathcal{P}^\pm(\omega)$ represent the frequency dependent amplitudes of the outbound and inbound waves to be determined from the boundary conditions, as for the uniform horn.

Now that we have seen examples of the Webster horn equation Eq. 5.4, we can appreciate the importance of the gradient $\nabla()$, divergence $\boldsymbol{\nabla}\cdot()$, and the Laplacian $\nabla^2()$ in formulating partial differential equations (Sect. 1.5.1, p. 95).

Next we systematically introduce the three vector operators $\nabla()$, $\boldsymbol{\nabla}\cdot()$ and $\boldsymbol{\nabla}\times()$ in terms of the integral definitions. In Section 1.5.3 we gave the differential definition, which is what one uses when computing. However a better definition based on the *integral definitions*, gives much more insight into the inner-workings of these operators.

1.5.4   Lec 38: Integral forms of $\nabla()$, $\boldsymbol{\nabla}\cdot()$ and $\boldsymbol{\nabla}\times()$

There are two definitions for each operator, the point or differential form, and the integral limit form. The integral form gives an intuitive view of what the operator does, and in the limit, converges to the differential form. These two definitions are discussed next.

The divergence and the curl each have fundamental theorems: Gauss' Law (divergence theorem) and Stokes' Law (curl theorem). Without the use of these very fundamental vector calculus theorems, Maxwell's equations cannot be understood. The history of these important theorems is discussed in the caption of Fig. 1.11, p. 45.

Gradient: $\boldsymbol{E} = -\nabla\phi(x,y,z)$ [V/m]

As briefly summarized on page 96, the differential definition of the gradient maps $\mathbb{R}^1 \underset{\nabla}{\mapsto} \mathbb{R}^3$. For example, the electric field strength is the gradient of the voltage

$$\boldsymbol{E}(\boldsymbol{x}) = -\left[\partial_x,\ \partial_y,\ \partial_z\right]^T \phi(\boldsymbol{x}) = -\left[\frac{\partial\phi}{\partial x},\ \frac{\partial\phi}{\partial y},\ \frac{\partial\phi}{\partial z}\right]^T(\boldsymbol{x}).$$

The negative sign is optional.

The gradient is the slope of the tangent plane of the potential $\phi(\boldsymbol{x})$ at $\boldsymbol{x}$ pointing in the direction of the maximum slope. The gradient (i.e., $\boldsymbol{E}(\boldsymbol{x})$) is $\perp$ of the constant potential contours of $\phi(\boldsymbol{x})$ (the iso-potential lines), since derivatives along any iso-contours must be zero.

As a specific example, consider the paraboloid $z = 1 - (x^2 + y^2)$ as the potential, with iso-potential circles of constant z that have radius of zero at $z = 1$, and unit radius at $z = 0$. The negative gradient

$$\boldsymbol{E}(\boldsymbol{x}) = -\nabla z(x,y) = 2(x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + 0\hat{\mathbf{k}})$$

is $\perp$ to the circles of constant radius (constant z), thus points in the direction of the radius.

If one were free-fall skiing this surface, they would be the first one down the hill. Normally skiers try to stay close to the iso-clines (not in the direction of the gradient), so they can stay in control. If you ski an iso-cline, you must walk, since there is no pull due to gravity.

The integral definition of the gradient: The gradient may be defined in terms of the integral

$$\nabla\phi(x,y,z) \equiv \lim_{|\mathcal{S}|\to 0} \left\{ \frac{\int_{\mathcal{S}} \phi(x,y,z)\,\hat{\mathbf{n}}\,d|\mathcal{S}|}{||\mathcal{S}||} \right\}, \tag{1.123}$$

over a closed surface $\mathcal{S}$, having area $|\mathcal{S}|$ and volume $||\mathcal{S}||$, centered at (x,y,z)(Greenberg, 1988, p. 773).[70]. Here $\hat{\mathbf{n}}$ is the unit vector perpendicular to the surface $\mathcal{S}$

$$\hat{\boldsymbol{n}} = \frac{\nabla\phi}{||\nabla\phi||}.$$

5   The dimensions of Eq. eq:GradInt are in the units of the potential times the area, divided by the volume, as needed for a gradient (e.g., [Volts/m]).

The natural way to define the surface and volume is to place the surface on the iso-potential surfaces, forming either a cube or pill-box shaped volume. As the volume goes to zero, so must the area. One must avoid *irregular volumes* such that the area is finite as the volume goes to zero (Greenberg, 1988, 10 footnote p. 762).

A well known example is the potential

$$\phi(x,y,z) = \frac{Q}{\epsilon_o\sqrt{x^2+y^2+z^2}} = \frac{Q}{\epsilon_o R} \quad \text{[Volts]}$$

around a point charge Q [SI Units of Coulombs]. The constant ϵ_o is the *permittivity* [Farad/m^2].

How does this work?   To better understand what Eq. 1.123 means, consider a three-dimentional Taylor series expansion of the potential in $\boldsymbol{x}$ about the limit point $\boldsymbol{x}_o$ To compute the *higher order terms* (HOT) one needs the *Hessian* matrix[71]

$$\phi(\boldsymbol{x}) \approx \phi(\boldsymbol{x}_o) + \nabla\phi(\boldsymbol{x})\cdot(\boldsymbol{x}-\boldsymbol{x}_o) + \text{HOT}.$$

We could define the gradient using this relationship as

$$\nabla\phi(\boldsymbol{x}_o) = \lim_{\boldsymbol{x}\to\boldsymbol{x}_o} \frac{\phi(\boldsymbol{x})-\phi(\boldsymbol{x}_o)}{\boldsymbol{x}-\boldsymbol{x}_o}.$$

For this definition to apply, $\boldsymbol{x}$ must approach $\boldsymbol{x}_o$ along $\hat{\boldsymbol{n}}$.

The natural form for the surface $|\mathcal{S}|$ is to lie along the iso-potential surfaces as much as possible, so that the integral is a constant (the potential) times the area. The remainder of the surface must be 15 perpendicular to these iso-potential surfaces, in the direction of the gradient, or maximum change of the potential. The secret to the integral definition is in taking the limit. As the volume $||\mathcal{S}||$ shrinks to zero, the HOT terms are small, and the integral reduces to the first order term in the Taylor expansion, since the constant term integrates to zero. Such a construction is used in the proof of the Webster Horn equation (Appendix H; Fig. H.1, p. 194).

---

[70]See further discussions on pages Greenberg (1988, pp. 778, 791, 809)
[71]$H_{i,j} = \partial^2(\boldsymbol{x})\phi/\partial x_i\partial x_j$, which will exist if the potential is analytic in $\boldsymbol{x}$ at $\boldsymbol{x}_o$.
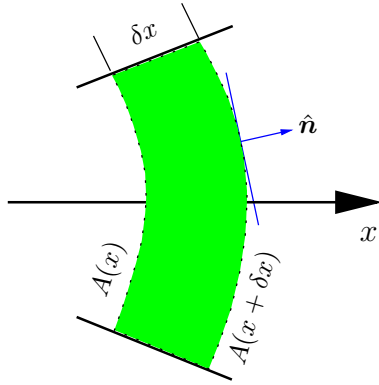
Figure 1.23: *Derivation of Gauss's Law: The divergence of the velocity $\nabla \cdot \mathbf{u}$, within δx, shown as the filled shaded region, is integrated over the closed volume. Then the divergence theorem is applied, transforming the integral to a surface integral normal to the surface, resulting in the difference of the two volume velocities $\delta \nu = \nu(x + \delta x) - \nu(x) = [\mathbf{u}(x + \delta x) \cdot \mathbf{A}(x + \delta x) - \mathbf{u}(x) \cdot \mathbf{A}(x)]$, as the limit of $\delta x \to 0$. Because the ends are in the direction of $\mathbf{u}$, the contribution to the volume velocity from the ends is zero.*

Divergence: $\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho \; [\mathbf{Col/m^3}]$

As briefly summarized on page 97, the differential definition of the gradient which maps $\mathbb{R}^3 \underset{\nabla}{\mapsto} \mathbb{R}^1$ is

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} \equiv [\partial_x, \, \partial_y, \, \partial_z] \cdot \boldsymbol{D} = \left[\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} \right] = \rho(x, y, z)$$

The divergence is a direct measure of the flux (flow) of vector field it operates on ($\boldsymbol{D}$), coming from $\boldsymbol{x}$. A vector field is said to be *incompressible* if the divergence of that field is zero. It is therefore *compressible* when the divergence is non-zero, Compared to air, water is considered to be incompressible. However at very low frequencies, air can also be considered as incompressible. Thus the definition of compressible depends on the wavelength in the medium, so the terms must be used with some awareness of the circumstances.

  If we restrict ourselves to the electrical case, then at this point it is helpful to think of the physics of the electron, a negatively charged particle that is a single point in space. Of course this it is nonsense to consider an electron to be a point, but given that it is so small (it is so small it is difficult to estimate its size). This size is called the Lorentz radius, and is estimated to be 2.810^{-15} [m]. One could summarize the Lorentz radius as follows: *Here lie many unsolved problems in physics.* More specifically, at dimensions of the Lorentz radius, what exactly is the structure of the electron?

  Ignoring the difficulties, if one integrates the charge density of the electron over the Lorentz radius and places the total charge at a single point, then one may make a grossly over simplified model of the electron. For example the *Electric Displacement* (flux density) around a point charge ($\boldsymbol{D} = \epsilon_o \boldsymbol{E}$)

$$\boldsymbol{D} = -\epsilon_o \nabla \phi(R) = -Q \nabla \left\{ \frac{1}{R} \right\} = -Q \, \delta(R).$$

This is a formula taught in many classic texts, but one should remember how crude a model of an electron it is. But it does describe the electric flux in an easily remembered form. However, computationally, it is less nice, due to the delta function. A limitation of this model is that the electron has a magnetic dipole moment (aka, *spin*), which a simple point charge does not capture. When placed in a magnetic field, due to the dipole, the electron will align itself with the field.

Divergence and Gauss' Law

Like the gradient, the divergence of a vector field may be defined as the surface integral of a *compressible* vector field, as a limit as the volume enclosed by the surface goes to zero. As for the case of the gradient, for this definition to make sense, the surface $\mathcal{S}$ must be a closed, defining volume $\mathcal{V}$. The difference is that the surface integral is over the normal component of the a vector field being operated on. Specifically (Greenberg, 1988, p. 762-763)

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = \lim_{|\mathcal{S}| \to 0} \left\{ \frac{\int_{\mathcal{S}} \boldsymbol{D} \cdot \hat{\mathbf{n}} \, d|\mathcal{S}|}{||\mathcal{S}||} \right\} = \rho(x, y, z). \tag{1.124}$$

As with the case of the gradient we have defined the surface as $\mathcal{S}$, its area as $|\mathcal{S}|$ and the volume within as $||\mathcal{S}||$. Here $\hat{\mathbf{n}}$ is a unit vector normal to the surface $\mathcal{S}$. The limit as the volume goes to zero, defines the total flux across the surface. Thus the surface integral is a measure of the total flux $\perp$ to the surface. It is helpful to compare this formula with that for the gradient Eq. 1.123.

5 **Gauss' Law:**   The above definitions resulted in a major breakthrough in vector calculus, and the first[72] fundamental theorem of vector calculus, *Gauss' Law*[73]

> *The surface integral over the normal component of the flux (i.e., the total flux) is equal to the volume integral over the divergence of the flux.*

For the electrical case this is equivalent to the observation that the total flux across the surface is equal to the net charge enclosed by the surface. Since the volume integral over charge density $\rho(x, y, z)$ is total charge enclosed Q_{enc} [Col]

$$Q_{enc} = \iiint_{\mathcal{V}} \boldsymbol{\nabla}\cdot\boldsymbol{D}\,dV = \iint_{\mathcal{S}} \boldsymbol{D}\cdot\hat{\boldsymbol{n}}\,dA. \tag{1.125}$$

It trivially follows that when $Q_{enc} = 0$. That is when the surface integral over the normal component
10 of $\boldsymbol{D}(\boldsymbol{x})$, and the volume integral the charge density $\boldsymbol{\nabla}\cdot\boldsymbol{D} = \rho(\boldsymbol{x})$, are both zero.

Hopefully this result makes common sense, and can be grasped intuitively.

Integral definition of the Curl: $\boldsymbol{\nabla}\times\boldsymbol{H} = \boldsymbol{C}$ [Amps]

As briefly summarized on page 97 (p. 97), the differential definition of the Curl maps $\mathbb{R}^3 \underset{\nabla\times}{\mapsto} \mathbb{R}^3$. For example, the curl of the *magnetic field strength* $\boldsymbol{H}(\boldsymbol{x})$ is equal to the total current $\boldsymbol{C}$

$$\boldsymbol{\nabla}\times\boldsymbol{H} \equiv \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \partial_x & \partial_y & \partial_z \\ H_x & H_y & H_z \end{vmatrix} = \boldsymbol{C}.$$

As we shall see in Sect. 1.5.6 (p. 110), the curl and the divergence are both key when writing out Maxwell's equation four equations. Without a full understanding of these two differential operators
15 $(\boldsymbol{\nabla}\cdot(), \boldsymbol{\nabla}\times())$, there is ho hope of understanding Maxwell's basic result, typically viewed as the most critical elements of mathematical physics, and the starting point for Einstein's relativity theories. Some will say that quantum mechanics fall outside the realm of MEs, but this at least open to debate, if not hotly debated.

The curl is a measure of the rotation of a vector field. If this were water, it would correspond to
20 the angular momentum of the water, such as water going down the drain, as in a whirl pool, or with wind, a tornado. A spinning top is another an excellent example, given a spinning solid body. While a top (aka gyro-scope) will fall over if not spinning, once it is spinning, it can stability stand on its pointed tip. These structure are stable due to conservation of angular momentum: Once something is spinning, it will continue to spin.
25 By way of example, when $\boldsymbol{H} = -y\hat{\mathbf{i}} + x\hat{\mathbf{j}} + 0\hat{\mathbf{k}}$ then $\boldsymbol{\nabla}\times\boldsymbol{H} = 2\hat{\mathbf{k}}$, thus has a constant *rotation*; when $\boldsymbol{H} = 0\hat{\mathbf{i}} + 0\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$ then $\boldsymbol{\nabla}\times\boldsymbol{H} = \boldsymbol{0}$ has a curl of zero, thus has zero rotation thus is said to be *irrotational*. There are rules that precisely govern when a vector field is rotational versus irrotational, and compressible versus incompressible. These classes are dictated by Helmholtz's Theorem, the third fundamental theorem of vector calculus (Eq. 1.128, p. 108).

---

[72]As best I know, the three fundamental theorems of vector calculus, Gauss' Law, the Stokes' Law and Helmholtz's theorem (aka decomposition) are not ordered.

[73]aka, Gauss's Law

$$\hat{\mathbf{n}} \perp dA$$
$$\mathcal{S} \text{ Area (open)}$$
$$\boldsymbol{\nabla} \times \boldsymbol{H} \equiv \lim_{|\mathcal{S}| \to 0} \left\{ \frac{\int_{\mathcal{S}} \hat{\boldsymbol{n}} \times \boldsymbol{H} \, d|\mathcal{S}|}{||\mathcal{S}||} \right\}$$

$$\mathcal{I}_{enc} = \iint_{\mathcal{S}} (\boldsymbol{\nabla} \times \boldsymbol{H}) \cdot \hat{\boldsymbol{n}} \, dA = \oint_{\mathcal{B}} \boldsymbol{H} \cdot d\boldsymbol{l} \quad \text{[Amps]}$$

$$dl \quad \mathcal{B} \text{ Boundary}$$
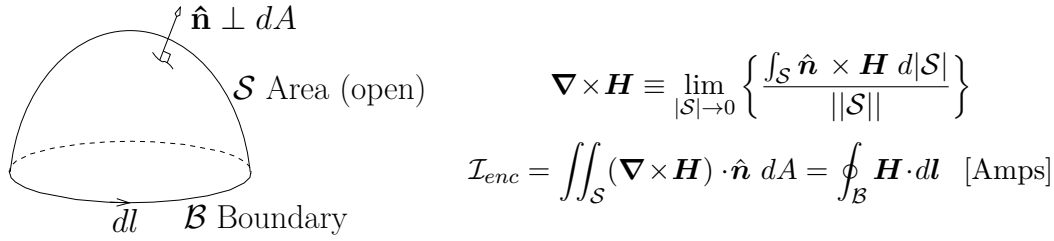
Figure 1.24: *The integral definition of the curl is similar to that of the divergence (Greenberg, 1988, p. 774), except that 1) the integration is over the plane tangent to the surface, defined as the cross product of $\hat{\boldsymbol{n}}$, the normal to the surface, crossed with a current density of $\hat{\boldsymbol{n}} \times \boldsymbol{H}$ [Amps/m^2], and 2) the surface is open, giving it a boundary $\mathcal{B}$, along the open edge. As with the divergence, which lead to Gauss' Law, this definition leads to a second fundamental theorem of vector calculus:* Stokes' Law *(aka the* Curl *theorem).* There is a serious problem here, since an open area does not have any enclosed volume (Greenberg, 1988, p. 814). Greenberg (1988, p. 824) trys to resolve this, but perhaps fails.

Curl and Stokes Law: As in the cases of the gradient and divergence, the curl also may be written in integral form, allowing for the physical interpretation of its meaning.

Surface integral definition of $\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{C}$ where the current $\boldsymbol{C}$ is $\perp$ to the rotation plane of $\boldsymbol{H}$. As shown in Fig. 1.24, Stokes law states that the *open surface integral* over the normal component of the curl of the magnetic field strength ($\hat{\boldsymbol{n}} \cdot \dot{\boldsymbol{\nabla}} \times \boldsymbol{H}$ [Amps/m^2]) is equal to the line integral $\oint_{\mathcal{B}} \boldsymbol{H} \cdot d\boldsymbol{l}$ along the boundary $\mathcal{B}$. Namely, Stokes' Law is

$$\mathcal{I}_{enc} = \iint_{\mathcal{S}} (\boldsymbol{\nabla} \times \boldsymbol{H}) \cdot \hat{\boldsymbol{n}} \, dA = \oint_{\mathcal{B}} \boldsymbol{H} \cdot d\boldsymbol{l} \quad \text{[Amps]}, \tag{1.126}$$

namely

> *The line integral of $\boldsymbol{H}$ along the open surface's boundary $\mathcal{B}$ is equal to the total current enclosed $\mathcal{I}_{enc}$ [Amps].*   5

Summing it up: As mentioned earlier (Fig. 1.11, p. 45), the history of the discovery and proof of this theorem was both complex and interesting.

Since integration is a linear process (sums of smaller elements), one may tile, or tessellate the surface, breaking it up into a sum over smaller surfaces, and their boundaries, the sum over which is equal to the integral over the original boundary. This is an important concept, which leads to the proof   10
of Stokes' Law.

The integral formulations of Gauss' and Stokes' Laws use $\hat{\boldsymbol{n}} \cdot \boldsymbol{D}$ and $\boldsymbol{H} \times \hat{\boldsymbol{n}}$ in the integrands. The key distinction between the two laws naturally follows from the properties of the scalar ($\boldsymbol{A} \cdot \boldsymbol{B}$) and vector ($\boldsymbol{A} \times \boldsymbol{B}$) products, as discussed in Sect. 1.3.4, p. 55, and detailed in Fig. 1.12, p. 57. To fully appreciate the differences between Gauss' and Stokes' Laws, these two types of vector products must   15
be mastered.

1.5.5 Lec 39 Second-order operators: Terminology

Besides the above first order vector derivatives, second order combinations exist, the most common being the scalar Laplacian $\boldsymbol{\nabla} \cdot \boldsymbol{\nabla}() = \nabla^2()$ (Table 1.8, p. 96; Appendix H.4, p. 195).

There are other important second-order combinations of ∇, enough that we need a memory aid to   20
remember them. Thus I define mnemonics DoC, DoG, CoG CoC and GoD as follows:

1. DoG: Divergence of the Gradient ($\boldsymbol{\nabla} \cdot \boldsymbol{\nabla} = \nabla^2$), i.e., Laplacian

2. DoC: Divergence of the Curl ($\boldsymbol{\nabla} \cdot \boldsymbol{\nabla} \times$),

3. CoG: Curl of the Gradient ($\boldsymbol{\nabla} \times \nabla$),

4. CoC: Curl of the curl ($\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times$), and

5. GoD: Gradient of the Divergence ($\boldsymbol{\nabla}^2$), or the vector Laplacian.

Two of these, DoC and CoG, are special because they are always zero,

$$\boldsymbol{\nabla} \times \nabla \phi = 0; \ \ \nabla \cdot \boldsymbol{\nabla} \times \boldsymbol{A} = 0,$$

a property that makes them useful in proving the Fundamental theorem of vector calculus (Helmholtz decomposition) (Eq. 1.128, p. 108). A third key vector identity CoC may be expanded as

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{A} = \nabla(\boldsymbol{\nabla} \cdot \boldsymbol{A}) - \boldsymbol{\nabla}^2 \boldsymbol{A}, \tag{1.127}$$

thus defining the vector Laplacian (GoD) $\boldsymbol{\nabla}^2$.

5   When using second order differential operators one must be careful with the order of operations, which can be subtle in a few cases. Most of this is common sense. For example, don't operate on a scalar field with $\boldsymbol{\nabla} \times$, and don't operate on a vector field with ∇. The vector Laplacian GoD must not be though of as $\nabla(\boldsymbol{\nabla} \cdot \boldsymbol{A})$, rather it acts as the Laplacian on each vector component $\nabla^2(\boldsymbol{A}) = \nabla^2 A_x \hat{\mathbf{i}} + \nabla^2 A_y \hat{\mathbf{j}} + \nabla^2 A_z \hat{\mathbf{k}}$.

10   **Helmholtz's decomposition**

We may now restate everything defined above in terms of two types of vector fields that decompose every analytic vector field. The *irrotational fields* is define as one that is "curl free." An *incompressible field* is one that is "diverge free." According to Helmholtz's decomposition, every analytic vector field may be decomposed into independent rotational and a compressible components. Another name for *Helmholtz*

15 *decomposition* is the *Fundamental theorem of vector calculus* (FTVC). Gauss' and Stokes' Laws, along with Helmholtz's decomposition, form the three key fundamental theorems of vector calculus. Images of Helmholtz and Kirchhoff are provided in Fig. 1.25, p. .109.

    The term *magnetic solenoidal field* is one that is generated by a solenoidal coil, is an excellent approximation, uniform inside the coil. As a result, the curl of a solenoidal field is, to a good approx-

20 imation, zero. I recommend you know this term (it is widely used), but suggest the preferred term *irrotational*, Strictly speaking the term "solenoidal field" only applies to a magnetic field produced by a solenoid, thus the term is specific to that case.

    The decomposition of differentiable vector fields: This theorem is easily stated (and proved), but less easily appreciated. A physical description facilitates: *Every vector field may be split into two*

25 *independent parts: rotation and dilation.* We have seen this same idea appear in vector algebra, where the vector and cross products of two vectors are perpendicular (Fig. 1.12), p. 57. Also think of linear and angular momentum, which are independent in that they represent different ways of absorbing and delivering kinetic energy, leading to independent degrees of freedom. Thus the idea of the linear and rotational parts being independent, is a common theme, rooted in geometry. In the same sense, a vector

30 field may be split into a dilation and rotational parts, which are independent (but can interact under certain conditions). An object with mass can be moving along a path, and independently be rotating. The two modes of motion define different types of kinetic energy, transnational and rotational. In some real sense, Helmholtz decomposition quantifies these degrees of freedom, one DoF for translation and three DoFs for rotation. Each eigenmode of vibration can be viewed as a DoF.

    Helmholtz's decomposition is expressed as the linear sum of a scalar potential $\phi(x, y, z)$ (think voltage) and a vector potential (think magnetic vector potential). Specifically

$$\boldsymbol{E}(\boldsymbol{x}, s) = -\nabla \phi(\boldsymbol{x}, s) + \nabla \times \boldsymbol{A}(x, s), \tag{1.128}$$

where ϕ is the scalar and $\boldsymbol{A}$ is the vector potential, as a function of the Laplace frequency s. Of course this decomposition is general (not limited to the electro-magnetic case). It applies to linear fluid vector fields, which includes most liquids and air. When the rotational and dilation become coupled, this relation must break down.[74]
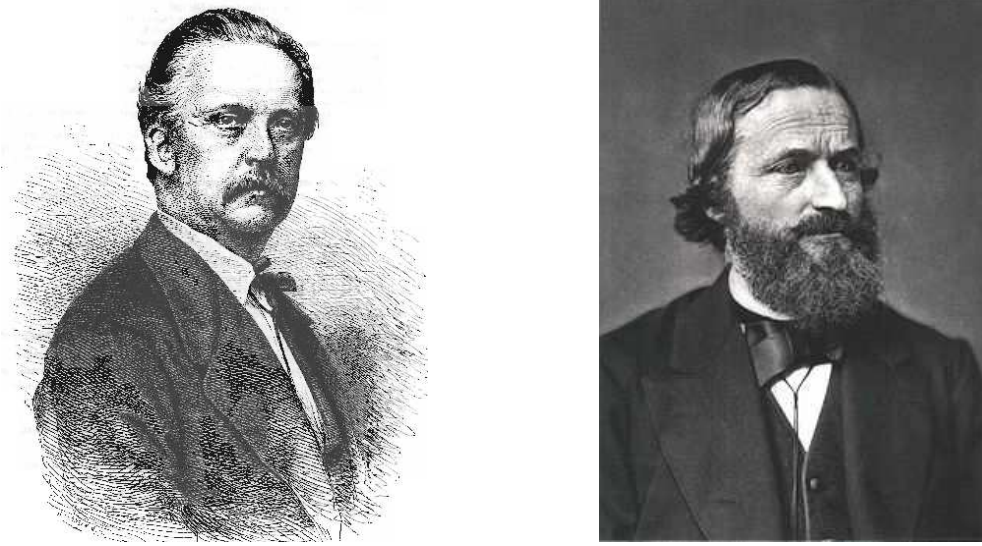
Figure 1.25: *Left: von Helmholtz portrait taken from the English translation of his 1858 paper "On integrals of the hydrodynamic equations that correspond to Vortex motions" (in German) (Helmholtz, 1978). Right: Gustav Kirchhoff*

To show how this relationship splits the vector fields $\boldsymbol{E}$ into its two parts, we need DoC and CoG, the two key vector identities that are always zero for analytic fields: the *curl of the gradient* (CoG)

$$\boldsymbol{\nabla} \times \nabla \phi(\boldsymbol{x}) = 0, \tag{1.129}$$

and the *divergence of the curl*[75] (DoC)

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{A}) = 0. \tag{1.130}$$

The above identities are easily verified by working out a few specific examples, based on the definitions of the three operators, gradient, divergence and curl, or in terms of the operator's integral definitions, defined in Sect. 1.5.4, p. 103. The identities have a physical meaning, as stated above: every vector field may be split into its transnational and rotational parts. If $\boldsymbol{E}$ is the electric field [V/m], ϕ is the voltage and $\boldsymbol{A}$ is the induced rotational part, induced by a current. We shall explore this in our discussion of Maxwell's equations in Sect. 1.5.6 and Chapter 5.

By applying these two identities to Helmholtz's decomposition, we can better appreciate the theorems significance. It is a form of proof actually, once you have satisfied yourself that the vector identities are true. In fact one can work backward using a physical argument, that rotational momentum (rotational energy) is independent of the transnational momentum (transnational energy). Once these forces are made clear, the meaning of the vector operations all take on a very well defined meaning, and the mathematical constructions, centered around Helmholtz's theorem, begins to provide some common-sense meaning. One could conclude that the physics is simply related to the geometry via the scalar and vector product.

Specifically, if we take the divergence of Eq. 1.128, and use the DoG

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \boldsymbol{\nabla} \cdot \{-\nabla \phi + \overset{0}{\cancel{\boldsymbol{\nabla} \times \boldsymbol{A}}}\} = -\boldsymbol{\nabla} \cdot \nabla \phi = -\nabla^2 \phi,$$

---

[74]The nonlinear Navier–Stokes equations may be an example.

[75]Helmholtz was the first person to apply mathematics in modeling the eye and the ear (Helmholtz, 1863a).

since the DoG zeros the vector potential $\boldsymbol{A}(x, y, z)$. If instead we use the CoG

$$\boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{\nabla} \times \{\underbrace{-\boldsymbol{\nabla}\phi}^{0} + \boldsymbol{\nabla} \times \boldsymbol{A}\} = \boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{A} = \nabla(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \boldsymbol{\nabla}^2 \boldsymbol{A},$$

since the CoG zeros the scalar field $\phi(x, y, z)$, and followed with the used of GoD.

Field: $\boldsymbol{v}(\boldsymbol{x}, t)$	Compressible $\boldsymbol{\nabla}\cdot\boldsymbol{v} \neq 0$	Incompressible $\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0$
Rotational $\boldsymbol{\nabla}\times\boldsymbol{v} \neq 0$	$\boldsymbol{v} = \nabla\phi + \boldsymbol{\nabla}\times\boldsymbol{\omega}$ Vector wave Eq. (EM) $\boldsymbol{\nabla^2}\boldsymbol{v} = \frac{1}{c^2}\ddot{\boldsymbol{v}}$	$\boldsymbol{v} = \boldsymbol{\nabla}\times\boldsymbol{w}$ Lubrication theory Boundary layers
Irrotational Conservative $\boldsymbol{\nabla}\times\boldsymbol{v} = 0$	Acoustics $\boldsymbol{v} = \nabla\psi$ $\nabla^2\varrho(\boldsymbol{x}, t) = \frac{1}{c^2}\ddot{\varrho}(\boldsymbol{x}, t)$	Statics $\nabla^2\phi = 0$ Laplace's Eq. $(c \to \infty)$

Figure 1.26: *Table showing the four possible classifications of scalar and vector potential fields, rotational/irrotational. compressible/incompressible. Rotational fields are generated by on the vector potential (e.g., $\boldsymbol{A}(\boldsymbol{x}, t)$), while compressible fields are generated by a the scalar potentials (e.g., voltage $\phi(\boldsymbol{x}, t)$, (velocity ψ, pressure $\varrho(\boldsymbol{x}, t)$), temperature $T(\boldsymbol{x}, t)$).*

The four categories of fluid flow: The following is a summary of the four cases for fluid flow, as summarized in Fig. 1.26:

1,1 Compressible and rotational fluid (general case): $\nabla\phi \neq 0$, $\boldsymbol{\nabla}\times\boldsymbol{w} \neq 0$. This is the case of wave propagation if a medium where viscosity can not be ignored, as in the case of acoustics close to the boundaries, where viscosity contributes losses (Batchelor, 1967).

1,2 Incompressible, but rotational fluid (Lubrication theory): $\boldsymbol{v} = \boldsymbol{\nabla}\times\boldsymbol{w} \neq 0, \boldsymbol{\nabla}\cdot\boldsymbol{v} = 0, \nabla^2\phi = 0$. In this case the flow is dominated by the walls, the viscosity and heat transfer introduce shear. This is typical of lubrication theory.

2,1 Fluid compressible irrotational flow (acoustics): $\boldsymbol{v} = \nabla\phi$, $\boldsymbol{\nabla}\times\boldsymbol{w} = 0$. Here losses (viscosity and thermal diffusion) are small (assumed to be zero). One may define a velocity potential ψ, the gradient of which gives the air particle velocity. For a irrotational fluid $(\boldsymbol{\nabla}\times\boldsymbol{v} = 0)$ (Greenberg, 1988, p. 826), thus $\boldsymbol{v} = -\nabla\phi$. This is the case of the conservative field, where $\int \boldsymbol{v} \cdot \hat{\boldsymbol{n}}\, dR$ only depends on the end points, and $\oint_{\mathcal{B}} \boldsymbol{v} \cdot \hat{\boldsymbol{n}}\, dR = 0$. When a fluid may be treated as having no viscosity, it is typically assumed to be irrotational, since it is the viscosity that introduces the shear (Greenberg, 1988, p. 814). A *fluid's angular velocity* is $\Omega = \frac{1}{2}\boldsymbol{\nabla}\times\boldsymbol{v} = 0$, thus irrotational fluids have zero angular velocity $(\Omega = 0)$.

2,2 Incompressible and irrotational fluid (statics): $\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0$ and $\boldsymbol{\nabla}\times\boldsymbol{v} = 0$ thus $\boldsymbol{v} = \nabla\phi$ and $\nabla^2\phi = 0$. An example of such a case is water in a small space at low frequencies, where the wavelength is long compared to the size of the container, the fluid may be treated as incompressible. When $\boldsymbol{\nabla}\times\boldsymbol{v} = 0$, the effects of viscosity may be ignored, as it is the viscosity that creates the shear leading to rotation. This is the case of modeling the cochlea, where losses are ignored and the quasi-static limit is justified.

In summary, each of the cases is some sort of approximation that best applies in the low frequency limit. This is why it is called *quasi-static*, meaning low, but not zero frequency, where the wavelength is large compared with the dimensions (e.g., diameter).

Thanksgiving Holiday 11/19–11/27 2016

1.5.6   Lec 40 Maxwell's Equations: The unification of electricity and magnetism

Once you have mastered the three basic vector operations, the gradient, divergence and curl, you are ready to appreciate Maxwell's equations. Like the vector operations, these equations may be written in

integral or vector form. A very important difference is, that with Maxwell's equations, we are dealing with well defined physical quantities. The scalar and vector fields take on meaning, and units. Thus to understand these important equations, one must master the names of the four fields E, H, B, D.

Figure 1.27: *A solenoid is a uniform coil of wire. When a current is passed through the wire, a uniform magnetic field intensity H is created. From a properties point of view, this coil is indistinguishable from a permanent bar magnet, having north and south poles. Depending on the direction of the current, one end of a finite solenoidal coil is the north pole of the magnet, and the other end is the south pole. The uniform field inside the coil is called* solenoidal, *a confusing synonym for irrotational. Figure taken from Wikipedia.*

Field strengths E, H: As summarized in Fig. 1.28 there are two field strengths, the electric E, with units of [Volts/meter] and the magnetic H having units of [Amps/meter]. Their the ratio $|E|/|H|$ is in [ohms].

To understand the meaning of E, if two conducting plates are place 1 [m] apart, with 1 [V] across them, the electric field is $E = 1$ [V/m]. If a charge (i.e., and electron) is placed in an electric field, it feels a force $F = qE$, where q is the magnitude of the charge [Col].

To understand the meaning of H, consider the solenoid made of wire, as shown in Fig. 1.27, which carries a current of 1 [Amp]. The magnetic field H inside such a solenoid is uniform and is pointed along the long axis, with a direction that depends on the polarity of the applied voltage (i.e., direction of the current in the wire).

Symbol	Name	Units	Maxwell's Eq.
E	EF: Electric Field strength	[Volts/m]	$\nabla \times E = -\partial_t B$
$D = \epsilon_o E$	ED: Electric Displacement (flux density)	[Col/m^2]	$\nabla \cdot D = \rho$
H	MF: Magnetic Field strength	[Amps/m]	$\nabla \times H = \partial_t D$
$B = \mu_o H$	MI: Magnetic Induction (flux density)	[Webers/m^2]	$\nabla \cdot B = 0$

Figure 1.28: *The variables of Maxwell's equations have names (e.g., EF, MI) and units (in square brackets [SI Units]). The units are required to gain a full understanding of each of the four variable and their corresponding equation. For example, Eq. EF has units [Volts/m]. When you integrate E from $x = a, b$, you obtain the voltage difference between those two points. The* speed of light in vacuo *is $c = 3 \times 10^8 = 1/\sqrt{\mu_o \epsilon_o}$ [m/s], and the* characteristic resistance of light *$r_o = 377 = \sqrt{\mu_o/\epsilon_o}$ [Ω] (i.e., ohms). The dot over a vector is shorthand for the partial with respect to time (i.e., $\dot{B} = \partial B/\partial t$).*

Flux: Flux is a flow, such as the mass flux of water flowing in a pipe [kg/s], driven by a force (pressure drop) across the ends of the pipe, or the heat flux in a thermal conductor, having a temperature drop across it (i.e., a window or a wall). The flux is the same as the flow, be it current, mass or heat. In Maxwell's equations there are also two fluxes, the electric flux D, and the magnetic flux B. The flux density units for D [Amps/m^2] (flux in [Amps]) and the magnetic flux is B [Tesla] (flux in [Webers]).

Maxwell's equations

As shown in Fig. 1.28, Maxwell's equations consist of two curl equations, operating on the field strengths EF E and MF H, and two divergence equations, operating of the field fluxes ED D and MI B. Stokes' Law may be applied to the curl equations and Gauss' Law may be used on the divergence equations. This should be logically obvious.

Exercise: Explain how Stokes' Law may be applied to $\nabla \times E = -\dot{B}$, and explain what it means. Sol: Integrating the left side of equation EF over an open surface results in an voltage (emf) induced in the loop closing the boundary $\mathcal{B}$ of the surface

$$\phi_{\text{induced}} = \iint_{\mathcal{S}} (\nabla \times E) \cdot \hat{n}\, dA = \oint_{\mathcal{B}} E \cdot dl \quad [\text{Volts}].$$

The emf (electromagnetic force) is the same as the Thévenin source voltage induced by the rate of change of the flux. Integrating the right side of equation EF over the same open surface results in the source of this induced voltage, the rate of change of the flux [Webers]

$$\phi_{\text{induced}} = -\frac{\partial}{\partial t} \iint_{\mathcal{S}} \boldsymbol{B} \cdot \hat{\boldsymbol{n}} \, dA = \dot{\psi} \quad [\text{Webers/s}].$$

The area integral on the left is in [W/m^2] resulting in the total flux crossing normal to the surface ψ [W].

If we apply Gauss' Law to the divergence equations, we find the total flux crossing the closed surface.

Exercise: Apply Gauss' Law to equation ED and explain what it means in physical terms. Sol: The area of the normal component of $\boldsymbol{D}$ is equal to the volume integral over the charge density, thus Gauss' Law says that the total charge within the volume Q_{enc}, found by integrating the charge density $\rho(\boldsymbol{x})$ over the volume $\mathcal{V}$ is equal to the normal component of the flux $\boldsymbol{D}$ crossing the surface $\mathcal{S}$

$$Q_{enc} = \iiint_{\mathcal{V}} \boldsymbol{\nabla} \cdot \boldsymbol{D} \, dV = \iint_{\mathcal{S}} \boldsymbol{D} \cdot \hat{\boldsymbol{n}} \, dA.$$

5

Summary: Maxwell's four equations relate the field strengths to the flux densities. There are two types of variables, field strengths ($\boldsymbol{E}, \boldsymbol{H}$) and flux densities ($\boldsymbol{D}, \boldsymbol{B}$). There are two classes, electric ($\boldsymbol{E}, \boldsymbol{D}$) and magnetic $\boldsymbol{H}, \boldsymbol{B}$. One might naturally view this as a 2x2 matrix, with rows being electric and magnetic strengths, and columns being electric and magnetic and flux densities, defining a total of four variables:

	Strength	Flux
Electric	$\boldsymbol{E}$ [V/m]	$\boldsymbol{D}$ [Col/m^2]
Magnetic	$\boldsymbol{H}$ [A/m]	$\boldsymbol{B}$ [W/m^2]

Applying Stokes' curl law to the forces induces a Thévein voltage (emf) or Norton current source. Applying Gauss' divergence law to the flows, gives the total charge enclosed. The magnetic charge always zero since $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$, because magnetic mono-poles do not exist. However magnetic dipoles do exist, as in the example of the electron which contains a magnetic dipole.

