A Dialogue on Numerical Methods

David C. Wilson

October 17, 2007

The School of Athens, by Raphael (1483-1520)

Fresco Vatican, Stanza della Segnatura, Rome

Contents

Remarks by the Author

Topics and Clientele

> Keep the interests of the students in mind and the rest will work itself out.–Bill Harris, NSF

The goal of this set of notes is to present mathematical topics selected from numerical analysis, which are suitable for a semester course at the upper level undergraduate level. The topics have been organized thematically under the headings of root finding and approximation theory. The discussion of root finding techniques includes the square root method of Archimedes/Heron, the method of Newton/Raphson, the bisection method, and the contraction mapping theorem. The discussion of approximation theory includes the topics of Taylor's Theorem, polynomial approximation, least squares, Fourier Series, splines, and wavelets. The Pythagorean Theorem and the concept of orthogonality provide a unifying overarching theme which appears throughout. The topics have been selected with the idea that they will be particularly relevant for students in computer science, electrical engineering, and computer engineering.

Since engineering students are typically inexperienced, untrained, and uninterested in formal mathematics, the subject of numerical methods has a sad reputation for being a dull, difficult, and irrelevant requirement for graduation. In the numerous times I have taught this course, I have not infrequently encountered the attitude: "This is my last math course–hopefully." In particular, I have found teaching a course on numerical methods a pedagogical challenge because students lack the

required mathematical training to appreciate the discussions. In one class, I noticed that one of my engineers was visibly resistant to the proof of a key theorem. On further questioning it became evident that he saw no justification for his time being wasted in such an exercise. For some reason, I finally asked "What is the difference between a definition and a theorem?" His response was "Aren't they the same?" I was startled to think that a student, who had passed three semesters of Calculus as well as semester courses in Linear Algebra and Differential Equations could make such a statement. Even the teachings of Euclid were beyond this fellow. Unfortunately, he is not alone. Since that experience, I now regularly confront such issues on the first day of class by asking the following basket of questions:

1. "Why do we have definitions and theorems?"
2. "What is a conditional sentence?"
3. "What is the structure of a theorem?"
4. 'What is the difference between the way a mathematician and a statistician uses the word hypothesis?"
5. "What is a mathematical system?"
6. "Why should anyone care?" (This question is the most important!)

I try to answer these questions by giving short expositions on basic propositional logic and the ramifications of Euclid's famous 5^{th} Postulate. After one such introduction, a computer science student, a native of Southeast Asia, stated she was shocked by the remedial level of the discussion. She left and never returned.

In case you are thinking I am prejudiced against the engineering students, let me mention that my math majors also have deficiencies when taking more applied courses. One extremely bright and talented student (also from Southeast Asia) earned an almost perfect score on every exam. However, when asked to write five lines of computer code to approximate the square root of a number, she was helpless. In

general, the engineers complain about the theory and clamor for more projects, while the mathematics students thrive on the theory and wish the projects were not a part of the course. Thus, I have found that the instructor of an applied mathematics course should be alert to the differing needs of the students, while at the same time not getting derailed repairing too many deficits.

In my experience, the single most important reason students find numerical analysis dull, boring, and difficult is their lack of skill and knowledge in Logic, Geometry, and Linear Algebra. A second reason is their inability to connect the theory with some aspect of their expected future employment. The "Interview" has been included in an effort to address these issues. For students who have been away from mathematics for a long time, I have included many other brief reviews throughout the notes.

While the focus of the discussion is on the mathematics, the goal is to present a readable account of the thought behind the theory in a manner that will be appreciated by a large subset of the students. The approach is to present the material as a historical progression of ideas motivated by key examples and easy-to-understand special cases. Hopefully, this approach will help neutralize negative attitudes and better meet the needs of the students.

A Brief History of the Dialogue Format

Mathematics is written for mathematicians. – Nicolaus Copernicus

With a quick glance through the these notes, the reader will immediately notice that they are written in a dialogue format. Surely the author must be joking. Why would anyone waste his/her time writing a mathematics textbook in dialogue format? Why would anyone waste hard-earned money purchasing such a volume? Galileo as a central character in the discussion? However, that is exactly what is offered: an allegorical presentation of real mathematical ideas.

Let us begin our defense by noting that numerous books from antiquity were

written as dialogues. Plato (427-347 B.C.E.) wrote virtually all his works in this dramatic style. In his "Apology," he dramatizes the fatal conflict between Socrates and his enemies Meletus, Anytus, and Lycon, who had accused him of "corrupting the youth." For this crime, Socrates receives the ultimate punishment. In his "Allegory of the Cave," Plato tries to clarify the concepts of intellect, belief, and knowledge. In this dialogue, he chains prisoners in an underground cave, where they see only shadows cast on the wall in front of them and hear only echoes from behind. This allegory dramatizes the fundamental human conflict that we can never know reality. His commentaries on ethics, politics, astronomy, and mathematics were also written as dialogues.

In 1632, Galileo (1564-1642) published his "Dialogue Concerning the Two Chief World Systems: Ptolemaic and Copernican," [3] where he dramatizes the scientific conflict between two different mathematical models of the solar system. Simplicio, his spokesman for the Aristotle/Ptolemaic earth-centric view of the universe, plays the role of a foil to Salviati, who advocates the Copernican view that the sun is the center of the solar system. A third character, Sagredo, plays the role of the forward looking aristocrat, who considers both sets of arguments, but consistently ends up siding with Salviati. In the narrative, Salviati presents observations of the ocean tides, the moons of Jupiter, and the phases of Venus as evidence that the Earth moves. The main reason for his use of the dialogue format was to present the case for the Copernican view while pretending to be impartial. Of course, this ruse failed to protect him from the wrath of the Inquisition of Pope Urban VIII (1568-1644). On 22 June 1633, he was found guilty of heresy and sentenced to house arrest for the remainder of his life.

In 1638, Galileo published a second dialogue "Dialogues Concerning Two New Sciences," [5]. In this work he again presents the same three characters in a four day discussion of fundamental concepts in two key areas of modern Physics. The focus of the discussion for the first two days is on the strength of materials. The focus for the second two days is on the behavior of a falling object. While Galileo's style is

again engaging, the style of this second volume is more mathematically challenging than the first. Much of the writing is in an definition, theorem, proof format, where the reader is subjected to numerous difficult mathematical arguments. (Most of these discussions are geometric in nature.) On the first day, he even considers several of the paradoxes of infinitesimals and infinity, which arise in his discussion of strength of such materials as copper wire, glass, marble, and rope. At the beginning of the fourth day, his Proposition I concludes that the path of a falling object describes a parabola. Later in the same day, his Proposition VIII asserts the familiar physics/calculus fact that a projectile fired from a cannon at a 45 degree angle will travel farther than when fired at any other angle. While much of the complexity of these arguments can be reduced if armed with a knowledge of modern calculus, the discussions remain fresh to this day. For example, on the second day Salviati argues that a giant cannot be arbitrarily sized in the same proportion as a smaller creature unless the bones are made from a stronger material. Thus, real physical reasons exist that explain why the largest mammals reside in the great oceans of the world.

A number of modern authors have also employed a dialogue format in their mathematical writings. In 1895, Lewis Carroll (1838-1898) published "What the Tortoise Said to Achilles," where the discussion elucidates the subtleties of the logical argument of modus ponens. In particular, he addresses the logical problem of self-referencing. (The easiest example of self-referencing is to consider the truth or falsity of the statement: "I am lying." Think about it.) In 1963-64, the Hungarian mathematician/economist/historian Imre Lakatos (1922-1974) published four articles entitled "Proofs and Refutations." (The articles were published as a book in 1976 [8].) In this small set of dialogues, the author creates a classroom setting through conversations between a teacher and a small group of students. The teacher is named Teacher and the students are named Alpha, Beta, Gamma, etc. Through their interactions the reader is drawn into the world of mathematical rigor. The concepts of axioms, definitions, and theorems are discussed through a question/answer format, where the focus of the mathematics is Euler's famous theorem that $V - E + F = 2$ for any

polyhedral 2−sphere, where V, E, F denote the number of vertices, edges, and faces, respectively. While mathematical rigor, logic, proof, examples and counterexamples (i.e. refutations) are central, Lakatos teaches the process of formulating carefully worded definitions and theorems so that ambiguity or vagueness are removed. As the discussion shows, if you are sloppy or careless with your wording, a counterexample to what you had expected may be lurking nearby. Alfred Renyi (1921-1970) was one of the outstanding Hungarian mathematicians and statisticians of the 20^{th} Century. He even has an institute constructed in his honor. In 1965, he published "Dialogues on Mathematics," [8] where Socrates, Archimedes, King Hieron II, and Galileo are featured discussing such subjects as "pure versus applied mathematics." On occasion, he even performed these works with his daughter. His best known quote is "A mathematician is a machine for converting coffee into theorems." (Another Hungarian, Paul Erdös, has also received credit for this quote.) In his 1974 dialogue "Surreal Numbers," [7] Donald Knuth strands two ex-students, Bill and Alice, on an isolated beach. Bored and lonesome, they find happiness in mathematics (and a touch of romance) through a highly rigorous discussion of the properties of the real number system. In 1979, Douglas Hofstadter expanded on Lewis Carroll's discussion of of self-referencing in his highly popular Pulitzer Prize winning book "Gödel, Escher, Bach" [4], where he makes connections between a myriad of subjects including logic, art, music, computer programming, the nature of language, the nature of thought, the replication of our genetic code, Turing machines, artificial intelligence, and free will. Dialogues between Achilles, the Tortoise, the Anteater, the Crab, and Charles Babbage interlace this book of ideas. Most recently, Keith Kendig has written the book "Conics" [6], where a Teacher, a Philosopher, and a Student uncover the properties of the conics through an engaging and readable dialogue. The Philosopher is looking for unity and beauty, the student loves stories, and the teacher provides the details. Along the way, questions are asked and mathematical discoveries are made.