The wave equation: When Maxwell's equations are combined the vector wave equation results. There is the famous story that Maxwell introduce the displacement current $\dot{\boldsymbol{D}}$ to make the equations symmetric, allowing him to predict the formula for the speed of light. The displacement current was not verified by the experiments of Faraday and Ampere.

Taking the curl of Eq. EF [Maxwell's Electric Field strength equation for the $\boldsymbol{\nabla} \times \boldsymbol{E}$ (Fig. 1.28)], we obtain in the electric vector wave equation

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{E} = -\boldsymbol{\nabla} \times \dot{\boldsymbol{B}} = -\mu_o \boldsymbol{\nabla} \times \dot{\boldsymbol{H}} = -\mu_o \epsilon_o \ddot{\boldsymbol{E}} \tag{1.131}$$

Using the GoD identity (Eq. 1.127) then gives

$$\boldsymbol{\nabla}^2 \boldsymbol{E} - \frac{1}{\epsilon_o} \nabla(\boldsymbol{\nabla} \cdot \boldsymbol{D}) = \mu_o \epsilon_o \ddot{\boldsymbol{E}} \tag{1.132}$$

In a charge free region $\boldsymbol{\nabla} \cdot \boldsymbol{D} = 0$. Since $1/c^2 = \mu_o \epsilon_o$, the *vector wave equation* results:

$$\boldsymbol{\nabla}^2 \boldsymbol{E} = \frac{1}{c^2} \ddot{\boldsymbol{E}}. \tag{1.133}$$

It is important to distinguish the vector Laplacian from the scalar Laplacian. As was discussed in Table 1.8 (p. 96), Recall the d'Alembert solution of the scalar wave equation (Eq. 1.67, p. 79)

$$\boldsymbol{E}(\boldsymbol{x}, t) = \boldsymbol{f}(\boldsymbol{x} - ct) + \boldsymbol{g}(\boldsymbol{x} + ct),$$

where $\boldsymbol{f}, \boldsymbol{g}$ are arbitrary vector fields.

In a like manner one may derive the wave equation in terms of $\boldsymbol{H}$

$$\boldsymbol{\nabla}^2 \boldsymbol{H} = \frac{1}{c^2} \ddot{\boldsymbol{H}}. \tag{1.134}$$

This equation does not have the restriction that there is no free charge, because Eq. MI is zero. Thus both $\boldsymbol{E}, \boldsymbol{H}$ obey the wave equation (thus they are locked together in space-time if we assume no free charge). For an excellent treatment of this topic, read Sommerfeld (1952).

1.5.7   Lec 41 The Quasi-static approximation   5

There are a number of assumptions and approximations that result in special cases, many of which are classic. These manipulations are all done at the differential equation level, by making assumptions that change the basic equations that are to be solved. These approximations distinct from assumptions made while solving a specific problem.

A few and important examples include   10

1. *In vacuo* waves (free-space scalar wave equation)

2. Expressing the vector wave equation in terms of scalar and vector potentials

3. Quasi-statics

   (a) scalar wave equation
   (b) Kirchhoff's low-frequency lumped approximation (LRC networks)   15
   (c) Transmission line equations (Telephone and Telegraph equations)

One of the very first insights into wave propagation was due to Huygens (c1640) (Fig. 1.17).

Quasi-statics and it implications: The term *quasi-statics* (Postulate P9, p. 73) is an approximation, used to reduce a a partial differential equations to a scalar (one-dimensional) equation (Sommerfeld, 1952). Quasi-statics is a way of reducing a three dimensional problem to a one–dimensional   20
problem. So that it is not miss-applied, it is important to understand the nature of this approximation, which is goes to the heart of transmission line theory. The quasi-static approximation states that the wavelength λ is greater than the dimensions of the size of the object Δ (e.g., $\lambda \gg \Delta$). The best known example, Kirchhoff's current and voltage laws, KCL and KVL, almost follow from Maxwell's equations given the quasi-static approximation (Ramo et al., 1965). These laws state that the sum of the currents   25
at a node must be zero (KCL) and the some of the voltages around a loop must be zero (KCL).

These well known laws the analogue of Newton's laws of mechanics. The sum of the forces at a point is the analogue of the sum of the voltages. Voltage ϕ is the force potential, since the electric field $\boldsymbol{E} = -\nabla \phi$. The sum of the currents is the analogue of the vector sum of velocities at a point is zero.

The acoustic wave equation describes how the scalar field pressure $p(\boldsymbol{x}, t)$, the vector force density   30
potential ($f(\boldsymbol{x}, t) = -\nabla p(\boldsymbol{x}, t)$ [N/m^2]), propagates in three dimensions. (The net force is the integral of the pressure gradient over an area.) If the wave propagation is restricted to a pipe (e.g., organ pipe), or to a string (e.g., an guitar or lute), the transverse directions may be ignored, due to the quasi-static approximation. What needs to be modeled by the equations is the wave propagation along the pipe (string). Thus we may approximate the restricted three-dimensional wave by a one-dimensional wave.   35

However if we wish to be more precise about this reduction in geometry ($\mathbb{R}^2 \to \mathbb{R}$), we need to consider the quasi-static approximation, as it makes assumptions about what is happening in the other directions, and quantifies the effect ($\lambda \gg \Delta$). Taking the case of wave propagation in a tube, say the ear canal, there is the main wave direction, down the tube. But there is also wave propagation in the transverse direction, perpendicular to the direction of propagation. As shown in Table 3.1 (p. 149), the key statement of the quasi-static approximation is that the wavelength in the transverse direction is much larger that the radius of the pipe. This is equivalent to saying that the radial wave reaches the walls and is reflected back, in a time that is small compared to the distance propagated down the pipe. Clearly the speed of sound down the pipe and in the transverse direction is the same if the medium is homogeneous (i.e., air or water). Thus the sound reaches the walls and is returned to the center line in a time that the axial wave traveled about 1 diameter along the pipe. So if the distance traveled is several diameters, the radial parts of the wave have time to come to equilibrium. So the question one must ask is, what are the conditions of such an equilibrium. The most satisfying answer to this is to look at the internal forces on the air, due to the gradients in the pressure.

The pressure $p(x, y, z, t)$ is a potential, thus its gradient is a force density $\mathbf{f}(x, y, z, t) = -\nabla p(x, y, z, t)$. What this equation tells us is that as the pressure wave approaches that of a plane wave, the radial (transverse) forces go to zero. If the tube has a curvature, or a change in area, then there will be local forces that create radial flow. But after traveling a few diameters, these forces will come to equilibrium and the wave will return to a plane wave. The internal stress caused by a change is area must settle out very quickly. There is a very important caveat however: it is only at low frequencies that the plane wave can dominate. At frequencies such that the wavelength is very small compared to the diameter, the distance traveled between reflections is much greater than a few diameters. Fortunately the frequencies where this happens are so high that they play no role in frequencies that we care about. This effect is referred to as *cross-modes* which imply some sort of radial standing waves. In fact such modes exist in the ear canal, but on the eardrum where the speed of sound is much slower that that of air. Because of the slower speed, the ear drum has cross-modes, and these may be seen in the ear canal pressure. Yet they seem to have a negligible effect on our ability to hear sound with good fidelity. The point here is that the cross modes are present, but we call upon the quasi-static approximation as a justification for ignoring them, to get closer to the first-order physics.

Quasi-statics and Quantum Mechanics

It is important to understand the meaning of Planck's constant h, which appears in the relations of both photons (light "particles") and electrons (mass particles). If we could obtain a handle on what exactly Planck's constant means, we might have a better understanding of quantum mechanics, and physics in general. By cataloging the *dispersion relations* (the relation between the wavelength $\lambda(\nu)$ and the frequency ν), for between electrons and photons, this may be attainable.

Basic relations from quantum mechanics for photons and electrons include:

1. Photons (mass=0, velocity $= c$)

    (a) $c = \lambda\nu$: The speed of light c is the product of its wavelengths λ times its frequency ν. This relationship is only for mono-chromatic (single frequency) light.

    (b) The speed of light is
    $$c = \frac{1}{\sqrt{\mu_o \epsilon_o}} = 0.3 \times 10^6 \quad \text{[m/s]}$$

    (c) The *characteristic resistance* of light $r_o = \sqrt{\mu_o/\epsilon_o} = |E|/|H| = 377$ [ohms] is defined as the magnitude of the ratio of the electric $\boldsymbol{E}$ and magnetic $\boldsymbol{H}$ field, of a plane wave *in-vacuo*.

    (d) $E = h\nu$: the photon energy is given by Planck's constant $h \approx 6.623 \times 10^{-34}$ [m^2 kgm/s], times the frequency (or bandwidth) of the photon)

2. Electrons (mass $= m_e$, velocity $V = 0$):

(a) $E_e = m_e c^2 \approx 0.91 \cdot 10^{-30} \cdot 0.3^2 \cdot 10^{12} = 8.14 \times 10^{-20}$ [J] is the *electron rest energy* (velocity $V = 0$) of every electron, of mass $m_e = 9.1 \times 10^{-31}$ [kgm], where c is the speed of light.

(b) $p = h/\lambda$: The momentum p of an electron is given by Planck's constant h divided by the wavelength of an electron λ. It follows that the bandwidth of the photon is given by

$$\nu_e = \frac{E_e}{h}$$

and the wavelength of an electron is

$$\lambda_e = \frac{h}{p_e}.$$

One might reason that QM obeys the quasi-static (long wavelength) approximation. If we compare the velocity of the electron V to the speed of light c, then we see that

$$c = E/p \gg V = E/p = mV^2/V$$

Conjecture on photon energy:

Photons are seen as quantized because they are common generator by atoms, which produce light-particles having the difference in two energy (quantum, or eigen-states) levels. The relation $E = h\nu$ does not inherently depend on ν being a fixed frequency. Planck's constant h is the EM energy density over frequency, and $E(\nu_o)$ is the integral over frequency

$$E(\nu_o) = h \int_{-\nu_o}^{\nu_o} d\nu = 2h\nu_o.$$

When the photon is generated by an atom ν_o is quantized by the energy level difference that corresponds to the frequency (energy level difference) of the photon jump. ⁵

1.5.8   Lec 42: Review for Final Exam

Summary

Mathematics began as a simple way of keeping track of how many things there were. But eventually physics and mathematics evolved together as tools to help us navigate our environment, not just physically around the globe, but how to solve daily problems such as food, water and waste management, ¹⁰ understand the solar system and the stars, defend ourselves, use tools of war, etc.

Based on the historical record of the abacus, one can infer that people precisely understood the concept of counting, addition, subtraction and multiplication (recursive addition).

There is some evidence that the abacus, a simple counting tool, formalizing the addition of very large numbers, was introduced to the Chinese by the Romans, where it was used for trade. ¹⁵

However this working knowledge of arithmetic did not to show up in written number systems. The Roman numerals were not useful for doing calculations done on the abacus. The final answer would then be expressed in terms of the Roman number system.

According to the known written record, the number zero (null) had no written symbol until the time of Brahmagupta (628 CE). One should not assume the concept of zero was not understood simply ²⁰ because there was no symbol for it in the Roman Numeral system. Negative numbers and zero would be obvious when using the abacus. Numbers between the integers would be represented as *rational numbers* $\mathbb{Q}$ since any number may be approximated with arbitrary accuracy with rations numbers.

Mathematics is the science of formalizing a repetitive method into a set of rules, and then generalizing it as much as possible. Generalizing the multiplication and division algorithm, to different types ²⁵ of numbers, becomes increasingly more complex as we move from integers to rational numbers, irrational numbers, real and complex numbers and ultimately, vectors and matrices. How do you multiply two vectors, or multiply and divide one matrix by another? Is it subtraction as in the case of two

numbers? Multiplying and dividing polynomials (by long division) generalizes these operations even further. Linear algebra is further important generalization, fallout from the fundamental theorem of algebra, and essential for solving the generalizations of the number systems.

Many of the concepts about numbers naturally evolved from music, where the length of a string (along with its tension) determined the pitch (Stillwell, 2010, pp. 11, 16, 153, 261). Cutting the string's length by half increased the frequency by a factor of 2. One forth of the length increases the frequency by a factor of 4. One octave is a factor of 2 and two octaves a factor of 4 while a half octave is $\sqrt{2}$. The musical scale was soon factored into rational parts. This scale almost worked, but did not generalize (sometimes known as the *Pythagorean comma*[76]), resulting in today's *well tempered scale*, which is based on 12 equal geometric steps along one octave, or 1/12 octave ($\sqrt[12]{2} \approx 1.05946 \approx 18/17 = 1 + 1/17$).

But the concept of a *factor* was clear. Every number may be written as either a sum, or a product (i.e., a repetitive sum). This led the early mathematicians to the concept of a prime number, which is based on a unique factoring of every integer. At this same time (c5000 BCE), the solution of a second degree polynomial was understood, which lead to a generalization of factoring, since the polynomial, a sum of terms, may be written in factored form. If you think about this a bit, it is sort of an amazing idea, that needed to be discovered (Stillwell, 2010, p.). This concept lead to an important string of theorems on factoring polynomials, and how to numerically describe physical quantities. Newton was one of the first to master these tools with his proof that the orbits of the planets are ellipses, not circles. This lead him to expanding functions in terms of their derivatives and power series. Could these sums be factored? The solution to this problem led to calculus.

So mathematics, a product of the human mind, is a highly successful attempt to explain the physical world. All aspects of our lives were impacted by these tools. Mathematical knowledge is power. It allows one to think about complex problems in increasingly sophisticated ways. An equation is a mathematical sentence, expressing deep knowledge. Witnessed $E = mc^2$ and $\nabla^2 \psi = \ddot{\psi}$.

Reading List: The above concepts come straight from mathematical physics, as developed in the 17th–19th centuries. Much of this was first developed in acoustics by Helmholtz, Stokes and Rayleigh, following in Green's footsteps, as described by Lord Rayleigh (1896). When it comes to fully appreciating Green's theorem and reciprocity, I have found Rayleigh (1896) to be a key reference. If you wish to repeat my reading experience, start with Brillouin (1953, 1960), followed by Sommerfeld (1952); Pipes (1958). Second tier reading contains many items Morse (1948); Sommerfeld (1949); Morse and Feshbach (1953); Ramo et al. (1965); Feynman (1970); Boas (1987). A third tier might include Helmholtz (1863a); Fry (1928); Lamb (1932); Bode (1945); Montgomery et al. (1948); Beranek (1954); Fagen (1975); Lighthill (1978); Hunt (1952); Olson (1947). It would be a mistake to ignore other massive physics writings by stalwart authors, J.C. Slater[77] and Landau and Lifshitz,[78] and their impressive series of mathematical physics books.

You must enter at a level that allows you to understand. Successful reading of these books critically depends on what you already know. A rudimentary (high school) level of math comprehension must be mastered first. Read in the order that helps you best understand the material.

Without a proper math vocabulary, mastery is hopeless. I suspect that one semester of college math can bring you up to speed. This book is my attempt to present this level of understanding.

---

[76]https://en.wikipedia.org/wiki/Pythagorean_comma

[77]https://en.wikipedia.org/wiki/John_C._Slater

[78]https://www.amazon.com/Mechanics-Third-Course-Theoretical-Physics/dp/0750628960

Chapter 2

Number Systems: Stream 1

Chapters 2-5 contain advanced materal that go beyond Chapter 1. You should start with Chapter 1, and once mastered, then move to the rest of the book.

This chapter (2) is devoted to *Number Systems* (Stream 1), starting with the counting numbers $\mathbb{N}$. In this chapter we delve more deeply into the details of the topics of Chapter 1, Lectures 4-9.

2.1 Week 2

In Section 1.2.3 we explore in more detail the two fundamental theorems of prime numbers, working out a sieve example, and explore the logarithmic integral $Li(N)$ which approximates the density of primes $\rho_k(N)$ up to prime N.

The topics of Section 1.2.4 consider the practical details of computing the *greatest common divisor* (GCD) of two integers m, n (Matlab's routine `l=gcd(m,n)`), with detailed examples and comparing the algebraic and matrix methods. Homework assignments will deal with these two methods. Finally we discuss the relationship between coprimes and the GCD. In Section 1.2.5 we defined the Continued Fraction algorithm (CFA), a method for finding rational approximations to irrational numbers. The CFA and GCD are closely related, but the relation needs to be properly explained. In Section 1.2.6 we derive *Euclid's formula*, the solution for the Pythagorean triplets (PT), based on Diophantus's *chord–tangent* method. This method is used many times throughout the course notes, first for computing Euclid's formula for the PTs, then for finding a related formula in Section 1.2.7 for the solutions to Pell's equation, and finally for finding the mapping from the complex plane to the extended complex plane (the Riemann sphere).

Finally in Section 1.2.8 the general properties of the *Fibonacci sequence* is discussed. This equation is a special case of the second order digital resonator (well known in digital signal processing), so it has both historical and practical application for engineering. The general solution of the Fibonacci is found by taking the Z-transform and finding the roots, resulting in an eigenvalue expansion (Appendix D).

2.1.1 Lec 4 Two theorems on primes

Theorem 1: *Fundamental Theorem of Arithmetic*

Factoring integers: Every integer $n \in \mathbb{N}$ has a unique factorization (Stillwell, 2010, p. 43) (Eq. 1.3, p. 32).

Cofactors: Integers 2312 and 2313 are said to be *coprime*, since they have no common factors. *Coprimes* may be identified via the *greatest common divisor*:

$$gcd(a, b) = 1$$

using the *Euclidean algorithm* (Stillwell, 2010, p. 41).

Theorem 2: *Prime Number Theorem*

The primes are a random field since there is no way to predict when the next prime will occur. Thus one needs to use statistical methods to characterize their density. Based on a sample of approximately 3 million primes, Gauss showed empirically that the average total number of primes less than N is

$$\sum_{n=1}^{N} \delta_n \sim \frac{N}{\ln N}. \tag{2.1}$$

These primes were obtained by manual calculations "as a pastime" in 1792-3 (Goldstein, 1973).

Define $\delta_n = 1$ if n is a prime, and zero otherwise.[1]

It follows that the *average density of primes* is $\rho_\pi(N) \sim 1/\ln N$, thus

$$\rho_\pi(N) \equiv \frac{1}{N} \sum_{n=1}^{N} \delta_n \approx \frac{1}{N} Li(N) \equiv \frac{1}{N} \int_2^N \frac{d\xi}{\ln \xi}, \tag{2.2}$$

where $Li(N)$ is the *offset logarithmic integral* (Stillwell, 2010, p. 585). The primes are distributed as $1/\ln(n)$ since the average total number of primes is proportional to the *logarithmic integral $Li(n)$* (Goldstein, 1973; Fine, 2007).

Add plot.

Here is a Matlab/Octave code that tests this formula:

```
%Computes density of primes from average density

NP=1e6; % 10^6 primes
p=primes(NP); %compute primes
delta=zeros(1,NP); delta(p)=1; %put 1 at each prime
PI=cumsum(delta); %Number of primes vs N
rho=PI./(1:NP); %estimate of the density of primes= PI(N)./N
%
figure(1)
semilogy(rho); %plot of density vs number of primes
title('Density of primes vs. N'); ylabel('\rho(N)'); xlabel('N'); grid on
```

From the Prime Number Theorem it is clear that the density of primes is large (they are not scarce). As best I know there are no methods to find primes other than by the sieve method (Section 1.2.3, p. 31). If there is any good news it is that they only need to be computed once, and saved. In practical applications this may not help much, given their large number. But I suspect they could easily be saved on a modern disk (e.g., 1 TB).

Not surprisingly, playing with primes has been a popular pastime of mathematicians. Perhaps this is because those who have made inroads, providing improved understanding, have become famous.

2.1.2   RSA public-private keys

Internet security depends on a public-private key system, called RSA[2] which is built on the difficulty of factoring large primes. When a forward coputation is easy, such as multiplying two primes, is easy, and the inverse problem (factoring a product into two primes) is hard, it is called a trap door.

What makes RSA work is a trap door. Internet security is not be as important as global warming, but its pretty important. Global warming has no "end game." Internet security does, redesign the internet.

Lets consider the problem of factoring two primes, given the product $N = \pi_k \pi_l$, $k \neq l$. Lets further assume we have list of all the primes $\{\pi_1, ... \pi_m\}$ including $N < \pi_m$. To factor N we must divide N by

---

[1]You may view δ_n for the first 100 numbers with the one-line Matlab/Octave command `stem(isprime(1:100))`

[2]`https://en.wikipedia.org/wiki/RSA_(cryptosystem)`

each prime less than N. With intellegent sampling, on average. we would find the prime factor in $N/2$ tries. In theory, given primes π_n up to $n = N$, the density $\rho_\pi(N)$ could help one search for a particular prime of known size N, by estimating how many primes there are in the neighborhood of N. The cost of factroing any integer $\in \mathbb{N}$ would be related to the magnitude of $N = \pi_k \pi_l$, where N is the size of the private key, assuming a simple product key. Worst case an exhaustive search would take $m - 1$ divides of the form N/π_k $k = 1 \cdots m - 1$. If one knew approximately where the factor is $\in \mathbb{P}$ (i.e., k of π_k), and knows one of the primes, the search could be done with an exhaustive search, by addition. The number of primes less that N may be computed using Gauss' formula for the density of primes.

So the question remains: How does internet security actually work? Unfortunately at this time I can not give you a proper summary. But it is more complex than I have let on. The full answer requires a proper course in number theory and internet security, beyond what is presented here. You can read more about this at `https://en.wikipedia.org/wiki/RSA_(cryptosystem)#Operation`.

The power of this security device is based on the relatively high density of primes $\mathbb{P} \subset \mathbb{N}$, which is addressed by the Prime Number Theorem.

Here is an RSA public key:

```
AAAAB3NzaC1yc2EAAAADAQABAAAABAQCsyfp9uDXubR/ukXMpVgK2eEttQcJeEIMg6mC7qjRa2hrhFCV1+r9YSdqcKdkkefHL+
HI3VAYyThE94m4CmGFbPM8xpjNyH1nQw6hUdhhOtBqnCg517WDzpEqqIQCPe2ljMs+3Y1Tb05hNMv9W6SqzFCNmtJr11VXkcU
XnrbB/bbIMs6AGGnViyLup+8rSNlqW1oH89wQcHGMnOh7P5Y/bJdgYWzX/9t5MoTNeuq7jghnP0J6aqUAfTXAq00SBKXyUM8i
3HbQqP8nOlEHkDKyrfb72MnBrXuVl5m+B3qFkhO+infLaqrzqoyIEOEBWq46W5CKvldAv+DTIF0cS5YRl
```

This strings of characters represents a prime number π_{pub}, the public key.

Here is an RSA private key:

```
MIIEpAIBAAKCAQEArMn6fbg17mOf7pFzKVYCtnhLbUHCXhCDIOpgu6oOWtoa4RQldfq/WEnanCnZJHnxy/hyN1QGMk4RPeJuA
phhWzzPMaYzch9ZOMOoVHYYTrQapwoOde1g86RKqiEAj3tpYzLPt2NU29OYTTL/VukqsxQjZrSa9dVV5HFF562wf22yDLOgBh
p1Ysi7qfvKOjZaltaB/PcEHBxjJ9Iez+WP2yXYGFs1//beTKEzXrqu44IZz9CemqlAH01wKtNEgSl8lDOotx2OKj/J9JRB5Ay
sq32+9jJwa17lZeZvgd6hZITvop3y2qq86qMiBNBAVquOluQir5XQL/gOyBdHEuWEZQIDAQABAoIBAQCPB9sMurIaHcbzCv/T
VdmFBmsZk/e7OdvyIVHVxvzzv9vEOZMQ6HIAOZJ29sVCxpb7izuwxSMj/S+UlJOQFOdySRISPUwUyhLlFzF+SQuyCrCNWWvC/
q7kwDj2P79hBn2kmekMG3OhTLSkIvTvErA/8Qmk+RfVK7WgFowRTdvufplSJCjrB1HcLIXd6tQxC5gCHLbhtJjYOYmW2xBg1f
n1qXAGObVXpOp9Vr/kOy5uH/dS1Zp9dcySn6HnOG+XAEs4bhYwSqZqcgsKZxiza2r3B8UUE6NjeqSVLCki5uMyXEqxbFwHy+K
vjm8OzPgLwk2sODe+o/3DKYIOXCdNmIaBAoGBANOIHcQATwJO/ImcgmLkTKFbNm8Xlojw4bKIFTRG8jPvssRQjWsAc6ozac73
D7Ri5YcDeOYNg7Zyujml7hImo8NM9v53sfepCtqC4Ixv7o2snPXDSyrsOF898IgOGlO4/IEZAPwybVVg26Jn2mgtBWpj0R4vG
AylpyX+j9crPMdRAoGBANEc3dYjJSai3oYEwBOFsBKU81a2xonsQi5+cDgHc/ugr6taWuinmSoj82MbeyKgLdyA6Vb4EO3HqR
jBRGhs44szFy9FhexBN2apAbw30a6/+MRPkAWKfDDIYW1oOtUf7nyvpuKcT5a6OnsuVJiFLzfw3mSElqin8f5xDFA2yU7VAoG
AFJ14YSiyiofCMNsa7e5X9YXBKVSALDpllJOklWKwqhUbM/yxu5Lgst6t2ijD1A2rVtwkxS7d/kLyGTotmIFBhoAStjwOERfd
YexNXAJwmkgPR/hPdBBvcgDEVO4gpakqmp5INUvoOeBNp2c9ptdOZqjxsWJfkqP2RUKo3AxvsPECgYB8+cObN2X8sVpMvooSz
lNsx5inOpVt2MW1A9wNoGaw/FgBEP9EDOZ9SP8tPD9bcl/2aRUL5TekRK/LG7bGsk1L5Dz3A5YGnPN/Z4Kh+rpWtmT27pYkuZ
9/iEKyIJkIxwzBeDlFNiZa70zcUdQ9U1vlEw/XA2cHZVg7uVSinBuOAQKBgQDGcNLEXgYEFZoIMQ3Qwzvyk9ghVSv1us2ulmg
C2MBXiInjDzyQ2ZMOnFx64YPsUjUpxl7lM4UIB6sFfXPFqtS7KygFwbqOVBORAQ/xC1EdKBcrUK/kO9JGqhvSnqE1hS6NYMJF
n4hHNt3OIhDO9JHQfvzKoCfA1r+OeZSynIDqrA==
```

This much longer string of character represents a second prime number π_{pri}, the private key.

If I gave you the ascii string that represents the product of the two prime numbers $\pi_k \pi_l$, along with π_k, can you find π_l? Yes, divide one by the other, with remainder zero. If I ask you to find π_l given $\pi_k \pi_l$, you would need to call the NSA, or be some sort of expert on factoring primes. Given $\pi_k \pi_l$, for some large values of k and l, the GCD is not helpful, nor the CFA. While the factoring can be done (it is not impossible), it is a *very* expensive calculation.

Why not use Matlab's `factor(N)` routine to find the factors? This is where *cost* comes in. If the primes are large enough, even the NSA cannot do it in a lifetime. The numbers used in RSA are too large for Matlab's routine to deliver an answer. In fact, even the largest computer in the world (such as the University of Illinois' super computer (NCSA Water) cannot do this computation. But, for a carefully chose set of primes, not to large for the NCSA to factor them, they can decrypt a message, but you cannot. This is why they pay for these massive super-computers.

As best I know you need to try one prime at a time to see if the divisor gives you a zero remainder. There are no hints, such as $l \approx k$. Nothing to help but to grind your way through all the known primes. And if the primes are too large to be to be on the list of all known primes, then you have the additional burden of generating a list of primes long enough to contain the ones your looking for. You can be

sure that the NSA (or whoever has the fastest computer) has a longer list than you do, and its getting longer every day.

What you are given is the public key (prime π_{pub}) (the shorter one), and you need to guess the private key π_{pri} and decrypt the message I send you with the product $\pi_{pub} \cdot \pi_{pri}$.

2.1.3   Lec 5 Greatest common divisor (GCD)

Division with rounding method: This method implements $\gcd(a, b)$. In matrix form we have $(m, n, k \in \mathbb{N})$

$$\begin{bmatrix} m_{k+1} \\ n_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -\lfloor \frac{m}{n} \rfloor \end{bmatrix} \begin{bmatrix} m_k \\ n_k \end{bmatrix}. \tag{2.3}$$

This starts with $k = 0, m_0 = a, n_0 = b$. With this method there is no need to test if $n_n < m_n$, as it is built into the proceedure. The method uses the *floor function* $\lfloor x \rfloor$, which finds the integer part of x ($\lfloor x \rfloor$ rounds toward $-\infty$). Following each step we will see that the value $n_{k+1} < m_{k+1}$.

The method terminates when $n_{k+1} = 0$ with $\gcd(a, b) = m_{k+1}$. The previous values m_k, n_k are the solutions to Bézout's identity (`gcd(n,m)=1`, namely $m_k m_0 + n_k n_0 = 1$), since the terminal state and the GCD of a, b is $m - n\lfloor m/n \rfloor = 0$, for which $n = gcd(a, b)$.

Below is 1-line vectorized code that is much more efficient than the direct matrix method:

```
k = gcd(m,n) %entry point: input m,n; output k=gcd(m,n)
M=[abs(m),abs(n)]; %init M
 while M(2) ~=0 % < n*eps to ''almost work'' with irrational inputs
  M = [M(2) - M(1)*floor(M(2)/M(1)); M(1)]; %M = [M(1); M(2)] with M(1)<M(2)
 end
```

With a minor extension in the test for "end," this code can be made to work with irrational inputs (e.g., $(n\pi, m\pi)$).

This method calculates the number of times $n < m$ must subtract from m using the floor function. This operation is the same as the mod function.[3] Specifically

$$n_{k+1} = m_k - m_k \operatorname{modulo} n_k, \tag{2.4}$$

so that the output is the definition of the remainder of modular arithmetic. This would have been obvious to anyone using an abacus, which explains why it was discovered so early.

Note that the next value of m ($M(1)$) is always less than n ($M(2)$), and must remain greater or equal to zero. This one-line vector operation is then repeated until the remainder ($M(1)$) is 0. The gcd is then n ($M(2)$). When using irrational numbers, this still works except the error is never exactly zero, due to IEEE 754 rounding. Thus the criterion must be that the error is within some small factor times the smallest number (which in Matlab is the number `eps` $= 2.220446049250313 \times 10^{-16}$, as defined in the IEEE 754 standard.)

Thus without factoring the two numbers, Eq. 2.4 recursively simply finds the gcd. Perhaps this is best seen with some examples.

The GCD is an important and venerable method, useful in engineering and mathematics, but, as best I know, is not typically taught in the traditional engineering curriculum.

Graphical meaning of the GCD: The Euclidean algorithm is very simple when viewed graphically. In Fig. 2.1 we show what is happening as one approaches the threshold. After reaching the threshold, the two number must be swapped, which is addressed by upper row of Eq. 2.3.

Multiplication is simply recursive addition, and finding the gcd takes advantage of this fact. Lets take a trivial example, (9,6). Taking the difference of the larger from the smaller, and writing multiplication as sums, helps one see what is going on. Since 6=3*2, this difference may be written two

---

[3] `https://en.wikipedia.org/wiki/Modulo_operation`
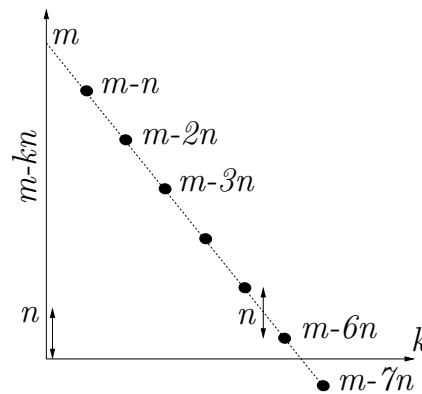
Figure 2.1: *The Euclidean Algorithm recursively subtracts n from m until the remainder $m - kn$ is either less than n or zero. Note that this is similar to* mod(m,n). *The GCD recursively computes mod, then swaps m, n so that $n < M$, and repeats, until it finds the GCD. For the case depicted here the value of k that renders the remainder less than n is $k = 6$. If one more step were taken ($k = 7$) the remainder would become negative. By linear interpolation we can find that $m - kn = 0$ when $k = floor m/n$. For this example $k = 6.5$, thus $k = 6$. thus $6 = floor(m/n) < n$. This is nonlinear arithmetic, which is the key to the GCD.*

different ways

$$9 - 6 = (3 + 3 + 3) - (3 + 3) = 0 + 0 + 3 = 3,$$

or

$$9 - 6 = (3 + 3 + 3) - (2 + 2 + 2) = 1 + 1 + 1 = 3.$$

Written out the first way, it is 3, because subtracting (3+3) from (3+3+3) leaves 3. Written out in the second way, each 3 is matched with a -2, leaving 3 ones, which add to 3. Of course the two decompositions must yield the same result because $2 \cdot 3 = 3 \cdot 2$. Thus finding the remainder of the larger number minus the smaller yields the gcd of the two numbers.

Coprimes: When the gcd of two integers is 1, the only common factor is 1. This is of key importance   5
when trying to find common factors between the two integers. When 1=gcd(m,n) they are said to be
coprime, which can be written as $m \perp n$. By definition, the largest common factor of coprimes is 1.
But since 1 is not a prime ($\pi_1 = 2$), they have no common primes.

Generalizations of GCD: The GCD may be generalized in several significant ways. For example   **How about 3x3**
what is the GCD of two polynomials? To answer this question one must factor the two polynomials to   10
identify common roots. This will be discussed in more detail in Section 3.2.2 (p. 134).

2.1.4 Lec 6 Continued Fraction Expansion (CFA)

Continued Fractions and circuit theory: One of the most powerful generalizations of the CFA
seems to be the expansion of a function of a complex variable, such as the expansion of an impedance
$Z(s)$, as a function of complex frequency s, as described in Fig. 2.2 and Eq. 2.5. This is especially   15
interesting in that it leads to a physical interpretation of the impedance in terms of a transmission
line (horn), a structure well know in acoustics as having a variable area $A(x)$ as function of the range
variable x.

The CFA expansion is of great importance in circuit theory, where it is equivalent to an infinitely
long segment of transmission line, composed of series and shunt impedance elements. Thus such a   20
cascade network composed of 1 ohm resistors, has an input impedance of $(1 + \sqrt{5})/2 \approx 1.6180$ [ohms]
Eq. 1.5 (p. 37).

The CFA may be extended to monomials in s. For example consider the input impedance of a
cascade L-C transmission line as shown in Fig. 2.2. The input impedance of this transmission line is

given by a continued fraction expansion of the form

$$Z_{in} = sL + \cfrac{1}{sC + \cfrac{1}{sL + \cfrac{1}{sC + \cfrac{1}{\cdots}}}} \qquad =: [sL., sC, sL, sC, \cdots]. \tag{2.5}$$

In some vague way, Eq. 2.5 is reminiscent of a Taylor series expansion about $s = 0$, yet very different. In the limit, as the frequency goes to zero ($s \to 0$), the impedance of the inductors go to zero, and that of the capacitors go to ∞. In physical terms, the inductors become short circuits, while the capacitors become open circuits. The precise relation may be quantified by the use of composition, described in Fig. 1.13 (p. 61). Specifically

$$\begin{bmatrix} P_1 \\ U_1 \end{bmatrix} = \begin{bmatrix} 1 & sL \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & sL \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \begin{bmatrix} 1 & sL \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \begin{bmatrix} P_2 \\ -U_2 \end{bmatrix}. \tag{2.6}$$

It seems possible that this is the CFA generalization of the Taylor series expansion, built on composition. If we were to do the algebra we would find that $\mathcal{A}(s), \mathcal{B}(s), \mathcal{C}(s), \mathcal{D}(s)$ (i.e., Sections 1.3.7 (pp. 61, 141)) are ratios of polynomials having rational expansions as Taylor series. This seems like an important observation, that should have applications beyond the engineering literature (Campbell, 1903; Brillouin, 1953; Ramo et al., 1965). Note that Brillouin (1953) credits Campbell (1903).
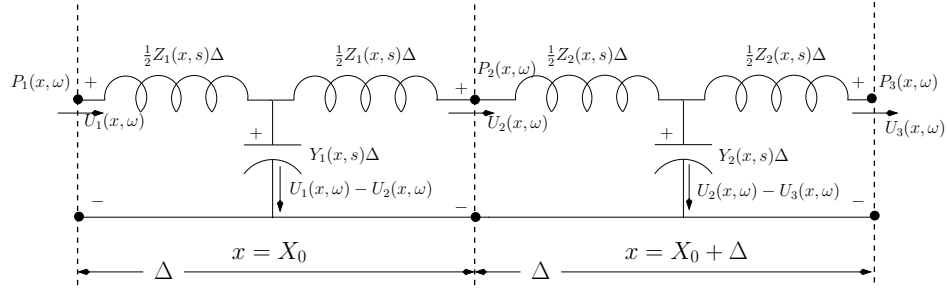
Figure 2.2: *This transmission line (TL) is known as a low-pass filter wave-filter (Campbell, 1922), composed of series inductors having an impedance $Z_k = sL$ and shunt admittance $Y_k = sC$. The characteristic resistance of the TL is $r_0 = \sqrt{Z_k/Y_k}$ and the wave velocity is $\kappa = \sqrt{Z_k Y_k} = s/\sqrt{LC} = s/c$. Thus the wave velocity is $c = 1/\sqrt{LC}$. The length Δ [m] defines the physical size of each cell. For long wavelengths the wave-filter acts as a pure delay. But when the frequency increases above a cutoff frequency $f_c > c/\Delta$, the wavelength $\lambda = c/f$ is less than the size of a section Δ, and the system response becomes very high-order low-pass filter.*

Figure 2.6 is a piece of wire having a delay determined by its velocity and the length, in units of the cells, each of length Δ. There are two basic parameters that characterize a transmission line, the characteristic resistance $r_0 = \sqrt{sL/sC} = \sqrt{L/C}$, and wave number

$$\kappa = \sqrt{sL \cdot sC} = s\sqrt{LC} = s/c_o,$$

since $c_o = \sqrt{LC}$. Each of these is a constant as $\Delta \to 0$, and in that limit the waves travel as

$$f(t - x/c_o) = e^{-\kappa x} e^{-st},$$

with a wave resistance ($r_0 = \sqrt{L/C}$ [Ω]). The total delay $T = N\Delta/c_o$[s], where N is the number of sections.

Since the CFA has a physical representation as a transmission line, as shown in Fig. 2.2, it can be of high utility for the engineer.[4] The theory behind this will be discussed in greater detail in Chapter 5. If you're ready to jump ahead, read the important book Brillouin (1960) and the collected works of Campbell (1937).

---

[4]Continued fraction expansions of functions are know in the circuit theory literature as a *Cauer* synthesis (Van Valkenburg, 1964b).

2.2 Week 3

2.2.1 Lec 7 Pythagorean triplets (PTs) and Euclid's formula

Pythagorean triplets (PTs) have many applications in architecture and scheduling, which explains why they are important. For example, if one wished to construct a triangle made from bricks having a perfect $90°$ angle, then the materials need to be squared off as shown in Fig. 1.8 (p. 38). The lengths of the sides need to satisfy PTs.

Derivation of Euclid's formula: Euclid's formula (Eq. 1.6, p. 38) provides integer solutions to the Pythagorean theorem (Eq. 1.1, p. 15). The solution method, said to be due to Diophantus, is call a *chord/tangent* method (Stillwell, 2010, p. 48). The method composes (Section 1.3.4, p. 55) a line and a circle, where the line defines a chord within the circle (its not clear where the tangent line might go). The slope of the line (i.e., denoted t) is then taken to be rational, allowing one to determine integer solutions of the intersections points. This solution for *Pythagorean triplets* $[a, b, c]$ is known as *Euclid's formula* (Eq. 1.6, p. 38) (Stillwell, 2010, p. 4–9, 222).

The derivation methods of Diophantus have been lost, but Fermat and Newton figured out what Diophantus must have done (Stillwell, 2010, p. 7, 49, 218). Since Diophantus worked before algebra was invented, he described all the equations in prose (Stillwell, 2010, p. 93).

Derivation of Euclid's formula: The derivation is outlined in Fig. 2.3. Starting from two integers $[p > q \in \mathbb{N}]$, composing a line having a rational slope $t = p/q > 0$, with a circle (Stillwell, 2010, p. 6), provides the formula for the Pythagorean triplets.

Euclid's formula for Pythagorean triplets $[a, b, c]$ 5.2.4

Euclidean Proof:
1) $2\phi + \eta = \pi$
2) $\eta + \Theta = \pi$
3) $\therefore \phi = \Theta/2$
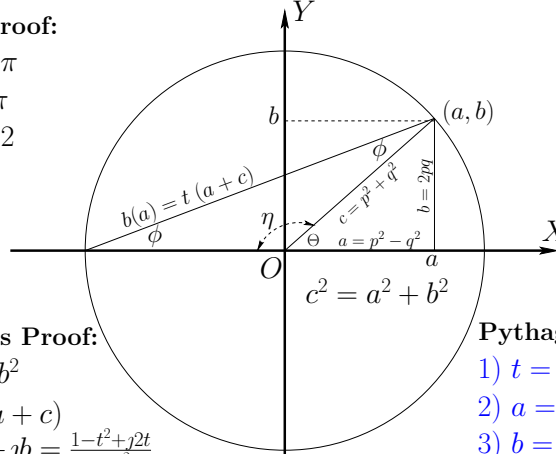
$c^2 = a^2 + b^2$

Diophantus's Proof:
1) $c^2 = a^2 + b^2$
2) $b(a) = t(a + c)$
3) $\zeta(t) \equiv a + \jmath b = \frac{1 - t^2 + \jmath 2t}{1 + t^2}$
4) $\zeta = |c|e^{i\theta} = |c|\frac{1 + it}{1 - it} = |c|(\cos(\theta) + i\sin(\theta))$

Pythagorean triplets:
1) $t = p/q \in \mathbb{Q}$
2) $a = p^2 - q^2$
3) $b = 2pq$
4) $c = p^2 + q^2$

Figure 2.3: *Derivation of Euclid's formula for the Pythagorean triplets $[a, b, c]$, based on a composition of a line, having a rational slope $t = p/q \in \mathbb{N}$, and a circle $c^2 = a^2 + b^2$, $[a, b, c] \in \mathbb{N}$. This analysis is attributed to Diophantus (Di·o·phan'·tus) (250 CE), and today such equations are called Diophantine (Di·o·phan'·tine) equations. PTs have applications in architecture and scheduling, and many other practical problems.*

The construction starts with a circle and a line, which is terminated at the point $(-1, 0)$. The slope of the line is a free parameter t. By composing the circle and the line (i.e., solving for the intersection of the circle and line), the formula for the intersection point (a, b) may be determined in terms of t, which will then be taken as the rational slope $t = p/q \in \mathbb{Q}$.

In Fig. 2.3 there are three panels, two labeled "Proofs." The *Euclidean Proof* shows the angle relationships of two triangles, the first an isosceles triangle formed by the chord, having slope t and two equal sides formed from the radius of the circle, and a second right triangle having its hypotenuse as the radius of the circle and its right angle vertex at $(a, 0)$. As shown, it is this smaller right triangle

that must satisfy Eq. 1.1. The inner right triangle has its hypotenuse c between the origin of the circle (O) to the point (a, b). Side a forms the x axis and side b forms the y ordinate. Thus by construction Eq. 1.1 must be obeyed.

The formula for the line is $b(a) = t(a + c)$, which goes through the points $(-c, 0)$ and (a, b). Composing the line with the circle gives

$$a^2 + (t(a + c))^2 = c^2$$
$$a^2 + t^2(a^2 + 2ac + c^2) = c^2$$
$$(1 + t^2)a^2 + 2ct^2 a + c^2(t^2 - 1) = 0$$

This last equation is a quadratic equation in a. In some sense it is not really a quadratic equation, since we know that $a = -c$ is a root. Dividing by $1 + t^2$

$$a^2 + \frac{2ct^2}{1 + t^2} a + \frac{c^2(t^2 - 1)}{1 + t^2} = 0,$$

makes it easier to complete the square, delivering the roots:

$$\left(a + \frac{ct^2}{1 + t^2}\right)^2 - \left(\frac{ct^2}{1 + t^2}\right)^2 + \frac{c^2(t^2 - 1)}{1 + t^2} = 0$$

$$\left(a + \frac{ct^2}{1 + t^2}\right)^2 - \frac{c^2 t^4}{(1 + t^2)^2} + \frac{c^2(t^2 - 1)(t^2 + 1)}{(1 + t^2)^2} = 0$$

$$\left(a + \frac{ct^2}{1 + t^2}\right)^2 - \frac{\cancel{c^2 t^4} + c^2(\cancel{t^4} - 1)}{(1 + t^2)^2} = 0$$

$$\left(a + \frac{ct^2}{1 + t^2}\right)^2 = \left(\frac{c}{1 + t^2}\right)^2$$

The second to last equation simplifies (magic happens) because the known root $a = -c$ is embedded in the result.

Taking the square root gives the two roots

$$a_\pm + \frac{ct^2}{1 + t^2} = \pm\frac{c}{1 + t^2}$$
$$(1 + t^2)a_\pm = -ct^2 \pm c = -c(t^2 \mp 1)$$
$$a_\pm = -c\frac{t^2 \mp 1}{1 + t^2}.$$

The known root is $a_+ = -c$, because when the sign is $+$, the numerator and denominator terms cancel.

The root we have been looking for is a_-

$$a_- = c\frac{1 - t^2}{1 + t^2},$$

which allows us to solve for b_-

$$b_- = \pm\sqrt{c^2 - a_-^2}$$

$$= \pm c\sqrt{1 - \left(\frac{1 - t^2}{1 + t^2}\right)^2}$$

$$= \pm c\sqrt{\frac{(1 + t^2)^2 - (1 - t^2)^2}{(t^2 + 1)^2}}$$

$$= \pm\frac{2ct}{t^2 + 1}.$$

Therefore the coordinates (a, b), the intersection point of the line and circle, are

$$(a(t), b(t)) = c \frac{[1 - t^2, 2t]}{1 + t^2}.$$

To obtain the Pythagorean triplets, as given in Fig. 2.3 (p. 123) and Eq. 1.6 (p. 38), set $t = p/q$, assuming $p > q \in \mathbb{N}$, and simplify.

Complex roots: Defining the root as a complex number $\zeta(\Theta) \equiv a + b\jmath$ forces $a \perp b$ (i.e., forces the right triangle) and gives us polar coordinates, as defined by the figure as the Euclidean Proof

$$\zeta(\Theta) = |c|e^{\Theta\jmath} = |c|\left(\cos(\Theta) + \jmath\sin(\Theta)\right).$$

This naturally follows since

$$\zeta = |c|e^{\jmath\Theta(t)} = |c|\frac{1 - t^2 + 2t\jmath}{1 + t^2} = |c|\frac{(1 + \jmath t)(1 + \jmath t)}{(1 + t\jmath)(1 - t\jmath)} = (q + p\jmath)\sqrt{\frac{q + \jmath p}{q - p\jmath}}.$$

Examples of PTs include $a = 2^2 - 1^2 = 3$, $b = 2 \cdot 2 \cdot 1 = 4$, and $c = 2^2 + 1^2 = 5$, $3^2 + 4^3 = 5^2$.

Defining $p = q + N$ $(p, q, N \in \mathbb{N})$ gives slightly better parametric representation of the answers, as the pair (q, N) are a more systematic representation than (p, q), because of the condition $p > q$, so the general properties of the solutions are expressed more naturally. Note that $b + c$ must always be a perfect square since $b + c = (p + q)^2 = (2q + N)^2$, as first summarized by Fermat (Stillwell, 2010, p. 212).

2.2.2 Lec 8 Pell's Equation

Eigenvalue solution to Pell's equation: To provide a full understanding of what was known to the Pythagoreans about irrational numbers, it is helpful to provide the full solution to this recursive matrix equation.

As shown in Fig. 1.10 (p. 41), (x_n, y_n) may be written as a power series of the 2x2 matrix A. The well known modern approach to find A^n is to *diagonalize* the matrix as detailed in Appendix C (p. 175). For the 2x2 matrix case, this is relatively simple. The final result written out in detail for the general solution (x_n, y_n) (Appendix D, p. 179):

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} = \jmath^n \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}^n \begin{bmatrix} 1 \\ 0 \end{bmatrix} = E \begin{bmatrix} \lambda_+^n & 0 \\ 0 & \lambda_-^n \end{bmatrix} E^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{2.7}$$

The eigen-values are $\lambda_\pm = \jmath(1 \pm \sqrt{2})$ while the eigen-matrix (Eq. C.6) and its inverse are

$$E = \frac{1}{\sqrt{3}} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.8165 & 0.8165 \\ 0.5774 & -0.5774 \end{bmatrix}, \qquad E^{-1} = \frac{\sqrt{3}}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} \\ 1 & -\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.6124 & 0.866 \\ 0.6124 & -0.866 \end{bmatrix}$$

The relative "weights" on the two eigen-solutions are equal, as determined by

$$E^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{\sqrt{3}}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} \\ 1 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{\sqrt{3}}{2\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We still need to prove that

$$\frac{x_n}{y_n} \underset{\infty}{\to} \sqrt{N}.$$

This follows intuitively from Pell's equation, since as $(x_n, y_n) \to \infty$, the difference between x_n^2 and $2y_n^2$, the (± 1) becomes negligible.

Given the development of linear algebra c19th century, this may be evaluated by eigenvector diagonalization.[5]

---

[5]https://en.wikipedia.org/wiki/Transformation_matrix#Rotation

Pell's Equation and irrational numbers: Since the eigenvalues of Eq. 1.8 ($\lambda_\pm = 1 \mp \sqrt{N} \notin \mathbb{N}$), solutions to Pell's equation raised the possibility that all numbers are not rational. This discovery of irrational numbers forced the jarring realization that the Pythagorean dogma "all is integer" was wrong. The significance of irrational numbers was far from understood.

⁵ # WEEK 4

2.3   Week 4

2.3.1   Lec 9 Fibonacci Numbers

The Fibonacci sequence is famous in number theory. It is said that the sequence commonly appears in
¹⁰ physical systems. Fibonacci numbers are related to the "golden ratio" $(1 + \sqrt{5})/2$.

But from a mathematical point of view, the Fibonacci sequence does not seem special. It is generated by a linear recursion relationship, where the next number is the sum of the previous two (Eq. 1.9, p. 41)

$$x_{n+1} = x_n + x_{n-1}. \tag{2.8}$$

The term *linear* means that the principle of superposition holds [P1 (linear/nonlinear) of Section 1.3.12, p. 72) and 3.5.1, p. 145]. To understand the meaning of this we need to explore the z-transform, the discrete-time version of the Laplace transform. We will return to this in Chapter 4 (p. 151).

A related linear recurrence relation is to define the next output x_{n+1} be the average of the previous two

$$x_{n+1} = \frac{x_n + x_{n-1}}{2}.$$

In some ways this relationship, which has no famous name, is more useful than the famous Fibonacci
¹⁵ recursion, since it perfectly removes oscillations of the form -1^n (it is a 2-sample *moving average*, a trivial form of low-pass filter). Unlike the Fibonacci sequence, which diverges, it seems that biology prefers divergent sequences.

Equation 2.8 may be written as a 2x2 matrix relationship. If we define $y_{n+1} = x_n$ then Eq. 2.8 is equivalent to (Eq. 1.10, p. 41)

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix}. \tag{2.9}$$

The first equation is $x_{n+1} = x_n + y_n$ while the second is $y_{n+1} = x_n$, which is the same as $y_n = x_{n-1}$. Note that the Pell 2x2 recursion is similar in form to the Fibonacci recursion. This removes mystique
²⁰ from both equations.

The most general 2d order recurrence relationships is

$$x_{n+1} = -bx_n - cx_{n-1},$$

with constants $b, c \in \mathbb{R}$.

In the matrix diagonalization of the Pell equation we found that the eigenvalues were $\lambda_\pm = 1 \mp \sqrt{N}$, and the two solutions turned out to be powers of the eigenvalues. The solution to the Fibonacci recursion may similarly be expressed in terms of a matrix. These two cases may thus be reduced by the same
²⁵ 2x2 eigenvalue solution method.

The eigenvalues of the Fibonacci matrix are

$$det \begin{bmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{bmatrix} = \lambda^2 - \lambda - 1 = (\lambda - 1/2)^2 - (1/2)^2 - 1 = (\lambda - 1/2)^2 - 5/4 = 0$$

thus $\lambda_\pm = \frac{1 \pm \sqrt{5}}{2} = [1.618, -0.618]$.

<div style="border:1px solid">

General properties of the Fibonacci numbers[a]

$$x_n = x_{n-1} + x_{n-2}$$

- This is a 2-sample *moving average* difference equation with an unstable pole

- $x_n = [0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \cdots]$, assuming $x_0 = 0, x_1 = 1$:

- Analytic solution (Stillwell, 2010, p. 194): $\sqrt{5}\, x_n \equiv \left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n \to \left(\frac{1+\sqrt{5}}{2}\right)^\infty$

  - $\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = \frac{1+\sqrt{5}}{2}$
  - Example: $34/21 = 1.6190 \approx \frac{1+\sqrt{5}}{2} = 1.6180$ 0.10% error

- Matlab's $\texttt{rat}(1 + \sqrt{5}) = 3 + 1/(4 + 1/(4 + 1/(4 + 1/(4)))) =: [3; 4, 4, 4, \cdots]$

[a]https://en.wikipedia.org/wiki/Fibonacci_number

</div>

Figure 2.4: Properties of the Fibonacci numbers (Stillwell, 2010, p. 28).

Note that this 2x2 equation is similar to Pell's equation, suggesting that an eigenfunction expansion of Eq. 1.10 may be used to analyze the sequence, as shown in Section 2.3.1 (p. 126) (Stillwell, 2010, 192). It is also related to a $10,000 prize, that was eventually solved.[6]

2.3.2 Lec 10 Exam I

---

[6]http://www.nytimes.com/1988/08/30/science/intellectual-duel-brash-challenge-swift-response.html

Chapter 3

Algebraic Equations: Stream 2

3.1 Week 4

3.1.1 Lec 11 Algebra and geometry as physics

Before Newton could work out his basic theories, algebra needed to be merged with Euclid's early quantification of geometry. The key to putting geometry and algebra together is the Pythagorean theorem (Eq. 1.1), which is both geometry and algebra. To make the identification with geometry the sides of the triangle needed to be viewed as a length. This is done by recognizing that the area of a square is the square of a length. Thus a geometric proof requires one to show that the area of the square $A = a^2$ plus the area of square $B = b^2$ must equal the area of square $C = c^2$. There are many such constructions that show $A + B = C$ for the right triangle. It follows that in terms of coordinates of each vertex, the length of c is given by

$$c = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \tag{3.1}$$

with $a = x_2 - x_1$ and $b = y_2 - y_1$ (Fig. 1.8, p. 38). Thus Eq. 1.1 is both an algebraic and a geometrical   5
statement. This is not immediately obvious.

Analytic geometry is based on coordinates of points, with the length given by Eq. 3.1. Geometry treats lines as lengths without specifying the coordinates (Eq. 1.1). Algebra gave a totally new view to the quantification of geometrical lengths by introducing a coordinate system. This resulted in an entire new way to work with conic sections, which were now explained in terms of equations having   10
coordinate systems. When viewed through the lens of algebra, Eq. 1.1 is a circle having radius c. Complex numbers provide an equivalent representations, since if $z = x + y\jmath$, the unit circle is $z = e^{\jmath\theta}$ and $|z|^2 = x^2 + y^2$. Here we explore the relationships between points, represented as coordinates, describing geometrical objects. We shall do this with simple examples from analytic geometry.

For example, in terms of the geometry, the intersection of two circles can occur at two points, and   15
the intersection of two spheres gives a circle. These ideas may be verified using algebra, but in a very different, since the line can traverse through the circle, like a piece of thread going through the eye of a needle. In such cases the intersections are complex intersections.

For each of these problems, the lines and circles may intersect, or not, depending on how they are drawn. Yet we now know that even when they do not intersect on the sheet of paper, they still have   20
an intersection, but the solution is $\in \mathbb{C}$. Finding such solutions require the use of algebra rather than geometry. These ideas were in the process of being understood, first by Fermat and Descartes, then by Newton, followed by the Bernoulli family and Euler.

Complex analytic functions: A very delicate point, that seems to have been ignored for centuries, is that the roots of $P_n(x)$ are, in general, complex, namely $x_k \in \mathbb{C}$. It seems a mystery that complex   25
numbers were not accepted once the quadratic equation was discovered, but they were not. Newton called complex roots *imaginary*, presumably in a pejorative sense. The algebra of complex numbers

is first attributed to Bombelli in 1575, more than 100 years before Newton. One can only begin to imagine what Bombelli learned from Diophantus, following his discovery of Diophantus' Arithmetic, that he discovered in the Vatican library (Stillwell, 2010, p. 51).

It is interesting that Newton was using power series with fractional degree, thus requiring multi-valued solutions, much later to be known as *branch cuts* (c1851). These topics will be explored in Section 3.1.1.

When the argument is complex, analytic functions takes on an entirely new character. For example Euler's identity (1748) with $z = x + y\jmath \in \mathbb{C}$ results in $e^z \in \mathbb{C}$ (Stillwell, 2010, p. 315)

$$e^z = e^x(\cos(y) + \jmath\sin(y)).$$

It should be clear that the complex analytic functions results in a new category of algebra, with no further assumptions beyond allowing the argument to be complex.

Prior to 1851 most of the analysis *assumed* that the roots of polynomials were real ($x_k \in \mathbb{R}$), even though there was massive evidence that they were complex ($x_n \in \mathbb{C}$). This is clearly evident in Newton's work (c1687): When he found a non-real root, he ignore it (Stillwell, 2010, pp. 115-7). Euler (c1748) first derived the Zeta function as a function of real arguments $\zeta(x)$ with $\zeta, x \in \mathbb{R}$. Cauchy (c1814) broke this staid thinking with his analysis of complex analytic functions, but it was Riemann thesis (c1851), when working with Gauss (1777-1855), which had a several landmark breakthroughs. In this work Riemann introduced the *extended complex plane*, which explained the *point* at infinity. He also introduced *Riemann sheets* and *Branch cuts*, which finally allowed mathematics to better describe the physical world (Section 1.4.2).

Once the argument of an analytic function is complex, for example an impedance $Z(s)$, or the Riemann Zeta function $\zeta(s)$, The development of complex analytic functions led to many new fundamental theorems. Complex analytic functions have poles and zeros, branch cuts, Riemann sheets and can be analytic at the point at infinity. Many of these properties were first worked out by Augustin-Louis Cauchy (1789-1857), who drew heavily on the much earlier work of Euler, expanding Euler's ideas into the complex plane (Chapter 4).

Systems of equations

We don't need to restrict ourselves to polynomials in one variable. In some sense $y(x) = ax + b$ is already an equation in two variables. Moving beyond such a simple example, we can work with the equation for a circle, having radius r

$$y^2 + x^2 = r^2,$$

which is quadratic in two variables. Solving for roots $y(x_r) = 0$ ($y^2(x_r) = r^2 - x_r^2 = 0$) gives $(r - x_r)(r + x_r)$, which simply says that when the circle crosses the $y = 0$ line at $x_r = \pm r$.

This equation may also be factored as

$$(y - x\jmath)(y + x\jmath) = r^2,$$

as is easily demonstrated by multiplying out the two monomials. This does not mean that a circle has complex roots. A root is defined by either $y(x_r) = 0$ or $x(y_r) = 0$.

Writing the conic as a 2d degree polynomial gives

$$P_2(x) = ax^2 + bx + c,$$

with $y^2(x) = -P_2(x)$. Setting this equal to zero and completing the square (Eq. 1.20, p. 45), gives the equation for the roots

$$\left(x_\pm + \frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2 + \frac{c}{a} = 0,$$

or

$$x_\pm = -\frac{b}{2a} \pm \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}}.$$

The polynomial in factored form is

$$y^2 = -\left(x - \frac{b}{2a}\right)^2 + \left(\frac{b}{2a}\right)^2 - \frac{c}{a},$$

is a conic section, and becomes a circle with $a = 1$, $b = 0$ and $c = -r^2$.

3.2   Week 5

3.2.1   Lec 12 The physics behind complex analytic expressions: linear vs. nonlinear

The question we address here is "When do multi-variable complex analytic expressions appear in physics?" The most common example comes from the solution of the wave equation (Eq. 1.14) in three dimensions. Such cases arise in wave-guide problems, semiconductors, plasma waves, or for acoustic wave propagation in crystals (Brillouin, 1960) and the earth's mantel (e.g., seismic waves, earthquakes, etc.). The solutions to these problems are based on the *eigenfunction* for the *vector wave equation* (see Chapter 5),

$$P(s, \mathbf{x}) = e^{st}e^{-\boldsymbol{\kappa}\cdot\mathbf{x}}, \tag{3.2}$$

where vector $\mathbf{x} = [x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}] \in \mathbb{R}^3$ points in the direction of the wave, unit vectors $[\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}]$ are $\in \mathbb{R}^3$ and $s = \sigma + \omega\jmath \in \mathbb{C}$ [rad] is the Laplace frequency. The vector function $\boldsymbol{\kappa}(s) \in \mathbb{R}^3$ is the complex vector *wave number*, which describes the propagation of a plane wave of radian frequency ω, in the $\mathbf{x}$ direction.   5

Just as the frequency $s = \sigma + \omega\jmath$ must be complex, it is important to allow the *wave number* function[1] to be complex,[2] because in general it will have a real part, to account for losses as the wave propagates. While it is common to assume there are no losses, in reality this assumption cannot be correct. In many cases it is an excellent approximation (e.g., even the losses of light in-vacuo are not zero) that gives realistic answers. But it is important to start with a notation that accounts for the most general situation, so that when losses must be accounted for, the notation need not change. With this in mind, we take the vector wave number to be complex

$$\boldsymbol{\kappa} = \mathbf{k}_r + \mathbf{k}\jmath,$$

where vector expression for the *lattice vector* is the imaginary part of $\boldsymbol{\kappa}$

$$\Im\boldsymbol{\kappa} = \mathbf{k} = \frac{2\pi}{\lambda_x}\hat{\mathbf{i}} + \frac{2\pi}{\lambda_y}\hat{\mathbf{j}} + \frac{2\pi}{\lambda_z}\hat{\mathbf{k}}, \tag{3.3}$$

is the vector wave number for three dimensional lossless plane-wave solutions ($\mathbf{k}_r = 0$).

Equation Eq. 3.2 is linear in $\mathbf{x}$. If one takes the derivative with respect to either time or space,

$$\frac{\partial}{\partial t}e^{st}e^{-\boldsymbol{\kappa}\cdot\mathbf{x}} = se^{st}e^{-\boldsymbol{\kappa}\cdot\mathbf{x}}, \qquad \frac{\partial}{\partial x}e^{st}e^{-\boldsymbol{\kappa}\cdot\mathbf{x}} = \frac{2\pi}{\lambda_x}e^{st}e^{-\boldsymbol{\kappa}\cdot\mathbf{x}}, \qquad \nabla e^{st}e^{-\boldsymbol{\kappa}\cdot\boldsymbol{x}} = \boldsymbol{\kappa}e^{st}e^{-\boldsymbol{\kappa}\cdot\boldsymbol{x}}$$

we find the eigenvalue of that derivative.   10

The units of $\boldsymbol{\kappa}$ are reciprocal length [m^{-1}] since $\boldsymbol{\kappa} \cdot \mathbf{x}$ has units of radians. When there are losses $\kappa_r(s) = \Re\kappa(s)$ must be a function of frequency, due to the physics behind these losses. In many important cases, such as loss-less wave propagation in semiconductors, $\boldsymbol{\kappa}(\mathbf{x})$ is a function of direction and position (Brillouin, 1960).

Add figure expl ing of $[\lambda_x, \lambda_y, \lambda_z$

---

[1]This function has many names in the literature, all of which are confusing. It is called the *wave number*, *propagation constant* and the Brillouin zone *dispersion function* (Brillouin, 1953, Ch. 1). However, its neither a number nor constant.

[2]In fact $\boldsymbol{\kappa}(s)$ is a complex analytic function of the Laplace frequency s.

When the eigenfunction Eq. 3.2 is applied to the wave equation, a quadratic (degree 2) algebraic expression results, known as the *dispersion relation*. The three dimensional dispersion relation

$$\left(\frac{s}{c}\right)^2 = \boldsymbol{\kappa} \cdot \boldsymbol{\kappa} = \left(\frac{2\pi}{\lambda_x}\right)^2 + \left(\frac{2\pi}{\lambda_y}\right)^2 + \left(\frac{2\pi}{\lambda_z}\right)^2 = k_x^2 + k_y^2 + k_z^2 \tag{3.4}$$

is a complex analytic algebraic relationship in four variables, frequency s and the three complex *lattice wave numbers*. This represents a three-dimensional generalization of the well know relation between wavelength and frequency, i.e., $f\lambda = c$. For lossless plane waves propagating in free space, $|\boldsymbol{\kappa}(s)| = \pm|s/c|$, where the sign accounts for the direction of the plane wave.

5   This scalar relation ($f\lambda = c$) was first deduced by Galileo in the 16th century and was then explored further by Mersenne a few years later.This relationship would have been important to Newton when formulating the wave equation, which he needed to estimate the speed of sound. We shall return to this in Chapters 4 and 5.

Inner product space: Another important example of algebraic expressions in mathematics is Hilbert's
10   generalization of Eq. 1.1, known as the Schwarz inequality, shown in Fig. 3.1. What is special about this generalization is that it proves that when the vertex is 90°, the length of the leg is minimum.

Vectors may be generalize to have ∞ dimensions: $\vec{U}, \vec{V} = [v_1, v_2, \cdots, v_\infty])$. The inner product (i.e., dot product) between two such vectors generalizes the finite dimentional case

$$\vec{U} \cdot \vec{V} = \sum_{k=1}^{\infty} u_k v_k.$$

As with the finite case, the *norm* $||\vec{U}|| = \sqrt{\vec{U} \cdot \vec{U}} = \sqrt{\sum u_k^2}$ is the dot product of the vector with itself, defining the *length* of the infinite component vector. Obviously there is an issue of convergence, if the norm for the vector to have a finite length.

It is a somewhat arbitrary requirement that $a, b, c \in \mathbb{R}$ for the Pythagorean theorem (Eq. 1.1). This seems natural enough since the sides are lengths. But, what if they are taken from the complex numbers, as for the lossy vector wave equation, or the lengths of vectors in $\mathbb{C}^n$? Then the equation generalizes to

$$\mathbf{c} \cdot \mathbf{c} = ||\mathbf{c}||^2 = \sum_{k=1}^{n} |c_k|^2,$$

15   where $||\mathbf{c}||^2 = (\mathbf{c}, \mathbf{c})$ is the inner (dot) product of a vector $\mathbf{c}$ with itself, where $|c_k|$ is the magnitude the complex c_k. As before, $||\mathbf{c}|| = \sqrt{||\mathbf{c}||^2}$ is the *norm* of vector $\mathbf{c}$, akin to a length.
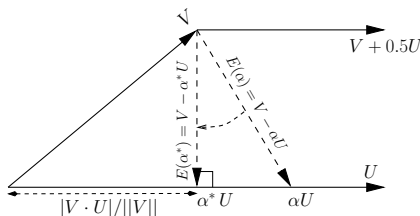
Figure 3.1:   *The* Schwarz inequality *is related to the shortest distance (length of a line) between the ends of the two vectors.* $||U|| = \sqrt{(U \cdot U)}$ *as the dot product of that vector with itself.*

Schwarz inequality   The *Schwarz inequality*[3] says that the magnitude of the inner product of two vectors is less than or equal to the product of their lengths

$$U \cdot V \leq ||U|| \, ||V||.$$

---

[3]A simplified derivation is provided in Sect. 1.3.4 (p. 55).

This may be simplified by normalizing the vectors to have unit length ($\widehat{U} = U/||U||, \widehat{V} = V/||V||$), in which case $-1 < \widehat{U} \cdot \widehat{V} \leq 1$. Another simplification is to define the scaler product in terms of the direction cosine

$$\cos(\theta) = |\widehat{U} \cdot \widehat{V}| \leq 1.$$

A proof of the Schwarz inequality is as follows: From these definitions we may define the minimum difference between the two vectors as the perpendicular from the end of one to the intersection of the second. As shown in Fig. 3.1, $U \perp V$ may be found by minimizing the length of the vector difference:

$$\min_{\alpha} ||V - \alpha U||^2 = ||V||^2 + 2\alpha V \cdot U + \alpha^2 ||U||^2 > 0$$
$$0 = \partial_\alpha (V - \alpha U) \cdot (V - \alpha U)$$
$$= V \cdot U - \alpha^* ||U||^2$$
$$\therefore \quad \alpha^* = V \cdot U/||U||^2.$$

The *Schwarz inequality* follows:

$$I_{\min} = ||V - \alpha^* U||^2 = ||V||^2 - \frac{|U \cdot V|^2}{||U||^2} > 0$$

$$0 \leq |U \cdot V| \leq ||U|| \, ||V||$$

An important example of such a vector space includes the definition of the *Fourier Transform*, where we may set

$$U(\omega) = e^{-\omega_0 \jmath t} \qquad V(\omega) = e^{\omega \jmath t} \qquad U \cdot V = \int_\omega e^{\jmath \omega t} e^{-\jmath \omega_0 t} \frac{d\omega}{2\pi} = \delta(\omega - \omega_0).$$

It seems that the Fourier transform is a result that follows from a minimization, unlike the Laplace transform that follows from causal system parameters. This explains the important differences between the two, in terms of their properties (Unlike the LT, the FT is not complex analytic). We also explored this topic in Lecture 1.3.10 (p. 66).

Power vs. power series, linear vs. nonlinear

Another place where equations of second degree appear in physical applications is in energy and power calculations. The electrical power is given by the product of the voltage $v(t)$ and current $i(t)$ (or in mechanics as the force times the velocity). For example if we define $\mathcal{P} = v(t)i(t)$ to be the power $\mathcal{P}$ [Watts], then the *total energy* [Joules] at time t is (Van Valkenburg, 1964a, Chapter 14)

$$\mathcal{E}(t) = \int_0^t v(t)i(t)dt.$$

From this observe that the power is the rate of change of the total energy

$$\mathcal{P}(t) = \frac{d}{dt}\mathcal{E}(t).$$

Ohm's Law and impedance: The ratio of voltage over the current is call the *impedance* which has units of [Ohms]. For example given a resistor of $R = 10$ [ohms],

$$v(t) = R\,i(t).$$

Namely 1 amp flowing through the resistor would give 10 volts across it. Merging the linear relation due to Ohm's law with the definition of power, shows that the instantaneous power in a resistor is quadratic in voltage and current

$$\mathcal{P} = v(t)^2/R = i(t)^2 R. \tag{3.5}$$

Note that Ohm's law is linear in its relation between voltage and current whereas the power and energy are *nonlinear*.

Ohm's Law generalizes in a very important way, allowing the impedance (e.g., resistance) to be a linear complex analytic function of complex frequency $s = \sigma + \omega\jmath$ (Kennelly, 1893; Brune, 1931b). Impedance is a fundamental concept in many fields of engineering. For example:[4] Newton's second law $F = ma$ obeys Ohm's law, with mechanical impedance $Z(s) = sm$. Hooke's Law $F = kx$ for a spring is described by a mechanical impedance $Z(s) = k/s$. In mechanics a "resistor" is called a *dashpot* and its impedance is a positive and real constant.[5]

Kirchhoff's Laws KCL, KVL: The laws of electricity and mechanics may be written down using Kirchoff's Laws current and voltage laws, (KCL, KVL), which lead to linear systems of equations in the currents and voltages (velocities and forces) of the system under study, with complex coefficients having positive real parts.

Points of major confusion are a number of terms that are misused, and overused, in the fields of mathematics, physics and engineering. Some of the most obviously abused terms are *linear/nonlinear*, *energy*, *power*, *power series*. These have multiple meanings, which can, and are, fundamentally in conflict.

Transfer functions (Transfer matrix):   The only method that seems to work, to sort this out, is to cite the relevant physical application, in specific contexts. The most common touch point is a physical system that has an input $x(t)$ and an output $y(t)$. If the system is linear, then it may be represented by its *impulse response* $h(t)$. In such cases the system equation is

$$y(t) = h(t) \star x(t) \leftrightarrow Y(\omega) = H(s)|_{\sigma=0}\, X(\omega),$$

namely the convolution of the input with the impulse response gives the output. From Fourier analysis this relation may be written in the real frequency domain as a product of the Laplace transform of the impulse response, evaluated on the $\omega\jmath$ axis and the Fourier transform of the input $X(\omega) \leftrightarrow x(t)$ and output $Y(\omega) \leftrightarrow y(t)$.

Mention ABCD Transfer matrix

If the system is nonlinear, then the output is not given by a convolution, and the Fourier and Laplace transforms have no obvious meaning.

The question that must be addressed is why is the power said to be nonlinear whereas a power series of $H(s)$ said to be linear. Both have powers of the underlying variables. This is massively confusing, and must be addressed. The question will be further addressed in Section 3.5.1 in terms of the system postulates of physical systems.

Whats going on? The domain variables must be separated from the codomain variables. In our example, the voltage and current are multiplied together, resulting in a nonlinear output, the power. If the frequency is squared, this is describing the degree of a polynomial. This is not nonlinear because it does not impact the signal output, it characterizes the Laplace transform of the system response.

3.2.2   Lec 13 Root classification of polynomials

Root classification for polynomials of Degree * = 1–4 (p.102);
Quintic (* = 5) cannot be solved: Why?
Fundamental Thm of Algebra (d'Alembert, $\approx$1760)

Add intro & merge convolution discussions.

---

[4]In acoustics the pressure is a potential, like voltage. The force per unit area is given by $f = -\nabla p$ thus $F = -\int \nabla p\, dS$. Velocity is analogous to a current. In terms of the velocity potential, the velocity per unit area is $v = -\nabla\phi$.

[5]`https://en.wikipedia.org/wiki/Impedance_analogy`

Convolution

As we discussed in Chapter 1, given the roots, the construction of higher degree polynomials, is greatly assisted by the convolution method. This has physical meaning, and gives insight into the problem of factoring higher order polynomials. By this method we can obtain explicit relations for the coefficients of any polynomial in terms of its roots.

Extending the example of Section 1.3.2, let's find the relations for the cubic. For simplicity, assume that the polynomial has been normalized so that the lead x^3 term has coefficient 1. Then the cubic in terms of its roots $[a, b, c]$ is a convolution of three terms

$$[1, a] \star [1, b] \star [1, c] = [1, a + b, ab] \star [1, c] \tag{3.6a}$$

$$= [1, a + b + c, ab + c(a + b), abc]. \tag{3.6b}$$

Working out the coefficients for a *quartic* gives

$$[1, a + b + c, ab + c(a + b), abc] \star [1, d] = \tag{3.7}$$

$$[1, a + b + c + d, d(a + b + c) + c(a + b) + ab, d(ab + ac + bc) + abc, abcd]. \tag{3.8}$$

It is clear what is going on here. The coefficient on x^4 is 1 (by construction). The coefficient for x^3 is the sum over the roots. The x^2 term is the sum over all possible products of pairs of roots, The linear term x is the sum over all triple products of the four roots, and finally the last term (a constant) is the product of the four roots.

In fact this is a well known, a frequently quoted result from the mathematical literature, and trivial to show given an understand of convolution. If one wants the coefficients for the quintic, it is not even necessary to use convolution, as the pattern (rule) for all the coefficients is now clear.

You can experiment with this numerically using Matlab's convolution routine `conv(a,b)`. Once we start studying Laplace and Fourier transforms, convolution becomes critically important because multiplying an input signal in the frequency domain by a transfer function, also a function of frequency, is the same a convolution of the time domain signal with the inverse Laplace transform of the transfer function. So you didn't need to learn how to take a Laplace transform, and then learn convolution. We have learned convolution first independent of the Fourier and Laplace transforms.

When the coefficients are real, the roots must appear as conjugate pairs. This is an important symmetry.

For the case of the quadratic we have the relations between the coefficients and the roots, found by completing the square. This required isolating x to a single term, and solving for it. We then proceeded to find the coefficients for the cubic and quartic case, after a few lines of calculation. For the quartic

$$a_4 = 1$$
$$a_3 = a + b + c + d$$
$$a_2 = d(a + b + c) + c(a + b) + ba$$
$$a_1 = d(ab + ac + bd) + abc$$
$$a_0 = abcd$$

These relationships are algebraically nonlinear in the roots. From the work of Galois, for $N \geq 5$, this system of equations is impossible to invert. Namely, given a_k, one may not determine the four roots $[a, b, c, d]$ analytically. One must use numeric methods.

To gain some insight, let us look at the problem for $N = 2$, which has a closed form solution:

$$a_2 = 1$$
$$a_1 = a + b$$
$$a_0 = ab$$

We must solve for $[a, b]$ given twice the mean, $2(a+b)/2$, and the square of the geometric mean $\left(\sqrt{ab}\right)^2$. Since we already know the answer (i.e, the quadratic formula). The solution was first worked out by the Babylonians (2000 BCE) Stillwell (2010, p. 92). It is important to recognize that for physical systems, the coefficients a_k are real. This requires that the roots come in conjugate pairs ($b = a^*$), thus $ab = |a|^2$

5 and $a + b = 2\Re a$, which makes the problem somewhat more difficult, due to the greater symmetry.