The inspiration behind the dialogue format set forth in these notes is Dava Sobel's book "Galileo's Daughter," [10]. While most books on Galileo (1564-1642) provide

an account of his scientific achievements and/or his political problems, the focus of Sobel's book is his relationship with his eldest daughter, Virginia (1600-1633). While Galileo had two other children, Virginia was probably his favorite. She was bright, beautiful, serious, and passionately devoted to her father. Since she was illegitimate (as were his other two children), marriage was problematic. Thus, at the age of 16 she followed the respectable alternative of the times by dutifully taking vows as Suor Maria Celeste at the convent of St. Mateo in Padova, Italy. (The name Celeste is derived from celestial and is probably an indirect reference to Galileo's astronomical discoveries.) Life at the convent was dominated by prayer, never ending chores, and grinding poverty. Despite their separation and difficult circumstances, the father and daughter adored each other. She provided him with aid and comfort when he was ill and wrote him continually during their extended separations. In return, Galileo never failed to respond to her requests for money. Sobel speculates that this dutiful daughter may have assisted in the preparation of his dialogue "Two Chief World Systems." One can only wonder what she might have achieved if she had been more fortunate in her birthright.

A downside to the dialogue format is a lack of economy. Since mathematics lives perfectly well in its own sparse setting, the experienced instructor or reader may find the conversational style not only unnecessary, but also distracting and irritating. If this is the case, simply move on to a new topic. The author has no intention that someone would teach word-for-word what is written in these notes. What is written here contains too much of one individual instructors own classroom style.

Cultural Impacts on Pedagogy

> We note that a huge body of evidence attests to the fact that a society's values are passed from generation to generation through a process of transmission which may be vertical (from parents) or oblique (from others in the prior generation) and involves a psychological internalization of

values. –Karl Marx

How does society optimize the transfer of mathematical knowledge and skills from one generation to the next? While the educators, politicians, and media have spent inordinate quantities of time, thought, and cash trying to address this issue, my view is that the answers lie in the culture of the community, the reward system for those involved, and the method of delivery. Needless to say these three forces are not unrelated.

If a community values finance, fashion, and football more than mathematics and science, then guess what? The resources and talent of the community will flow into those more preferred areas. Sometimes political events change the behavior of a community. Before the rise of the Nazis, mathematics training in America was almost nonexistent at every level. With the immigration of prominent scientists to the United States in the 1930's, interest in mathematics began to rise. In 1957 the Russians changed science forever by launching Sputnik. This event provided the impetus for educators to launch advanced science and mathematics courses in high schools throughout the United States. The "New Math" was part of this Cold War effort to catch up. In 1962, John F. Kennedy's push to land a man on the moon created an excitement that boosted the production of PhD mathematicians to never before seen levels. The study of mathematics in America was transformed from being worst to first. Students and young faculty now came from all over the world to study in America. Unfortunately, only a short time later the excitement began ebbing back to the historical mean. In the 1970's, the concern became: How are we going to find employment for all these mathematicians? In the 1990's, the concern refocused to: Why does a kid in a far-off land perform better on standardized math exams than those in America? Recently, I quizzed a number of (excellent) Chinese graduate students on this issue. I asked whether or not their mothers pushed them to excel. Their response was that not only did their parents insist they study hard, but the expectation was uniform among their friends so negotiation was not part of the equation. When they performed well, they were rewarded. Their parents had also given

them a choice: They could study or they could work. In a culture where education was a privilege, not a right and where drudge labor was the norm, the connection was clear. Thus, parents, prestige, and profit combined to create an environment where they became driven by internal forces. My students from Eastern Europe, Russia, India, and South America are driven by similar pressures. In all these cultures, math is easy when compared with the alternatives.

So what incentives are available for motivating students in today's world? While the excitement of the space race and the new math have evaporated and the economies of the world are doing reasonably well, a plethora of new gadgets, technologies, and issues have exploded in their place. Calculators are everywhere. Imaging Science is a field that permeates medicine and the military. Environmental (e.g. global warming), public safety (e.g. hurricane tracking), and public health issues (e.g. the spread of AIDS) abound. These new areas all require appropriately chosen numerical methods and models. Since engineers enjoy projects that impact society, a focus of this dialogue is to connect the abstract mathematical ideas to as many applications as possible.

Pedagogy as a Process

> Knowing something for oneself or for communication to an expert colleague is not the same as knowing it for explanation to a student. –Hyman Bass

While mathematicians are expected to write in a definition-theorem-proof style that is clear, rigorous, and lean, I have found few undergraduate (or even graduate) students, who can retain much from this style of information transfer. Instead, I prefer to present modern mathematics as a naturally unfolding "Socratic process," where simple questions and observations lead to fundamental insights. The key is to formulate and answer clearly stated questions, which get to the heart of the problem. If you "Begin with the easiest problem you don't understand," then the solution to one

problem often leads to new questions and new answers which lead to new solutions. Simple observations evolve into ever more general and abstract concepts. These abstract general results become more accessible and easier to understand. The dialogue format provides a mechanism which can be used to capture this spirit of discovery. The question "What does it mean for a technique to work?" leads to a precise definition of the rules of the game. In my experience, students typically find definitions an unnecessary and pedantic annoyance. A mathematicians attitude is that you can't play the game until you have a precise statement of the rules. The question "Does the technique always work?" frequently leads to examples demonstrating a negative answer. These examples lead to the question "When does the technique work?" The response of the mathematician is to formulate a theorem or proposition, which provides exact conditions when a positive result can be guaranteed. The question "Can the method be generalized?" may lead to a technique that can be applied to a wider range of problems. Once a generalization has been formulated the process repeats itself.

The Contraction Mapping Theorem of Stephan Banach (1906-1960) is a notable example of this evolution from simple to abstract. Without reference to the ancient Archimedes/Heron square root algorithm and the Newton/Raphson root finding technique 1700 years later, this theorem lacks seems to emerge from nowhere. Approximation theory provides a second progression of ideas, where the topics presented include: polynomials, Fourier, splines, and wavelets. In each case, orthogonality (or lack thereof) is fundamental to the success (or failure) of the technique. Since orthogonality is nothing but a fancy way of saying perpendicular, the Pythagorean Theorem is at the heart of the discussion. The fact that root finding and approximation took several thousand years to unfold indicates the richness of the ideas underlying the techniques. Our approach is to use this rich history to drive the discussion. Armed with an understanding of this mathematical process, the hope is that the reader should be better able to evaluate, select, and apply numerical methods in their own endeavors.

While not as important as the development of mathematical ideas, I find that students also enjoy mathematical gossip. By introducing cartoon versions of some the great contributors to mathematics, I am hoping the reader can begin to appreciate some of their quirky personalities. Probably my favorite story is Fourier's personal interest in the heat equation. In short, after an enjoyable visit to sunny Egypt with Napoleon in 1798, Fourier returned to the miserable rain and snow of Grenoble's winter, where he turned up the heat in his apartment to the highest setting. Thus stimulated, he developed stable methods for solving the heat equation. Such anecdotes lead to the questions: "Who cares?" and "Why would anyone be interested in solving these types of problems?" George Polya (1887-1985) also endorses this "journalistic" approach to pedagogy when he remarks that your five best friends are What, Why, Where, When, and How [9]. I would also add Who. Thus, the mathematical ideas are embedded in an interactive discussion of the background, significance, and historical context of the subject. In my experience, I have found that my engineering and medical students find this approach an agreeable alternative to the more traditional one, where they are stuffed with facts, formulas, and techniques like the overfed goose headed for the dinner table as "paté de foie gras."