Once you have solved this problem, feel free to attempt the cubic case. Again, the answer is known, after thousands of years of searching. The solution to the cubic is given in (Stillwell, 2010, pp. 97-9), as discovered by Cardano in 1545. According to Stillwell "The solution of the cubic was the first clear advance in mathematics since the time of the Greeks." The ability to solve this problem required

10 algebra, and the solutions were complex numbers. The denial of complex numbers was, in my view, the main stumbling block in the progress of these solutions. For example, how can two parallel lines have a solution? Equally mystifying, how can a circle and a line, that do not intersect, have intersections? From the algebra we know that they do. This was a basic problem that needed to be overcome. This story is still alive,[6] because the cubic solution is so difficult.[7] One can only begin to imagine how much

15 more difficult the quartic is, solved by Cardano's student Ferrair, and published by Cardano in 1545. The impossibility of the quintic was finally resolved in 1826 by Able (Stillwell, 2010, p. 102).

Finally with these challenges behind them, Analytic Geometry, relating of algebra and geometry, via coordinate systems, was born.

3.2.3   Lec 14: Development of Analytic Geometry

20 **Composition, Intersection and Gaussian elimination:**

When the impedance load, as a function of wave penetration depth x is uniform, then as the wave P^+ travels, then there are no reflections. In this case $\Gamma(s, x = 0) = 0$, and the impedance is equal to r_o. Given variations in the properties of the medium, say after a delay of $\tau = x/c$, backward propagated waves are returned to the input. This results in a change in the load impedance, both in real and

25 imaginary parts. As long as $|\Gamma(s, x)| \leq 1$, the real part of $Z(s, x = 0 \geq 0$. This is further discussed in Sect. 1.5.3 (p. 100).

As a counter example consider the well known problem in geometry: the intersection of a plane with a cone, which leads to the conic sections: the circle, hyperbola, ellipse and parabola, along with some degenerate cases, such as the intersection of two straight lines[8]. If we stick to such 3-dimensional

30 objects, we can write equations in the three variables $[x, y, z]$, and be sure that they each represent some physical geometry. For example $x^2 + y^2 + z^2 = r_o^2$ is a sphere of radius r_o.

The geometry and the algebra do not always seem to agree. Which is correct? In general the geometry only looks at the real part of the solution, unless you know how to tease out the complex solutions. However the roots of every polynomial can be in $\mathbb{C}$, so we may not ignore the imaginary

35 roots, as Newton did. There is an important related fundamental theorem, known as Bézout's theorem, that address this case, as described at the end of Section 1.3.4, 59.

Composition of polynomials: Bézout's theorem

When we multiply two polynomials, the resulting degree is the sum of the degrees of the individual polynomials (Section 1.3.3, p. 54). However when we compose two polynomials, of degree m and n,

40 the resulting degree is the *product* of the degrees. For example given two polynomials of degree n and m, $f_n(x)$ and $g_m(x)$, their composition is defined as $f_n \circ g_m = f(g(x))$. For example, the composition of a line with a line, is a line. The composition of a line with a circle, is a circle, and the composition of a circle with a circle is a quartic.

---

[6]M. Kac, *How I became a mathematician." American Scientist (72), 498–499.*

[7]`https://www.google.com/search?client=ubuntu&channel=fs&q=Kac+\%22how+I+became+a\%22+1984+pdf&ie=utf-8&oe=utf-8`

[8]Such problems were first studied algebraically and Descartes (Stillwell, 2010, p. 118) and Fermat (c1637).

This is known as Bézout's theorem, which states the number of roots of the composition of two functions is determined by the product of their degrees. The power of Bézout's theorem is that it applies to functions of more than one variable. Composition of functions is also known as the *construct of equations* (Stillwell, 2010, p. 118). It was proved by Bézout (1779), who showed how one must count all the roots, including multiple roots, and roots at infinity (Stillwell, 2010, p. 295).

5
Totally rework

Exercises:

1. **Two Circles:** The composition of 2 independent circles give 4 points:

$$x^2 + y^2 = 1, \qquad (x-1)^2 + y^2 = 1$$

Two circles intersect at 2 point: How does this work on using algebra?

Here we subtract the two equations, eliminating y

$$x^2 - (x-1)^2 = x^2 - (x^2 - 2x + 1) = 2x - 1 = 0$$

Thus $x = 1/2$. Verifying gives

$$(1/2)^2 + y^2 = 1, \qquad (1/2 - 1)^2 + y^2 = 1$$

or

$$y^2 = 1 - 1/4 = 3/4, \qquad y^2 = 1 - 1/2 = 1/2$$

or

$$y = \pm\sqrt{3/4}, \qquad y = \pm\sqrt{1/2}.$$

In summary, there is only one value of $x = 1/2$ but there are two values of y, which is independent of the sign, since y appears in the equations as y^2. Thus there are *four* solutions: $(1/2, \pm\sqrt{3/4})$ and $(1/2, \pm\sqrt{1/2})$.

10

2. **Composition of a line and a circle:** clearly state the problem

Let the circle is $x^2 + y^2 = 1$ and the line be $x + y = 1$. Graphically it is easy to show that the solutions are $(0, 1)$ and $(1, 0)$. This problem has already been addressed with the search for Pythagorean triplets. In Fig. 2.3 we derive the real integer solutions for the line from $-1, 0$ to x, y intersection the circle. The values of a, b, c are presented as Proof III in that figure. But what can we say about the angles Θ? The formula for the angles is

$$e^{1\jmath\Theta} = \frac{1 + t\jmath}{1 - t\jmath}$$

with $t = p/q \in \mathbb{Q}$.

If we solve this equation for $t\jmath$ we find

$$(1 - t\jmath)e^{1\jmath\Theta} = 1 + t\jmath$$
$$(e^{1\jmath\Theta} - 1) = t\jmath(e^{1\jmath\Theta} + 1)$$
$$t\jmath = \frac{e^{1\jmath\Theta} - 1}{e^{1\jmath\Theta} + 1} = \frac{e^{1\jmath\Theta/2} - e^{-1\jmath\Theta/2}}{e^{1\jmath\Theta/2} + e^{-1\jmath\Theta/2}}$$

3. **Problem 3:** Suppose we have two circles

$$y^2 + x^2 = 1 \qquad\qquad\qquad y^2 + x^2 = 1 \qquad\qquad (3.9)$$

There are an infinite number of solutions since the circles lie on top of each other. All values of x, y are a solution. If

$$y^2 + x^2 = 1 \qquad\qquad\qquad y^2 + x^2 = 2 \qquad\qquad (3.10)$$

there are no solutions, as the circles never touch. When

$$y^2 + x^2 = 1 \qquad\qquad\qquad y^2 + (x-1)^2 = 1 \qquad\qquad (3.11)$$

there are two intersection points. How can we find them?

We may remove y from these equations by subtracting them, which results in

$$(x-1)^2 - x^2 = 0$$

which is easily solved since $x^2 - 2x + 1 - x^2 = 0$, $x = 1/2$. We can then put this back into the first equation to find $y^2 + 1/4 = 1$ or $y = \pm\sqrt{3/4}$. This is easily verified.

What if the two circles are

$$y^2 + x^2 = 1 \qquad\qquad\qquad y^2 + (x-1\jmath)^2 = 1 \qquad\qquad (3.12)$$

Removing y as before gives $(x - 1\jmath)^2 - x^2 = 0$ or $\cancel{x^2} - 2\jmath x + 1 - \cancel{x^2} = 0$. Thus $x = 1\jmath/2$. Putting this back in the first (or second) equation gives $y^2 - 1/2 = 1$, or $y = \pm\sqrt{5}/2$. Thus the final solution is $(x_o, y_o) = 1\jmath/2, \pm 5/2$.

What can we take away from this exercise? The geometry fails us, at least as far as we know how to view it, in real coordinates. However using algebra we can determine the complex solutions. It is no wonder that Newton, and many others, were fooled by such calculations, and called such results imaginary. But as we shall learn, only by understanding such equations, can we fully appreciate the physics that is built upon them.

4. **Problem 4:** This problem is similar to problem 3 of two circles, but is in 3 dimensions with the intersection of two spheres. A unit radius sphere is given by

$$x^2 + y^2 + z^2 = 1$$

If we take $[x, y, z] \in \mathbb{R}$ we can easily picture the intersection of two spheres. If they just touch, the intersection is a single point. If they overlap, the solution is a circle, and if one is inside the other, there are no (real) solutions.

Next is an example of a polynomial in three variables, with no linear terms. We can generalize it by a shift in space

$$(x-1)^2 + (y+2)^2 + (z-3)^2 = 2^2.$$

Here we shift x by 1, y by -2 and z by 3, and scaled the radius from 1 to 2. Clearly this equation has the same geometry (its still a sphere), in factored form. Expanding it out gives

$$x^2 - 2x + y^2 + 2y + z^2 - 3z + 10 = 0.$$

Clearly the factored format is much easier to interpret. It should be clear how find the factored form given the general quadratic polynomial in the three variables (use Eq. 1.20 three times). Another interesting question is "Why are there no ZY, ZY, xyz terms?

What if $[x, y, z] \in \mathbb{C}$? Does this make physical sense? The answer is yes, if the variables are not the spatial coordinates but something called the *complex wave number*. More on this later.

3.3 Week 6

3.3.1 Lec 15 Gaussian Elimination of linear equations

Example problems using Gaussian Elimination: Gaussian Elimination is valid for nonlinear systems of equations. Till now we have emphasized the reduction of linear systems of equations.

Problem 1: Two lines *in a plane* either intersect or are parallel, in which case they are said to meet at ∞. Does this make sense? The two equations that describe this may be written in matrix form as $Ax = b$, which written out as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{3.13}$$

The intersection point x_0, y_0 is given by the solution two these two equations

$$\begin{bmatrix} x_1 \\ x_1 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \tag{3.14}$$

where $\Delta = a_{11}a_{22} - a_{12}a_{21}$ is the determinant of matrix A (Matlab's `det(A)` function).   5

It is useful to give an interpretation of these two equations. Each row of the 2x2 matrix defines a line in the (x, y) plane. The top row is

$$a_{11}x + a_{12}y = b_1.$$

Normally we would write this equation as $y(x) = \alpha x + \beta$, where α is the slope and β is the *intercept* (i.e., $y(0) = \beta$). In terms of the elements of matrix A, the slope of the first equation is $\alpha = -a_{11}/a_{12}$ while the slope of the second is $\alpha = -a_{21}/a_{22}$. The two slopes are equal (the lines are parallel) when $-a_{11}/a_{12} = -a_{21}/a_{22}$, or written out

$$\Delta = a_{11}a_{22} - a_{12}a_{21} = 0.$$

Thus when the determinate is zero, the two lines are parallel and there is no solution to the equations.

This 2x2 matrix equation is equivalent to a 2^d degree polynomial. If we seek an eigenvector solution $[e_1, e_2]^T$ such that

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \lambda \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{3.15}$$

the 2x2 equation becomes singular, and λ is one of the roots of the polynomial. One may proceed by merging the two terms to give

$$\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{3.16}$$

Clearly this new matrix has no solution, since if it did, $[e_1, e_2]^T$ would be zero, which is nonsense. If it has no solution, then the determinant of the matrix must be zero. Forming this determinate gives

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

thus we obtain the following quadratic equation for the roots $\lambda_\pm$ (eigenvalues)

$$\lambda_\pm^2 - (a_{11} + a_{22})\lambda_\pm + \Delta = 0.$$

When $\Delta = 0$, one eigenvalue is zero while the other is $a_{11} + a_{22}$, which is known as the *trace* of the matrix.

In summary: Given a "linear" equation for the point of intersection of two lines, we see that there must be two points of intersection, as there are always two roots of the quadratic *characteristic polynomial*. However the two lines only intersect at one point. Whats going on? What is the meaning of this second root?

Some simple examples will help. The eigenvalues depend on the relative slopes of the lines, which in general can become complex. The intercepts are dependent on **b**. Thus when the RHS is zero, the eigenvalues are irrelevant. This covers the very simple examples. When one eigenvalue is real and the other is imaginary, more interesting things are happening since the slope of one line is real and the slope of the other is pure imaginary. The lines can intersect in the real plane, and again in the complex plane.

Lets try an example of two lines, slopes of 1 and 2: $y_1 = x + a$ and $y_2 = 2x + b$. In matrix form Let

$$\begin{bmatrix} 1 & -1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \tag{3.17}$$

The determinate is $\Delta = -1$, thus the solution is

$$\begin{bmatrix} y_0 \\ x_0 \end{bmatrix} = -1 \begin{bmatrix} -2 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2a - b \\ a - b \end{bmatrix}. \tag{3.18}$$

Thus the two real lines having slopes of 1 and 2 having intercepts of a and b, meet $(x_0, y_0) = (2a-b, a-b)$. We may verify by substituting $x = a - b$ into the starting equations $y_1 = (a - b) + a = 2a - b$ and $y_2 = 2(a - b) + b = 2a - b$, which each $2a - b$.

While there is a unique solution, there are two eigenvalues, given by the roots of

$$(1 - \lambda_\pm)(-2 - \lambda_\pm) + 1 = 0.$$

If we transfer the sign from one monomial to the other

$$(-1 + \lambda_\pm)(2 + \lambda_\pm) + 1 = 0$$

and reorder for simplicity

$$(\lambda_\pm - 1)(\lambda_\pm + 2) + 1 = 0$$

we obtain the quadratic for the roots

$$\lambda_\pm^2 + \lambda_\pm - 1 = 0.$$

Completing the square gives

$$(\lambda_\pm + 1/2)^2 = 3/4.$$

or

$$\lambda_\pm = -1/2 \pm \sqrt{3}/2.$$

The question is, what is the relationship between the eigenvalues and the final solution, if any? Maybe none. The solution (x_0, y_0) is reasonable, and its not clear that the eigenvalues play any useful role here, other than to predict there is a second solution. I'm confused.

Two lines in 3-space: In three dimensions

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{3.19}$$

Each row of the matrix describes a plane, which is said to be linear in the unknowns (x, y, z). Thus the system of linear equations represents three planes, which must intersect at one point. If two planes are parallel, there is no real solution. In this case the intersection by the third plane generates two parallel lines.

As in the 2x2 case, one may convert this linear equation into a cubic polynomial by setting the determinant of the matrix, with $-\lambda$ subtracted from the diagonal, equal to zero. That is, $det(A - \lambda I) = 0$. Here I is the matrix with 1 on the diagonal and zero off the diagonal.

Simple example: As a simple example, let the first plane be $z = 0$ (independent of x, y), the second parallel plane be $z = 1$ (independent of (x, y)) and the third plane be $x = 0$ (independent of y, z). This results in the system of equations

$$\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & 0 & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{3.20}$$

Writing out the three equations we find $a_{13}z = 0$, $a_{23}z = 1$, and $a_{31}x = 0$. Note that $det(A) = 0$ (we need to learn how to compute the 3x3 determinant). This means the three planes never intersect at one point. Use Matlab to find the eigenvalues.

3.3.2 Lec 16 Matrix composition: Bilinear and ABCD transformations

The Transmission matrix 5

A transmission matrix is a 2x2 matrix that characterizes a 2-port circuit, one having an input and output voltage and current, as shown in Fig. 1.13. The input is the voltage and current V_1, I_1 and the output is the voltage and current $V_2, -I_2$, with the current always defined to flow into the port. For any such a linear network, the input-output relations may be written in a totally general way as

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix}.$$

In Section 1.3.7 we showed that a cascade of such matrices is composition. We shall show below that the justification of this relationship is based on the composition of bilinear transformations.

Expanding Eq. 5.8 into its individual equations demonstrates the linear form of the relations

$$V_1 = A(s)V_2 - B(s)I_2 \qquad\qquad I_1 = C(s)V_2 - D(s)I_2,$$

quantifying the relationship between the input voltage and current to its output voltage and current.

Define $H(s) = V_2/V_1$ as the *transfer function*, as the ratio of the output voltage V_2 over the input voltage V_1, under the constraint that the output current $I_2 = 0$. From this definition $H(s) = 1/A(s)$. 10

In a similar fashion we may define the meaning of all four functions as

$$A(s) \equiv \left.\frac{V_1}{V_2}\right|_{I_2=0} \qquad\qquad B(s) \equiv -\left.\frac{V_1}{I_2}\right|_{V_2=0} \tag{3.21}$$

$$C(s) \equiv \left.\frac{I_1}{V_2}\right|_{I_2=0} \qquad\qquad D(s) \equiv -\left.\frac{I_1}{I_2}\right|_{V_2=0} \tag{3.22}$$

From Eq. 5.8 one may compute any desired quantity, specifically those quantities defined in Eq. 3.22, the open circuit voltage transfer function $(1/A(s))$, the short-circuit transfer current $(1/D(s))$ and the two transfer impedances $B(s)$ and $1/C(s)$.

In the engineering fields this matrix composition is called the *Transmission matrix*, also known as the ABCD method. It is a powerful method that is easy to learn and use, that gives important insights 15 into transmission lines, and thus even the 1 dimensional wave equation.

Derivation of ABCD matrix for example of Fig. 1.13 (p. 61).

The derivation is straight forward by the application of Ohm's Law, as shown in Section 1.3.7.

The convenience of the ABCD matrix method is that the output of one is identically the input of the next. Cascading (composing) the results for the series inductor with the shunt compliance leads to the 2x2 matrix form that precisely corresponds to the transmission line CFA shown in Fig. 2.2,

$$\begin{bmatrix} V_n(s) \\ In(s) \end{bmatrix} = \begin{bmatrix} 1 & sL_n \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_{n+1}(s) \\ -I_{n+1}(s) \end{bmatrix}. \tag{3.23}$$

This matrix relation characterizes the series mass term sL_n. A second equation maybe be used for the shunt capacitance term $sY_n(s)$

$$\begin{bmatrix} V_n(s) \\ In(s) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ sC_n & 1 \end{bmatrix} \begin{bmatrix} V_{n+1}(s) \\ -I_{n+1}(s) \end{bmatrix}. \tag{3.24}$$

The positive constants $L_n, C_n \in \mathbb{R}$ represent the series mass (inductance) and the shunt compliance (capacitance) of the mechanical (electrical) network. The integer n indexes the series and shunt sections, that are composed one following the next.

A more detailed analysis shows that for the case of no losses, the wave velocity is

$$c = 1/\sqrt{LC}. \tag{3.25}$$

Also the *wave number* is

$$\kappa = \sqrt{ZY} = \sqrt{sL \cdot sC} = s\sqrt{LC} = \frac{s}{c}. \tag{3.26}$$

Finally the *characteristic resistance* is

$$r_o = \sqrt{Z/Y} = \sqrt{sL/sC} = \sqrt{L/C}. \tag{3.27}$$

All of these based on a unit lenth Δ_x as shown in Fig. 2.2 (p. 122).

5   **Matrix composition and the bilinear transform:**   Now that we have defined the composition of two functions, we will use it to define the *Möbius* or *bilinear* transformation. Once you understand how this works, hopefully you will understand why it is the unifying element in many important engineering problems.

The bilinear transformation is given by

$$w = \frac{a + bz}{c + dz}$$

This takes one complex number $z = x + iy$ and transforms it into another complex number $w = u + iv$. This transformation is *bilinear* in the sense that its linear in both the input and output side of the equation. This may be seen when written as

$$(c + dz)w = a + bz,$$

since this relation is linear in the coefficients $[a, b, c, d]$. An important example is the transformation between impedance $Z(s)$ and reflectance $\Gamma(s)$,

$$\Gamma(s) = \frac{Z(s) - r_0}{Z(s) + r_0},$$

which is widely used in transmission line problems. In this example $w = \Gamma, z = Z(s), a = -r_0, b = $
10 $1, c = r_0, d = 1$.

If we define a second bilinear transformation (this could be the transformation from reflectance back to impedance)

$$r = \frac{\alpha + \beta w}{\gamma + \delta w},$$

and then compose the two something astray wrt arguments

$$w \circ r = \frac{a + b\,r}{c + d\,r} = \frac{a(\gamma + \delta w) + b(\alpha + \beta w)}{c(\gamma + \delta w) + d(\alpha + \beta)w} = \frac{a\gamma + b\alpha + (a\delta + b\beta)w}{c\gamma + d\alpha + (c\delta + d\beta)w},$$

something surprising happens. The composition $w \circ r$ may be written in matrix form, as the product of two matrices that represents each bilinear transform. This may be seen as true by inspecting the coefficients of the composition $w \circ r$ (shown above) and the product of the two matrices

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} (a\gamma + b\alpha) & (a\delta + b\beta) \\ (c\gamma + d\alpha) & (c\delta + d\beta) \end{bmatrix}.$$

The the power of this composition property of the bilinear transform may be put to work solving important engineering problems, using *transmission matrices*.

3.3.3 Lec 17 Introduction to the Branch cut and Riemann sheets

Branch cuts are required to preserve the single-valued nature of complex analytic functions. When an analytic function is multi-valued, some method needs to be devised to allow the multi-valued complex analytic function to be expanded as a Taylor series, which is necessarily single-valued. It follows that each single-valued sheet must have a different expansion, valid out to the nearest pole (or singularity). We shall explain these ideas with the simplest case, the double-valued square root function $w(z) = \pm\sqrt{z}$, as shown in Fig. 1.18 (p. 82).

3.4 Week 7

3.4.1 Lec 18 Complex analytic mappings (domain coloring)

When one uses complex analytic functions it is helpful to understand their properties in the complex plane. In this sections we explore several well-known functions using domain coloring, discussed in some detail in Section 1.3.9, p. 65. For the following figures the coordinate systems are defined by $s = \sigma + \omega\jmath = \sqrt{x^2 + y^2}e^{\jmath\theta}$ and $w = u + v\jmath = \sqrt{u^2 + v^2}e^{\jmath\psi}$.

For the first example (Fig. 1.19) $w(s) = s^2$ and its inverse $s(w) = \sqrt{w}$ are shown. On the left the red region, corresponding to $0°$ [degrees], appears at both 0 and 180 ($u = \pm 1$) in the w plane. This is because in polar coordinates $s^2 = |s|^2 e^{2\theta\jmath}$ where θ is the angle of $s = |s|e^{2\theta\jmath}$. The square causes the phase to rotate twice around for once around the s plane. Namely the angle is doubled, and the magnitude squared. Due to the faster changing phase in w, there are two red regions, one when $\theta = 0$ and the second at $\theta = \pi$ ($\angle w(s) = 2\theta$). The black spot is dilated due to the squaring of the radius (expanding it).

On the right the $\sqrt{w} = \sqrt{|w|}e^{\phi/2}$. Because the angle of w is divided by two, it takes twice as much phase (in w) to cover the same angle. Thus the red region ($0°$) is expanded. We barely see the violet $90°$ and yellow $-90°$ angles. There is a *branch cut* running from $w = 0$ to $w = \infty$. As the branch cut is crossed, the function switches *Riemann sheets*, going from the top sheet (shown here) to the bottom sheet (not shown). Figure 1.18 in Section 3.3.3 depicts what is going on with these two sheets, and show the branch cut from the origin (point O) to ∞. In this depiction the first sheet ($+\sqrt{z}$) is on the bottom, while the second sheet ($\sqrt{z}$) is on top. For every value of z there are two possible outcomes, $\pm\sqrt{z}$, represented by the two sheets.

Two more examples are given in Fig. 3.2 to interpret the two complex colorized mappings $w = \cos(\pi s)$ (left) and the Bessel function $J_0(\pi z)$. Note how the white and black contour lines are always perpendicular where they cross, just as in the calibration plots for the x and y axes, shown in Fig. 1.15 in Section 1.3.9 (p. 65).

Along the σ axis the $\cos(\pi x$ is the periodic with a period of π. The dark spots are at the zeros. at $\pm\pi/2, \pm 3\pi/2, \ldots$. When we stray off the $\omega\jmath = 0$ axis, the function either goes to zero (black) or ∞ (white). This behavior carries the same π periodicity as it has along the $\omega = 0$ line. On the right is the Bessel function $J_0(\pi z)$, which is similar to $\cos(\pi z)$, except the zeros are distorted away from the orgin. These figure are worthy of careful study to develop an intuition for complex functions of complex variables. In Section 1.3.9 we shall explore more complex mappings, and in greater detail.
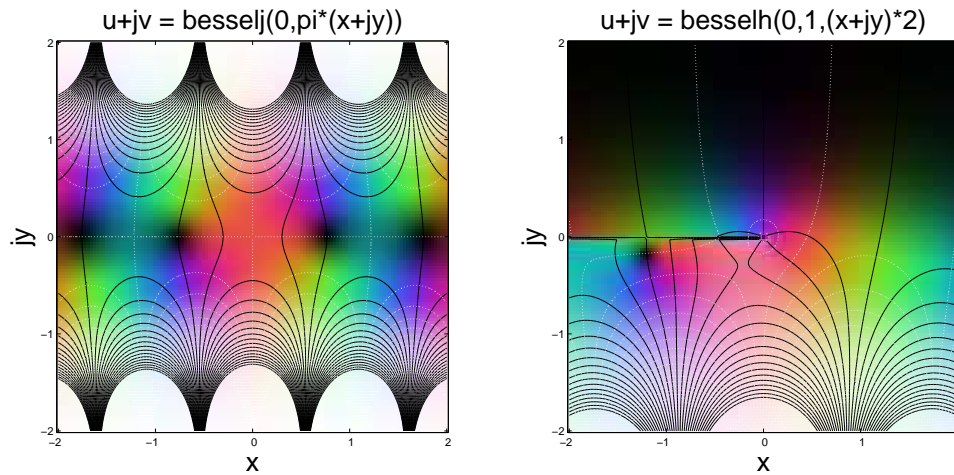
Figure 3.2: *On the left is the Bessel function $J_0(\pi z)$, which is similar to $\cos(\pi z)$, except the zeros are distorted away from $s = 0$ by a small amount, due to the cylindrical geometry. On the right is the related Hankel function $H_0^{(1)}(w^2)$. The Bessel and Hankel functions are solution to the wave equation in cylindrical coordinates with different boundary conditions. The zeros in the function are the places where the pinned boundary condition is satisfied (where the string is restricted, by the boundary, from moving). The Hankel function $H_1^{(1)}(\pi z^2)$ has a branch cut, and a zero poking out near $z = -1.2 + 0.1\jmath$.*
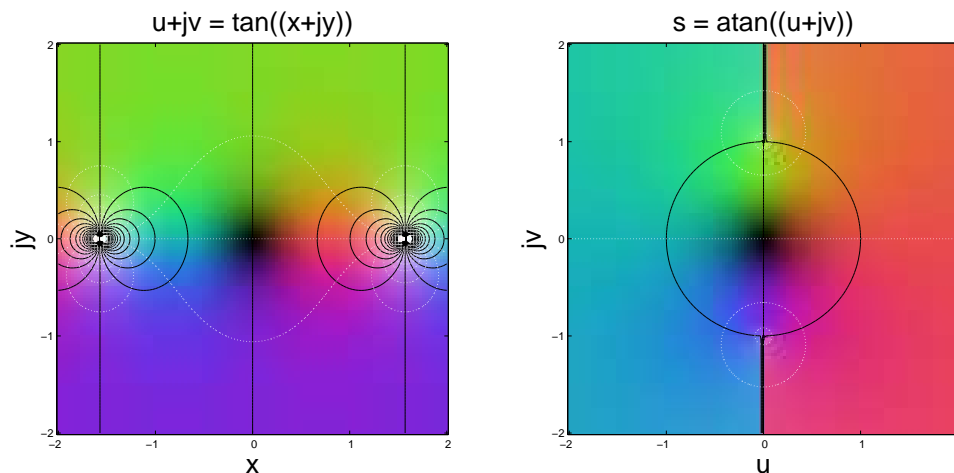
Figure 3.3: *On the left is the function $w(z) = \tan(z)$. On the right is the inverse $s = \tan^{-1}(w)$. Of special interest is* `zviz atan(i*Z) (i/2)*log((1+Z)./(1-Z)).`

In the third example (Fig. 3.3) we show $w = \tan(z)$ and its inverse $z = \tan^{-1}(w)$. The tangent function has zeros where $\sin(z)$ has zeros (e.g., at $z = 0$) and poles where $\cos(z)$ is zero (e.g., at $\pm\pi/2$). The inverse function $s = \operatorname{atan}(w)$ has a zero at $w = 0$ and branch cuts eliminating from $z = \pm\pi$.

The command `zviz besselj(0,pi*Z) besseli(0,j*pi*Z)` gives exactly the same plot, demon-
strating that $I_0(z) = J_0(z\jmath)$.

It is fun, easy and interesting to study polynomials, say of degree 5 and 4 (i.e, with one zero removed), to demonstrate the Fundamental Theorem of Algebra.

3.4.2   Lec 19 Signals and Systems: Fourier vs. Laplace Transforms

Signals and Systems: Fourier vs. Laplace Transforms **AE-3**

3.4.3   Lec 20 Role of Causality and the Laplace Transform

Role of Causality and the Laplace Transform:

$u(t) \leftrightarrow 1/s$ (LT)
$2\tilde{u}(t) \equiv 1 + \text{sgn}(t) \leftrightarrow 2\pi\delta(\omega) + 2/\jmath\omega$ (FT)

3.5 Week 8

3.5.1 Lec 21 The 9 postulates of System of algebraic Networks

Phyical system obey very important rules, that follow from the physics. It is helpful to summarize these physical restrictions in terms of postulates, presented in terms of a taxonomy, or catagorization method, of the fundmental properties of physical systems. Nine of these are listed below. These nine come from a recently published paper (Kim and Allen, 2013). It is possible that given time, others could be added.

A taxonomy of physical systems comes from a systematic summary of the laws of physics, which includes at least the nine basic network postulates, described in Section 1.3.12. To describe each of the network postulates it is helpful to begin with the 2-port transmission (aka ABCD, chain) matrix representation, discussed in Section 3.3.2 (p. 141).

Add figures as sh
notebook (p. 3
ples of these pos
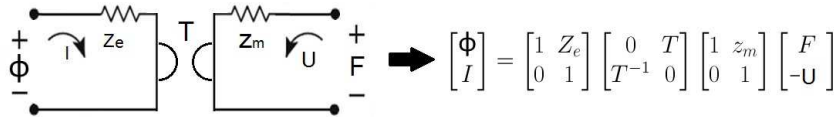
Figure 3.4: *A schematic representation of a 2-port ABCD electro-mechanic system using Hunt parameters $Z_e(s)$, $z_m(s)$, and $T(s)$: electrical impedance, mechanical impedances, and transduction coefficient (Hunt, 1952; Kim and Allen, 2013). Also $V(f)$, $I(f)$, $F(f)$, and $U(f)$ are the frequency domain voltage, current, force, and velocity respectively. Notice how the matrix method 'factors' the 2-port model into three 2×2 matrices. This allows one to separate the physical modeling from the algebra. It is a standard impedance convention that the flows $I(f), U(f)$ are always defined into the port. Thus it is necessary to apply a negative sign on the velocity $-U(f)$ so that it has an outward flow, to feed the next cell with an inward flow. Replace Φ with V.*

As a specific example we show the 2-port *transmission matrix* for an acoustic transducer (loudspeaker), shown in Fig. 3.4, characterized by the equation

$$\begin{bmatrix} \Phi_i \\ I_i \end{bmatrix} = \begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix} \begin{bmatrix} F_l \\ -U_l \end{bmatrix} = \frac{1}{T} \begin{bmatrix} z_m(s) & z_e(s)z_m(s) + T^2 \\ 1 & z_e(s) \end{bmatrix} \begin{bmatrix} F_l \\ -U_l \end{bmatrix}. \tag{3.28}$$

This equation comes from the product of the three 2x2 matrices representing each of the three elements of the figure.

This figure represents the motor of the loudspeaker (not the box that it comes in). The system consists of three elements, the electrical input impedance $Z_e(s)$, a gyrator, which is similar to a transformer, but relates current to force, and an output mechanical impedance $Z_m(s)$. This circuit describes what is needed to fully characterize it operation, from electrical input to mechanical (acoustical) output.

The input is electrical (voltage and current) $[\Phi_i, I_i]$ and the output (load) are the mechanical (force and velocity) $[F_l, U_l]$. The first matrix is the general case, expressed in terms of four unspecified functions $A(s), B(s), C(s), D(s)$, while the second matrix is for the specific example of Fig. 3.4. The four entries are the electrical driving point impedance $Z_e(s)$, the mechanical impedance $z_m(s)$ and the transduction $T = B_0 l$ where B_0 is the magnetic flux strength and l is the length of the wire crossing the flux. Since the transmission matrix is anti-reciprocal, its determinate $\Delta_T = -1$, as is easily verified.

Other common transduction examples of cross-modality transduction include current–thermal (thermoelectric effect) and force–voltage (piezoelectric effect). These systems are all reciprocal, thus the transduction has the same sign.

Impedance matrix

These nine postulates describe the properties of a system having an input and an output. For the case of an electromagnetic transducer (Loudspeaker) the system is described by the 2-port, as show in

Fig. 3.4. The electrical input impedance of a loudspeaker is $Z_e(s)$, defined by

$$Z_e(s) = \left.\frac{V(\omega)}{I(\omega)}\right|_{U=0}.$$

Note that this *driving-point impedance* must be causal, thus it has a Laplace transform and therefore is a function of the complex frequency $s = \sigma + j\omega$, whereas the Fourier transforms of the voltage $V(\omega)$ and current $I(\omega)$ are functions of the real radian frequency ω, since the time-domain voltage $v(t) \leftrightarrow V(\omega)$ and the current $i(t) \leftrightarrow I(\omega)$ are signals that may start and stop at any time (they are not typically causal).

The corresponding 2-port *impedance matrix* for Fig. 3.4 is

$$\begin{bmatrix}\Phi_i \\ F_l\end{bmatrix} = \begin{bmatrix}z_{11}(s) & z_{12}(s) \\ z_{21}(s) & z_{22}(s)\end{bmatrix}\begin{bmatrix}I_i \\ U_l\end{bmatrix} = \begin{bmatrix}Z_e(s) & -T(s) \\ T(s) & z_m(s)\end{bmatrix}\begin{bmatrix}I_i \\ U_l\end{bmatrix}. \tag{3.29}$$

Such a description allows one to define *Thévenin parameters*, a very useful concept used widely in circuit analysis and other network models from other modalities.

The impedance matrix is an alternative description of the system, but with generalized forces $[\Phi_i, F_l]$ on the left and generalized flows $[I_i, U_l]$ on the right. A rearrangement of the equations allows one to go from the ABCD to impedance set of parameters (Van Valkenburg, 1964b). The electromagnetic transducer is anti-reciprocal (P6), $z_{12} = -z_{21} = T = B_0 l$.

Additional or modified postulates

The postulates must go beyond postulates P1-P6 defined by Carlin and Giordano (Section 1.3.12, p. 72), when there are interaction of waves and a structured medium, along with other properties not covered by classic network theory. Assuming QS, the wavelength must be large relative to the medium's lattice constants. Thus the QS property must be extended to three dimensions, and possibly to the cases of an-isotropic and random media.

Causality: P1   As stated above, due to causality the negative properties (e.g., negative refractive index) must be limited in bandwidth, as a result of the Cauchy-Riemann conditions. However even causality needs to be extended to include the delay, as quantified by the d'Alembert solution to the wave equation, which means that the delay is proportional to the distance. Thus we generalize P1 to include the space dependent delay. When we wish to discuss this property we denote it *Einstein causality*, which says that the delay must be proportional to the distance x, with impulse response $\delta(t - x/c)$.

Linearity: P2   The wave properties of may be non-linear. This is not restrictive as most physical systems are naturally nonlinear. For example, a capacitor is inherently nonlinear: as the charge builds up on the plates of the capacitor, a stress is applied to the intermediate dielectric due to the electrostatic force $F = qE$. In a similar manner, an inductor is nonlinear. Two wires carrying a current are attracted or repelled, due to the force created by the flux. The net force is the product of the two fluxes due to each current.

In summary, most physical systems are naturally nonlinear, it's simply a matter of degree. An important counter example is a amplifier with negative feedback, with very large open-loop gain. There are, therefore, many types of non-linear, instantaneous and those with memory (e.g., hysteresis). Given the nature of P1, even an instantaneous non-linearity may be ruled out. The linear model is so critical for our analysis, providing fundamental understanding, that we frequently take P1 and P2 for granted.

Passive/Active: P3 This postulate is about conservation of energy and Otto Brune's *positive Real* (PR aka physically realizable) condition, that every passive impedance must obey. Following up on the earlier work of his primary PhD thesis advisor Wilhelm Cauer (1900-1945) and Ernst Guillemin, along with Norbert Weiner and Vannevar Bush at MIT, Otto Brune mathematically characterized the properties of every PR 1-port driving point impedance (Brune, 1931a). 5

When the input resistance of the impedance is real, the system is said to be *passive*, which means the system obeys conservation of energy. The real part of $Z(s)$ is positive if and only if the corresponding reflectance is less than 1 in magnitude. The definition of the reflectance of $Z(s)$ is defined as a bilinear transformation of the *impedance*, normalized by its *surge resistance* r_0 (Campbell, 1903)

$$\Gamma(s) = \frac{Z(s) - r_0}{Z(s) + r_0}.$$

The surge resistance is defined in terms of the inverse Laplace transform of $Z(s) \leftrightarrow z(t)$, which must have the form

$$z(t) = r_0\delta(t) + \zeta(t),$$

where $\zeta(t) = 0$ for $t < 0$. It naturally follows that $\gamma(t) \leftrightarrow \Gamma(s)$ is zero for negative and zero time, namely $\gamma(0) = 0, t \leq 0$. at

Given any linear PR impedance $Z(s) = R(\sigma, \omega) + jX(\sigma, \omega)$, having real part $R(\sigma, \omega)$ and imaginary part $X(\sigma, \omega)$, the impedance is defined as being PR (Brune, 1931a) if and only if

$$R(\sigma \geq 0, \omega) \geq 0. \tag{3.30}$$

That is, the real part of any PR impedance is non-negative everywhere in the right half s plane ($\sigma \geq 0$). This is a very strong condition on the complex analytic function $Z(s)$ of a complex variable s. This condition is equivalent to any of the following statements (Van Valkenburg, 1964a): 10

1. There are no poles or zeros in the right half plane ($Z(s)$ may have poles and zeros on the $\sigma = 0$ axis).

2. If $Z(s)$ is PR then its reciprocal $Y(s) = 1/Z(s)$ is PR (the poles and zeros swap).

3. If the impedance may be written as the ratio of two polynomials (a limited case, related to the quasi-statics approximation, P9) having degrees N and L, then $|N - L| \leq 1$. 15

4. The angle of the impedance $\theta \equiv \angle Z$ lies between $[-\pi \leq \theta \leq \pi]$.

5. The impedance *and its reciprocal* are *complex analytic* in the right half plane, thus they each obey the Cauchy Riemann conditions there.

Energy and Power: The PR (positive real or Physically realizable) condition assures that every impedance is *positive-definite* (PD), thus guaranteeing conservation of energy is obeyed (Schwinger and Saxon, 1968, p.17). This means that the total energy absorbed by any PR impedance must remain positive for all time, namely

$$\mathcal{E}(t) = \int_{-\infty}^{t} v(t)i(t) \, dt = \int_{-\infty}^{t} i(t){\star}z(t) \, i(t) \, dt > 0,$$

where $i(t)$ is *any* current, $v(t) = z(t) \star i(t)$ is the corresponding voltage and $z(t)$ is the real causal impulse response of the impedance, e.g., $z(t) \leftrightarrow Z(s)$ are a Laplace Transform pair. In summary, if 20 $Z(s)$ is PR, $\mathcal{E}(t)$ is PD.

As discussed in detail by Van Valkenburg, any rational PR impedance can be represented as a *partial fraction expansion*, which can be expanded into first-order poles as

$$Z(s) = K\frac{\Pi_{i=1}^{L}(s - n_i)}{\Pi_{k=1}^{N}(s - d_k)} = \sum_n \frac{\rho_n}{s - s_n}e^{j(\theta_n - \theta_d)}, \tag{3.31}$$

where ρ_n is a complex scale factor (residue). Every pole in a PR function has only simple poles *and* zeros, requiring that $|L - N| \leq 1$ (Van Valkenburg, 1964b).

Whereas the PD property clearly follows P3 (conservation of energy), the physics is not so clear. Specifically what is the physical meaning of the specific constraints on $Z(s)$? In many ways, the impedance concept is highly artificial, as expressed by P1-P7.

When the impedance is not rational, special care must be taken. An example of this is the semi-inductor $M\sqrt{s}$ and semi-capacitor $K/\sqrt{s}$ due, for example, to the *skin effect* in EM theory and viscous and thermal losses in acoustics, both of which are frequency dependent boundary-layer diffusion losses (Vanderkooy, 1989). They remain positive-real but have a branch cut, thus are double valued in frequency.

Real time response: P4 The impulse response of every physical system is real, vs. complex. This requires that the Laplace Transform have conjugate-symmetric symmetry $H(s) = H^*(s^*)$, where the $*$ indicates conjugation (e.g., $R(\sigma, \omega) + X(\sigma, \omega) = R(\sigma, \omega) - X(\sigma, -\omega)$).

Time invariant: P5 The meaning of *time-invariant* requires that the impulse response of a system does not change over time. This requires that the system coefficients of the differential equation describing the system are constant (independent of time).

Rayleigh Reciprocity: P6 Reciprocity is defined in terms of the unloaded output voltage that results from an input current. Specifically

$$\begin{bmatrix} z_{11}(s) & z_{12}(s) \\ z_{21}(s) & z_{22}(s) \end{bmatrix} = \frac{1}{C(s)} \begin{bmatrix} A(s) & \Delta_T \\ 1 & D(s) \end{bmatrix}, \tag{3.32}$$

where $\Delta_T = A(s)D(s) - B(s)C(s) = \pm 1$ for the reciprocal and anti-reciprocal systems respectively. This is best understood in term of Eq. 3.29. The off-diagonal coefficients $z_{12}(s)$ and $z_{21}(s)$ are defined as

$$z_{12}(s) = \left.\frac{\Phi_i}{U_l}\right|_{I_i=0} \qquad\qquad z_{21}(s) = \left.\frac{F_l}{I_i}\right|_{U_l=0}.$$

When these off-diagonal elements are equal $[z_{12}(s) = z_{21}(s)]$ the system is said to obey *Rayleigh reciprocity*. If they are opposite in sign $[z_{12}(s) = -z_{21}(s)]$, the system is said to be *anti-reciprocal*. If a network has neither of the reciprocal or anti-reciprocal characteristics, then we denote it as *non-reciprocal* (McMillan, 1946). The most comprehensive discussion of reciprocity, even to this day, is that of Rayleigh (1896, Vol. I). The reciprocal case may be modeled as an ideal transformer (Van Valkenburg, 1964a) while for the anti-reciprocal case the generalized force and flow are swapped across the 2-port. An electromagnetic transducer (e.g., a moving coil loudspeaker or electrical motor) is anti-reciprocal (Kim and Allen, 2013; Beranek and Mellow, 2012), requiring a gyrator rather than a transformer, as shown in Fig. 3.4.

Reversibility: P7 A second 2-port property is the *reversible/non-reversible* postulate. A reversible system is invariant to the input and output impedances being swapped. This property is defined by the input and output impedances being equal.

Referring to Eq. 3.32, when the system is *reversible* $z_{11}(s) = z_{22}(s)$ or in terms of the transmission matrix variables $\frac{A(s)}{C(s)} = \frac{D(s)}{C(s)}$ or simply $A(s) = D(s)$ assuming $C(s) \neq 0$.

An example of a non-reversible system is a transformer where the turns ratio is not one. Also an ideal operational amplifier (when the power is turned on) is non-reversible due to the large impedance difference between the input and output. Furthermore it is *active* (it has a power gain, due to the current gain at constant voltage) (Van Valkenburg, 1964b).

Generalizations of this lead to group theory, and *Noether's theorem*. These generalizations apply to systems with many modes whereas quasi-statics holds when operate below a cutoff frequency (Table

3.1), meaning that like the case of the transmission line, there are no propagating transverse modes. While this assumption is never exact, it leads to highly accurate results because the non-propagating evanescent transverse modes are attenuated over a short distance, and thus, in practice, may be ignored (Montgomery et al., 1948; Schwinger and Saxon, 1968, Chap. 9-11).

We extend the Carlin and Giordano postulate set to include (P7) Reversibility, which was refined by Van Valkenburg (1964a). To satisfy the reversibility condition, the diagonal components in a system's impedance matrix must be equal. In other words, the input force and the flow are proportional to the output force and flow, respectively (i.e., $Z_e = z_m$).

Spatial invariant: P8 The characteristic impedance and wave number $\kappa(s, x)$ may be strongly frequency and/or spatially dependent, or even be negative over some limited frequency ranges. Due to *causality*, the concept of a negative group velocity must be restricted to a limited bandwidth (Brillouin, 1960). As is made clear by Einstein's theory of relativity, all materials must be strictly causal (P1), a view that must therefore apply to acoustics, but at a very different time scale. We first discuss generalized postulates, expanding on those of Carlin and Giordano.

The QS constraint: P9 When a system is described by the wave equation, delay is introduced between two points in space, which depends on the wave speed. When the wavelength is large compared to the delay, one may successfully apply the *quasi-static approximation*. This method has wide-spread application, and is frequency used without mention of the assumption. This can lead to confusion, since the limitations of the approximation may not be appreciated. An example is the use of QS in Quantum Mechanics. The QS approximation has wide spread use when the signals may be accurately approximated by a band-limited signal. Examples include KCL, KVL, wave guides, transmission lines, and most importantly, impedance. The QS property is not mentioned in the six postulates of Carlin and Giordano (1964), thus they need to be extended in some fundamental ways.

When the dimensions of a cellular structure in the material are much less than the wavelength, can the QS approximation be valid. This effect can be viewed as a *mode filter* that suppresses unwanted (or conversely enhances the desired) modes (Ramo et al., 1965). QSs may be applied to a 3 dimensional specification, as in a semiconductor lattice. But such applications fall outside the scope of this text (Schwinger and Saxon, 1968).

Although I have never seen the point discussed in the literature, the QS approximation is applied when defining Green's theorem. For example, Gauss's Law is not true when the volume of the container violates QS, since changes in the distribution of the charge have not reached the boundary, when doing the integral. Thus such integral relationships assume that the system is in quasi steady-state (i.e., that QS holds).

Table 3.1: *There are several ways of indicating the quasi-static (QS) approximation. For network theory there is only one lattice constant a, which must be much less than the wavelength (wavelength constraint). These three constraints are not equivalent when the object may be a larger structured medium, spanning many wavelengths, but with a cell structure size much less than the wavelength. For example, each cell could be a Helmholtz resonator, or an electromagnetic transducer (i.e., an earphone).*

Measure	Domain
$ka < 1$	Wavenumber constraint
$\lambda > 2\pi a$	Wavelength constraint
$f_c < c/2\pi a$	Bandwidth constraint

Formally, QS is defined as $ka < 1$ where $k = 2\pi/\lambda = \omega/c$ and a is the cellular dimension or the size of the object. Other ways of expressing this include $\lambda/4 > a$, $\lambda/2\pi > a$, $\lambda > 4a$ or $\lambda > 2\pi a$. It is not clear if it is better to normalize λ by 4 (quarter-wavelength constraint) or $2\pi \approx 6.28 > 4$, which is

more conservative by a factor of $\pi/2 \approx 1.6$. Also k and a can be vectors, e.g., Eq. 1.68, p. 79, Section 1.4.1.

Schelkunoff may have been the first to formalize this concept (Schelkunoff, 1943) (but not the first to use it, as exemplified by the Helmholtz resonator). George Ashley Campbell was the first to use the concept in the important application of a wave-filter, some 30 years before Schelkunoff (Campbell, 1903). These two men were 40 years apart, and both worked for the telephone company (after 1929, called AT&T Bell Labs) (Fagen, 1975).

There are alternative definitions of the QS approximation, depending on the geometrical cell structure. The alternatives are outlined in Table 3.1.

Summary

A transducer converts between modalities. We propose the general definition of the nine system postulates, that include all transduction modalities, such as electrical, mechanical, and acoustical. It is necessary to generalize the concept of the QS approximation (P9) to allow for guided waves.

Given the combination of the important QS approximation, along with these space-time, linearity, and reciprocity properties, a rigorous definition and characterization a system can thus be established. It is based on a taxonomy of such materials, formulated in terms of material and physical properties and in terms of extended network postulates.

3.5.2 Lec 22 Exam II (Evening)

Chapter 4

Ordinary Differential Equations: Stream 3a

4.0.1   Lec 32 Brune impedance: Working with experimental data

L32: Detailed examples of the Inverse LT:
1) Modulation, 2) Translation, 3) convolution, 4) periodic functions
Tables of common LTs

Estimating c_k from numerical data: Assuming the poles of $Z(s)$ have been accurately computed (e.g., assuming $D(s)$ has been numerically factored using `P=root(D)`), the partial fraction expansion defines a linear system of equations in the residues c_k, uniquely determined by $N(s)$. One problem that needs to be addressed is that we don't know $Z(s)$, rather we know $Z(\omega j)$. One could possibly work with

$$\hat{c}_k = \lim_{\omega j \to s_k} (s - s_k) Z(\omega j),$$

but it is not clear how well this would work, if at all. Poles far from the ωj could be a problem. Another possibility would be to work in the time-domain. A third possibility would be to use the impedance zeros (roots of $N(s)$, poles of the admittance $Y(s)$).

Generalized impedance: The Brune impedance is a restricted definition of impedance, more generally defined via the *reflectance* $\Gamma(s)$.

1. The bilinear (Möbius) transformation is useful for the *generalized impedance* (i.e., extending the Brune impedance)

$$Z(s) = r_o \frac{1 + \Gamma(s)}{1 - \Gamma(s)}, \qquad \Gamma(s) = \frac{1 - Z(s)/r_o}{1 + Z(s)/r_o}, \tag{4.1}$$

where $Z(s)$ is the generalized impedance, having *characteristic resistance* r_o, such that

$$z(t) = r_o \delta(t) + \zeta(t), \tag{4.2}$$

then the *reflectance function* $\Gamma(s)$ has the properties

$$\Gamma(s) = \frac{\mathcal{P}_-}{\mathcal{P}_+} \leftrightarrow \gamma(t \le 0) = 0 \quad \text{and} \quad |\Gamma(j\omega)| \le 1. \tag{4.3}$$

Here the most general form of impedance is defined by the reflectance function. This brings the definition close to the physics of an impedance. The reflectance is determined by the internal scattering of the forward going pressure wave $\mathcal{P}_+$, which is induced by variations in the area function $A(x)$ of the *Webster horn equation.* When the area function is constant (e.g., $A(x) = A_o$), $\Gamma(s) = 0$, and there are no internal reflections.

A second, and important way that internal reflections are generated, is by propagation losses in the medium. In the case of acoustics a thin layer at the surface of the horn introduces viscous and thermal losses, modifying the propagation function $\kappa(x, s)$, causing the propagation to be dispersive (i.e., frequency dependent). This will be discussed in greater detail in Chapter 4.

5   When $|\Gamma(\jmath\phi(\omega)| = 1$ ($\phi(\omega)$ is the system phase and $\tau(\omega) = -\partial\phi/\partial\omega$ is called the group delay), the system is loss-less and all the poles and zeros lie on the $\jmath\omega$ axis (i.e., Foster Theorem applies). When the poles and zeros lie off the $\jmath\omega$ axis, this reflectance ($|\Gamma| = 1$) is called *all–pass*, which means it is a strictly causal dispersive delay. This is the case of a loss-less dispersive system. Semi-conductors are in this class, as are super-conductors. These systems are not Brune impedances.

10   **Examples:**

Important examples of generalized impedance include the semi-inductor $Z(s) = l_o\sqrt{s}$ and semi-capacitor $Z(s) = 1/c\sqrt{s}$, which correspond to evanescent wave penetration into the magnetic core of a transformer, or the losses in the acoustic wave, due to viscosity and thermal effects in acoustics.
L 34: Exam III
15   Thanksgiving Holiday 11/19–11/27 2016

Chapter 5

Vector Calculus: Stream 3b

5.1 Stream 3b

5

5.1.1 **Lec 35** Scalar vs. Vector wave equations

Scalar vs. Vector wave equations

L 36 The Webster Horn equation; WKB method   10
   A real-world example of large delay, where the branch-cut placement is critical

L 37 Gradient, divergence, curl, scalar Laplacian **VC-1**

L 38 More on the divergence and curl: Stokes' and Gauss' Laws; Vector Laplacian

   – Thanksgiving Holiday 11/19–11/27 2016   15

5.2 Stream 3b

5.2.1 **Lec 36** The Webster Horn equation

The quasi-static approximation: There is an approximation that must play a subtle role, that of quasi-statics across the radius. In the typical application of the horn, the wavelength is much greater than the diameter of the horn. When this condition is violated higher order modes can play a significant   20
role. In fact there is a frequency where the radius will be equal to a quarter wavelength. When this happens, the quasi-static approximation fails, and cross-modes will become significant. I have not seen any discussion of this in the context of Eq. 1.106.

Role of admittance: Since every admittance is *causal*, it must be a function of Laplace frequency s, and thus has a Laplace transform (poles and zeros). We shall maintain the distinction that functions   25
of ω are Fourier transforms, not analytic in $j\omega$, while functions of s correspond to Laplace transforms, which *are* necessarily analytic in s, in the *right half plane* (RHP) *region of convergence* (ROC) (i.e., a

causal function). This distinction is critical, since we typically describe impedance $Z(s)$ and admittance $Y(s)$ in terms of their poles and zeros, as analytic functions.[1]

In this document we shall consider four different cases of $A(x)$, as summarized in Table 5.1. Given an area function $A(x)$, each horn has a distinct wave equation, and thus a distinct solution.

Table 5.1: *Table of horns and their properties discussed in this document. The horn axis is defined by x, the radius at x is $r(x)$, the area is $A(x) \equiv 2\pi r^2(x)$, $F(x)$ is the coefficient on $\mathcal{P}_x$ and $\kappa(s) \equiv s/c$, where c is the speed of sound. A dimensionless range variable may be defined as $x \equiv (\xi - \xi_0)/(L - \xi_0)$, with ξ the linear distance along the horn, where $x = \xi_0$ to L corresponding to $x = 0$ to 1. The horn's primitive solutions are $\mathcal{P}^\pm(x, s) \leftrightarrow \varrho^\pm(x, t)$. When $\pm$ is indicated, the outbound solution corresponds to the negative sign. The function $E(t)$ is discussed on page 191. The last column is the radiation admittance normalized by $A(x)/\rho c$, for a simplified expression.*

#D	Name	radius	Area/A_0	$F(x)$	$\mathcal{P}^\pm(x,s)$	$\varrho^\pm(x,t)$	$Y_{rad}^\pm/y_0$
1D	plane	1	1	0	$e^{\mp\kappa(s)x}$	$\delta(t \mp x/c)$	1
2D	parabolic	$\sqrt{x/x_0}$	x/x_0	$1/x$	$H_0^\pm(-j\kappa(s)x)$	—	$-jxH_1^\pm/H_0^\pm$
3D	conical	x	x^2	$2/x$	$e^{\mp\kappa(s)x}/x$	$\delta(t \mp x/c)/x$	$1 \pm c/sx$
EXP	exponential	e^{mx}	e^{2mx}	$2m$	$e^{-(m\pm\sqrt{m^2+\kappa^2})x}$	$e^{-mx}E(t)$	—

5 **Goals of the study:** A primary focus of this study is to identify and characterize the physical significance of the reactance (imaginary) component of the loss-less horn's radiation admittance $Y_{rad}(s)$ (and impedance $Z_{rad}(s)$). Of course for the case of the uniform horn (i.e., plane-waves), the input (driving-point) admittance $Y_{rad} = y_0$ (impedance) must be real and positive (zero imaginary part).[2] When $A(x)$ is not constant the radiation admittance is always complex (Salmon, 1946a,b; Olson, 1947;

10 Morse, 1948; Leach, 1996). Interestingly, in some special cases, Y_{rad} is purely reactive (the real part is zero), as for the exponential horn below the cut-off frequency (see Fig. 5.1). We will see that the Horn equation generalizes Strum-Louville theory to situations where due to the choice of $A(x)$, a separation of variables fails. This has interesting and wide-reaching consequences.

As shown by[3] Youla (1964); Sondhi and Gopinath (1971), there is a unique relationship between the

15 reactive part of the input "driving point" admittance Y_{rad} and the horn area function $A(x)$. Determining $A(x)$ from $Y_{rad}(0, s)$ is known as *the inverse problem*.

Alternate forms of the Webster Horn Equation

Equation 5.4 is equivalent to the traditional second-order Webster horn equation in the pressure. To see this take the partial derivative with respect to x (abbreviated ∂_x) of the Newton pressure equation, giving $\mathcal{P}_{xx} + \mathscr{L}_x\mathcal{V} + \mathscr{Z}(s)\mathcal{V}_x = 0$. Next use the Newton and Hooke equations once again, to remove the velocity, obtaining the *Webster wave equation* for the Laplace Transform pair $\varrho(x,t) \leftrightarrow \mathcal{P}(x,\omega)$ for pressure

$$\mathcal{P}_{xx} - (\mathscr{L}_x/\mathscr{L})\,\mathcal{P}_x = \frac{s^2}{c^2}\mathcal{P} \leftrightarrow \frac{1}{c^2}\varrho_{tt}. \tag{5.1}$$

This is the more traditional form of the Webster equation (Webster, 1919; Morse, 1948; Morse and Feshbach, 1953; Pierce, 1981).

---

[1]When an analytic function of complex variable s includes the pole it is called a *Laurent series* in s. For example, the impedance of a capacitor C is $Z_c(s) = 1/sC$, which is analytic in s everywhere other than $s = 0$. The capacitor has a voltage time response given by the integral of the current, i.e., $v(t) = \frac{1}{C}\int^t i(t)dt = \frac{1}{C}u(t) \star i(t)$, where $u(t)$ is the Heaviside step function and $\star$ represents convolution.

[2]Here we are only considering loss-less horns, i.e., those that satisfy Eq. 5.4.

[3]The publication of Youla (1964) seems easier to understand.

General solutions of the Horn Equation

In the mathematical literature the general solution of equation of this form are known as *Strum-Liouville* (SL) problems, the solution of which may be obtained by *integration by parts*, following the multiplication by an "integration factor" $\sigma(x)$:

$$\frac{1}{\sigma}\partial_x(\sigma\mathcal{P}_x) \equiv \mathcal{P}_{xx} + \partial_x \ln \sigma(x) \; \mathcal{P}_x(x) \tag{5.2}$$

By a direct comparison of the above to Eq. 5.1, we see that $\sigma_x/\sigma \equiv -\mathscr{Z}_x/\mathscr{Z}$. Thus $\partial_x \ln(\sigma)/ \equiv -\partial_x \ln(\mathscr{Z}) = \partial_x \ln(A)$, or

$$\sigma(x) \equiv A(x). \tag{5.3}$$

One may conclude the very interesting observation that *the integration factor is physically the area function*. Physically this makes sense. It explains, for example, why the integration factor in the Sturm-Liouville theory must be strictly positive. We are not aware of this observation in the existing literature. While it seems likely that Morse (1948) must have been aware of this connection, to our knowledge he did explicitly state so in writing.

 Following this approach we put Eq. 5.3 into Eq. 5.2 to obtain the Webster Horn equation (Morse, 1948, p. 269)

$$\frac{1}{A(x)}\frac{\partial}{\partial x}\left(A(x)\frac{\partial\varrho}{\partial x}\right) = \frac{1}{c^2}\frac{\partial^2\varrho}{\partial t^2}. \tag{5.4}$$

 In summary, if one starts with the 3D wave equation, and by the application of Gauss' law, transforms it into a "1 dimensional" (i.e., single range variable) equivalent horn equation, then a Sturm-Louville-like equation results. This may be integrated by the application of the area function as the integration factor.

Primitive solutions $\varrho^{\pm}(x,t)$

For each choice of area function used in Eq. 5.4 (or Eq. 1.106), there are two causal *primitive solutions* of the homogeneous (i.e., undriven) equation, identified as an *outbound* (right-traveling) and *inbound* (left-traveling) wave, denoted as the Laplace transform pair

$$\mathcal{P}^{\pm}(x,s) = \int_0^{\infty} \varrho^{\pm}(x,t)e^{-st}dt$$

(aka, $\varrho^{\pm}(x,t) \leftrightarrow \mathcal{P}^{\pm}(x,s)$), normalized such that $\mathcal{P}^{\pm}(x_0,s) = 1$, where x_0 defines the input excitation location, as show in Fig. 1.22. This construction is equivalent to the seperation of time and place variables.

Examples: The primitive solutions for the four horns are summarized in Table 5.1.
 For the *uniform horn* the two primitive solutions are the two plane waves

$$\delta(t \mp x/c) \leftrightarrow P^{\pm}(x,\kappa) = e^{\mp\kappa(s)x}.$$

These two directed wave solutions are functions of Laplace frequency s, since they must be causal. They may be viewed as the impulse response of a semi-infinite section of horn, namely the causal solutions of Eq. 5.4, driven at their input by an impulse at $t = 0$. It is a convention that these primitive solutions are normalized to 1 at the input ($x = 0$ and $x = L$ in this example).
 For the *spherical geometry* the two waves are

$$\mathcal{P}^{\pm} = e^{\mp\kappa r}/r$$

where the $+$ is the "outbound" wave and the $-$ is the inbound wave. As we shall see, when the area is variable, each of the primitive solutions has local reflections due to this variation. This gives rise to a reactive mass-component in the radiation impedance.

Figure 5.1: *Throat acoustical resistance r_A and acoustical reactance x_A, frequency characteristics of infinite parabolic, conical, exponential, hyperbolic and cylindrical horns having a throat area of 1 square centimeter. Note how the "critical" frequency (defined here as the frequency where the reactive and real parts of the radiation impedance are equal) of the horn reduces dramatically with the type of horn. For the uniform horn, the reactive component is zero, so there is no cutoff frequency. For the parabolic horn (1), the cutoff is around 3 kHz. For the conical horn (2), the cutoff is at 0.6 [kHz]. For the exponential horn the critical frequency is around 0.18 [kHz], which is 16 times smaller than for the parabolic horn. For each horn the cross-sectional area of the parabolic, conical exponential and hyperbolic horns is defined as 100 square centimeters at a distance of 100 centimeters from the throat (Olson, 1947, p. 101).*
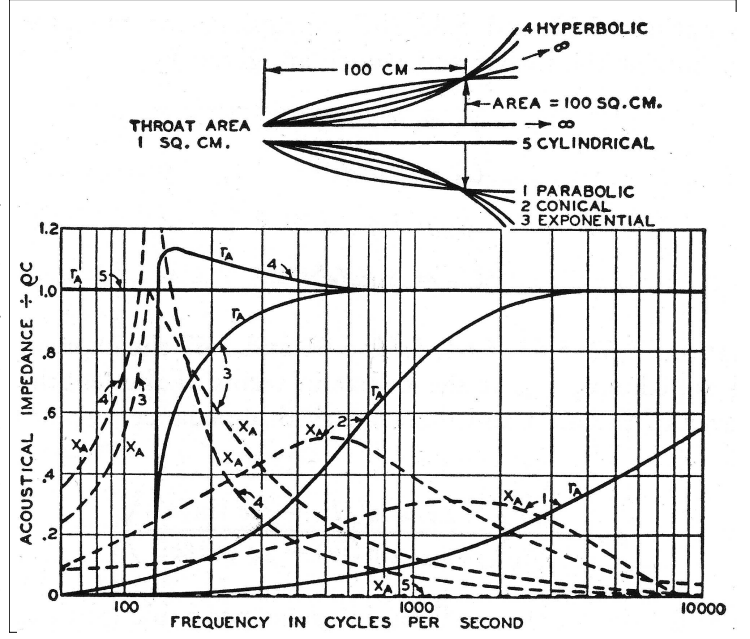
The case of the *exponential horn*

$$\mathcal{P}^{\pm}(x) = e^{-mx}e^{\mp j\sqrt{\omega^2-\omega_c^2}\,x/c}, \tag{5.5}$$

which is of special interest because the radiation impedance is purely reactive below the horn's cutoff frequency ($\omega < \omega_c = mc$) where the reflection coefficient has magnitude 1 (no energy can radiate from an open horn below this frequency.

Def $y_o = 1/r_o$

Characteristic Admittance y_0**:** A second key definition is the wave *characteristic admittance $y_0(x)$* defined as the square root of the *ratio* of $\mathcal{Z}$ and $\mathcal{Y}$

$$y_0(x) \equiv \sqrt{\frac{\mathcal{Y}(x,s)}{\mathcal{Z}(x,s)}} = \frac{\kappa}{\mathcal{Z}} = \frac{A(x)}{\rho_o c}, \tag{5.6}$$

which depends specifically on $A(x)$, but not on frequency s. Based on physical requirements that the admittance must be positive, only the positive square root is allowed.

Since Eq. 5.4 is loss less, $y_0(x)$ must be real (and positive). When losses are introduced, $y_0(s,x)$ and $\kappa(s)$ must be complex function of the Laplace frequency s. While such cases are interesting, and realistic, they are outside of the traditional loss-less Webster horn equation formulation.

Radiation Admittance: The *radiation admittance* at x is defined as the admittance looking into a semi-infinite horn (Fig. 5.1)

$$Y_{rad}^{\pm}(x,s) \equiv \pm\frac{\mathcal{V}^{\pm}(x,\omega)}{\mathcal{P}^{\pm}(x,\omega)} = \mp\frac{1}{\mathcal{Z}(x,s)}\frac{\mathcal{P}_x^{\pm}}{\mathcal{P}^{\pm}}\bigg|_x = \mp\frac{y_0(x)}{\kappa(s)}\frac{\partial\ln\mathcal{P}^{\pm}}{\partial x}\bigg|_x. \tag{5.7}$$

In general, $Y_{rad}^{\pm}$ depends on the direction of the velocity, but the real part of the radiation impedance must always be positive. Care must be taken when fixing the signs to obey these conventions. It is helpful to always define the volume velocity $\mathcal{V}^{\pm}(x,s)$ into the port.

Typically the velocity and pressure are functions of frequency ω, not complex frequency s, since they need not be causal functions. However the ratio of the two, defines an admittance, which necessarily is causal, and therefore is necessarily a function of s. Since the primitive solutions must be causal and stable functions, they must be analytic functions of s for $\sigma > 0$.

Surge admittance: Every radiation admittance may be written as $y_{rad}(t) \leftrightarrow Y_{rad}(s)$ which may be further split into a real *surge admittance*[4] $y_o\delta(t)$ (Campbell, 1922), and a causal *remainder admittance* $y_r(t)$, as[5]

$$y_{rad}(t) = y_o\delta(t) + y_r(t).$$

Alternatively this may also be written as the sum of an impedance surge and remainder components $z_{rad}(t) = r_o\delta(t) + z_r(t)$. These functions are inverses of each other in the convolution sense, namely $y_{rad}(t) \star z_{rad}(t) = \delta(t)$, which follows from $Z_{rad}(s)Y_{rad}(s) = 1$. Any function having a causal inverse is said to be *minimum phase* thus every impedance and admittance must be minimum phase. The remainder characterizes the dispersive component of the impedance, thus when it is zero, the impedance is purely real (e.g., the reactance is zero).

Wave admittance $Y(x, s)$: The driving-point *wave admittance*, defined as the ratio of the volume velocity $\mathcal{V}(x, s)$ to the average pressure $\mathcal{P}(x, s)$ at any point x along the range axis, may be interpreted as follows: If the horn were split at any point x (i.e., a node is defined), the pressure at the two throats are the same ($\mathcal{P}^+ = \mathcal{P}^-$). The velocities are defined as into the port, thus $\mathcal{V}^+ = -\mathcal{V}^-$. Due to this definition of the flow into the port, the total velocity is the difference of these two driving point node velocites ($\mathcal{V} = \mathcal{V}^+ - \mathcal{V}^-$). Consistent with Kirchhoff's laws, the wave admittance as the sum of the two radiation admittances

$$Y(x, s) \equiv Y_{rad}^+(x, s) + Y_{rad}^-(x, s).$$

ABCD Transmission matrix: The transmission matrix is useful for computing the cascade of several system, such as a horn driven by a Thévenin system and loaded by the radiation impedance or a cascade of several horns. The solution of a horn having *finite length* may be expressed in terms of a 2-port ABCD matrix, that relates the pressure and volume velocity at the input and output ports (the two ends) of the horn ($x = 0$ and $x = L$)

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = \begin{bmatrix} \mathcal{A}(s) & \mathcal{B}(s) \\ \mathcal{C}(s) & \mathcal{D}(s) \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}. \tag{5.8}$$

Note that $\mathcal{A}(s) \equiv \left.\frac{\mathcal{P}_0}{\mathcal{P}_L}\right|_{\mathcal{V}_L=0}$ must not be confused with the horn area $A(x)$ (note the different font).

By definition, the output velocity $\mathcal{V}_L$, of an ABCD matrix is *out of* the port, hence the negative sign, since $\mathcal{V}_0, \mathcal{V}_L$ are defined into their respective ports (Orfanidis, 2009). When the system is reversible, $\mathcal{A} = \mathcal{D}$, *reciprocal* when $\Delta_T \equiv \mathcal{A}\mathcal{D} - \mathcal{B}\mathcal{C} = 1$, and *anti-reciprocal* when $\Delta_T = -1$. With the trivial exception of the uniform horn, all horns are non-reversable and reciprocal.

Boundary Conditions: The pressure and velocity at any point x be written in terms of a super-position of the *two homogeneous solutions* $\mathcal{P}^+(x, s)$ and $\mathcal{P}^-(x, s)$ (aka, *primative solutions*) of Eq. 5.4 since by the d'Alembert principle $\mathcal{P} = \mathcal{P}^+ + \mathcal{P}^-$ and $\mathcal{V} = \mathcal{V}^+ - \mathcal{V}^-$. The formula for the impedance at any point x is given in terms of the forward and retrograde waves

$$Z_{rad}(x) \equiv \frac{\mathcal{P}(x, \omega)}{\mathcal{V}(x, \omega)} = \frac{\mathcal{P}_+}{\mathcal{V}_+}\left(\frac{1 + \mathcal{P}^-/\mathcal{P}^+}{1 - \mathcal{V}^-/\mathcal{V}^+}\right) = r_o\frac{1 + \Gamma(s, x)}{1 - \Gamma(x, s)}, \tag{5.9}$$

Here the *reflectance* $\Gamma(x, s)$ is defined as the ratio of the forward and retrograde traveling waves

$$\Gamma(x, s) \equiv \frac{\mathcal{P}^-(x, \omega)}{\mathcal{P}^+(x, \omega)} = \frac{\mathcal{V}^-(x, \omega)}{\mathcal{V}^+(x, \omega)}. \tag{5.10}$$

---

[4]Since it is real it would best be called a *surge conductance*.
[5]It seem obvious that $y_o \equiv y_0$?

At any point x there are two reflectances, for the forward waves and retrograde waves. The radiation admittance $Y_{rad}(x, s)$ is the sum of a forward $Y_{rad}^+(x, s)$ and retrograd $Y_{rad}^-(x, s)$ radiation admittances (see above).

By a rearrangement of the above terms we may recover the definition of the characteristic impedance

$$r_o(x) = \frac{\mathcal{P}^+(x, \omega)}{\mathcal{V}^+(x, \omega)} = \frac{\mathcal{P}^-(x, \omega)}{\mathcal{V}^-(x, \omega)}, \tag{5.11}$$

where $r_o(x) = 1/y_o(x) = \rho_o c/A(x)$ (Eq. 1.111)

5   One must carefully in the definition the area $A(x)$. This area is *not* the cross-sectional area of the horn, rather it is the wave-front area, as discussed in Appendix H.[6]

In matrix notation this superposition may be written as

$$\begin{bmatrix} \mathcal{P}(x) \\ \mathcal{V}(x) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ y_0(x) & -y_0(x) \end{bmatrix} \begin{bmatrix} \mathcal{P}^+(x) \\ \mathcal{P}^-(x) \end{bmatrix}. \tag{5.12}$$

Complex-analytic nature of $\Gamma(s)$ and $Z_{rad}(s)$:   A *very* important assumption has been made here when expressing the complex reflectance $\Gamma(s)$ as a function of the complex frequency $s = \sigma + j\omega$, even though it is defined by the ratio of two functions of real (radian) frequency ω. This *assumption* is based on the fact that, like the impedance, the reflectance must be *causal*. Namely $\gamma(t) \leftrightarrow \Gamma(s)$ is zero for $t < 0$. The same may be assumed of the time-domain impedance $\zeta(t) \leftrightarrow Z_{rad}(s)$. That $\gamma(t)$ and $\zeta(t)$ are causal is require by the physics. It follows that both $\Gamma(s)$ and $Z_{rad}(x)$ are *complex analytic* functions of s, which means they must have a Taylor series expansion in s everywhere in the right-half s plane ($\sigma > 0$). This follow from the *Cauchy's integral theorem* (aka *The fundamental theorem of complex calculus.*), which is a special case of the *Cauchy's integral formula* (aka, *The Residue theorem.*).

The forward and retrograde waves may be any function of frequency ω, since they depend on the source pressure (or velocity) at the input to the horn. The reflectance is a transfer function (thus the source term cancels), that only depends on the impedance (or reflectance) looking into the system (at any position x).

As an alternative way to derive $\Gamma(s, x)$, invert Eq. 5.12

$$\begin{bmatrix} \mathcal{P}^+(x) \\ \mathcal{P}^-(x) \end{bmatrix} = \frac{1}{2y_0(x)} \begin{bmatrix} y_0(x) & 1 \\ y_0(x) & -1 \end{bmatrix} \begin{bmatrix} \mathcal{P} \\ \mathcal{V} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & \mathcal{Z}(x) \\ 1 & -\mathcal{Z}(x) \end{bmatrix} \begin{bmatrix} \mathcal{P} \\ \mathcal{V} \end{bmatrix} \tag{5.13}$$

and form the ratio of reflected to incident waves

$$\Gamma(s) = \frac{\mathcal{P}^-}{\mathcal{P}^+} = \frac{\mathcal{P} - \mathcal{Z}\mathcal{V}}{\mathcal{P} + \mathcal{Z}\mathcal{V}} = \frac{Z_{rad} - \mathcal{Z}}{Z_{rad} + \mathcal{Z}}. \tag{5.14}$$

20   It is convenient to define the *normalized radiation impedance $Z_{rad}/\mathcal{Z}$* when working with this form of $\Gamma(s)$.

Given some experience with these two forms, $Z_{rad} \equiv 1/Y_{rad}$ and $\Gamma(s)$, one may quickly begin to appreciate the advantage of working with the refelctance over the radiation impedance/admittance (aka immittance). The impedance has so many different forms, each of which are complicated, whereas
25   the reflectance is easily understood, as it is closer to the physics.

Finite length Horns: For a horn of fixed length L these expressions may be rewritten in terms of renormalized primative waves. If we define the foward wave $\mathcal{P}^+(x)$ as launched from $x = 0$ and the

---

[6]I need to prove that the area is the same for the forward and backward waves. To the best of my knowlege, this has not been demonstrated. Note they both must satisfy Laplaces equation, thus making the two equal. But if the volume velocities are not equal for a given range r, conservation of energy is violated. It may be a calculus of variations problem, or better, related to the conformal map. Some work needs to be done to quantify this point.

retrograde wave $\mathcal{P}^-(x)$ as launched from $x = L$, we may also write the pressure and velocity in terms of the primitives.

$$\begin{bmatrix} \mathcal{P}(x) \\ \mathcal{V}(x) \end{bmatrix} = \begin{bmatrix} \mathcal{P}^+(x) & \mathcal{P}^-(x - L) \\ Y_{rad}^+ \mathcal{P}^+(x) & -Y_{rad}^- \mathcal{P}^-(x - L) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{5.15}$$

The *radiation admittance* $Y_{rad}^\pm(x, s)$ depends on the range x and Laplace frequency s. Coefficients $\alpha(\omega)$ and $\beta(\omega)$, which depend on frequency ω (but not x), are determined by the boundary condition at $x = L$. To find the four parameters $[\mathcal{A}(s), \mathcal{B}(s), \mathcal{C}(s), \mathcal{D}(s)]$ we evaluate the inverse of Eq. 5.15 at $x = L$, substitute this result into 5.15, and then evaluated the matrix product at $x = 0$.

For compactness we adopt the following simplified subscript notation: $\mathcal{P}_0^\pm \equiv \mathcal{P}^\pm(x = 0, s)$, $\mathcal{P}_L^\pm \equiv \mathcal{P}^\pm(x = L, s)$, i.e., $\mathcal{P}_L \equiv \mathcal{P}(x = L)$ and $\mathcal{V}_L \equiv \mathcal{V}(x = L)$. The normalization of the primitive solutions are taken as $\mathcal{P}_0^+ = 1$ and $\mathcal{P}_L^- = 1$.

To compute Δ_L we start with

$$-\Delta_T(x, s) = \mathcal{P}^+(x)\, Y_{rad}^-(x)\mathcal{P}^-(x - L) \,+\, \mathcal{P}_-(x - L)\, Y_{rad}^+(x)\mathcal{P}^-(x)$$

evaluated at $x = L$ (recall $Y_{rad} \equiv Y_{rad}^\pm(x = L)$, $P_0^+ = P_L^- = 1$) , giving

$$-\Delta_L = Y_{rad} \left[\mathcal{P}_L^+ \, \mathcal{P}_0^- + \mathcal{P}_0^- \, \mathcal{P}_L^- \right].$$

Thus

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \frac{-1}{\Delta_L} \begin{bmatrix} Y_{rad}^- \mathcal{P}^-(x - L) & \mathcal{P}^-(x - L) \\ Y_{rad}^+ \, \mathcal{P}^+(x) & -\mathcal{P}^+(x) \end{bmatrix}_L \begin{bmatrix} \mathcal{P}_L \\ \mathcal{V}_L \end{bmatrix}, \tag{5.16}$$

which when simplified is

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \frac{-1}{\Delta_L} \begin{bmatrix} Y_{rad}^-(L) & -1 \\ Y_{rad}^+(L)\, \mathcal{P}_L^+ & \mathcal{P}_L^+ \end{bmatrix}_L \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}. \tag{5.17}$$

This detailed set of computations results in the following:

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = \frac{-1}{\Delta_L} \begin{bmatrix} 1 & \mathcal{P}_{-L}^- \\ Y_{rad}^+(r_0) & -Y_{rad}^-(r_0)\mathcal{P}_{-L}^- \end{bmatrix}_0 \begin{bmatrix} Y_{rad}^-(r_L) & -1 \\ Y_{rad}^+(r_L)\, \mathcal{P}_L^+ & \mathcal{P}_L^+ \end{bmatrix}_L \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}. \tag{5.18}$$

The subscript to the right of each matrix indicates it is evaluated at $x = 0$ or $x = L$. Here $\mathcal{P}_{-L}^-$ is $\mathcal{P}^-(x - L)$ at $x = 0$. The sign of $\mathcal{V}_L$ must be negative to satisfy the definition of every ABCD matrix, that the output velocity (i.e., $-\mathcal{V}_L$) is out of the port.