In addition to presenting the theoretical ideas as a process, we have followed the lead of G. Polya in our discussion of examples and problems. In his book "How to Solve It," [9], he spells out a general four step process for solving a mathematic's problem:

1. understanding the problem,

2. devise a plan,

3. carry out the plan, and

4. look back and review what was done.

This process provides a student with a structure and framework for attacking a problem. Probably the best example of this approach is our treatment of limit problems,

where we insist students are able to know and apply the definition of a limit. In the problems we consider, the plan is always the same. Each solution requires three simple steps. While students argue that they should not be expected to know this skill, they soon find that they are far easier than the problems connected with real applications. As you will read many times in these notes: "Math is easy. It is life that is difficult."

Murphy's Law

What can go wrong, will go wrong. –Murphy

While logic and rigor are fundamental to the spirit of mathematics, computer scientists, engineers, and physicists turn to mathematics for techniques to mathematically analyze and model real-world phenomena. Students from these fields may enjoy the study of mathematics, but are driven by the needs of their particular application. Unfortunately, the curriculum has become so crowded that most instruction in these applied areas becomes "technique driven" rather than "process driven." In other words, the instructor presents the formulas and techniques, but hurries on to the next topic before discussing history, insights, or caveats associated with the method. However, in my experience, I have found Murphy's Law to be the one guiding principle that rules the study of numerical methods. In these notes, key examples have been provided to help the student identify the numerous tar pits that are forced on the subject. Hopefully, the student will develop a wariness when employing these and other techniques in their own investigations.

A Final Comment

And yet it moves. –Galileo

While Galileo's book, "Two Chief World Systems," contained thinly veiled political statements not in accord with the dogma of his times, the dialogue strategy failed

to keep him out of harm's way. For on 22 June 1633 the wrath of Pope Urban VIII descended upon him when the Holy Inquisition convicted him of heresy and subjected him to life imprisonment (later commuted to house arrest). If he had not been so famous and had not abjured himself, he might have been burned at the stake as was his predecessor, the heretic Giordano Bruno (1548-1600). It was not until 31 October 1992, after almost 13 years of investigation (including the testimony of Physicist Steven Hawking), that a commission appointed by Pope John Paul II admitted that "mistakes must be frankly recognized." And so it goes.

Acknowledgments

I doubt I would have ever taught a single mathematics class if it were not for the excellent education I received at Ithaca High School. My mathematics teachers for grades 9-12 were Miss Stenson, Miss DePew, and Miss Neighbour. (I have not included their first names because I have listed their names the way they appear in my old yearbooks. In a day when everyone is on a first name basis with their superiors, I find this formality refreshing.) In the 9^{th} grade, Miss Stenson taught us the quadratic formula. In the 10^{th} and 11^{th} grades, Miss DePew taught us Euclidean Geometry and Solid Geometry. Miss Neighbour taught us Analytic Geometry and 6 weeks of Calculus. While they were all excellent teachers, I adored Miss DePew. She always told us stories about Archimedes in the bathtub, the death of Archimedes, Napier and his bones, the history of Euclid's fifth postulate, and the young Gauss adding up the numbers from 1 to 100. She even tried to get us to discover the formula for the arithmetic series ourselves. (My recollection is that this experiment didn't work out too well.) She loved Geometry, where mathematical rigor was front and center. No sloppy thoughts were allowed in her class. She also had a fearsome intensity. When we did poorly on an exam, she did not hesitate to let us know. Fortunately for me, I sat near the back of the room and so could hide from her wrath. Of course, when we did well, her praise made you glow. In my 40 years of teaching several thousand students,

I have found only a handful with the training in the fundamentals of mathematics that equaled mine. She took her profession very seriously.

I have also extracted a multitude of photographs, quotes, and comments from the MacTutor History of Mathematics archive[2], which is based at the School of Mathematical and Computational Sciences at the University of St Andrews, Fife, Scotland. I found their database containing more than 1000 biographies of mathematicians to contain a gold mine of information.

Finally, I must also acknowledge my students, colleagues, friends, teachers, family, and assorted poets, who have unknowingly supplied much of the language that appears in these pages. I have stolen from them mercilessly.

Bibliography

[1] The California Council on Science and Technology Newsletter, *California Faces Critical Shortage of Math and Science Teachers,* http://www.ccst.us/news/2007/20070305TCPA.php, March 5, 2007

[2] *The MacTutor History of Mathematics archive,* http://www-history.mcs.st-andrews.ac.uk/history/index.html, May 2007.

[3] Galileo Galilei, *Dialogue on the Great World Systems in the Salusbury Translation,* Revised, Annotated, and with an Introduction by Giogio de Santillana, The University of Chicago Press, Chicago and London, 1953.

[4] Douglas R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid,* Vintage Books, New York, 1989.

[5] Stephen Hawking, *On the Shoulders of Giants: The Great Works of Physics and Astronomy,* pages 391-626, Running Press, Philadelphia, London, 2002.

[6] Keith Kendig, *Conics,* The Mathematical Association of America, 2005.

[7] D. E. Knuth, *Surreal Numbers,* Addison Wesley, 1974.

[8] Imre Lakatos, *Proofs and Refutations,* Cambridge University Press, Cambridge, New York, 1976.

[9] G. Polya, *How to Solve It,* Princeton University Press, First Princeton Science Library Edition, Princeton and Oxford, 1988.

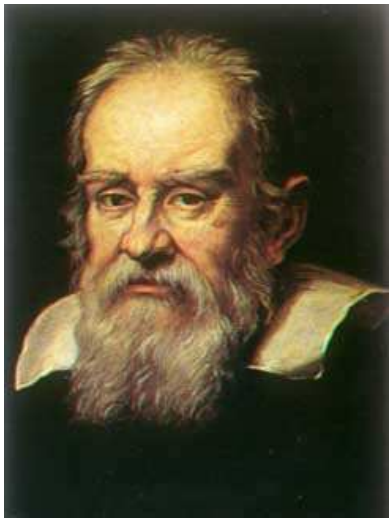
[10] Dava Sobel, *Galileo's Daughter,* Penguin Books, New York, 2000.

Part I

Day 1. The Interview

Chapter 1

Introductions



> The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in mathematical language, and the letters are triangles, circles and other geometrical figures, without which means it is humanly impossible to comprehend a single word. -Galileo Galilei (1564-1642)

The Setting:

The time is the present. Galileo sits at his desk absorbed in a manuscript. A small glass of Chianti rests nearby. Enter Virginia and Simplicio. Galileo looks up.

Galileo: And what brings you to my office?

Virginia: We are interested in learning more science and mathematics.

Galileo: I submit that the study of these subjects is a noble and worthy goal. Virginia, who is this young fellow with you?

Virginia: I would like you to meet my new friend Simplicio.

Galileo: I am pleased to meet you Mr. Simplicio. I am sure you have found Virginia to be a gracious lady with exquisite manners and charm. She is one of my favorites.

Simplicio: Indeed I do enjoy her company.

Galileo: And if I may ask, what career goals do you have?

Virginia: I am interested in teaching mathematics.

Simplicio: I would like to become more knowledgeable about important applications. An understanding of numerical methods seems to be a requirement for my future employment.

Galileo: Very interesting, but why?

Simplicio: I am not sure, but several prospective employers have mentioned data. It seems they are overloaded with data and having trouble making any sense of it. They recommended I discuss these issues with you. It seems you are the master of data.

Galileo: I am flattered. Others have not been so kind. It sounds like you have talked to someone, who requires a knowledge and skill in data acquisition, storage, and analysis techniques. Is that correct?

Simplicio: One company builds devices, which acquire and analyze signals for the military. One builds medical imaging equipment. One is in communications. One is in the business of compressing images.

Galileo: So, you are ready to journey through a mathematically rigorous study of these topics?

Simplicio: Unlike yourself, I do not enjoy the rigor of mathematics.

Galileo: I am sorry to hear that. I find the beauty, oder, and clarity of mathematical ideas a refreshing contrast to the sloppy thinking that surround us.

Chapter 2

Science, Models, and Applications

> From the same principles, I now demonstrate the frame of the System of the World.-Isaac Newton
>
> A job is death without dignity. –Dylan Thomas

Simplicio: While I have no objection to rigor for others, my reason for this visit is to learn techniques useful in my employment.
Galileo: Do I detect that "rigor" and "employment" are concepts separated by a void?
Simplicio: To be honest, I find mathematics to be difficult, boring, and irrelevant. I search for a job, where the pay is good and the work not too stressful.
Galileo: You are an honest man.
Simplicio: I always make an effort to be direct. What skills do we need?
Galileo: Over the ages, the ancient thinkers have developed numerical techniques to compute:

1. solutions to systems of linear equations,
2. solutions to systems of nonlinear equations,
3. derivatives

4. integrals,

5. eigenvalues and eigenvectors,

6. solutions to differential equations, and

7. solutions to partial differential equations.

While these methods are all useful, we are not going to have time to discuss them all. Choices must be made.

Simplicio: Which skills would an employer prefer?

Galileo: The big picture is that all these techniques are useful in setting up and solving mathematical models of physical phenomena. In short, these techniques are joined as the computational component of the scientific method. This simple, but severe test can be summarized as repeated iterations of the following procedure:

1. observational and/or experimental data is acquired,

2. a mathematical/statistical model is formulated, and

3. the model and the data are tested for agreement.

The reason for this process is to make predictions, which help answer the questions "when," "where," or "how much." Interestingly, sometimes the data comes first and stimulates the search for a model. The data I collected on the motion of a falling body showed that the motion can be modeled by a quadratic equation. Johannes Kepler (1571-1630) demonstrated that Tycho Brahe's data forced the conclusion that the orbit of Mars is an ellipse. Soon after, Isaac Newton proved that both these models can be explained as consequences of his laws of motion. This tour de force is unmatched in the history of science. On the other hand, sometimes the theory comes first. Albert Einstein's special theory of relativity wasn't confirmed by data until more than a decade after the discovery. In both scenarios, confirmation of agreement is key. Each time new data is acquired, the accuracy of the model is reevaluated. If one model provides better agreement and predictions than another, then it is preferred.

This process is ongoing. While the process is imperfect, it is better than all its competitors. Needless to say, some models have greater predictive value than others. Aristotle asserted that the earth is the center of the universe. The epicycle model of Ptolemy (Claudius Ptolemaeus, 87-150) was based on this assumption. For centuries, the church accepted this view as dogma. Even though this model provided reasonably accurate predictions for the motion of the planets, the Newton/Kepler model is easier to understand and provides a clear explanation for such anomalies as the apparent retrograde motion of Mars.

Simplicio: The method seems to be intelligently designed.

Virginia: Only if you play by the rules.

Galileo: We now have successful models for the motion of the planets, the motion of a pendulum, the motion of a spring, fluid flow, the nature of electricity and magnetism, the nature of waves, and heat transfer. While many models are complicated, the best models are based on simple principles that you sure are correct. Our confidence in many of these models is now so great we would be shocked if the unexpected happened. Every time you turn on one of your electronic gadgets, you are using the laws of electricity and magnetism.

Simplicio: What about hurricanes, floods, and beach erosion?

Galileo: The models for fluids are not as reliable as those for electricity. While you can criticize those making predictions based on less perfect models, you might think of them as an opportunity for employment. If you can accurately predict the future, you can make money. Better yet, you can begin to understand the world around you.

Virginia: You can also get into trouble.

Galileo: Sometimes my colleagues have been sloppy about their data. While my colleague Aristotle claimed the distance traveled by a falling body has a linear relationship with the time of flight, he never tested his ideas properly. My data shows the relationship is quadratic. In particular, if you double the time of flight, then the distance traveled will be quadrupled.

Simplicio: I guess data is important, but is an employer going to hire me to expound

on these already well-understood insights? Why would he care?

Galileo: The techniques of the ancient masters are embedded in the technology of the present. For example, Fourier series techniques used to solve partial differential equations are now being used in a multitude of applications including speech recognition, image analysis, and signal compression.

Simplicio: So where do numerical methods factor into this scenario?

Galileo: If you can model a problem by an equation or system of equations, then the goal of numerical analysis is to provide techniques to find the solution (or solutions). If your model is linear, then Linear Algebra is your tool of choice. Whenever possible, you should linearize your problem.

Simplicio: What do you do if your problem is not linear?

Galileo: If possible, you linearize your problem over a short period of time. The underlying concept in differential calculus is that the first derivative is the slope of the line that "best approximates" the curve. For us, the root finding method of Newton/Raphson is an example of a technique that repeatedly uses a linear approximation to solve a nonlinear problem.

Simplicio: OK, so what skills do I need to work in this area?

Galileo: If you find data fascinating, then I recommend you become versed in the following areas:

1. mathematics,
2. computer science,
3. statistics,
4. physics, and possibly
5. a biomedical area.

Virginia: I am worried about that computer science requirement. I have limited programming experience. My background in physics is a bit weak as well.

Galileo: You need to have enough computer skills to implement and test your own ideas. No one is going to do it for you. Otherwise, you will have no ability to test your ideas. You need to be comfortable with physics because different data acquisition devices employ different physical principles. A technique that produces accurate estimates for one modality may be useless when applied to signals or images acquired on another system. Any numerical method for analyzing data should be in sync with the device or method used to acquired it.

Simplicio: What about statistics? The only word that comes to mind is: boring, boring, boring. My view is:

> I know not χ,
> I know not square,
> Nor do I know,
> Why I should care.

Galileo: Maybe you should reconsider this attitude. Statisticians are the gatekeepers to a multitude of today's scientific questions because they provide us with tools for making sense of data. While the last century was the century of the hard sciences, the exciting new frontiers are now shifting to medical and biomedical applications. Imaging science will play a large role in these areas. Genomics with its terabytes of data may be a better example. In any case, anyone who has the ability to make sense of the mountains of data that is generated daily will be employable. In a word: Data, Data, Data!

Virginia: So that's why you mentioned biomedical applications?

Galileo: You got it.

Chapter 3

Topics for the Tutorial

He who does not understand motion, cannot understand Nature.-Galileo

Virginia: Good Sir, could you give us an overview of the topics you will be discussing in this tutorial?

Galileo: Certainly. The two main themes will be root finding and approximation theory. Since root finding has a long and distinguished history, we will begin with this theme. The task of finding a root is equivalent to that of solving a system of nonlinear equations.

Simplicio: Could you remind me about roots?

Galileo: A root of a function is a point $x = r$, where the graph of function crosses the $x-$axis. The official definition is:

Definition 3.0.1. *If* $f(x) : [a, b] \to \Re$, *is a function and* $f(r) = 0$, *then* $x = r$ *is a root.*

Simplicio: Why would I care?

Galileo: If you recall from your study of Calculus, the problem of maximizing and/or minimizing a function $f(x) : [a, b] \to \Re$ is at the heart of a multitude of applications. The strategy is to compute the first derivative $f'(x)$ at each critical point $x = r$. The maximum of the function $y = f(x)$ on the interval $[a, b]$ will equal the maximum of the values $f(a), f(b), f(r_1), f(r_2), \ldots, f(r_n)$, where $r_1, r_2, \ldots, r_n$ is the list of all the

critical points for $f(x)$. A similar statement is true for computing the minimum of the function. The beauty of this strategy is that an infinite problem has been reduced to a finite one.

Simplicio: Forgive me, but it has been a long time since I have suffered through Calculus. What is a critical point?

Virginia: A critical point of a function is a point $x = r$, where the graph of the first derivative crosses the $x-$axis. In other words, a location where the function has a horizontal tangent line. The precise definition is:

Definition 3.0.2. *If $f(x) : [a, b] \to \Re$, is a differentiable function and $f'(r) = 0$, then $x = r$ is a critical point for $f(x)$.*

Galileo: Very good. Note that the critical point always lies in the domain of the function.

Simplicio: And why should I care about critical points?

Galileo: If a company can represent their profits by a function, then they can maximize their profits by simply computing this function at all the critical points. The largest value will be the maximum of the function. A similar statement holds for minimizing their costs.

Simplicio: I must admit that I am having a bit of trouble visualizing this situation.

Galileo: How about the example of the parabola? Calculus is nothing more than the recognition that concepts such as velocity and acceleration associated with the motion of a falling body can be generalized to arbitrary functions. If you understand the parabola, you are a long way home.

Simplicio: Sounds good.

Galileo: If $f(x) = ax^2 + bx + c$, then the first derivative is $f'(x) = 2ax + b$. The critical point $x = r$ is commuted by solving the equation $f'(x) = ax + b = 0$. As an expert in Algebra, you immediately recognize that the critical point is $r = x = -\frac{b}{a}$ and the critical value is $f(r) = f(-\frac{b}{a}) = a(-\frac{b}{a})^2 + b(-\frac{b}{a}) + c = -2\frac{b}{a} + c$. In the special case of a falling body, I found that the height can be modeled by the formula $s(t) = -\frac{1}{2}gt^2 + v_0t + s_0$, where $g = -32\frac{ft}{sec^2} = -9.8\frac{m}{sec^2}$, v_0 denotes the initial velocity,

and s_0 denotes the initial height. Since this curve is concave down, the highest point of the flight of the ball will occur when the velocity equals zero. Since the velocity is the first derivative of the height function, the critical point will occur when $v(t) = s'(t) = -gt + v_0 = 0$ or $t = \frac{v_0}{g}$.

Virginia: If you toss the ball in a downward direction, then the initial velocity is negative. In this case, the maximum value of $f(x)$ will occur at time $t = 0$.