The relationship of β/α has special significance because it specifies the ratio of the reflected wave amplitude $\beta(\omega)$ in terms of the incident wave amplitude $\alpha(\omega)$. This ratio is know as the *reflectance*

$$\gamma_L(t) \leftrightarrow \Gamma_L(s) \equiv \frac{\beta}{\alpha}. \tag{5.19}$$

It has a critical role in the theory of horns, as we shall see as it is determined by the relative rate of change of the impedance (i.e., area) with range (i.e., $d\ln(\mathscr{Z})/dx$).

Impedance Matrix: For a finite section of horn, the 2×2 impedance matrix (a generalization of Ohm's Law) may expressed in terms of the ABCD matrix elements (Van Valkenburg, 1964b) as

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{P}_L \end{bmatrix} = \frac{1}{\mathcal{C}(s)} \begin{bmatrix} \mathcal{A}(s) & \Delta_T \\ 1 & \mathcal{D}(s) \end{bmatrix} \begin{bmatrix} \mathcal{V}_0 \\ \mathcal{V}_L \end{bmatrix}. \tag{5.20}$$

Note that $\Delta_T = 1$ since the horn must be *reciprocal* (Morse and Feshbach, 1953; Hunt, 1982; Pierce, 1981).

While the Transmission (ABCD) matrix is convenient when modeling, the impedance matrix (and its inverse, the *admittance matrix*) are useful when one makes experimental measurements. For example

$$Y_0|_{\mathcal{V}_L=0} \equiv \frac{\mathcal{C}(s)}{\mathcal{A}(s)} \qquad \text{and} \qquad Y_L|_{\mathcal{V}_0=0} \equiv \frac{\mathcal{C}(s)}{\mathcal{D}(s)}$$

are the *unloaded input admittances* of the horn looking into the two ports (Eq. 5.7). These admittances are typically easily measured experimentally, given access to the endpoints.

In section **??** we work out these relationships for the trivial case of the 1D horn (Goldsmith and Minton, 1924; Olson, 1947).

5.2.2   Lec 37 Partial differential equations of physics

Scalar wave equations and the Webster Horn equation; WKB method
Example of a large delay, where a branch-cut placement is critical (i.e., phase unwrapping)
    L 37: Partial differential equations of Physics
Scalar wave equation and its solution in 1 and 3 Dimensions

5.2.3   Lec 38 Gradient, divergence and curl vector operators

L 38: Vector dot and cross products $A \cdot B, A \times B$
Gradient, divergence and curl vector operators

WEEK 14   37.14.0

L 39 Gradient, divergence and curl: Gauss's (divergence) and Stokes's (curl) theorems

L 40 J.C. Maxwell unifies Electricity and Magnetism with the formula for the speed of light
    Basic definitions of E, H, B, D
    O. Heaviside's (1884) vector form of Maxwell's EM equations and the *vector wave equation*
    How a loud-speaker works

L 41 *The Fundamental Thm of vector calculus*
    Incompressable and *Irrotational* fluids and the two defining vector identities

5.3   Thanksgiving Holiday 11/19–11/27 2016

Thanksgiving Vacation: 1 week of rest

5.4   Vector Calculus

5.4.1   Lec 39 Geometry of Gradient, divergence and curl vector operators

Geometry of Gradient, divergence and curl vector operators

Lec 39: Review of vector field calculus 39.14.2

5.4.2 Lec 40 Introduction to Maxwell's Equation

L 40: J.C. Maxwell unifies Electricity and Magnetism with the formula for the speed of light
Basic definitions of E, H, B, D
O. Heaviside's (1884) vector form of Maxwell's EM equations and the *vector wave equation*
How a loud-speaker works.

5.4.3 Lec 41 The Fundamental theorem of Vector Calculus

L 41: *The Fundamental Thm of vector calculus*
Incompressible and *Irrotational* fluids and the two defining vector identities

WEEK 15 40.15.0 5

L 42 Quasi-static approximation and applications:
The Kirchoff's Laws and the *Telegraph wave equation*, starting from Maxwell's equations The
telegraph wave equation starting from Maxwell's equations
Quantum Mechanics 10

L 43 Last day of class: Review of Fund Thms of Mathematics:
Closure on Numbers, Algebra, Differential Equations and Vector Calculus,
The Fundamental Thms of Mathematics & their applications:
Theorems of Mathematics; Fundamental Thms of Mathematics (Ch. 9); Normal modes vs. eigen-
states, delay and quasi-statics; 15

– Reading Day
VC-1 Due

5.5 Kirchhoff's Laws

5.5.1 Lec 42: The Quasi-static approximation and applications

L 42: The Kirchhoff's Laws and the *Telegraph wave equation*, starting from Maxwell's equations 20
Quantum Mechanics

The quasi-static approximation failure at high frequencies: At high frequencies the quasi-
static approximation must break down. Thus at higher frequencies we need to consider other significant
physics of the system, known as *higher order modes*. A further complication is that at higher frequencies,
damping becomes an issue. 25

In acoustics viscosity and thermal effects are typically ignored, by assuming that wave propagation
is irrotationa, thus is described by the scalar wave equation. It turns out that these two loss mechanisms
are related (Rayleigh, 1896). But to understand why is complex. Helmholtz, with help from Kirchhoff,
explained this interaction, and independently published them between 1863 (Helmholtz, 1863b) and
1868 (Kirchhoff, 1868b). Their collective theory is summarized by Lord Rayleigh (Rayleigh, 1896),
and experimentally verified by Warren P. Mason (Mason, 1928). The nature of the correction is that
the wave number is extended to be of form

$$\kappa(s) = \frac{s + \beta_0\sqrt{s}}{c_0}, \tag{5.21}$$

where the forwarded P_- and backward P_+ pressure waves propagate as

$$P_\pm(s,x) = e^{-\kappa(s)x}, e^{-\overline{\kappa}(s)x} \tag{5.22}$$

with $\overline{\kappa}(s)$ the complex conjugate of $\kappa(s)$, and $\Re\kappa(s) > 0$. The term $\beta_0\sqrt{s}$ effects both the real and imaginary parts of $\kappa(s)$. The real part is a frequency dependent loss and the imaginary part introduces a frequency dependent velocity (Mason, 1928).

The frequency where the loss-less part equals the lossy part is an important parameter of the system. This frequency is $s_0 + \beta_0\sqrt{s_0} = 0$, or $\sqrt{s_0} = \beta_0$ or $f_0 = \beta^2/2\pi$.

Assuming air at 23.5° [C], $c_0 = \sqrt{\eta_0 P_0/\rho_0} \approx 344$ [m/s] is the speed of sound, $\eta_0 = c_p/c_v = 1.4$ is the ratio of specific heats, $\mu_0 = 18.5 \times 10^{-6}$ [Pa-s] is the viscosity, $\rho_0 \approx 1.2$ [kgm/m^2] is the density, $P_0 = 10^5$ [Pa] (1 atm).

The constant $\beta_0 = P\eta'/2S\sqrt{\rho_0}$

$$\eta' = \sqrt{\mu_0}\left[1 + \sqrt{5/2}\left(\eta_0^{1/2} - \eta_0^{-1/2}\right)\right]$$

is a thermodynamic constant, P is the perimeter of the tube and S the area (Mason, 1928).

For a cylindrical tube having radius $R = 2S/P$, $\beta_0 = \eta_0'/R\sqrt{\rho}$. To get a feeling for the magnitude of β_0 consider a 7.5 [mm] tube (i.e., the average diameter of the adult ear canal). Then $\eta' = 6.6180 \times 10^{-3}$ and $\beta_0 = 1.6110$. Using these conditions the wave-number cutoff frequency is $1.611^2/2\pi = 0.4131$ [Hz]. At 1 kHz the ratio of the loss over the propagation is $\beta_0/\sqrt{|s|} = 1.6011/\sqrt{2\pi 10^3} \approx 2\%$. At 100 [Hz] this is a 6.4% effect.[7]

Mason shows that the wave speed drops from 344 [m/s] at 2.6 [kHz] to 339 [m/s] at 0.4 [kHz], which is a 1.5% reduction in the wave speed. In terms of the losses, this is much larger effect.   At 1 [kHz] the loss is 1 [dB/m] for a 7.5 [mm] tube. Note that the loss and the speed of sound vary inversely with the radius. As the radius approaches the boundary layer thickness (the radial distance such that the loss is e^{-1}), the effect of loss dominates.

In Section 5.5.1 we shall look at some simple problems where we use the quasi-static effect and derive the Kirchhoff voltage and current equations, starting from Maxwell's equations.

5.5.2 Lec 43: Last day of class: Review of Fund Thms of Mathematics

L 43: Closure on Numbers, Algebra, Differential Equations and Vector Calculus,
The Fundamental Thms of Mathematics & their applications:
Theorems of Mathematics; Fundamental Thms of Mathematics (Ch. 9)
Normal modes vs. eigen-states, delay and quasi-statics;
Reading Day

Properties

Closure: Properties of fields of Maxwell's Equations   **39.14.6**

The variables have the following names and defining equations:

Symbol	Equation	Name	Units
E	$\nabla \times \mathbf{E} = -\dot{\mathbf{B}}$	Electric Field strength	[Volts/m]
D	$\nabla \cdot \mathbf{D} = \rho$	Electric Displacement (flux density)	[Col/m^2]
H	$\nabla \times \mathbf{H} = \dot{\mathbf{D}}$	Magnetic Field strength	[Amps/m]
B	$\nabla \cdot \mathbf{B} = 0$	Magnetic Induction (flux density)	[Weber/m^2]

In vacuo $\boldsymbol{B} = \mu_0\boldsymbol{H}$, $\boldsymbol{D} = \epsilon_0\boldsymbol{E}$, $c = \frac{1}{\sqrt{\mu_0\epsilon_0}}$ [m/s], $r_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} = 377$ [Ω].

---

[7]`/home/jba/Mimosa/2C-FindLengths.16/doc.2-c_calib.14/m/MasonKappa.m`

Vector field properties

Closure: Summary of vector field properties

- Notation:
$$\boldsymbol{v}(x,y,z) = -\nabla\phi(x,y,z) + \boldsymbol{\nabla}\times\boldsymbol{w}(x,y,z)$$

- Vector identities:
$$\boldsymbol{\nabla}\times\nabla\phi = 0; \ \nabla\cdot\boldsymbol{\nabla}\times\mathbf{w} = 0$$

Field type	Generator:	Test (on $\boldsymbol{v}$):
Irrotational	$\boldsymbol{v} = \nabla\phi$	$\nabla\times\boldsymbol{v} = 0$
Rotational	$\boldsymbol{v} = \boldsymbol{\nabla}\times\boldsymbol{w}$	$\nabla\times\boldsymbol{v} = \boldsymbol{J}$
Incompressible	$\boldsymbol{v} = \nabla\times\boldsymbol{w}$	$\nabla\cdot\boldsymbol{v} = 0$
Compressible	$\boldsymbol{v} = \nabla\phi$	$\nabla\cdot\boldsymbol{v} = \rho$

- Source density terms: Current: $\boldsymbol{J}(x,y,z)$, Charge: $\rho(x,y,z)$
  - Examples: $\boldsymbol{\nabla}\times\boldsymbol{H} = \dot{\boldsymbol{D}}(x,y,z)$, $\boldsymbol{\nabla}\cdot D = \rho(x,y,z)$

Fundamental Theorem of integral Calculus

Closure: Fundamental Theorems of integral calculus

1. $f(x) \in \mathbb{R}$ (Leibniz Integral Rule): $F(x) = F(a) + \int_a^x f(x)dx$

2. $f(s) \in \mathbb{C}$ (Cauchy's formula): $F(s) = F(a) + \int_a^s f(\zeta)d\zeta$
   –When integral is independent of path, $F(s) \in \mathbb{C}$ obeys CR conditions
   –Contour integration inverts causal Laplace transforms

3. $\boldsymbol{F} \in \mathbb{R}^3$ (Helmholtz Formula): $\boldsymbol{F}(x,y,z) = -\nabla\phi(x,y,z) + \boldsymbol{\nabla}\times\boldsymbol{A}(x,y,z)$
   –Decompose $\boldsymbol{F}(x,y,z)$ as *compressible* and *rotational*

4. Gauss' Law (Divergence Theorem): $Q_{enc} = \iiint \boldsymbol{\nabla}\cdot\boldsymbol{D}\,dV = \iint_{\mathcal{S}} \boldsymbol{D}\cdot\hat{\boldsymbol{n}}\,dA$
   –Surface integral describes enclosed compressible sources

5. Stokes' Law (Curl Theorem): $\mathcal{I}_{enc} = \iint (\boldsymbol{\nabla}\times\mathbf{H})\cdot\hat{\boldsymbol{n}}\,dA = \oint_{\mathcal{B}} \mathbf{H}\cdot d\boldsymbol{l}$
   –Boundary vector line integral describes enclosed rotational sources

6. Green's Theorem ... Two-port boundary conditions
   –*Reciprocity* property (*Theory of Sound*, Rayleigh, J.W.S., 1896)

Closure: Quasi-static (QS) approximation   **39.14.9**

- Definition: $ka \ll 1$ where a is the size of object, $\lambda = c/f$ wavelength

- This is equivalent to $a \ll \lambda$ or

- $\omega \ll c/a$ which is a low-frequency approximation

- The QS approximation is widely used, but infrequently identified.

- All *lumped parameter models* (inductors, capacitors) are based on QS approximation as the lead term in a Taylor series approximation.

Appendix A

Notation

A.1 Number systems

The notation used in this book is defined in this appendix so that it may be quickly accessed.[1] Where the definition is sketchy, Page numbers are provided where these concepts are fully explained, along with many other important and useful definitions. For example $\mathbb{N}$ may be found on page 22.

A.1.1 Units

Strangely, or not, math does not contain the concept of units. This makes it distinct from physics, where almost everything has a unit attached. Presumably this makes mathematics more general (i.e., abstract). But for the engineering mind, this is not good, as it necessarily means that the physical meaning has been removed, by design. We shall stick to SI units when ever possible. Spatial coordinates are quoted in meters [m], and time in seconds [s]. Angles in degrees have no units, whereas radians have units of inverse-seconds $[\text{s}^{-1}]$.

A.1.2 Symbols and functions

We use ln as the log function base e, log as base 2, and π_k to indicate the kth prime (e.g., $\pi_1 = 2, \pi_2 = 3$).

When working with Fourier $\mathcal{F}$ and Laplace $\mathcal{L}$ transforms, lower case symbols are in the time domain while upper case indicates the frequency domain, as $f(t) \leftrightarrow F(\omega)$. An important exception are Maxwell's equations, because they are so widely used as upper case bold letters (e.g., $\boldsymbol{E}(\boldsymbol{x}, \omega)$). It seems logical to change this to conform to lower case, with $\boldsymbol{e}(\boldsymbol{x}, t) \leftrightarrow \boldsymbol{E}(\boldsymbol{x}, \omega)$ as the preferred notation.

A.1.3 Greek letters

The Greek letters used in this text include (at least) $\alpha, \beta, \gamma, \delta, \epsilon, \kappa, \rho, \xi, \omega, \sigma, \phi, \psi, \zeta$, and upper-case $\Gamma, \Xi, \Phi, \Psi, \Delta, \Omega$. Many of these are commonly associated with a physical meaning. For example, ω is the radian frequency, ρ is typically a density. ϕ, ψ are commonly used to indicate angles of a triangle, and $\zeta(s)$ is the Riemann zeta function. Many of these are so well established it makes no sense to define new terms, so we will not do that.

Likely you do not know all of these Greek letters, commonly used in mathematics. Some of them are pronounced in strange ways. The symbol ξ is pronounced "c," ζ is "zeta," β is "beta," and χ is "kie" (rhymes with pie). I will assume you know how to pronounce the others, which are more phonetic in English. One advantage of learning LaTeX is that all of these math sympols are built in, and thus more easily learned, once you have adopted this powerful open-source math-oriented word-processing system (e.g., used to write this book).

---

[1] https://en.wikipedia.org/wiki/List_of_mathematical_symbols_by_subject#Definition_symbols

Table A.1: Double-bold notation for the types of numbers. (#) is a page number. Symbol with an exponent denote the dimensionality. Thus $\mathbb{R}^2$ represents the real plane. An exponent of 0 denotes point, e.g., $\jmath \in \mathbb{C}^0$.

Symbol (p. #)	Genus	Examples	Counter Examples
$\mathbb{N}$ (22)	Counting	1,2,17,3, 10^{20}	0, -10, 5j
$\mathbb{P}$ (22)	Prime	2,17,3, 10^{20}	0, 1, 4, 3^2, 12, -5
$\mathbb{Z}$ (22)	Integer	-1, 0, 17, 5j, -10^{20}	$1/2, \pi$, $\sqrt{5}$
$\mathbb{Q}$ (23)	Rational	2/1, 3/2, 1.5, 1.14	$\sqrt{2}$, $3^{-1/3}$, π
$\mathbb{F}$ (23)	Fractional	1/2, 7/22	2/1, $1/\sqrt{2}$
$\mathbb{I}$ (23)	Irrational	$\sqrt{2}$, $3^{-1/3}$, π, e	Vectors
$\mathbb{R}$ (23)	Reals	$\sqrt{2}$, $3^{-1/3}$, π	$2\pi j$
$\mathbb{C}$ (24)	Complex	$1, \sqrt{2}j$, $3^{-j/3}$, π^j	Vectors

A.1.4   Double-Bold notation

Table A.1 indicates the symbol followed by a page number indicationg where it is discussed, and the Genus (class) of the number type. For example, $\mathbb{N} > 0$ indicates the infinite set of *counting numbers* $\{1, 2, 3, \cdots \}$, not including zero. Starting from any counting number, you get the next one by adding 1. Counting numbers are also know as the *Cardinal numbers.*

   We say that a number is in the set with the notation $3 \in \mathbb{N} \in \mathbb{R}$, which is read as "3 is in the set of counting numbers, which in turn in the set of real numbers," or in vernacular language "3 is a real counting number."

   Prime numbers ($\mathbb{P} \in \mathbb{N} > 1$) are taken from the counting numbers, do not include 1, and cannot be factored.

   The signed integers $\mathbb{Z}$ include 0 and negative integers. Rational numbers $\mathbb{Q}$ are historically defined to include $\mathbb{Z}$, a somewhat inconvient definition, since the more interesting class are the *fractionals* $\mathbb{F}$, a subset of rationals $\mathbb{F} \in \mathbb{Q}$ that exclude the integers (i.e., $\mathbb{F} \perp \mathbb{Z}$). This is a useful definition because the rationals $\mathbb{Q} = \mathbb{Z} \cup \mathbb{F}$ are formed from the union of integers and fractionals.

   Irrational numbers $\mathbb{I}$ are very special: They are formed by taking a limit of fractionals, as tne numerator and denominator $\to \infty$, and apporach a limit point. It follows that irrational numbers must be approximated by fractionals.

   The reals ($\mathbb{R}$) include complex numbers ($\mathbb{C}$) having a zero imaginary part (i.e., $\mathbb{R} \subset \mathbb{C}$).

   The *cardinality* of a set is denoted by taking the absolute value (e.g., $|\mathbb{N}|$).

Classification of numbers: From the above definitions there exists a natural heiertical structure of numbers:
$$\mathbb{P} \in \mathbb{N}, \quad \mathbb{Z} : \{\mathbb{N}, 0, -\mathbb{N}\}, \quad \mathbb{F} \perp \mathbb{Z}, \quad \mathbb{Q} : \mathbb{Z} \cup \mathbb{F}, \quad \mathbb{R} : \mathbb{Q} \cup \mathbb{I} \subset \mathbb{C}$$

1. The primes are a subset of the counting numbers: $\mathbb{P} \subset \mathbb{N}$.

2. The signed integers $\mathbb{Z}$ are composed of $\pm\mathbb{N}$ and 0, thus $\mathbb{N} \subset \mathbb{Z}$.

3. The fractionals $\mathbb{F}$ do not include of the signed integers $\mathbb{Z}$.

4. The rationals $\mathbb{Q} = \mathbb{Z} \cup \mathbb{F}$ are the union of the signed integers and fractionals

5. Irrational numbers $\mathbb{I}$ have the special property $\mathbb{I} \perp \mathbb{Q}$.

6. The reals $\mathbb{R} : \mathbb{Q}, \mathbb{I}$ are the union of rationals and *irrationals* $\mathbb{I}$

7. Reals $\mathbb{R}$ may be defined as a subset of those complex numbers $\mathbb{C}$ having zero imaginary part.

A.2 Complex vectors and the impedance matrix

Vectors as columns of ordered sets of scalars. When we write then out in text, we typically use row notation, with the *transpose* symbol:

$$[a, b, c]^T = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

This is strictly to save space on the page. The above example is said to be a 3 *dimensional* vector, because it has three components. With rare exceptions, vectors are rows.

The transpose of a vector (a "so called" *row-vector*) are typically used as *weights*. A vector *dot product* is defined between weights and vectors, resulting in a scalar. When the elements are complex ($\in \mathbb{C}$), the transpose is defined as the complex conjugate of the elements. In such complex cases the transpose conjugate is denoted with a † rather than T. For example

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 1 + 2 + 3 = 6.$$

Vectors are also frequency written using a bold font (e.g., $\boldsymbol{a}^T = \begin{bmatrix} 2\jmath & -3\jmath & 1 \end{bmatrix} \in \mathbb{C}$.) For this case the dot product is

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^\dagger \boldsymbol{b} = a_1^* b_1 + a_2^* b_2 + a_3^* b_3.$$

Such a construction is useful when $\boldsymbol{a}$ and $\boldsymbol{b}$ are related by an impedance matrix

$$\boldsymbol{V}(s) = \boldsymbol{Z}(s)\boldsymbol{I}(s).$$

For example, the impedance of a mass is ms and a capacitor is $1/sC$. When given a system of equations (a mechanical or electrical circuit) one may define a matrix impedance.

GIVE MORE EXAMPLES

Complex power: In this special case, the *complex power* $\mathcal{P}(s) \in \mathbb{R}(s)$ is defined, in the complex frequency domain (s), as

$$\mathcal{P}(s) = \boldsymbol{I}^\dagger(s)\boldsymbol{V}(s) = \boldsymbol{I}^\dagger(s)\boldsymbol{Z}(s)\boldsymbol{I}(s).$$

The complex power must be positive, with $\mathcal{L}^{-1} p(t) \leftrightarrow \mathcal{P}$.

GIVE EXAMPLES

Norm of a vector: The dot product of a vector with itself is called the *norm* of $\boldsymbol{a}$, designated as

$$||\boldsymbol{a}|| = \sqrt{\boldsymbol{a}^\dagger \boldsymbol{a}} \geq 0$$

which is always non-negative.

A.2.1 Vectors in $\mathbb{R}^3$

The case of three-dimensions is special, allowing definitions that are not defined in more dimensions. A vector in $\mathbb{R}^3$ labels the point having the coordinates of that vector.

Dot product in $\mathbb{R}^3$: The dot $\boldsymbol{B} \cdot \mathbb{C} = ||\boldsymbol{B}|| \, ||\boldsymbol{C}|| \cos(\theta)$, and $\cos(\theta)$ is called the *direction cosine* between $\boldsymbol{B}$ and $\boldsymbol{C}$.

Norm of a vector in $\mathbb{R}^3$: The norm of a vector is the dot product of it with itself

$$||\boldsymbol{A}|| = \sqrt{\boldsymbol{A} \cdot \boldsymbol{A}}$$

Euclidian distance between two points in $\mathbb{R}^3$:   The dot product of the difference between two vectors $(\boldsymbol{A} - \boldsymbol{B}) \cdot (\boldsymbol{A} - \boldsymbol{B})$ is the Euclidian distance between the points they define

$$||\boldsymbol{A} - \boldsymbol{B}|| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}.$$

Cross product: and cross product $\boldsymbol{A} \times \boldsymbol{B} = ||\boldsymbol{A}|| \, ||\boldsymbol{B}|| \sin(\theta)$ are defined between the two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$.

The cross product gives the area of the trapazoid (diamond) outlined by the two vectors (Fig. 1.12).

The triple product: This is defined between three vectors as

$$\boldsymbol{A} \cdot (\boldsymbol{B} \times \boldsymbol{C}) = det \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix},$$

also defined in Fig. 1.12. This may be indicated without the use of parentheses, since there can be no other meaningful interpretation. However for rigor, parentheses should be used.

A.3   Matrices

Unfortunately when working with matrices, the role of the weights and vectors can change, depending on the context. A useful way to view a matrix is as a set of column vectors, weighted by the elements of the column-vector of weights multiplied from the right. For example

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1M} \\ a_{21} & a_{22} & a_{32} & \cdots & a_{3M} \\ & & \ddots & & \\ a_{N1} & a_{N2} & a_{N3} & \cdots & a_{NM} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_M \end{bmatrix} = w_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{21} \\ \cdots \\ a_{N1} \end{bmatrix} + w_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \\ \cdots \\ a_{N2} \end{bmatrix} \ldots w_M \begin{bmatrix} a_{1M} \\ a_{2M} \\ a_{3M} \\ \cdots \\ a_{NM} \end{bmatrix},$$

where the weights are $\left[w_1, w_2, \ldots, w_M. \right]^T$

Another way to view the matrix is a set of row vectors of weights, each of which re applied to the vector $[w_1, w_2, \cdots, W_M]^T$.

The determinant of a matrix is denoted either as $det\,\boldsymbol{A}$ or simply $|\boldsymbol{A}|$, as in the absolute value. The inverse of a square matrix is $\boldsymbol{A}^{-1}$ or $inv\,\boldsymbol{A}$. If $|A| = 0$, the inverse does not exist. $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}$.

Matlab's notional convention for a row-vector is $[a, b, c]$ and a column-vector is $[a; b; c]$. A prime on a vector takes the complex conjugate transpose. To suppress the conjugation, place a period before the prime. The : argument converts the array into a column vector, without conjugation. A tacit notation in Matlab is that *vectors* are columns and the index to a vector is a row vector. Matlab defines the notation $1\!:\!4$ as the "row-vector" $[1, 2, 3, 4]$, which is unfortunate as it leads users to assume that the default vector is a row. This can lead to serious confusion later, as Matlab's default vector is a column. I have not found the above convention explicitly stated, and it took me years to figure this out for myself.

When writing a complex number we shall adopt $1_\jmath$ to indicate $\sqrt{-1}$. Matlab/Octave prefer this as well, as its explicit.

Units are SI; Angles in degrees [deg] unless otherwise noted. The units for π are always in radians [rad]. Ex: $\sin(\pi), e^{\jmath 90°} e^{\jmath \pi/2}$.

A.4   Periodic functions

Fourier series tells us that periodic functions are discrete in frequency, with frequencies given by nT_x, where T_s is the sample period ($T_s = 1/2F_{max}$ and $F_{min} = F_{max}/\text{NFT}$).

This concept is captured by the *Fourier series*, which is a frequency expansion of a periodic function. This concept is quite general. Periodic in frequency implies discrete in time. Periodic and discrete in time requies periodic and discrete in frequency (the case of the DFT). The modulo function *mmodn* is periodic with period n.

A periodic function may be conveniently indicated using double-parentheses notation. This is sometimes known as modular arithmetic. For example

$$f((t))_T = f(t) = f(t \pm kT),$$

is periodic on $t, T \in \mathbb{R}$ with a period of T and $k \in \mathbb{Z}$. This notation is useful when dealing with Fourier series of periodic functions.

When a discrete valued (e.g., time) sequence is periodic we use square brackets

$$f[[n]]_N = f[n] = f[n \pm kN],$$

with $n, k, N \in \mathbb{Z}$ and period N. This notation will be used with discrete-time signals that are periodic, such as the case of the DFT.

A.5   Differential equations vs. Polynomials

A polynomial has *degree N* defined by the largest power. A quadratic equation is degree 2, and a cubic has degree 3. We shall indicate a polynomial by the notation

$$P_N(z) = z^N + a_{N-1}z^{N-1} \cdots a_0.$$

Always normalize the polynomial so that $a_N = 1$. This will not change the roots, defined by Eq. 1.18 (p. 44). The coefficient on z^{N-1} is always the sum of the roots z_n ($a_{N-1} = \sum_n^N z_n$), and a_0 is always their product ($a_0 = \prod_n^N z_n$).

Differential equations have *order* (polynomials have degree). If a second order differential equation is Laplace transformed (Lec. 1.3.11, p. 70), one is left with a degree 2 polynomial. For example:

$$\frac{d^2}{dt^2}y(t) + b\frac{d}{dt}y(t) + cy(t) = \alpha \left(\frac{d}{dt}x(t) + \beta x(t) \right) \leftrightarrow \tag{A.1}$$

$$(s^2 + bs + c)Y(s) = \alpha(s + \beta)X(s). \tag{A.2}$$

$$\frac{Y(s)}{X(s)} = \alpha\frac{s + \beta}{s^2 + bs + c} \equiv H(s) \leftrightarrow h(t). \tag{A.3}$$

Using the same argument as for polynomials, the lead coefficent must always be 1. The coefficent $\alpha \in \mathbb{R}$ is called the *gain*. The complex variable s is the *Laplace frequency*. Discuss half-derivatives, etc.

The ratio of the output $Y(s)$ over the input $X(s)$ is called the system *transfer function $H(s)$*. When $H(s)$ is the ratio of two polynomials in s, the transfer function is said to be *bilinear*, since it is linear in both the input and output. The roots of the numerator are called the *zeros* and those of the denominator, the *poles*. The inverse Laplace transform of the transfer function is called the system *impulse response*, which describes the system's output signal $y(t)$ for any given input signal $x(t)$, via convolution (i.e., $y(t) = h(t) \star x(t)$).

Appendix B

Linear algebra of 2x2 matrices

B.1 Notation

Definitions:

1. *Scalar*: A number, e.g. $\{a, b, c, \alpha, \beta, \cdots\} \in \{\mathbb{Z}, \mathbb{Q}, \mathbb{I}, \mathbb{R}, \mathbb{C}\}$

2. *Vector*: A quantity having direction as well as magnitude, often denoted by a bold-face letter with an arrow, $\boldsymbol{x}$. In matrix notation, this is typically represented as a single row $[x_1, x_2, x_3, \ldots]$ or single column $[x_1, x_2, x_3 \ldots]^T$ (where T indicates the transpose). In this class we will typically use column vectors. The vector may also be written out using unit vector notation to indicate direction. For example: $\boldsymbol{x}_{3\times 1} = x_1\hat{\mathbf{i}} + x_2\hat{\mathbf{j}} + x_3\hat{\mathbf{k}} = [x_1, x_2, x_3]^T$, where $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ are unit vectors in the x, y, z cartesian directions (here the vector's subscript 3×1 indicates its dimensions). The type of notation used may depend on the engineering problem you are solving.

3. *Matrix:* $A = \begin{bmatrix} \boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \cdots, \boldsymbol{a}_M \end{bmatrix}_{N\times M} = \{a_{n,m}\}_{N\times M}$, can be a non-square matrix if the number of elements in each of the vectors (N) is not equal to the number of vectors (M). When $M = N$, the matrix is square. It may be inverted if its determinant $|A| = \prod \lambda_k \neq 0$ (where λ_k are the eigenvalues).

   We shall only work with 2×2 and 3×3 square matrices throughout this course.

4. *Linear system of equations*: $A\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{x}$ and $\boldsymbol{b}$ are vectors and matrix A is a square.

   (a) *Inverse:* The solution of this system of equations may be found by finding the inverse $\boldsymbol{x} = A^{-1}\boldsymbol{b}$

   (b) *Equivalence:* If two systems of equations $A_0\boldsymbol{x} = \boldsymbol{b}_0$ and $A_1\boldsymbol{x} = \boldsymbol{b}_1$ have the same solution (i.e., $\boldsymbol{x} = A_0^{-1}\boldsymbol{b}_0 = A_1^{-1}\boldsymbol{b}_1$), they are said to be equivalent.

   (c) *Augmented matrix:* The first type of augmented matrix is defined by combining the matrix with the right-hand-side. For example, given the linear system of equations $A\boldsymbol{x} = \boldsymbol{y}$

   $$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

   then the augmented matrix is

   $$A|y = \begin{bmatrix} a & b & y_1 \\ c & d & y_2 \end{bmatrix}$$

   A second type of augmented matrix may be used for finding the inverse of a matrix (rather than solving a specific instance of linear equations $Ax = b$). In this case the augmented matrix is

   $$A|I = \begin{bmatrix} a & b & 1 & 0 \\ c & d & 0 & 1 \end{bmatrix}$$

Performing Gaussian elimination on this matrix, until the left side becomes the identity matrix, yields A^{-1}. This is because multiplying both sides by A^{-1} gives $A^{-1}A|A^{-1}I = I|A^{-1}$.

5. *Permutation matrix* (P): A matrix that is equivalent to the identity matrix, but with scrambled rows (or columns). Such a matrix has the properties $det(P) = \pm 1$ and $P^2 = I$. For the 2x2 case, there is only one permutation matrix:

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad P^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A permutation matrix P swaps rows or columns of the matrix it operates on. For example, in the 2x2 case, pre-multiplication swaps the rows

$$PA = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ \alpha & \beta \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ a & b \end{bmatrix},$$

whereas post-multiplication swaps the *columns*

$$AP = \begin{bmatrix} a & b \\ \alpha & \beta \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} b & a \\ \beta & \alpha \end{bmatrix}$$

For the 3x3 case there are $3 \cdot 2 = 6$ such matrices, including the original 3x3 identity matrix (swap a row with the other 2, then swap the remaining two rows).

6. *Gaussian elimination (GE) operations* G_k: There are 3 types of elementary row operations, which may be performed without fundamentally altering a system of equations (e.g. the resulting system of equations is *equivalent*). These operations are (1) swap rows (e.g. using a permutation matrix), (2) scale rows, or (3) perform addition/subraction of two scaled rows. All such operations can be performed using matrices.

For lack of a better term, we'll describe these as 'gaussian elimination' or 'GE' matrices.[1] We will categorize any matrix that performs only elementary row operations (but any number of them) as a 'GE' matrix. Therefore, cascade of GE matrices is also a GE matrix.

Consider the GE matrix

$$G = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$$

(a) Pre-multiplication scales and adds the *rows*

$$GA = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} a & b \\ \alpha & \beta \end{bmatrix} = \begin{bmatrix} a & b \\ a - \alpha & b - \beta \end{bmatrix}$$

The result is a Gaussian elimination operation.

(b) Post-multiplication adds and scales *columns*.

$$AG = \begin{bmatrix} a & b \\ \alpha & \beta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} a - b & b \\ \alpha - \beta & \beta \end{bmatrix}$$

Here the second column is subtracted from the first, and placed in the first. The second column is untouched. **This operation is *not* a Gaussian elimination**. Therefore, to

---

[1]The term 'elementary matrix' may also be used to refer to a matrix that performs an elementary row operation. Typically, each elementary matrix differs from the identity matrix by one single row operation. A cascade of elementary matrices could be used to perform Gaussian elimination.

put Gaussian elimination operations in matrix form, we form a cascade of **pre-multiply** matrices.

Here $\det(G) = 1$, $G^2 = I$, which won't always be true if we scale by a number greater than 1. For instance, if $G = \begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix}$ (scale and add), then we have $\det(G) = 1$, $G^n = \begin{bmatrix} 1 & 0 \\ n \cdot m & 1 \end{bmatrix}$.

B.1.1 Gaussian elimination exercises

Find the solution to the following 3x3 matrix equation $Ax = b$ by Gaussian elimination. Show your intermediate steps. You can check your work at each step using Matlab.

$$\begin{bmatrix} 1 & 1 & -1 \\ 3 & 1 & 1 \\ 1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 9 \\ 8 \end{bmatrix}.$$

1. Show (i.e., verify) that the first GE matrix G_1, which zeros out all entries in the first column, is given by

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Identify the elementary row operations that this matrix performs. Sol: Operate with GE matrix on A

$$G_1[A|b] = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 3 & 1 & 1 & | & 9 \\ 1 & -1 & 4 & | & 8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 0 & -2 & 4 & | & 6 \\ 0 & -2 & 5 & | & 7 \end{bmatrix}$$

It scales the first row by -3 and adds it to the second row, and scales the first row by -1 and adds it to the third row.

2. Find a second GE matrix, G_2, to put G_1A in upper triangular form. Identify the elementary row operations that this matrix performs. Sol:

$$G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

which scales the second row by -1 and adds it to the third row. Thus we have

$$G_2G_1[A|b] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 3 & 1 & 1 & | & 9 \\ 1 & -1 & 4 & | & 8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 0 & -2 & 4 & | & 6 \\ 0 & 0 & 1 & | & 1 \end{bmatrix}$$

3. Find a third GE matrix, G_3, which scales each row so that its leading term is 1. Identify the elementary row operations that this matrix performs. Sol:

$$G_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which scales the second row by -1/2. Thus we have

$$G_3G_2G_1[A|b] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 3 & 1 & 1 & | & 9 \\ 1 & -1 & 4 & | & 8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 0 & 1 & -2 & | & -3 \\ 0 & 0 & 1 & | & 1 \end{bmatrix}$$

4. Finally, find the last GE matrix, G_4, that subtracts a scaled version of row 3 from row 2, and scaled versions of rows 2 and 3 from row 1, such that you are left with the identity matrix $(G_4 G_3 G_2 G_1 A = I)$. Sol:

$$G_4 = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus we have

$$G_4 G_3 G_2 G_1 [A|b] = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & | & 1 \\ 3 & 1 & 1 & | & 9 \\ 1 & -1 & 4 & | & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

5. Solve for $[x_1, x_2, x_3]^T$ using the augmented matrix format $G_4 G_3 G_2 G_1 [A|b]$ (where $[A|b]$ is the augmented matrix). Note that if you've performed the preceding steps correctly, $x = G_4 G_3 G_2 G_1 b$. Sol: From the preceding problems, we see that $[x_1, x_2, x_3]^T = [3, -1, 1]^T$

B.2 Inverse of the 2x2 matrix

We shall now apply Gaussian elimination to find the solution $[x_1, x_2]$ for the 2x2 matrix equation $Ax = y$ (Eq. 1.39, left). We assume to know $[a, b, c, d]$ and $[y_1, y_2]$. We wish to show that the intersection (solution) is given by the equation on the right.

Here we wish to prove that the left equation has an inverse given by the right equation:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad\qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

How to take inverse:
1) Swap the diagonal, 2) change the signs of the off-diagonal, and 3) divide by Δ.