Galileo: Good point. I should have mentioned that we are assuming $v_0 > 0$. While the critical points are easy to find for this problem, real-world problems require much more general techniques. We will focus our discussion on the Newton/Raphson, bisection, and Contraction Mapping Theorem techniques. The Newton/Raphson method is based on finding the root $x = r$ for the linear function $y = f(x) = mx + b$. Since $r = -\frac{b}{m}$, the problem is not too difficult. Right?

Simplicio: These remarks help, but why are we discussing several different methods for finding roots? Why not simplify the discussion and just focus on one method?

Galileo: Each has its place. Our discussions will be driven by such questions as: Does the method always work? Which converges faster? Unfortunately, with numerical techniques, you don't always get clear winners. We will often find that the application drives the choice of technique.

Simplicio: And why would I care about the Contraction Mapping Theorem?

Galileo: This theorem is an elegant generalization of the method of Archimedes/Heron and Newton/Raphson. While these extensions are easy to understand in retrospect, they took 2000 years to unfold.

Simplicio: Do I need elegance?

Galileo: This theorem can be used to solve linear systems of equations, non-linear equations, and differential equations. It is even used to generate fractal pictures and compress images. In other words, it can be used to solve a multitude of different types of problems. In its most basic form, the technique is easy to understand, can be implemented in only a few lines of computer code, and always works. I call that elegant and I appreciate it when I find it.

Simplicio: I like the idea of compressing images.

Virginia: I too have enjoyed the beautiful snowflake example.

Galileo: While we won't have time to discuss fractals, we will lay the foundation so you can study that subject on your own.

Virginia: Are these all the topics we will cover?

Galileo: The second theme of our tutorial is approximation theory, where we will discuss the topics of Taylor's Theorem, polynomial approximation, Fourier Series, cubic splines, and wavelets. These methods are useful if you would like to approximate a function $f(x)$ by a function with certain desirable properties. For example, given the function $f(x) = \sin(x)$, we would like to approximate its value at a particular point $x = x_0$. We can do this with a Taylor polynomial of the form $p_1(x) = x, p_3(x) = x - \frac{1}{6}x^3, p_5(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5$, etc. Since polynomials are easy to compute and the method always converges to the correct answer, Taylor's Theorem is a great place to start. Taylor's Theorem provides a fundamental tool for the numerical approximation of first and second derivatives. Virtually any problem involved with rates of change requires the estimation of velocity or acceleration. The formulas we will derive are used everywhere in differential equations, partial differential equations, and signal and image processing.

Simplicio: What's next?

Galileo: After Taylor's Theorem, we turn to a second technique for approximating functions by polynomials. The advantage of this method is we use a sampling of the values of the function at scattered points rather than the values of the function and its derivatives at one particular point.

Simplicio: So?

Galileo: Typically, when we are given a set of data points, we are not given any information about the derivatives so Taylor's Theorem cannot be applied. Thus, we need a new technique.

Simplicio: OK.

Galileo: This topic also provides an excellent entry point into the modeling of data.

Since we usually have more data than we know what to do with, we usually try to reduce the data to a form that is easy to understand. Straight lines and parabolas are often a good place to begin. The technique that gets us there is linear least squares. While least squares is usually associated with straight line approximations, it can also be used to approximate data with a parabola of the form $p_2(x) = a_0 + a_1 x + a_2 x^2$. Our falling body problem is a good example, where a parabolic fit works. In 1958, Charles Keeling (1928-2005) began the collection of data measuring the concentration of carbon dioxide in the atmosphere. These measurements have been made monthly ever since he began this effort. When least squares is used to fit a parabolic curve to this data, the fit is excellent. A current political issue is whether or not the rising concentration of this gas causes global warming. Just because the fit is good, doesn't mean we can extrapolate out too many years. We shall see.

Simplicio: Interesting.

Virginia: Why would we worry about Fourier series?

Galileo: Fourier made his mark in mathematics by recognizing that trigonometric approximations produce much more accurate results than polynomial ones when solving the heat equation. We will discuss that famous Runge example, which shows that high degree polynomials are evil.

Simplicio: Good and evil in a mathematics class?

Galileo: If you are an engineer making a calculation and your calculator gives you a stupid answer, then your attitude is that the device is evil.

Simplicio: Even I understand that.

Virginia: Why discuss polynomials at all?

Galileo: As we mentioned, linear and quadratic fits can often produce useful results. Least squares are used everywhere. However, probably the best reason is polynomial interpolation provides an excellent entry point to Fourier series. In fact, if you look at the subject properly, the discrete Fourier transform is exactly polynomial interpolation. Thus, if you understand polynomials, you are a long ways towards understanding Fourier. Better yet, waves and wavelike (i. e. periodic) motion are ev-

erywhere in nature. While the motion of the pendulum is the first one that comes to mind, light, radio, ocean, and sound waves are also examples. A wave with frequency ω can be written as a trigonometric function of the form $\cos(\omega(t-t_0))$. Fourier series are nothing but linear combinations of functions of the form $\cos(nx)$ and $\sin(nx)$. Not only are they perfectly designed for modeling waves, but they also have remarkable mathematical properties.

Simplicio: But I am not interested in their math properties.

Galileo: You should be. As it turns out, engineers love Fourier techniques because they are not only directly connected with wave phenomena, but because they are computationally stable. Thus, they can trust the answers. The fundamental reason for this trust takes us back to Pythagoras.

Simplicio: I can't wait.

Virginia: What about cubic splines?

Galileo: While they are not as useful in physics as Fourier series, they have the same stable characteristics as Fourier series but even better convergence properties for the first and second derivatives. This property is not necessarily true for Fourier series. Splines have another important property that Fourier series don't have. Namely, while functions like $\sin(x)$ and $\cos(x)$ oscillate up and down forever, splines equal zero outside some finite interval.

Simplicio: Why is that property important?

Galileo: When you compute a linear combination of a bunch of spline functions at a particular point x, you can ignore all the intervals not containing x. Typically, the point x will lie in no more than 5 intervals. Since splines are piecewise cubic polynomials, they are almost instantaneous to compute on each interval and lie in a small number of intervals, they are blazingly fast. For these reasons, they are often used in computer graphics and computer animations.

Simplicio: I will have to pay attention when we discuss that topic.

Galileo: You will enjoy the elegant theorems associated with splines as well.

Virginia: And finally, what are wavelets good for?

Galileo: Wavelets represent the best of all possible worlds. If you think about the name a minute, you realize that the word wavelet implies "little wave," which is exactly what they are. Wavelets oscillate like $\sin(x)$ and $\cos(x)$ so they are useful for modeling real physical phenomena. Like the trigonometric functions, they enjoy the benefits of Pythagoras and so are stable to compute. In addition, they have the same finiteness properties that splines have so they are fast to compute. Needless to say, wavelets are very popular and are used in a multitude of applications. In particular, Jean Morlet used them to search for intense, short term bursts in geologic sonography data. They are also used in a multitude of imaging applications including compression and analysis.

Simplicio: If wavelets are so great, why don't we skip the other topics?

Galileo: Because you would be lost and confused. We will try to let the story unfold so the ideas become more transparent.

Simplicio: So that's it?

Galileo: Since the heat equation and the wave equation gave rise to the popularity of Fourier series, we really are required to discuss partial differential equations. Since we know your limits, we will make the discussions as brief as possible. Since differential equations are also everywhere in Nature, we will mention those topics as well.

Simplicio: I never had a course in differential equations.

Galileo: He who does not understand motion, cannot understand Nature.

Simplicio: Maybe I should become a monk.

Galileo: You can run, but you cannot hide. Remember: Math is easy. Its life that is difficult. And young lady, why *are* you here?

Virginia: I find all this talk about data and applications quite exciting. Hopefully, this experience will make me a better teacher.

Galileo: If your students can see how mathematics connects with the real world, then maybe they will be more motivated.

Simplicio: Again, why would you want to teach?

Virginia: I enjoy the logic, clarity, and simplicity of mathematics. It all makes sense.

I enjoy interacting with young people. My material needs are few so I don't object to the low pay.

Galileo: (The phone rings. Galileo answers. After he mumbles "Yes." and "Hmmm." repeatedly, he gets up from his chair.) My benefactor feels I should return to my research. So ends my catechism.

Simplicio: One last question?

Galileo: Yes?

Simplicio: Every book on numerical methods I have looked at begins with a discussion of round-off errors. Why haven't you mentioned this topic?

Galileo: Round-off errors are a detail. The big picture comes first. (Galileo sips from his glass of wine and departs.)

Simplicio: What do you think? Should we enroll in this guy's tutorial or take someone else's class? All he talks about is definitions, theorems, and proofs. Nothing but math, math, math. Worse yet, he seems to be a preacher teacher. I am not sure I can handle it.