B.2.1 Derivation of the inverse of a 2x2 matrix

1. Step 1: normalize the first column to 1.

2. Step 2: subtract the top equation from the lower.

3. Step 3: express result in terms of the determinate $\Delta = ad - bc$.

$$\begin{bmatrix} 1 & \frac{b}{a} \\ 1 & \frac{d}{c} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{a}y_1 \\ \frac{1}{c}y_2 \end{bmatrix} \qquad \begin{bmatrix} 1 & \frac{b}{a} \\ 0 & \frac{d}{c} - \frac{b}{a} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{a} & 0 \\ -\frac{1}{a} & \frac{1}{c} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \begin{bmatrix} 1 & \frac{b}{a} \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{a} & 0 \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

4. Step 4: These steps give the solution for x_2: $x_2 = -\frac{c}{\Delta}y_1 + \frac{a}{\Delta}y_2$.

5. Step 5: Finally the top equation is solved for x_1: $x_1 = \frac{1}{a}y_1 - \frac{b}{a}x_2 \Rightarrow x_1 = [\frac{1}{a} + \frac{b}{a}\frac{c}{\Delta}]y_1 - \frac{b}{a}\frac{a}{\Delta}y_2$.

In matrix format, in terms of the determinate $\Delta = ab - cd$ becomes:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{a} + \frac{bc}{a\Delta} & -\frac{b}{\Delta} \\ -\frac{c}{\Delta} & \frac{a}{\Delta} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \frac{\Delta + bc}{a} & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Summary: This is a lot of algebra, that is why it is essential you memorize the formula for the inverse.

Appendix C

Eigenvector analysis

Here we show how to compute the eigenvalues and eigenvectors for the 2x2 Pell matrix

$$A = \begin{bmatrix} 1 & N \\ 1 & 1 \end{bmatrix}.$$

The analysis applies to any matrix, but since we are concentrated on Pell's equation, we shall use the Pell matrix, for $N = 2$. By using a specific matrix we can check all the equations below with Matlab, which I advise you to do.

The Matlab command `[E,D]=eig(A)` returns the eigenvector matrix E

$$E = [e_+, e_-] = \frac{1}{\sqrt{3}} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.8165 & -0.8165 \\ 0.5774 & 0.5774. \end{bmatrix}$$

and the eigenvalue matrix Λ (Matlab's D)

$$\Lambda \equiv \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix} = \begin{bmatrix} 1 + \sqrt{2} & 0 \\ 0 & 1 - \sqrt{2} \end{bmatrix} = \begin{bmatrix} 2.4142 & 0 \\ 0 & -0.4142 \end{bmatrix}.$$

The factor $\sqrt{3}$ on E normalizes each eigenvector to 1 (i.e., The Matlab command `norm([`$\sqrt{2}, 1$`])` gives $\sqrt{3}$).

In the following discussion we show how to determine E and D (i.e, Λ), given A.

Calculating the eigenvalue matrix (Λ): The matrix equation for E is

$$AE = E\Lambda. \tag{C.1}$$

Pre-multiplying by E^{-1} diagonalizes A, given the *eigenvalue matrix* (D in Matlab)

$$\Lambda = E^{-1}AE. \tag{C.2}$$

Post-multiplying by E^{-1} recovers A

$$A = E\Lambda E^{-1}. \tag{C.3}$$

Matrix power formula: This last relation is the entire point of the eigenvector analysis, since it shows that any power of A may be computed from powers of the eigen values. Specifically

$$A^n = E\Lambda^n E^{-1}. \tag{C.4}$$

For example, $A^2 = AA = E\Lambda \left(E^{-1}E \right) \Lambda E^{-1} = E\Lambda^2 E^{-1}$.

Equations C.1, C.2 and C.3 are the key to eigenvector analysis, and you need to memorize them. You will use them repeatedly throughout this course, and possibly for a long time after it is over.

Showing that $A - \lambda_{\pm}I_2$ is singular: If we restrict Eq. C.1 to a single eigenvector (one of $e_{\pm}$), along with the corresponding eigenvalue $\lambda_{\pm}$, we obtain a matrix equations

$$Ae_{\pm} = E_{\pm}\lambda_{\pm} = \lambda_{\pm}E_{\pm}$$

Note the important swap in the order of $E_{\pm}$ and $\lambda_{\pm}$. Since $\lambda_{\pm}$ is a scalar, this is legal (and critically important), since this allows us to remove (factored out) $E_{\pm}$

$$(A - \lambda_{\pm}I_2)E_{\pm} = 0. \tag{C.5}$$

This means that the matrix $A - \lambda_{\pm}I_2$ must be singular, since when it operates on $E_{\pm}$, which is not zero, it gives zero. It immediately follows that its determinant is zero (i.e., $|(A - \lambda_{\pm}I_2)| = 0$). This equation is used to uniquely determine the eigenvalues $\lambda_{\pm}$. Note the important difference between $\lambda_{\pm}I_2$ and Λ (i.e., $|(A - \Lambda)| \neq 0$).

Calculating the eigenvalues $\lambda_{\pm}$:   The eigenvalues $\lambda_{\pm}$ of A may be determined from $|(A - \lambda_{\pm}I_2)| = 0$

$$\begin{vmatrix} 1 - \lambda_{\pm} & N \\ 1 & 1 - \lambda_{\pm} \end{vmatrix} = (1 - \lambda_{\pm})^2 - N^2 = 0.$$

For our case of $N = 2$, $\lambda_{\pm} = (1 \pm \sqrt{2})$.[1]

Calculating the eigenvectors $e_{\pm}$:   Once the eigenvalues have been determined, they are substitute them into Eq. C.5, which determines the eigenvectors $E = \begin{bmatrix} e_+, e_- \end{bmatrix}$, by solving

$$(A - \lambda_{\pm})e_{\pm} = \begin{bmatrix} 1 - \lambda_{\pm} & 2 \\ 1 & 1 - \lambda_{\pm} \end{bmatrix} E_{\pm} = 0,$$

where $1 - \lambda_{\pm} = 1 - (1 \pm \sqrt{2}) = \mp\sqrt{2}$.

Recall that Eq. C.5 is singular, because we are using an eigenvalue, and each eigenvector is pointing in a unique direction (This is why it is singular). You might respectively suggest that this equation has no solution. In some sense you would be correct. When we solve for $E_{\pm}$, the two equations defined by Eq. C.5 *co-linear* (the two equations describe parallel lines). This follows from the fact that there is only one eigenvector for each eigenvalue.

Expecting trouble, yet proceeding to solve for $E = [e_1^+, e_2^+]^T$,

$$\begin{bmatrix} -\sqrt{2} & 2 \\ 1 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} e_1^+ \\ e_2^+ \end{bmatrix} = 0.$$

This gives two identical equations $-\sqrt{2}e_1^+ + 2e_2^+ = 0$ and $e_1^+ - \sqrt{2}e_2^+ = 0$. This is the price of an over-specified equation (the singular matrix is degenerate). The most we can determine is $e_+ = c \begin{bmatrix} -\sqrt{2}, 1 \end{bmatrix}^T$, where c is a constant. We can determine eigenvector direction, but not its magnitude.

Following *exactly* the same procedure for λ_-, the equation for e_- is

$$\begin{bmatrix} \sqrt{2} & 2 \\ 1 & \sqrt{2} \end{bmatrix} \begin{bmatrix} e_1^- \\ e_2^- \end{bmatrix} = 0.$$

In this case the relation becomes $e_1^- + \sqrt{2}e_2^- = 0$, thus $E_- = c \begin{bmatrix} \sqrt{2}, 1 \end{bmatrix}^T$ where c is a constant.

Normalization of the eigenvectors: The constant c may be determined by normalizing the eigenvectors to have unit length. Since we cannot determine the length, we set it to 1. In some sense the degeneracy is resolved by this normalization. Thus $c = 1/\sqrt{3}$, since

$$\left(\pm\sqrt{2}\right)^2 + 1^2 = 3 = 1/c^2.$$

---

[1]It is a convention to order the eigenvalues from largest to smallest.

Summary: Thus far we have shown

$$\boldsymbol{E} = [\boldsymbol{e}_+, \boldsymbol{e}_-] = \frac{1}{\sqrt{3}} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 1 & 1 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix} = \begin{bmatrix} 1 + \sqrt{2} & 0 \\ 0 & 1 - \sqrt{2} \end{bmatrix}.$$

Verify that $\Lambda = \boldsymbol{E}^{-1}\boldsymbol{A}\boldsymbol{E}$: To find the inverse of $\boldsymbol{E}$, 1) swap the diagonal values, 2) change the sign of the off diagonals, and 3) divide by the determinant $\Delta = 2\sqrt{2}/\sqrt{3}$ (see Appendix B)

$$\boldsymbol{E}^{-1} = \frac{\sqrt{3}}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} \\ -1 & \sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.6124 & 0.866 \\ -0.6124 & 0.866 \end{bmatrix}.$$

By definition for any matrix $\boldsymbol{E}^{-1}\boldsymbol{E} = \boldsymbol{E}\boldsymbol{E}^{-1} = \boldsymbol{I}_2$. Taking the product gives

$$\boldsymbol{E}^{-1}\boldsymbol{E} = \frac{\cancel{\sqrt{3}}}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} \\ -1 & \sqrt{2} \end{bmatrix} \cdot \frac{1}{\cancel{\sqrt{3}}} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{I}_2.$$

We wish to show that $\Lambda = \boldsymbol{E}^{-1}\boldsymbol{A}\boldsymbol{E}$

$$\begin{bmatrix} 1 + \sqrt{2} & 0 \\ 0 & 1 - \sqrt{2} \end{bmatrix} \cdot = \frac{\cancel{\sqrt{3}}}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} \\ -1 & \sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \cdot \frac{1}{\cancel{\sqrt{3}}} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 1 & 1 \end{bmatrix},$$

which is best verified with Matlab.

Verify that $\boldsymbol{A} = \boldsymbol{E}\Lambda\boldsymbol{E}^{-1}$: We wish to show that

$$\begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} = \frac{1}{\cancel{\sqrt{3}}} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 + \sqrt{2} & 0 \\ 0 & 1 - \sqrt{2} \end{bmatrix} \cdot \frac{\cancel{\sqrt{3}}}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} \\ -1 & \sqrt{2} \end{bmatrix},$$

which is best verified with Matlab (or Octave). All the above equations have been verified both with Matlab and Octave.

I suggest that you verify $\boldsymbol{E}\Lambda \neq \Lambda\boldsymbol{E}$ and $\boldsymbol{A}\boldsymbol{E} = \boldsymbol{E}\Lambda$ with Matlab. Here is the Matlab program which does this:

```
A = [1 2; 1 1]; %define the matrix
[E,D] = eig(A); %compute the eigenvector and eigenvalue matrices
A*E-E*D %This should be $\approx 0$, within numerical error.
E*D-D*E %This is not zero
```

Symbolic analysis of $\boldsymbol{E}$ and Λ: Using Matlab/Octave's `syms A B C D M L M=[A B;C D] [E,L]=eig(M)` the eigenvectors $\boldsymbol{E}_\pm$ are

$$\boldsymbol{e}_\pm = \begin{pmatrix} \frac{1}{2C}\left[(A-D) \mp \sqrt{(A-D)^2 + 4BC}\right] \\ 1 \end{pmatrix} \tag{C.6}$$

and eigenvalues are

$$\lambda_\pm = \frac{1}{2} \begin{pmatrix} (A+D) - \sqrt{(A-D)^2 + 4BC} \\ (A+D) + \sqrt{(A-D)^2 + 4BC} \end{pmatrix} \tag{C.7}$$

When $\boldsymbol{T}$ is symmetric $(A = D)$ these simplify to

$$\boldsymbol{E} = \begin{bmatrix} -\sqrt{\frac{B}{C}} & +\sqrt{\frac{B}{C}} \\ 1 & 1 \end{bmatrix} \qquad \Lambda = \begin{bmatrix} A - \sqrt{BC} & 0 \\ 0 & A + \sqrt{BC} \end{bmatrix} \tag{C.8}$$

Appendix D

Solution to Pell's Equation (N=2)

Section 2.2.2 (p. 125) showed that the solution $[x_n, y_n]^T$ to Pell's equation, for $N = 2$, is given by powers of Eq. 1.7. To find an explicit formula for $[x_n, y_n]^T$, one must compute powers of

$$\boldsymbol{A} = 1\jmath \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}. \tag{D.1}$$

We wish to find the solution to Pell's equation (Eq. 1.7), based on the recursive solution, Eq. 1.8 (p. 40). Thus we need is powers of A, that is A^n, which gives the a closed form expression for $[x_n, y_n]^T$. By the diagonalization of A, its powers are simply the powers of its eigenvalues. This diagonalization is called an *eigenvalue analysis*, a very general method rooted in linear algebra. This type of analysis allows us to find the solution to most of the linear the equations we encounter.

From Matlab with $N = 2$ the eigenvalues of Eq. D.1 are $\lambda_\pm \approx [2.4142\jmath, -0.4142\jmath]$ (i.e., $\lambda_\pm = 1\jmath(1 \pm \sqrt{2})$). The final solution to Eq. D.1 is given in Eq. 2.7 (p. 125). The solution for $N = 3$ is provided in Appendix D.1 (p. 179).

Once the matrix has been diagonalized, one may compute powers of that matrix as powers of the eigenvalues. This results in the general solution given by

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} = 1\jmath^n \boldsymbol{A}^n \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1\jmath^n \boldsymbol{E} \Lambda^n \boldsymbol{E}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The eigenvalue matrix D is diagonal with the eigenvalues sorted, largest first. The Matlab command `[E,D]=eig(A)` is helpful to find D and E given any A. As we saw above,

$$\Lambda = 1\jmath \begin{bmatrix} 1 + \sqrt{2} & 0 \\ 0 & 1 - \sqrt{2} \end{bmatrix} \approx \begin{bmatrix} 2.414\jmath & 0 \\ 0 & -0.414\jmath \end{bmatrix}.$$

D.1 Pell equation for N=3

This summarizes the solution of Pell's equation due to the Pythagoreans using matrix recursion, for the case of N=3. The integer solutions are shown in on the right. Note that $x_n/y_n \to \sqrt{3}$, in agreement with the Euclidean algorithm.[1] It seem likely that β_0 could be absorbed in the starting solution, and then be removed from the generating function, other than as the known factor β_0^n

Case of $N = 3$: $[x_0, y_0]^T = [1, 0]^T$, $\beta_0 = \jmath/\sqrt{2}$; Pell-3: $x_n^2 - 3y_n^2 = 1$; $x_n/y_n \xrightarrow[\infty]{} \sqrt{3}$

Try other trivial solutions such as $[-1, 0]^T$ and $[\pm\jmath, 0]^T$. Perhaps this can provide a clue to the proper value of β_0.

---

[1]The matlab program for generating this solution is `PellSol3.m`.

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \beta_0 \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad\qquad (1\beta_0)^2 - 3(1\beta_0)^2 = 1$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \beta_0^2 \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \beta_0^2 \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad\qquad \left(4\beta_0^2\right)^2 - 3\left(2\beta_0^2\right)^2 = 1$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \beta_0^3 \begin{bmatrix} 10 \\ 6 \end{bmatrix} = \beta_0^3 \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} \qquad\qquad \left(10\beta_0^3\right)^2 - 3\left(6\beta_0^3\right)^2 = 1$$

$$\begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \beta_0^4 \begin{bmatrix} 28 \\ 16 \end{bmatrix} = \beta_0^4 \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 10 \\ 6 \end{bmatrix} \qquad\qquad \left(28\beta_0^4\right)^2 - 3\left(16\beta_0^4\right)^2 = 1$$

$$\begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \beta_0^5 \begin{bmatrix} 76 \\ 44 \end{bmatrix} = \beta_0^5 \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 28 \\ 16 \end{bmatrix} \qquad\qquad \left(76\beta_0^5\right)^2 - 3\left(44\beta_0^5\right)^2 = 1$$

Appendix E

Laplace transforms

Given a *Laplace transform* ($\mathcal{L}$) pair $f(t) \leftrightarrow F(s)$, the frequency domain will always be upper-case [e.g. $F(s)$] and the time domain lower case [$f(t)$] and causal (i.e., $f(t < 0) = 0$). The definition of the forward transform ($f(t) \to F(s)$) is

$$F(s) = \int_{0^-}^{\infty} f(t)e^{-st}dt,$$

where $s = \sigma + j\omega$ is the complex Laplace frequency in [radians] and t is time in [seconds].

The *inverse Laplace transform* ($\mathcal{L}^{-1}$), $F(s) \to f(t)$ is defined as

$$f(t) = \frac{1}{2\pi j} \int_{\sigma_0 - j\infty}^{\sigma_0 + j\infty} F(s)e^{st}ds = \frac{1}{2\pi j} \oint_{\mathcal{C}} F(s)e^{st}ds$$

with $\sigma_0 > 0 \in \mathbb{R}$ is a positive constant.

As discussed in the lecture notes (Section 1.4.7, p. 72) we must the Cauchy Residue Theorem (CRT), requiring closure of the contour $\mathcal{C}$ at $\omega j \to \pm j\infty$

$$\oint_{\mathcal{C}} = \int_{\sigma_0 - j\infty}^{\sigma_0 + j\infty} + \int_{\subset_\infty}$$

where the path represented by '$\subset_\infty$' is a semicircle of infinite radius. For a causal, 'stable' (e.g. doesn't "blow up" in time) signal, all of the poles of $F(s)$ *must* be inside of the Laplace contour, in the left-half s-plane.
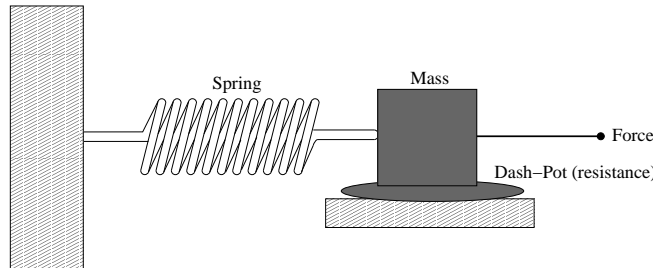
Figure E.1: *This three element mechanical resonant circuit consisting of a spring, mass and dash-pot (e.g., viscous fluid).*

Hooke's Law for a spring states that the force $f(t)$ is proportional to the displacement $x(t)$, i.e., $f(t) = Kx(t)$. The formula for a dash-pot is $f(t) = Rv(t)$, and Newton's famous formula for mass is $f(t) = d[Mv(t)]/dt$, which for constant M is $f(t) = Mdv/dt$.

The equation of motion for the mechanical oscillator in Fig. E.1 is given by Newton's second law; the sum of the forces must balance to zero

$$M\frac{d^2}{dt^2}x(t) + R\frac{d}{dt}x(t) + Kx(t) = f(t). \tag{E.1}$$

These three constants, the mass M, resistance R and stiffness K are all real and positive.. The dynamical variables are the driving force $f(t) \leftrightarrow F(s)$, the position of the mass $x(t) \leftrightarrow X(s)$ and its velocity $v(t) \leftrightarrow V(s)$, with $v(t) = dx(t)/dt \leftrightarrow V(s) = sX(s)$.

Newton's second law (c1650) is the mechanical equivalent of Kirchhoff's (c1850) voltage law (KCL), which states that the sum of the voltages around a loop must be zero. The gradient of the voltage results in a force on a charge (i.e., $F = qE$).

Equation E.1 may be re-expressed in terms of impedances, the ratio of the force to velocity, once it is transformed into the Laplace frequency domain.

The key idea that every impedance must be complex analytic and ≥ 0 for $\sigma > 0$ was first proposed by Otto Brune in his PhD at MIT, supervised by a student of Arnold Sommerfeld, Ernst Guilliman, an MIT ECE professor, who played a major role in the development of circuit theory. Brune's primary (non-MIT) advisor was Cauer, who was also well trained in 19th century German mathematics.[1]

---

[1]It must be noted, that Mac VanValkenburg, from the University of IL., was arguably more influential in circuit theory, during the same period. Mac's book are certainly more accessible, but perhaps less widely cited.

Appendix F

Transmission lines

In this problem, we will look at the transfer function of a two-port network, shown in Fig. F.1. We wish to model the dynamics of a freight-train having N such cars. The model of the train consists of masses connected by springs.

F.1 Transfer functions

The velocity *transfer function* for this system is defined as the ratio of the output to the input velocity. Consider the engine on the left pulling the train at velocity V_1 and each car responding with a velocity of V_n. Then

$$H(s) = \frac{V_N(s)}{V_1(s)}$$

is the frequency domain ratio of the last car having velocity V_N to V_1, the velocity of the engine, at the left most spring (i.e., coupler).
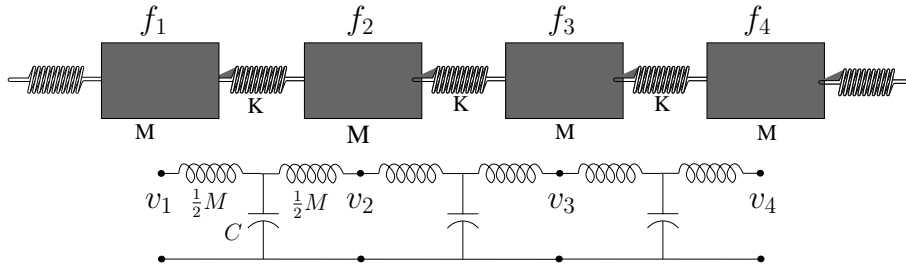
Figure F.1: *Depiction of a train consisting of cars, treated as a mass M and linkages, treated as springs of stiffness K or compliance $C = 1/K$. Below it is the electrical equivalent circuit, for comparison. The mass is modeled as an inductor and the springs as capacitors to ground. The velocity is analogous to a current and the force $f_n(t)$ to the voltage $v_n(t)$.*

F.2 The transmission matrix

Consistent with the figure, break the model into cells each consisting of three elements: a series inductor representing half the mass ($L = M/2$), a shunt capacitor representing the spring ($C = 1/K$), and another series inductor representing half the mass ($L = M/2$). Each cell is symmetric, making the model a cascade of identical cells.

To do: Use the ABCD method to find the matrix representation of Fig. F.1.

At each node define the force $f_n(t) \leftrightarrow F_n(\omega)$ and the velocity $v_n(t) \leftrightarrow V_n(\omega)$ at junction n.

1. Write the ABCD matrix $\boldsymbol{T}$ for a single cell, composed of series mass $M/2$, shunt compliance C and series mass $M/2$, that relates the input node 1 to node 2 where

$$\begin{bmatrix} F_1 \\ V_1 \end{bmatrix} = \boldsymbol{T} \begin{bmatrix} F_2(\omega) \\ -V_2(\omega) \end{bmatrix}$$

   Note that here the mechanical force F is analogous to electrical voltage, and the mechanical velocity V is analogous to electrical current.

$$\boldsymbol{T} = \begin{bmatrix} 1 & sM/2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \begin{bmatrix} 1 & sM/2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 + s^2MC/2 & (sM/2)(2 + s^2MC/2) \\ sC & 1 + s^2MC/2 \end{bmatrix} \qquad \text{(F.1)}$$

2. Assuming that $N = 2$ and that $F_2 = 0$ (two mass problem), find the transfer function $H(s) \equiv V_2/V_1$. From the results of the $\boldsymbol{T}$ matrix you determined above, find

$$H_{21}(s) = \left.\frac{V_2}{V_1}\right|_{F_2=0}$$

   From the lower equation we see that $V_1 = sCF_2 - (s^2MC/2 + 1)V_2$. Recall that $F_2 = 0$, thus

$$\frac{V_2}{V_1} = \frac{-1}{s^2MC/2 + 1} = \left(\frac{c_+}{s - s_+} + \frac{c_-}{s - s_-}\right).$$

   with $s_\pm = \pm\jmath\sqrt{\frac{2}{MC}}$ and $c_\pm = \pm\jmath/\sqrt{2MC}$.

3. Find $h_{21}(t)$, the inverse Laplace transform of $H_{21}(s)$.

$$h(t) = \oint_{\sigma_0-j\infty}^{\sigma_0+j\infty} \frac{e^{st}}{s^2MC/2 + 1} \frac{ds}{2\pi j} = c_+e^{-s_+t}u(t) + c_-e^{-s_-t}u(t).$$

   The integral follows from the CRT. The poles are at $s_\pm = \pm\jmath\sqrt{\frac{2}{MC}}$ and the residues are $c_\pm = \pm\jmath/\sqrt{2MC}$.

4. What is the input impedance $Z_2 = F_2/V_2$ if $F_3 = -r_0V_3$?

   Starting from T calculated above, find Z_2

$$Z_2(s) = \frac{F_2}{V_2} = T\begin{bmatrix} F_3 \\ -V_3 \end{bmatrix} = \frac{-(1 + s^2CM/2)r_0\cancel{V_3} - (sM/2)(2 + s^2CM/2)\cancel{V_3}}{-sCr_0\cancel{V_3} - (1 + s^2CM/2)\cancel{V_3}}$$

5. Simplify the expression for Z_2 with $N \to \infty$ by assuming that:
   1) $F_3 = -r_0V_3$ (i.e., V_3 cancels), 2) $s^2MC << 1$: 3) $r_0 = \sqrt{M/C}$

$$Z_2(s) = \frac{(1 + s^2CM/2)r_0 + (sM/2)(2 + s^2CM/2)}{sCr_0 + (1 + s^2CM/2)} \approx \frac{r_0 + sM}{sCr_0 + 1}$$

   Assumption 3 gives

$$Z_2 = \frac{sM + \sqrt{\frac{M}{C}}}{sC\sqrt{\frac{M}{C}} + 1} = r_0\frac{1 + s\sqrt{MC}}{1 + s\sqrt{MC}}$$

   The constant r_0 is called the *characteristic impedance.* This shows that below the cutoff frequency (approximation 2), the system approximates a transmission line.

6. State the ABCD matrix relationship between the first and Nth node, in terms of of the cell matrix.

$$\begin{bmatrix} F_1 \\ V_1 \end{bmatrix} = \boldsymbol{T}^N \begin{bmatrix} F_N(\omega) \\ -V_N(\omega) \end{bmatrix}$$

7. Given a $\boldsymbol{T}$ (ABCD) transmission matrix, the eigenvalues are and vectors are given in Appendix C of the Notes (p. 143), repeated here.

Eigenvalues:

$$\begin{bmatrix} \lambda_+ \\ \lambda_- \end{bmatrix} = \frac{1}{2} \begin{bmatrix} (A+D) - \sqrt{(A-D)^2 + 4BC} \\ (A+D) + \sqrt{(A-D)^2 + 4BC} \end{bmatrix}$$

Due to symmetry, $A = D$, this simplifies to $\lambda_\pm = A \mp \sqrt{BC}$ so that the eigen matrix is

$$\Lambda = \begin{bmatrix} A - \sqrt{BC} & 0 \\ 0 & A + \sqrt{BC} \end{bmatrix}$$

Eigenvectors: The eigenvectors are

$$\begin{bmatrix} \boldsymbol{E}_\pm \end{bmatrix} = \begin{bmatrix} \frac{1}{2C} \left[(A - D) \mp \sqrt{(A-D)^2 + 4BC} \right] \\ 1 \end{bmatrix}.$$

The determinant of E is

$$\Delta_E = -\frac{1}{2C} \sqrt{(A-D)^2 + 4BC}.$$

In the symmetric case when $A = D$

$$\boldsymbol{E} = \begin{bmatrix} -\sqrt{\frac{B}{C}} & +\sqrt{\frac{B}{C}} \\ 1 & 1 \end{bmatrix}, \qquad \boldsymbol{E}^{-1} = \frac{1}{2} \begin{bmatrix} -\sqrt{\frac{C}{B}} & 1 \\ +\sqrt{\frac{C}{B}} & 1 \end{bmatrix},$$

and

$$\Delta_E = -2\sqrt{\frac{B}{C}} = -2r_o.$$

There are two important invariants for the transmission lines, the wave number $k = \sqrt{BC} = s/c$, and the characteristic resistance $r_o = \sqrt{B/C}$.

Transfer functions of the transmission line: The velocity transfer function $H_{N,1} = \frac{V_N}{V_1}$ is given by

$$\begin{bmatrix} F_1 \\ V_1 \end{bmatrix} = \boldsymbol{T}^N \begin{bmatrix} F_N(\omega) \\ -V_N(\omega) \end{bmatrix}$$

along with the eigenvalue expansion

$$\boldsymbol{T}^N = E\Lambda^N E^{-1} = E \begin{bmatrix} \lambda_+^N & 0 \\ 0 & \lambda_-^N \end{bmatrix} E^{-1}.$$

Appendix G

2D parabolic horn

For 2D cylindrical waves the area function is $A(x) = A_0 x$, (horn radius $r \propto \sqrt{x}$) thus the Webster horn equation reduces to the cylindrical wave equation

$$\mathcal{P}_{xx}(x,\omega) + \frac{1}{x}\mathcal{P}_x(x,\omega) = \kappa^2(s)\mathcal{P}(x,\omega)$$

having Bessel function primitive solutions

$$\mathcal{P}^+(x,\omega) = J_0(\omega x/c) - iY_0(\omega x/c) = H_0^+(-j\kappa x)$$

and

$$\mathcal{P}^-(x,\omega) = J_0(\omega x/c) + iY_0(\omega x/c) = H_0^-(-j\kappa x),$$

where J_0 and Y_0 are the standard Bessel (and Neumann) functions, and $H_0^{\mp}(x)$ are the Hankel function of the first (-) and second (+) kind (respectively), all of order zero (indicated by the subscript) (Salmon, 1946a; Olson, 1947; Morse and Feshbach, 1953).

2D Radiation Admittance: Given a half-infinite section of parabolic horn, from Eq. 5.7[1]

$$Y_{rad}^{\pm}(x,s) = \mp\frac{y_0}{\kappa}\frac{\partial}{\partial x}\ln H_0^{\pm}(-j\kappa x) = \mp jy_0\frac{H_1^{\pm}}{H_0^{\pm}}. \tag{G.1}$$

2D ABCD Transmission matrix: Based on Eq. 5.18

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = \frac{-1}{\Delta_L}\begin{bmatrix} 1 & \mathcal{P}_{-L}^- \\ Y_{rad}^+ & -Y_{rad}^-\mathcal{P}_{-L}^- \end{bmatrix}_0 \begin{bmatrix} Y_{rad}^- & -1 \\ Y_{rad}^+\mathcal{P}_L^+ & \mathcal{P}_L^+ \end{bmatrix}_L \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}. \tag{G.2}$$

Verify the following

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = -\frac{1}{\Delta_L}\begin{bmatrix} 1 & H_0^-(x-L) \\ Y_{rad}^+ & -Y_{rad}^-H_0^-(x-L) \end{bmatrix}_{x=0} \begin{bmatrix} Y_{rad}^- & -1 \\ Y_{rad}^+H_0^+(x) & H_0^+(x) \end{bmatrix}_{x=L} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}, \tag{G.3}$$

where the subscript on each matrix indicates the value of x at which it is evaluated.

Impedance matrix:

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{P}_L \end{bmatrix} = \frac{1}{\mathcal{C}(s)}\begin{bmatrix} \mathcal{A}(s) & 1 \\ 1 & \mathcal{D}(s) \end{bmatrix}\begin{bmatrix} \mathcal{V}_0 \\ \mathcal{V}_L \end{bmatrix}. \tag{G.4}$$

---

[1]Note that

$$\frac{\partial \ln H_0^{\pm}(kx)}{\partial x} = -k\frac{H_1^{\pm}(kx)}{H_0^{\pm}(kx)}$$
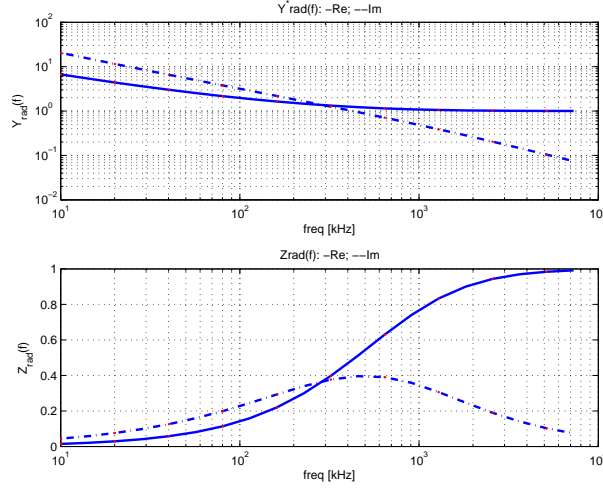
Figure G.1: *Real and imaginary parts of the radiation admittance Y_{rad}/y_0 (upper) and impedance (lower), as a function of frequency, for a infinite section of horn, as computed from Eq. G.1. The real and imaginary parts of the impedance are very similar to the results of Olson (1947) shown in Fig. 5.1. The real and imaginary parts of the admittance are much simpler to characterize: the imaginary part is a mass (i.e., is $\propto 1/s$) while the real part $\to 1$ above the critical frequency (where the real and imaginary parts are equal) at ≈ 0.3 [kHz].*

G.1   3D Conical Horn

For each horn we must find the natural normalization from the range variable to the normalize range variable x. For the conical horn the radius is proportional to the range variable r thus

$$A(r) = 4\pi \sin^2 (\Theta/2) \ r^2 \ \ [m^2].$$

The angle Θ is a measure of the solid (cone) angle. When $\Theta = \pi$ we have the case of the entire sphere so the solid angle is 4π [steradians] and the area is $4\pi r^2$. The formula for the area may be simplified by defining $A_\theta \equiv 4\pi \sin^2(\Theta/2) \, r_0^2 \ \ [m^2]$, resulting in the more convenient relation

$$A(r) = A_\theta \ (r/r_0)^2 \ \ [m^2].$$

This is a bit tricky because A_θ is not a constant since it depends on the place where the area was normalized, in this case r_0.

Using the conical horn area in *Eq.* 1.106 results in the spherical wave equation [Appendix H.4, (Olson, 1947; Morse, 1948; Benade, 1988)]

$$\mathcal{P}_{rr}(r,\omega) + \frac{2}{r}\mathcal{P}_r(r,\omega) = \kappa^2\mathcal{P}(r,\omega) \qquad\qquad\qquad (G.5)$$

(one must remember the steradian scale factor A_θ). Here $F(r) = \partial_r \ln A(r) = \frac{2}{r}$ (see Table 5.1).

Primitive solutions: We may now identify the two primitive d'Alembert solutions of Eq. G.5 (see Appendix H)

$$\mathcal{P}^\pm(r,s) = \mathcal{P}^\pm_\theta(r,s) \ e^{\mp\kappa(s)r}/r.$$

These are also known as the *zero order spherical Hankel functions* (i.e., $h_0^\pm(\kappa r)$).

We wish to normalize these primitive solution such that they are 1 at the source locations $r = r_0$ and $r = r_L$ (i.e., $\mathcal{P}^+(r_0,s) = 1$ and $\mathcal{P}^-(r_L,s) = 1$). The normalized primitive solutions are

$$\mathcal{P}^+(r,s) \equiv \frac{r_0}{r}e^{-\kappa(r-r_0)} \qquad\qquad \mathcal{P}^-(r,s) \equiv \frac{r_L}{r}e^{+\kappa(r-r_L)},$$

as may be easily verified by letting $r = r_0$ in $\mathcal{P}^+(r,s)$ and $r = r_L$ in $\mathcal{P}^-(r,s)$.
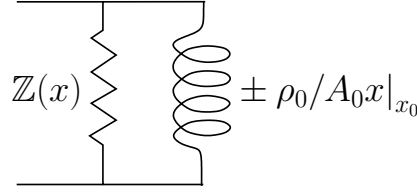
Figure G.2: *The equivalent circuit for the radiation impedance $Z_{rad}^+(s, r)$ for a conical is a parallel combination of the radiation resistance $\mathbb{Z}(r) = \rho_0 c/A_\theta(r)$ and a mass given by $M(r) = \pm\rho_0 r/A_\theta$. Note that the inbound primitive wave has a negative mass! This means that the inbound wave is **not** the time-reverse version of the outbound wave. I take this to be due to the different curvature of the wavefronts of the waves.*

Radiation Admittance: Given a half-infinite section of conical horn Eq. 5.7 gives

$$Y_{rad}^\pm(r, s) = \mp \frac{y_0(r)}{\kappa(s)} \frac{\partial \ln \mathcal{P}^\pm}{\partial r}\bigg|_{r_0, r_L} = y_0(r)\left(1 \pm \frac{c}{rs}\right) \leftrightarrow y_0(r)\left(\delta(t) \pm \frac{c}{r}u(t)\right), \quad \text{(G.6)}$$

Here are the gory details:

$$Y_{rad}^+(r_0, s) = -\frac{y_0}{\kappa}\frac{\partial \ln \mathcal{P}^-}{\partial r}\bigg|_{r_0} = \frac{y_0}{\kappa}\frac{\partial}{\partial r}(\kappa r + \ln r)\bigg|_{r_0} = y_0(r)\left(1 + \frac{c}{sr_0}\right)$$

$$Y_{rad}^-(r_L, s) = +\frac{y_0}{\kappa}\frac{\partial \ln \mathcal{P}^+}{\partial r}\bigg|_{r_L} = \frac{y_0}{\kappa}\frac{\partial}{\partial r}(\kappa r - \ln r)\bigg|_{r_L} = y_0(r)\left(1 - \frac{c}{sr_L}\right)$$

where $\kappa(s) = s/c$, $\delta(t)$ and $u(t)$ are the Dirac and Heaviside functions, respectively.

As shown in Fig. G.2 there are two additive admittance terms, a real part equal to the characteristic conductance $y_0(r)$, and a mass, equal to $\pm\mathbb{Z}r/c = \pm\rho_0 r/A_\theta$. For the reverse traveling primitive wave, these conventions result in *negative mass*, an interesting and perhaps surprising result. Since the admittances are summed, the impedances appear in parallel.

Figure 5.1 from Olson (1947, Fig. 5.3, p. 101) provides the real and imaginary impedance for each of the horns discussed here. The difference between Olson and the present work, is in our intent here to explain the physics behind the reactive term, in the time domain.