Virginia: You can always take the course with Professor Powertrip. You might prefer to be with all those engineers. It is probably more your style.

Simplicio: Not a chance. That guy is mean and will do whatever he can to make you feel stupid.

Virginia: How about Professor Poubelle's section?

Simplicio: At least he wouldn't expect much from us.

Virginia: While I am a bit worried about the computer projects and the applications, I have decided to enroll with the preachy guy.

Simplicio: Tonight is ladies night at the "Math and Music Bar." Interested?

Virginia: Are you serious? I have to study.

Simplicio: Tomorrow is another day, maybe.

Part II

Day 2. Background and Review

> Let no one ignorant of geometry enter here." -inscription above Plato's Academy

Galileo: You have returned?

Simplicio: While I am not yet certain this course of study is worth my time, I have decided to give your tutorial a try.

Galileo: My administrator will be pleased I have clients. This is good. In any case, be certain to pay your fees before you leave.

Simplicio: What?

Galileo: Don't you expect compensation for your labors?

Simplicio: I will have to discuss this problem with my father. What about her?

Galileo: She has been awarded a scholarship.

Virginia: Enough of this talk. Let's move on.

Galileo: I plan to begin our tutorial by presenting several proofs of the Pythagorean Theorem.

Simplicio: Why on Earth would you present a theorem we have seen in our youth?

Galileo: Recall from our first conversation that the computation of the square root is of fundamental importance in math, statistics, and engineering. The Linear Algebra version is at the heart of the success of Fourier series.

> The only prerequisite for this course is plenty between the ears.-Walter Rudin

Simplicio: What are the prerequisites for this tutorial?

Galileo: Since my funding requires that I sustain my research program, let me be brief. You only need to know one thing, but you have to figure it out.

Simplicio: (To Virginia) Is this guy serious? He speaks in tongues.

Galileo: OK, let me rephrase my response. To succeed in mathematics or science you need to develop the ability to solve a problem on your own. Most never get it.

Simplicio: But can I ask questions?

Galileo: The math gene is what separates you from the other primates so you have the talent. Do you really think that an employer is going to reward you with a high salary to implement well understood ideas? Unfortunately, mathematics is not a spectator sport. Just like an athletic competition, you have to put in the time and effort. I am not interested in passive learners who just say "feed me." I expect you to run up and down the field like everyone else. Otherwise, we will both be wasting our time. Attitude is everything.

Simplicio: How about if you just tell me what I need to know to survive this tutorial?

Galileo: Since we will not be discussing specific issues in physics and biology, you can learn those subjects on another occasion. While statistics is important and we will discuss the rudiments of least squares and classification, you will not need any training in statistics to follow our discussions. On the other hand, since one of the main goals of this tutorial will be to develop algorithms, you will definitely need to have basic skills in computer programming. If you don't, you will be helpless when asked to implement even the most rudimentary algorithm.

Simplicio: I can handle those requirements.

Virginia: I am worried.

Chapter 4

Geometry

There is no royal road to Geometry.–Euclid

Euclid alone hath seen beauty.–Emma Talley Shaw

Uncle Dave, Geometry is easy.–Carter McMillan

Simplicio: What mathematics prerequisites are required for this tutorial?

Galileo: A solid foundation in Euclidean Geometry is essential. You will find Pythagoras (569-475 B.C.E.) everywhere in our discussions.

Simplicio: Surely, you are joking Mr. Galileo. I found Euclid (325-270 B.C.E) dull, difficult, and irrelevant.

Virginia: Mr. Simplicio, I find that statement surprising. I loved Euclid with his points, angles, similar triangles, congruent triangles, the area formulas for a paral-

lelogram and rhombus, and ruler and compass constructions. I particularly enjoyed the careful and rigorous logic he used when presenting his axioms, postulates, and theorems. Side-angle-side was my favorite. He opened a whole new world for me.

Galileo: As you will see, a multitude of ideas from Geometry have inspired computational algorithms. Our first algorithm will be introduced by my colleague Archimedes (287-212 B.C.E.). He loved Geometry so much he had his formula for the volume of a sphere engraved on his tomb.

Simplicio: Whoever heard of using a ruler and compass to implement a mathematical technique on a computer? Side-angle-side? Give me a break.

Galileo: You will see.

4.1 The Pythagorean Theorem

> At its deepest level, reality is mathematical in nature.-Pythagoras
>
> There is geometry in the humming of the strings, there is music in the spacing of the spheres.-Pythagoras

Galileo: In the spirit of the ancients, we begin with the Pythagorean Theorem. I know you have seen it before.

Simplicio: It is a theorem I learned in geometry many years ago. Why would you begin our discussion with such an old theorem?

Galileo: Because the Pythagorean Theorem provides a unifying theme for this tutorial. In fact, it contains four important concepts that appear everywhere in modern mathematics. These concepts include:

1. distance,
2. roots,
3. irrational numbers,
4. orthogonality, and
5. projection.

Can you state the theorem?
Virginia: I remember it.

Theorem 4.1.1 (Pythagorean Theorem). *If the legs of a right triangle have lengths a and b and the hypotenuse has length c, then $c^2 = a^2 + b^2$.*

Galileo: We begin by making some easy observations about the theorem that should help to make these themes more transparent. First, since the length of the hypotenuse of a right triangle is the square root of the sum of the squares of the other two, it forms the basis for computing the distance between two points. In fact, the formula for the distance between two points $P(x_1, y_1)$ and $Q(x_2, y_2)$ in the plane is given by the formula:

$$dist(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

This rule is an immediate application of the Pythagorean Theorem. Note that we will begin our tutorial with a discussion of the Archimedes/Heron square root algorithm for approximating the square root of a number. As you will see, the ideas in this algorithm are embedded in a number of important modern techniques including Newton/Raphson and the Contraction Mapping Theorem. Also, while lengths and

distances may seem too easy, the concept of computing distances between points reappears in Linear Algebra, Fourier series, orthogonal polynomials, splines, and wavelets. We will revisit this idea repeatedly during our quest. Are wavelets new enough?

Simplicio: OK, OK.

Galileo: A key assumption in the Pythagorean Theorem is that one of the angles has to be a right angle. Without that assumption, the theorem is false. As we will see in our investigations, many numerical techniques fail badly. Engineers do not like being blind sided by a stupid result when they are in the middle of a project. They like methods that always produce accurate answers. The concept of orthogonality helps fulfill this wish.

Simplicio: I never heard of orthogonality before.

Galileo: Orthogonality is just a fancy way of saying right angle or perpendicular. In the Pythagorean Theorem, the two shorter sides of the triangle are assumed to be perpendicular (and thus orthogonal).

Simplicio: It looks easy from here.

Galileo: The fourth idea is that we can project the hypotenuse of the triangle onto either of the other two sides. Note that the length of the hypotenuse is greater than the length of either of the other two sides.

Simplicio: That's evident from the formula $c^2 = a^2 + b^2$.

Galileo: This desirable property is a consequence of our assumption that the angle opposite the hypotenuse is assumed to be a right angle. While not all projections have this wonderful property, Fourier does. Such projections are called orthogonal.

Virginia: Since I don't exactly understand Fourier series, I am not sure where you are going with this. In any case I find these ideas interesting.

Simplicio: So far, I like this discussion. Easy is good.

Galileo: I like to begin with easy examples. Can you prove this theorem of Pythagoras?

Simplicio: I fear it has evaporated from my cranium.

Galileo: Pythagoras of Samos ($ca.569 - ca.475$ B.C.E.) is often described as the first

pure mathematician. While he is an extremely important figure in the development of mathematics, we know very little about his mathematical achievements. Unlike many later Greek mathematicians, we have nothing of Pythagoras's writings. The society which he led was half religious and half scientific. His theorem has been claimed by both the Chinese and Babylonians at least 1000 years before his birth so maybe others deserve credit as well.

Virginia: Isn't it time we prove it?

Galileo: How about two proofs?

Proof. The Pythagorean Theorem

Proof 1:

After a cursory look at Figure 4.1, we see that the area of both squares equals $(a+b)^2$. Since the area of the square on the left is the sum of the square in the middle and 4 triangles, $A = c^2 + 4(\frac{1}{2}ab) = c^2 + 2ab$. Since the area of the square on the right is the sum of two squares and two rectangles, $A = a^2 + 2ab + b^2$. Thus, $A = c^2 + 2ab = a^2 + 2ab + b^2$. By subtracting the quantity $2ab$ from both sides of the equation, we arrive at the relation $c^2 = a^2 + b^2$.

Proof 2:

A second proof can be given using only the square on the left. Since the area of the large square is $(a+b)^2 = a^2 + 2ab + b^2$ and since the whole is equal to the sum of its parts, we see that $a^2 + 2ab + b^2 = c^2 + 4(\frac{1}{2}ab) = c^2 + 2ab$. Again, by subtracting $2ab$ from both sides of the equation, we find $c^2 = a^2 + b^2$. □

Galileo: That wasn't so bad was it?

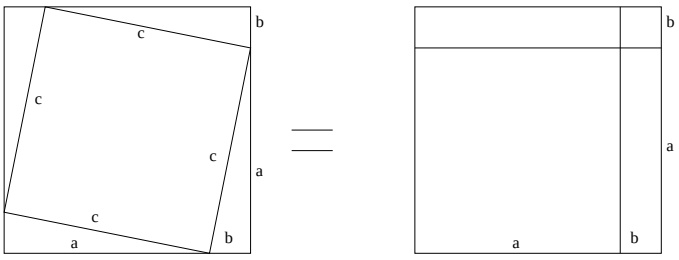


Figure 4.1: The Pythagorean Theorem

Simplicio: Even I can understand these proofs. What else did he do?

Galileo: Pythagoras led a remarkable life. In about 535 B.C.E Pythagoras visited Egypt, where he learned about their refusal to eat beans, wear even cloths made from animal skins, and their quest for purity. In 525 B.C.E. Cambyses II, the king of Persia, invaded Egypt. Pythagoras was captured and removed to Babylon. Eventually, he was allowed to leave and returned to Samos. In about 518 B.C.E. he left Samos and went to Croton in southern Italy, where he formed a mathematical/religious society. He and his followers believed that reality is mathematical in nature.

Simplicio: Really?

Galileo: They even believed that things are numbers and each number has its own personality.

Simplicio: Bizaar.

Galileo: They also believed that the Earth is a sphere at the center of the Universe and that every number should be rational.

Simplicio: Those ideas seem more reasonable.

Virginia: What happened when they discovered the quantity $\sqrt{2}$ is not a rational number?

Simplicio: They probably started eating beans again.

Galileo: And so it goes.

Exercise Set 4.1.

1. Prove the Pythagorean Theorem for three dimensions. In particular, if a, b, c represent the lengths of the sides of a rectangular box and d represents the length of the diagonal, then show that $d^2 = a^2 + b^2 + c^2$. (Hint: Apply the Pythagorean Theorem twice.)

4.2 Garfield's Proof of the Pythagorean Theorem

Ideas control the world.-James Garfield

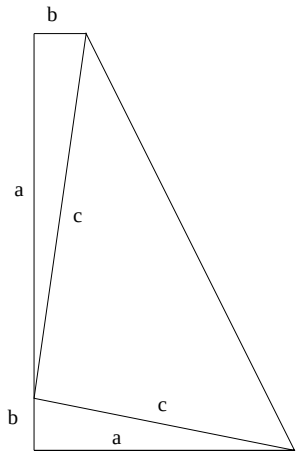


Figure 4.2: President Garfield's Proof of the Pythagorean Theorem

Galileo: While the Pythagorean theorem is of great interest to mathematicians, it even inspired President James Garfield to provide his own proof. Let's take a look.

Garfield: Instead of using a square, my proof based on the area of a trapezoid, where the two bases have lengths a and b and the height is $a + b$. A picture containing the idea of the proof is given in Figure 4.2.

Proof. If we compute the area of the trapezoid, we find:

$$\begin{aligned} A &= \frac{1}{2}(a+b)(a+b) \\ &= \frac{1}{2}(a^2+2ab+b^2) \\ &= \frac{1}{2}a^2+ab+\frac{1}{2}b^2 \end{aligned}$$

Now computing the same area as the sum of the areas of the three triangles that comprise the trapezoid we find:

$$\begin{aligned} A &= \frac{1}{2}ab+\frac{1}{2}ab+\frac{1}{2}c^2 \\ &= ab+\frac{1}{2}c^2 \end{aligned}$$

Setting these values for the area of the trapezoid equal to each other we find:

$$A = \frac{1}{2}a^2+ab+\frac{1}{2}b^2 = ab+\frac{1}{2}c^2.$$

Thus, by subtracting the quantity ab from both sides of the equation and multiplying both sides of the equation by 2 we have the desired result:

$$a^2+b^2=c^2.$$

□

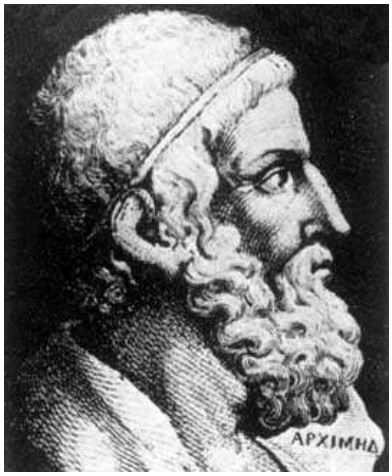
Simplicio: I don't see that his proof is much different from the two we just discussed. Dividing everything by two adds little to my understanding. He should have been shot.

Galileo: He was.

Exercise Set 4.2.

1. Investigate Alexander Graham Bell's role in trying to save President Garfield's life. What technology was used?

4.3 The Method of Archimedes/Heron



Archimedes (287-212 B.C.E.)

> Certain things first became clear to me by a mechanical method, although they had to be demonstrated by geometry afterwards because their investigation by the said method did not furnish an actual demonstration. But it is of course easier, when we have previously acquired by the method, some knowledge of the questions, to supply the proof than it is to find it without any previous knowledge.-Archimedes to Eratosthenes
>
> Noli turbare circulos meos. Do not disturb my circles! Last words. Sometimes reported as: Soldier, stand away from my diagram.-Archimedes

Simplicio: What are the topics for today's lesson?

Galileo: The first topic will be the Archimedes/Heron algorithm for computing the square root of a positive number. This technique is easy to understand, always works, and converges quickly. For an engineer this is the best of all possible worlds. To illustrate how the algorithm works, we will compute a number of examples such as $\sqrt{2}, \sqrt{3}$, and $\sqrt{5}$. These computations should increase your comfort zone.

Simplicio: Sounds like a plan.

Galileo: We now introduce one of the great masters of antiquity, Archimedes of Syracuse. He was one of the great mathematicians of all time, who wrote expositions solid geometry, pumps (the Archimedes' helix-shaped screw), floating bodies, the

center of gravity, and the area under a parabola. His proof of the formula for the volume of a sphere is a gem. If he had the ideas of modern algebra, he would have invented Integral Calculus. Professor Archimedes welcome to our tutorial.

Archimedes: I am glad to be here.

Galileo: Good sir, could you enlighten us on your method for computing square roots?

Archimedes: The underlying idea is quite simple: given a positive number K find two numbers a and b that are close together and have the property that $ab = K$. If the approximations are not good enough, then replace a by the average $\overline{a} = \frac{a+b}{2}$ and b by the product $\overline{b} = \frac{K}{\overline{a}}$. Note that $\overline{a} * \overline{b} = K$.

The square root method can now be implemented in the following steps:

Let $K > 0$ be a given real number.

Step 0. Begin the process by setting $a_0 = 1$ and $b_0 = K$.

Step 1. Set $a_1 = \frac{a_0+b_0}{2}$ and $b_1 = \frac{K}{a_1}$.

Step 2. Set $a_2 = \frac{a_1+b_1}{2}$ and $b_2 = \frac{K}{a_2}$.

Step n. Set $a_n = \frac{a_{n-1}+b_{n-1}}{2}$ and $b_n = \frac{K}{a_n}$.

Note that for each iteration n, we have the property that $a_n * b_n = K$.

Galileo: What can be more reasonable and elegant than computing the average of two numbers?

Simplicio: I like this method. It is easy to understand and easy to implement.

Archimedes: The algorithm can be simplified. In particular, if a_n is replaced by x_n and b_n is replaced by $\frac{K}{x_n}$, then the method becomes:

Let $K > 0$ be a given real number.

Step 0. Initialize the process by setting $x_0 = 1$.

Step 1. Set $x_1 = \frac{x_0+\frac{K}{x_0}}{2}$.

Step 2. Set $x_2 = \frac{x_1+\frac{K}{x_1}}{2}$.

Step n. Set $x_n = \frac{x_{n-1}+\frac{K}{x_{n-1}}}{2}$.

Simplicio: I like this version even better.

Example 4.3.1. *Galileo: In Figure 4.3 we have displayed the locations of the first three estimates on the real line.*

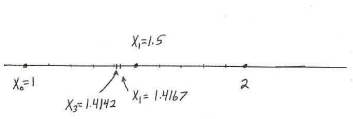


Figure 4.3: The First Three Estimates of $\sqrt{2}$

In Table 4.1 we have presented the first 6 *estimates of the square root of* 2 *when the initial guess is* $x_0 = 1$.

x_0	1.000000000000000
x_1	1.500000000000000
x_2	1.416666666666667
x_3	1.414215686274510
x_4	1.414213562374690
x_5	1.414213562373095
x_6	1.414213562373095

Table 4.1: Six Estimates of $\sqrt{2}$

Simplicio: Amazing!! After only 6 *iterations we have* 15 *digits of agreement. I like this algorithm.*

Galileo: What do you notice about the terms of the sequence? Do they increase or decrease?