3D ABCD Transmission matrix: For clarity the relations are re-derive "from scratch"

$$\begin{bmatrix} \mathcal{P}(r) \\ \mathcal{V}(r) \end{bmatrix} = \begin{bmatrix} \frac{r_0}{r}e^{-\kappa(r-r_0)} & \frac{r_L}{r}e^{+\kappa(r-r_L)} \\ Y_{rad}^+(r)\frac{r_0}{r}e^{-\kappa(r-r_0)} & -Y_{rad}^-(r)\frac{r_L}{r}e^{+\kappa(r-r_L)} \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \quad \text{(G.7)}$$

Evaluate at the boundary r_L

$$\begin{bmatrix} \mathcal{P}(r_L) \\ \mathcal{V}(r_L) \end{bmatrix} = \begin{bmatrix} \frac{r_0}{r_L}e^{-\kappa(r_L-r_0)} & \frac{r_L}{r_L}e^{+\kappa(r_L-r_L)} \\ Y_{rad}^+(r_L)\frac{r_0}{r_L}e^{-\kappa(r_L-r_0)} & -Y_{rad}^-(r_L)\frac{r_L}{r_L}e^{+\kappa(r_L-r_L)} \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad \text{(G.8)}$$

=

$$\begin{bmatrix} \mathcal{P}_L \\ \mathcal{V}_L \end{bmatrix} = -\begin{bmatrix} -\frac{r_0}{r_L}e^{-\kappa(r_L-r_0)} & -1 \\ -Y_{rad}^+(r_L)\frac{r_0}{r_L}e^{-\kappa(r_L-r_0)} & Y_{rad}^-(r_L) \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \quad \text{(G.9)}$$

and taking the inverse, gives the weights α, β and the boundary (load) impedance[2]

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}_{r_L} = \frac{-1}{\Delta_L}\begin{bmatrix} Y_{rad}^-(r_L) & -1 \\ Y_{rad}^+(r_L)\frac{r_0}{r_L}e^{-\kappa(r_L-r_0)} & +\frac{r_0}{r_L}e^{-\kappa(r_L-r_0)} \end{bmatrix}\begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix} \quad \text{(G.10)}$$

---

[2]The three red signs have been changed to assign $\mathcal{V}_L$ into the port.

where

$$-\Delta_L = \left[Y^-_{rad}(r_L) + Y^+_{rad}(r_L) \right] \frac{r_0}{r_L} e^{-\kappa(r_L - r_0)} = 2y_0(L) \frac{r_0}{r_L} e^{-\kappa r_L}.$$

Substituting Eq. G.10 back into Eq. G.7

$$\begin{bmatrix} \mathcal{P}(r) \\ \mathcal{V}(r) \end{bmatrix} = \frac{-1}{\Delta_L} \begin{bmatrix} \frac{r_0}{r} e^{-\kappa(r-r_0)} & \frac{r_L}{r} e^{+\kappa(r-r_L)} \\ Y^+_{rad}(r) \frac{r_0}{r} e^{-\kappa(r-r_0)} & -Y^-_{rad}(r) \frac{r_L}{r} e^{+\kappa(r-r_L)} \end{bmatrix} \begin{bmatrix} Y^-_{rad}(r_L) & -1 \\ Y^+_{rad}(r_L) \frac{r_0}{r_L} e^{-\kappa(r_L-r_0)} & +\frac{r_0}{r_L} e^{-\kappa(r_L-r_0)} \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}.$$
(G.11)

Evaluating this at the throat $r = r_0$ we obtain the final ABCD matrix in factored form[3]

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = \frac{-1}{\Delta_L} \begin{bmatrix} 1 & \frac{r_L}{r_0} e^{+\kappa(r_0-r_L)} \\ Y^+_{rad}(r_0) & -Y^-_{rad}(r_0) \frac{r_L}{r_0} e^{+\kappa(r_0-r_L)} \end{bmatrix} \begin{bmatrix} Y^-_{rad}(r_L) & -1 \\ Y^+_{rad}(r_L) \frac{r_0}{r_L} e^{-\kappa(r_L-r_0)} & +\frac{r_0}{r_L} e^{-\kappa(r_L-r_0)} \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}.$$
(G.12)

At this point it is not difficult to compute $\mathcal{A}(s), \mathcal{B}(s), \mathcal{C}(s), \mathcal{D}(s)$.

For reference, the general case (Eq. 5.18) is[4]

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = \frac{-1}{\Delta_L} \begin{bmatrix} 1 & \mathcal{P}^-_L \\ Y^+_{rad}(r_0) & -Y^-_{rad}(r_0)\mathcal{P}^-_L \end{bmatrix} \begin{bmatrix} Y^-_{rad}(r_L) & -1 \\ Y^+_{rad}(r_L)\,\mathcal{P}^+_L & \mathcal{P}^+_L \end{bmatrix} \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}.$$
(G.13)

Impedance matrix:

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{P}_L \end{bmatrix} = \frac{1}{\mathcal{C}(s)} \begin{bmatrix} \mathcal{A}(s) & 1 \\ 1 & \mathcal{D}(s) \end{bmatrix} \begin{bmatrix} \mathcal{V}_0 \\ \mathcal{V}_L \end{bmatrix}.$$
(G.14)

G.2   Exponential Horn

Starting from the basic transmission line equations with an area function given by $A(x) = A_0 e^{2mx}$, Eq.5.4 is

$$\frac{\partial^2 \mathcal{P}(x,\omega)}{\partial x^2} + 2m \frac{\partial \mathcal{P}(x,\omega)}{\partial x} = \kappa^2 \mathcal{P}(x,\omega),$$
(G.15)

thus $F(x) \equiv \frac{1}{A(x)} \frac{\partial A(x)}{\partial x} = 2m$ is a constant. Since Eq. G.15 is an ordinary constant coefficient differential equation, it has a closed form solution (Olson, 1947; Salmon, 1946a,b; Morse, 1948; Beranek, 1954; Leach, 1996; Beranek and Mellow, 2012). By the substitution $\mathcal{P}(x,\omega) = P(\kappa(s))e^{-\kappa_\pm(s)x}$, one may solve for the *characteristic roots* $\kappa_\pm(s) = m \pm \sqrt{m^2 + \kappa^2}$. Thus

$$\mathcal{P}^\pm(x) = e^{-mx} e^{\mp\sqrt{m^2+\kappa^2}x} = e^{-mx} e^{\mp j\sqrt{\omega^2-\omega_c^2}x/c},$$
(G.16)

which represent the horn's right (+) and left (-) traveling pressure waves. The dispersion diagram corresponding to the horn is shown in Fig. G.3. Note that the exponential horn has a dispersion diagram identical to the electron's wave properties in a semiconductor. *Simply by changing the flare function from conical to exponential, the horn impedance switches from a mass in parallel with a resistor, to a horn in cutoff, with conduction and stop band (an evanescent wave band region), as in a semiconductor.*

The outbound time domain solution of the exponential (exp) horn is $e^{-mx}E(t)$, with[5]

$$E(t) = \delta(t - x/c) + \frac{x}{c} \frac{J_1(\sqrt{t^2 - x^2/c^2})}{\sqrt{t^2 - x^2/c^2}} U(t - x/c) \leftrightarrow e^{-\sqrt{m^2+\kappa^2}x}.$$
(G.17)

---

[3]It is possible that this is useful in this format, say if $\beta = 0$.

[4]Notation changed to match that of Eq. G.12.

[5]See PowerWavesE paper for a more detailed expression.
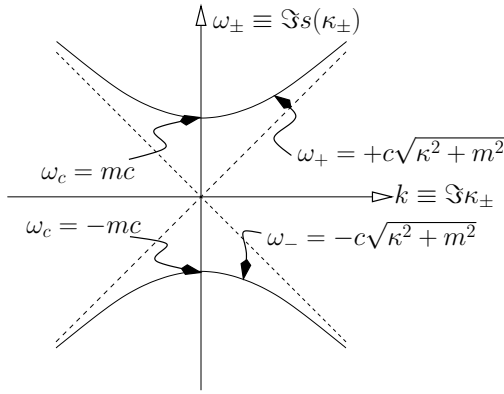
Figure G.3: *This shows the dispersion relation between the frequency and the wave number k for the exponential horn, with wave propagation disallowed below the critical frequency $\omega_c = mc$.*

Exp-Horn ABCD Transmission matrix:

From Eq. 18

$$\begin{bmatrix} \mathcal{P}_0 \\ \mathcal{V}_0 \end{bmatrix} = -\frac{1}{\Delta_L} \begin{bmatrix} 1 & \mathcal{P}^- \\ Y_{rad}^+ & -Y_{rad}^- \mathcal{P}^- \end{bmatrix}_0 \begin{bmatrix} Y_{rad}^- & -1 \\ Y_{rad}^+ \mathcal{P}^+ & \mathcal{P}^+ \end{bmatrix}_L \begin{bmatrix} \mathcal{P}_L \\ -\mathcal{V}_L \end{bmatrix}. \tag{G.18}$$

Radiation Admittance: Given a half-infinite section of the exponential horn (Salmon, 1946b; Leach, 1996), from Eq. 5.7

$$Y_{rad}^{\pm}(x_0, s) = \frac{y_0}{s} \left(\omega_c \mp \sqrt{\omega_c^2 + s^2} \right). \tag{G.19}$$

where $\omega_c \equiv mc$ is the horn *cutoff frequency*. At very high frequencies this approaches a real value of y_0. Below cutoff it is a purely reactive admittance.

Discussion: Since Eq. G.15 contains no viscous or other loss terms, the solution is always loss-less. The propagation function roots $\kappa_{\pm}(s)$ are imaginary when $\omega > \omega_c$, but change to a purely real value below the *cutoff frequency*, i.e., $\omega < \omega_c$.

At all frequencies the wave propagation is dispersive, meaning that the speed is frequency dependent. Above the cutoff frequency there is normal wave propagation. However the impedance of the wave changes, making the horn an "ideal" acoustic transformer. As shown in Fig. fig:DispersionDiagram, at very high frequencies ($\omega \geq \omega_c = mc$) the wave propagates without dispersion, but still with a decay, due to the exponential change in area.

When the diameter of the horn becomes greater than half a wavelength, higher order modes start to come into play, and the solution is no longer unique, as the quasistatic approximation totally breaks down. Any imperfection in the area will introduce cross modes, which can then propagate.

In regions where the diameter is less than a half wavelength, higher order modes will not propagate. This is best described by the quasistatic approximation. But the exact solution may be obtained by solving Eq. eq:WHEN locally for the pressure.

Below cutoff requires further analysis, as the wave solution is still causal, but *everywhere in phase*. To see this we may take the inverse Laplace transform of $P^+(x, s)$ to obtain the explicit wave behavior in the time domain. It is very useful to look at this case in both the time and frequency domains, in order to fully appreciate what is happening.

Appendix H

Derivation of the Webster Horn Equation

In this Appendix we transform the acoustic equations Eq. 1.99 (p. 99)

$$-\nabla \varrho(\mathbf{x}, t) = \rho_o \frac{\partial}{\partial t} \boldsymbol{u}(\mathbf{x}, t) \tag{H.1}$$

and 1.100 (p. 99)

$$-\nabla \cdot \boldsymbol{u} = \frac{1}{\eta_0 P_0} \frac{\partial}{\partial t} \varrho(\boldsymbol{x}, t) \tag{H.2}$$

into their equivalent integral form. This derivation is similar to that of Hanna and Slepian (1924); Pierce (1981, p. 360), with the difference that the present result is exact, not the common approximation.

Conservation of momentum: The first step is an integration the normal component of Eq. H.1 over the pressure iso-surface $\mathcal{S}$, defined by $\nabla p = 0$,

$$-\int_{\mathcal{S}} \nabla p(\mathbf{x}, t) \cdot d\mathbf{A} = \rho_0 \frac{\partial}{\partial t} \int_{\mathcal{S}} \mathbf{u}(\mathbf{x}, t) \cdot d\mathbf{A}. \tag{H.3}$$

The *average pressure* $\varrho(x, t)$ is defined by dividing by the total area

$$\varrho(x, t) \equiv \frac{1}{A(x)} \int_{\mathcal{S}} p(x, t) \, \hat{\boldsymbol{n}} \cdot d\boldsymbol{A}. \tag{H.4}$$

From the definition of the gradient operator (Greenberg, 1988, p. 769)

$$\frac{\partial p}{\partial x} = \nabla p \cdot \hat{\boldsymbol{n}}, \tag{H.5}$$

where $\hat{\boldsymbol{n}}$ is a unit vector perpendicular to the iso-pressure surface $\mathcal{S}$. This follows from the fact that pressure is constant $\perp$ to the flow, which is maximum in the direction of the gradient. Thus the left side of Eq. H.3 reduces to $\partial \varrho(x, t)/\partial x$.

The surface integral on the right side of Eq. H.3 defines the *volume velocity*

$$\nu(x, t) \equiv \int_S \mathbf{u}(x, t) \cdot d\boldsymbol{A}. \tag{H.6}$$

Thus Eq. H.3 reduces to

$$\frac{\partial}{\partial x} \varrho(x, t) = -\frac{\rho_0}{A(x)} \frac{\partial}{\partial t} \nu(x, t).$$

Taking the $\mathcal{L}$ gives

$$\frac{d}{dx} \mathcal{P}(x, \omega) = -s \frac{\rho_o}{A(x)} \mathcal{V}(x, \omega). \tag{H.7}$$

Figure H.1:   *Derivation of the horn equation using Gauss's Law: The divergence of the velocity $\nabla \cdot \mathbf{u}$, within δx, shown as the filled shaded region, is integrated over the enclosed volume. Next the divergence theorem is applied, transforming the integral to a surface integral normal to the surface of propagation. This results in the difference of the two volume velocities $\delta \nu = \nu(x+\delta x) - \nu(x) = [\mathbf{u}(x+\delta x) \cdot \mathbf{A}(x+\delta x) - \mathbf{u}(x) \cdot \mathbf{A}(x)]$.*
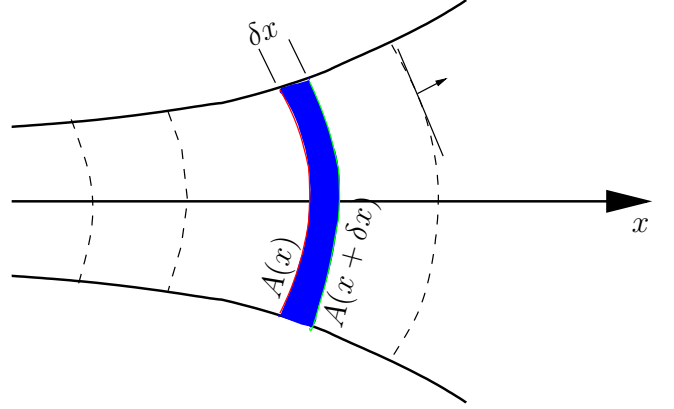
Conservation of mass: Integrating Eq. H.2 over the volume $\mathcal{V}$ gives

$$-\int_{\mathcal{V}} \nabla \cdot \boldsymbol{u} \, dV = \frac{1}{\eta_0 P_0} \frac{\partial}{\partial t} \int_{\mathcal{V}} p(\mathbf{x}, t) dV. \tag{H.8}$$

Volume $\mathcal{V}$ is defined by two iso-pressure surfaces between x and $x + \delta x$ (Fig. H.1, shaded blue). On the right hand side we use our definition for the *average pressure* (i.e., Eq. H.4), integrated over the thickness δx.

Applying Gauss' law to the left hand side[1], and using the definition of ϱ (on the right), in the limit $\delta x \to 0$, gives

$$\frac{\partial \nu}{\partial x} = -\frac{A(x)}{\eta_0 P_0} \frac{\partial \varrho}{\partial t}.$$

Transforming this to the frequency domain gives

$$\frac{\partial}{\partial x} \mathcal{V}(x, \omega) = -s \frac{A(x)}{\eta_0 P_0} \mathcal{P}(x, \omega). \tag{H.9}$$

Expressing Eq. H.7 and Eq. H.9 in matrix form, gives the desired result, Eq. 1.106 (p. 100).

H.1   The inverse problem

For each area function $A(x)$ there is a unique driving point input impedance $z_{rad}(t) \leftrightarrow Z_{rad}(s)$. We define $z_r(t)$ as the *impedance remainder* after normalizing by the surge impedance z_0 and then subtracting the initial impulse (delta function)

$$z_r(t) \equiv z_{rad}(t)/z_0 - \delta(t). \tag{H.10}$$

In the frequency domain $Z_r(s) = Z_{rad}(s)/z_0 - 1$. The remainder is related to the reactive portion of $Z(s)$.

For each remainder there is a corresponding unique $A(x)$. This relation is

$$f(a, \xi_0) + \frac{1}{2c} \int_{-a}^{a} z_r(|\xi_0/c - \xi/c|) f(\xi_0, \xi) d\xi = 1. \tag{H.11}$$

Once this integral equation has been solved for $f(x, x)$, the area function is $A(x) = f^2(x, x)$ (Sondhi and Gopinath, 1971; Sondhi and Resnick, 1983).

H.2   WKB method

This method is use in approximating the solution to the horn equation under the assumption that the reflections are minimal (no strong change in the area function with the range variable. Briefly, if

---

[1]As shown in Fig. H.1, we convert the divergence into the difference between two volume velocities, namely $\nu(x+\delta x) - \nu(x)$, and $\partial \nu / \partial x$ as the limit of this difference over δx, as $\delta x \to 0$

$$\frac{d^2}{dx^2}\Phi(x,s) = F(x)\,\Phi(x,s)$$

where the effective *wave number* is

$$F(x) = \frac{2m}{\hbar^2}\left[V(x) - E\right].$$

The WKB approximation gives the solution

$$\Phi(x,s) \approx C_0 \frac{e^{j\int \sqrt{F(x)}dx}}{\sqrt[4]{F(x)}}.$$

Picking the right branch of the square root is essential in the application of the WKB method, as is unwrapping the phase of each branch of the wave number.

This method is integrating the phase delay of the wave number as a function of the range variable. Thus its complex-analytic properties are essential to obtaining the correct answer (the wave number is not a single valued function of x and s).

H.3   Rydberg series

Fundamental the quantum mechanics is the Rydberg series, which describes the quantized energy levels of atoms

$$\nu_{n,m} = cRZ_n^2\left(\frac{1}{n^2} - \frac{1}{m^2}\right) \tag{H.12}$$

where $\nu_{n,m}$ are the possible eigen-frequencies, c is the speed of light, $R \approx 1.097 \times 10^7$ is the Rydberg constant, Z_n is the atomic number, along with positive integers n and m ($m > n$) which represent two *principal quantum numbers* that label all possible allowed atomic eigen-states. Integer n indicates the lowest (initial) atomic eigen-state while m is the higher state[2] When $n = 1$ the series is the Lyman series corresponding to Hydrogen ($Z_1 = 1$).

Given observed frequencies $\nu_{n,m}$ it is possible to find the area function that traps the photons into the Rydberg eigen-states.

One way to think of eigen-modes is to make an analogy to a piano string, or an organ pipe. In these much simpler systems, there is an almost constant delay, say τ due to a characteristic length, say $L = \tau c$, such that the eigen-modes are given by integer multiples of a half wavelength $\nu_n = nc/2L$. This assumes the endpoint boundary conditions are pinned displacement (i.e, zero velocity). For an organ pipe, closed at one end and open at the other the corresponding formula is multiples of a quarter wavelength $\nu_n = nc/4L$. In each case $\nu = n/\tau$ where $\tau = 2L/c$ is the round trip delay, thus $\nu = nc/2L$. We suggest looking at the Rydberg series in the say way, but with the very different eigen frequencies (Eq. H.12). Sommerfeld (1949, p. 201) makes a very interesting comment regarding Eq. H.12:

> This equation reduces to a simple mathematical the enigma of the spectral lines, with their finite cummulation point, the behavior of which differs so fundamentally from that of all mechanical systems.

H.4   Laplacian operator in N dimensions

To show that the Webster equation is in agreement with the wave equation in 2 and three dimensions, we need to express the Laplacian, and then determine $F(n)$ of Table 5.1. In general it may be shown that, in N dimensions (Sommerfeld, 1949, p. 227)[3]

$$\nabla_r^2 P \equiv \frac{1}{r^{N-1}}\frac{\partial}{\partial r}\left(r^{N-1}\frac{\partial P}{\partial r}\right)$$

---

[2] `http://en.wikipedia.org/w/index.php?title=Rydberg_formula&action=edit&section=2`.

[3] `http://en.wikipedia.org/wiki/Laplacian#Three_dimensions`

This may be expanded as

$$\frac{\partial^2 P}{\partial r^2} + F(r)\frac{\partial P}{\partial r},$$

where $F(r) = (N-1)/r$ is the same as in Table 5.1. In terms of the Webster Horn equation, $F(x) \equiv -\partial \ln \mathscr{Z}(x,s)/\partial x$. Thus we see there is a fundamental relation between $\mathscr{Z}$ and the dimensionality of the horn.

For each N, r^{N-1} is proportional to the area function $A(x)$. This generalizes to the Webster Horn equation Eq. 1.106. For the case of $N = 3$ (the conical Horn) there is a special relationship

$$\nabla_r^2 P \equiv \frac{1}{r^2}\partial_r r^2 \partial_r P = \frac{1}{r}\partial_{rr} rP$$

resulting in the general d'Alembert solutions

$$P^{\pm}(r,s) = e^{\mp\kappa(s)r}/r.$$

Appendix I

Filter classification

Let $z \equiv e^{sT}$ where T is the "sample period" at which data is taken (every T seconds). For example if $T = 22.7 = 1/44100$ seconds then the data is sampled at 44100 k [Hz]. This is how a CD player works with high quality music. Thus the *unit-time delay operator* z^{-1} as

$$\delta(t - T) \leftrightarrow e^{-sT}$$

I.1 Filter properties

Given the function

$$F(s) = \frac{(s+1)(s-1)}{(s+2)},$$

1. find the minimum phase $M(s)$ and all-pass $A(s)$ parts. The minimum phase part has all of its poles and zeros in the left half plane (LHP), while the all-pass part has its poles in the LHP and mirrored zeros in the RHP. Thus we place a removable pole zero pair symmetrically across from the RHP zero, and then write the expression as the product, that is $F(s) = M(s) \cdot A(s)$:

$$F(s) = \frac{(s+1)(s-1)}{(s+2)} \cdot \frac{s+1}{s+1} = \frac{(s+1)^2}{s+2} \cdot \frac{s-1}{s+1}$$

   Thus $M(s) \equiv \frac{(s+1)^2}{s+2}$ and $A(s) \equiv \frac{s-1}{s+1}$

2. Find the magnitude of $M(s)$ Take the real part of the log of M and then the anti-log. Thus $|M| = e^{\Re \ln M(s)}$

3. Find the phase of $M(s)$ In this case we use the imaginary part: $\angle M = \Im \ln M(s)$

4. Find the magnitude of $A(s)$ 1, by definition.

5. Find the phase of $A(s)$ $\angle A = \Im \ln(A)$

I.2 More questions

There are a number of question I would like to address in this report. Some of these are

1. Can we interpret the zeta function as a frequency domain quantity, and then inverse transform it into the time domain?

   The answer to this is yes, and the results are quite interesting.

2. Make a histogram of the entropy for the first million integers.

   This is a 5 minute job in Matlab. It goes something line this: `K=1e5; N=1:K; F=zeros(K,10);`
   `for n=1:K; f=factor(n); F(n,1:length(f))=f; end; hist(F);`

I.3   Entropy analysis

According to the *Fundamental theorem of arithmetic*, every integer may be uniquely written as the product of primes. Thus $n = \prod_k^{K_n} p_i^{\alpha_k}$ (e.g., $12 = 2^2 \cdot 3$ with $\alpha_2 = 2$ and $\alpha_3 = 1$). If one thinks of K_n as a random variable on n, then the constant K_n may be characterized by the concept of *entropy*. Thus each integer is associated with an entropy, defined by $\mathcal{H}_n = \sum^{K_n} \log_2(K_n)$.

Appendix J

Stillwell's Intersection vs. composition

```
--------------------------------------------------------------------------------------------
Page/c Group         Comments
6BCE   -------------------------------------------------------------------------------------
44     Pythagoreans  Pyth's used recursion to solve PE: y_n/x_n -> \sqrt{2}
46     Pythagoreans  2x2 recursion -> sqrt(2): possibly arose from Euclidean algorithm?
45     Greeks        Find sides of 1xsqrt(2) rectangle with Euclidean Alg => Matrix recurrence relation
46     Greeks        Other instances of PE occurred in Greek Math (anthyphairesis)
55     Pythagoreans  First use of 2x2 recurrence gave sols to Pell's Eq. (N=2): $x^2-2y^2= \pm1$


3BCE   -------------------------------------------------------------------------------------
47     Archimedes    Cattle problem: x^2 - 427,434 y^2 = 1


7CE    -------------------------------------------------------------------------------------
46     Indians       EA = "Pulverizer"
37     Bramagupta    "Methods for solving PE first discovered by Indian Mathematicians" (Ch 5, p.69)
69     Bramagupta    Rediscovered PE, both recurrence and composition methods
76     Bramagupta    Composition sol method of Pell's Eq (limited success, specific cases only)


12CE   -------------------------------------------------------------------------------------
78     Bh\^askara II Cyclic sol for N=61 $\in \P$ (1160CE)


17CE    ------------------------------------------------------------------------------------
46     Brouncker     CFA for sqrt{N} cite(Dickson 1920)
79     Fermat        Rediscovered N=61 case (hard case p. 80) (1657)
80     Fermat        Related to continued fractions and $sqrt{N}$ irrational case
                     The initial conditions determines the hidden large solutions
                     Discussion of CFA and Pell's Eq. "Method of finding 1" (Ex. 5.5.1).
       Newton        Works out Diophantus' chord/tangent methods (Stillwell p. 7, 49, 218)


18CE   -------------------------------------------------------------------------------------
46     Lagrange      First to prove why cyclic methods worked (Andrew Weil, 1984) (1768)
       --------------------------------------------------------------------------------------------
```

List of Figures

Bibliography

Batchelor, G. (1967). *An introduction to fluid dynamics.* Cambridge University Press, Cambridge, United Kingdom.

Benade, A. (1988). Equivalent circuits for conical waveguides. *J. Acoust. Soc. Am.*, 83(5):1764–9.

Beranek, L. L. (1954). *Acoustics.* McGraw–Hill Book Company, Inc., 451 pages, New York.

Beranek, L. L. and Mellow, T. J. (2012). *Acoustics: Sound Fields and Transducers.* Academic Press - Elsevier Inc., Waltham, MA.

Boas, R. (1987). *An invitation to Complex Analysis.* Random House, Birkhäuser Mathematics Series.

Bode, H. (1945). *Network Analysis and Feedback Amplifier Design.* Van Nostrand, New York.

Bombieri, E. (2000). Problems of the millennium: The Riemann hypotheis. *RiemannZeta-ClayDesc.pdf*, pages 1–11. Date unknown.

Boyer, C. and Merzbach, U. (2011). *A History of Mathematics.* Wiley.

Brillouin, L. (1953). *Wave propagation in periodic structures.* Dover, London. Updated 1946 edition with corrections and added appendix.

Brillouin, L. (1960). *Wave propagation and group velocity.* Academic Press. Pure and Applied Physics, v. 8, 154 pages.

Brune, O. (1931a). *Synthesis of a finite two-terminal network whose driving point impedance is a prescribed function of frequency.* PhD thesis, MIT.
`http://dspace.mit.edu/bitstream/handle/1721.1/10661/36311006.pdf`.

Brune, O. (1931b). Synthesis of a finite two-terminal network whose driving point impedance is a prescribed function of frequency. *J. Math and Phys.*, 10:191–236.

Calinger, R. S. (2015). *Leonhard Euler: Mathematical Genius in the Enlightenment.* Princeton University Press, Princeton, NJ, USA.

Campbell, G. (1903). On loaded lines in telephonic transmission. *Phil. Mag.*, 5:313–331. See Campbell22a (footnote 2): In that discussion, ... it is tacitly assumed that the line is either an actual line with resistance, or of the limit such with R=0.

Campbell, G. (1922). Physical theory of the electric wave filter. *Bell System Tech. Jol.*, 1(2):1–32.

Campbell, G. (1937). *The collected papers of George Ashley Campbell.* AT&T, Archives, 5 Reiman Rd, Warren, NJ. Forward by Vannevar Bush; Introduction by E.H. Colpitts.

Carlin, H. J. and Giordano, A. B. (1964). *Network Theory, An Introduction to Reciprocal and Nonreciprocal Circuits.* Prentice-Hall, Englewood Cliffs NJ.

Fagen, M., editor (1975). *A history of engineering and science in the Bell System — The early years (1875-1925)*. AT&T Bell Laboratories.

Feynman, R. (1970). *Feynman Lectures On Physics*. Addison-Wesley Pub. Co.

Fine (2007). *Number Theory: An Introduction via the Distribution of Primes*, chapter 4, pages 133–196. Birkhäuser Boston, Boston, MA.

Fletcher, N. and Rossing, T. (2008). *The Physics of Musical Instruments*. Springer New York.

Fry, T. (1928). *Probability and its engineering uses*. D. Van Nostrand Co. INC., Princeton NJ.

Goldsmith, A. N. and Minton, J. (1924). The performance and theory of loud speaker horns. *Radio Engineers, Proceedings of the Institute of*, 12(4):423–478.

Goldstein, L. (1973). A history of the prime number theorem. *The American Mathematical Monthly*, 80(6):599–615.

Gray, J. (1994). On the history of the Riemann mapping theorem. *Rendiconti Del Circolo Matematico di Palemrmo*, Serie II(34):47–94. Supplemento (34): 47–94, MR 1295591," Citation from `https://en.wikipedia.org/wiki/Riemann_mapping_theorem`.

Greenberg, M. D. (1988). *Advanced Engineering Mathematics*. Preentice Hall, Upper Saddle River, NJ, 07458.

Grosswiler, P. (2004). Dispelling the alphabet effect. *Canadian Journal of Communication*, 29(2).

Hanna, C. R. and Slepian, J. (1924). The function and design of horns for loud speakers. *American Institute of Electrical Engineers, Transactions of the*, XLIII:393–411.

Helmholtz, H. (1863a). *On the sensations of tone*. Dover (1954), 300 pages, New York.

Helmholtz, H. (1863b). Ueber den einfluss der reibung in der luft auf die schallbewegung. *Verhandlungen des naturhistoisch-medicinischen Vereins zu Heildelberg*, III(17):16–20. Sitzung vom 27.

Helmholtz, H. (1978). On integrals of the hydrodynamic equations that correspond to vortex motions. *International Journal of Fusion Energy*, 1(3-4):41–68. English translation 1978 of Helmholtz 1858 paper, translated by Uwe Parpart. The image is on page 41.

Hill, J. M. (2007). Laplace transforms and the riemann zeta function. *Integral Transforms and Special Functions*, 18(3):193–205. jhill@uow.edu.au.

Hunt, F. V. (1952). *Electroacoustics, The analysis of transduction, and its historical background*. The Acoustical Society of America, 260 pages, Woodbury, NY 11797.

Hunt, F. V. (1982). *Electroacoustics, The analysis of transduction, and its historical background*. The Acoustical Society of America, 260 pages, Woodbury, NY 11797.

Kennelly, A. E. (1893). Impedance. *Transactions of the American Institute of Electrical Engineers*, 10:172–232.

Kim, N. and Allen, J. B. (2013). Two-port network analysis and modeling of a balanced armature receiver. *Hearing Research*, 301:156–167.

Kim, N., Yoon, Y.-J., and Allen, J. B. (2016). Generalized metamaterials: Definitions and taxonomy. *J. Acoust. Soc. Am.*, 139:3412–3418.

Kirchhoff, G. (1868a). On the influence of heat conduction in a gas on sound propagation. *Ann. Phys. Chem.*, 134:177–193.

Kirchhoff, G. (1868b). On the influence of heat conduction in a gas on sound propagation. *Ann. Phys. Chem.*, 134:177–193.

Lamb, H. (1932). *Hydrodynamics.* Dover Publications, New York.

Leach, W. (1996). A 2-port analogous circuit and SPICE model for Salmon's family of acoustic horns. *J. Acoust. Soc. Am.*, 99(3):1459–1464.

Lighthill, S. M. J. (1978). *Waves in fluids.* Cambridge University Press, England.

Lin, J. (1995). The Needham Puzzle: Why the Industrial Revolution did not originate in China. In Christiansen, B., editor, *Economic Behavior, Game Theory, and Technology in Emerging Markets*, volume 43(2), chapter Economic Development and Cultural Change, pages 269–292. PNAS and IGI Global, 701 E Chocolate Ave., Hershey, PA. DOI:10.1073/pnas.0900943106 PMID:19131517, DOI:10.1086/452150.

Mason, W. (1928). The propagation characteristics of sound tubes and acoustic filters. *Phy. Rev.*, 31:283–295.

Mawardi, O. K. (1949). Generalized solutions of webester's horn theory. *J. Acoust. Soc. Am.*, 21(4):323–30.

McMillan, E. M. (1946). Violation of the reciprocity theorem in linear passive electromechanical systems. *Journal of the Acoustical Society of America*, 18:344–347.

Miller, D. A. (1991). Huygens's wave propagation principle corrected. *Optics Letters*, 16(18):1370–2.

Montgomery, C., Dicke, R., and Purcell, E. (1948). *Principles of Microwave Circuits.* McGraw-Hill, Inc., New York.

Morse, P. and Feshbach, H. (1953). *Methods of theoretical physics. Vol. I. II.* McGraw-Hill, Inc., New York.

Morse, P. M. (1948). *Vibration and sound.* McGraw Hill, 468 pages, NYC, NY.

Newton, I. (1687). *Philosophiae Naturalis Principia Mathematica.* Reg. Soc. Press.

Olson, H. F. (1947). *Elements of Acoustical Engineering (2^n Edition).* D. Van Nostrand Company, 539 pages.

Orfanidis, S. J. (2009). *Electromagnetic Waves and Antennas.* web, `http:www.ece.rutgers.edu/~orfanidi/ewa`.

Pierce, A. D. (1981). *Acoustics: An Introduction to its Physical Principles and Applications.* McGraw-Hill, 678 pages, New York.

Pipes, L. A. (1958). *Applied Mathematics for Engineers and Physicists.* McGraw Hill, NYC, NY.

Pleshko, P. and Plačz, I. (1969). Experimental observatons of Sommerfeld and Brillouin precursors in the microwave domain. *Phy. Rev. Let.*, 22(22):1201–4.

Ramo, S., Whinnery, J. R., and Van Duzer, T. (1965). *Fields and waves in communication electronics.* John Wiley & Sons, Inc., New York.

Rayleigh, J. W. (1896). *Theory of Sound, Vol. I 480 pages, Vol. II 504 pages.* Dover, New York.

Riemann, B. (1851). *Grundlagen für eine allgemeine Theorie der Funktionen einer veränderlichen complexen Grösse [An alternative basis for a general theory of the functions of complex variables].* PhD thesis, University of Göttingen.

Riemann, B. (1859). On the number of primes less than a given magnitude. *Monatsberichte der Berliner Akademie*, pages 1–10. Translated by David R. Wilkins-1998.

Salmon, V. (1946a). Generalized plane wave horn theory. *J. Acoust. Soc. Am.*, 17(3):199–211.

Salmon, V. (1946b). A new family of horns. *J. Acoust. Soc. Am.*, 17(3):212–218.

Schelkunoff, S. (1943). *Electromagnetic waves*. Van Nostrand Co., Inc., Toronto, New York and London. 6th edition.

Schwinger, J. S. and Saxon, D. S. (1968). *Discontinuities in waveguides : notes on lectures by Julian Schwinger*. Gordon and Breach, New York, United States.

Shaw, E. (1970). Acoustic horns with spatially varying density or elasticity. *J. Acoust. Soc. Am.*, 50(3):830–840.

Sommerfeld, A. (1949). *Partial differential equations in Physics, Lectures on Theoretical Physics, Vol. I*. Academic Press INC., New York.

Sommerfeld, A. (1952). *Electrodynamics, Lectures on Theoretical Physics, Vol. III*. Academic Press INC., New York.

Sondhi, M. and Gopinath, B. (1971). Determination of vocal-tract shape from impulse responses at the lips. *J. Acoust. Soc. Am.*, 49(6:2):1867–73.

Sondhi, M. and Resnick, J. (1983). The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis. *J. Acoust. Soc. Am.*, 73(3):985–1002.

Stillwell, J. (2002). *Mathematics and its history; Undergraduate texts in Mathematics; 2d edition*. Springer, New York.

Stillwell, J. (2010). *Mathematics and its history; Undergraduate texts in Mathematics; 3d edition*. Springer, New York.

Van Valkenburg, M. (1960). *Introduction to Modern Network Synthesis*. John Wiley and Sons, Inc., New York.

Van Valkenburg, M. (1964a). *Network Analysis second edition*. Prentice-Hall, Englewood Cliffs, N.J.

Van Valkenburg, M. E. (1964b). *Modern Network Synthesis*. John Weily & Sons, Inc., New York, NY.

Vanderkooy, J. (1989). A model of loudspeaker driver impedance incorporating eddy currents in the pole structure. *JAES*, 37(3):119–128.

Walsh, J. (1973). History of the Riemann mapping theorem. *The American Mathematical Monthly*, 80(3):270–276,. DOI: 10.2307/2318448.

Webster, A. G. (1919). Acoustical impedance, and the theory of horns and of the phonograph. *Proc. Nat. Acad. Sci.*, 5:275–282.

Youla, D. (1964). Analysis and synthesis of arbitrarily terminated lossless nonuniform lines. *IEEE Transactions on Circuit Theory*, 11(3):363–372.

Index

$\boldsymbol{\nabla} \times ()$, 160
$\boldsymbol{\nabla} \cdot ()$, 160
$\mathbb{F}$: See fractional number, 23
$\nabla ()$, 160
$\zeta(s)$ function: See Riemann zeta function, 91
`zviz.m`, see: plots, colorized, 144
`conv(A,B)`, 54
`polyval(P,X)`, 54
`residue(N,D)`, 54
`root(P)`, 54
1–port, 74
2–port, 74

abacus, 22, 115
admittance: see impedance, 63
analytic, 29, 31, 44, 47, 48, 50
analytic functions, 76
analytic geometry, 55, 58
analytic series, 47

Bézout's theorem, 136
Bardeen, 50
Beranek, 116
Bernoulli, Daniel, 45
Bernoulli, Jakob, 18
Bernoulli, Johann, 18, 75
Bethe, 50
Bhâskara II, 39
bilinear transformation, 54, 61, 65, 141, 142, 147, 169
Boas, 116
Bode, 116
Bombelli, 19
Brahmagupta, 16, 22, 39, 40, 115
branch cut, 81, 82, 143
Brillouin, 45, 50, 101, 116
Brillouin zones, 79, 131
Brune impedance, 89

Cantor, 23
Cauchy, 31, 45, 63, 81
Cauchy integral formula, 85
Cauchy integral theorem, 85
Cauchy residue theorem, 85

causality, 31, 67, 74, 145, 146
CFA, 36
Chebyshev, 30
chord–tangent methods, 63, 64, 117, 123
codomain, 49
colorized plots: see "plots, colorized", 93
complex analytic, 76, 80–82, 143
complex analytic functions, 63, 64, 66, 77, 81
complex analytic functions: see analytic, 49
complex analytic: see analytic, 49
Complex numbers, History of, 26
composition, 21, 30, 40, 42, 53, 55, 58, 61, 62, 64, 65, 136, 137
compressible, 110
conjugate variables, 101
conservation of energy, 14, 43, 147, 148
conservative field, 78
convolution, 51, 69, 87
coprime, 22, 23, 117
Curl $\boldsymbol{\nabla} \times ()$, 31, 107, 160

d'Alembert, 13, 18, 20, 44, 45, 50, 74, 75, 79, 98, 113
deconvolution, 22
degrees of freedom, 108
Descartes, 19, 20, 50, 74, 129
determinant, 57, 61, 97, 168
DFT, 67
Dicke, 116
Diophantus, 15, 16, 39
dispersion relation, 79
dispersive wave propagation, 44
Divergence $\boldsymbol{\nabla} \cdot ()$, 31, 107, 160
DoF: see degrees of freedom, 108
domain, 49
domain-coloring: see "plots, colorized", 66
dot product, 51
driving-point impedance, 146
DTFT, 67

eigen-values of d/dt, 88
eigenfunction expansion, 126
Einstein, 45, 75
elimination, 55