Simplicio: Looks to me like they decrease after the initial guess.

Galileo: Why not try a few exercises to see how the method works?

Virginia: Where did this algorithm come from? What inspired you?

Archimedes: Geometry is the key. Consider Figure 4.4 where we suppose $x^2 \approx K$ and we want to find a Δx such that $(x + \Delta x)^2 = K$.

Archimedes: Since Δx is small, Δx^2 is even smaller, so we can eliminate this shaded

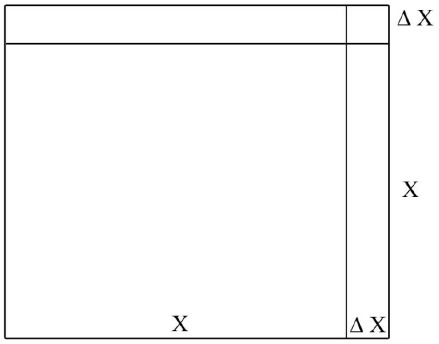


Figure 4.4: The Geometry Underneath the Square Root Algorithm

piece of the diagram. Doing so we find

$$\begin{aligned} K &= (x + \Delta x)^2 \\ &= x^2 + 2x\Delta x + \Delta x^2 \\ &\approx x^2 + 2x\Delta x, \end{aligned}$$

which implies

$$\Delta x \approx \frac{K - x^2}{2x}.$$

Thus,

$$x + \Delta x = x - \frac{x^2 - K}{2x}.$$

Rewriting $x + \Delta x$ as x_{n+1} and x as x_n, we arrive at the equation

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n^2 - K}{2x_n} \\ &= \frac{x_n + \frac{K}{x_n}}{2} \\ &= \frac{1}{2}x_n + \frac{1}{2}\frac{K}{x_n}, \end{aligned}$$

which is exactly the previously discussed method. In particular, the value of x_{n+1} is the average of x_n and $\frac{K}{x_n}$.

Simplicio: But I have one quick question. Will the algorithm eventually terminate or will we have to compute forever to get the exact answer?

Galileo: Note that if K is a rational number (i.e. the quotient of two integers), then each $x_1, x_2, \ldots, x_n$ must also be rational numbers. Thus, if $\sqrt{K} = x_n$, for some n, then $\sqrt{K}$ must also be rational. The bad news is that even our colleague Pythagoras noticed that the square root of 2 is irrational (i.e. not rational).

Virginia: Thus, if we start the process of approximating $\sqrt{2}$ with $x_0 = 1$, then every succeeding estimate x_n will be a rational number. And we are *forced* to make an infinite number of computations to get the exact answer.

Galileo: As we have already learned, the ancients found this knowledge quite upsetting and mystical. Archimedes do you have any other thoughts on this technique?

Archimedes: Note also that division by 2 in a calculator (or computer program) can be implemented as a bit shift. Thus, the only serious computation is the division $b_n = \frac{K}{x_n}$.

Simplicio: I like that observation.

Galileo: You can see that Archimedes is keeping up with current advances in technology.

Virginia: What is a bit shift?

Simplicio: Instead of representing a number base ten by a sequence of digits chosen from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, you represent a number base two by a sequence of digits from the set $\{0, 1\}$. For example, $6 = 2^2 + 1 * 2 + 0 = 110$. If you divide 6 by 2, you get $3 = 2 + 1 = 11$. In other words, to divide by 2 you simply drop the 0. A computer geek will say he has shifted the digits 110 one unit to 11.

Example 4.3.2. *Galileo: Let's use our algorithm to compute the square root of zero.*

Simplicio: Your kidding! Everyone in the room knows the answer. Why bother?

Galileo: I have an agenda. *Simplicio: In any case, it is easy. If* $K = 0$, *then* $x_{n+1} = x_n - \frac{x_n^2 - K}{2x_n} = \frac{1}{2}x_n$.

Galileo: If $x_0 = 1$, *then what is* x_6?

Simplicio: Since the value of the estimate at one step in the process is exactly half the estimate at the previous step, $x_6 = \frac{1}{2^6}$.

Galileo: How far is that from the final answer?

Virginia: Compared with the other examples we have just discussed, we are miles, no light years, from the final answer.

Galileo: How many iterations will we need to get 12 digits of accuracy?

Simplicio: Since $2^{10} \approx 1000$, *we observe that* $2^{40} \approx 1000^4 = 10^{12}$. *Thus,* $x_{40} \approx \frac{1}{10^{12}}$.

Virginia: Forty iterations is a lot more than six.

Simplicio: What's going on here?

Galileo: Think about it. We will return to this issue shortly. If you work the homework problems, you will see we have problems with very large numbers as well.

Simplicio: We were doing so well. Now I am worried.

Galileo: Before we leave the topic of computing square roots, we should observe the idea underneath this method is to "linearize" the problem. More specifically, when a problem is too difficult to solve in general, simply discard the higher order terms and solve the remaining linear part of the problem. With luck, the solutions to a sequence of simple linear problems will converge to the solution to the non-linear problem. We will see this strategy again with the method of Newton/Raphson.

Exercise Set 4.3.

1. Show that $\sqrt{2}$ is not a rational number.

2. Let $K = 5$ and $x_0 = 1$. Compute the first five iterations of the square root algorithm to estimate $\sqrt{5}$. What do you notice about the terms of the sequence? Do they increase or decrease? What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

3. Let $K = 10$ and $x_0 = 1$. Compute the first five iterations of the square root algorithm to estimate $\sqrt{10}$. What do you notice about the terms of the sequence? What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

4. Let $K = 100$ and $x_0 = 1$. Compute the first five iterations of the square root algorithm to estimate $\sqrt{100}$. What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

5. Let $K = 10,000$ and $x_0 = 1$. Compute the first five iterations of the square root algorithm to estimate $\sqrt{10,000}$. What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

6. Let $K = 1,000,000$ and $x_0 = 1$. Compute the first five iterations of the square root algorithm to estimate $\sqrt{1,000,000}$. What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? Compare the number of iterations require for this problem and when you approximated $\sqrt{2}$. Which is greater? (Make your computations with 10 digits of accuracy.)

7. Let $K = 0$ and $x_0 = 1$. Compute the first five iterations of the square root algorithm to estimate $\sqrt{0}$. What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? Compare the number of iterations require for this problem and when you approximated $\sqrt{2}$ and $\sqrt{1,000,000}$. (Make your computations with 10 digits of accuracy.)

Simplicio: These exercises would have been a drag without my trusty programmable calculator.

Galileo: While your calculator is fine for these problems it will be woefully inadequate for most real-life computations. Get used to idea of implementing your methods in computer software.

Simplicio: No problem.

Galileo: Note that these exercises were designed to stress the algorithm. By computing $\sqrt{K}$ for large and small numbers we are checking two important aspects of the algorithm. First, we are looking to see if we get the correct answers. Second, we are checking the rate of convergence. Both of these considerations will be addressed in future discussions.

Simplicio: I guess I had better redo these problems.

4.4 Two Applications of Square Roots

Figure 4.5: Heron of Alexandria ($ca.10 - ca.75$)

Galileo: While the Pythagorean Theorem provides one situation where the computation of a square root is needed, a couple of others should also be mentioned. You do remember the formula for computing the area of a triangle?

Simplicio: Of course, the area is simply one half the base times the height.

Galileo: OK, but would it not be more natural to have a formula, which produces the area in terms of the lengths of the three sides? This question is a natural one because the height may not be known.

Simplicio: I don't recall any such formula.

Galileo: Leave it to the ancient Greeks to not only have asked this question, but to have answered it as well. While Heron of Alexandria $(10 - 75)$ is frequently given credit for its discovery, the formula was already known to Archimedes of Syracuse (287-212 B.C.E.). For the area of a triangle whose sides have lengths: a, b, and c, the area is given by the formula:

$$A = \sqrt{s(s-a)(s-b)(s-c)},$$

where $s = \frac{a+b+c}{2}$ denotes the semi-perimeter. Note that the computation of a square root is required.

Note that as long as you know how to compute the square root of a number, the formula is straightforward to compute. Do either of you see why the formula might be useful?

Virginia: In fact good sir, I prefer this formula to the usual one given in Geometry because you frequently don't know the height of the triangle. This formula works great if you simply know the lengths of the three sides?

Simplicio: I like the formula, but how would anyone have ever thought of it?

Galileo: While I can't answer that question, always remember that those ancient fellows were smart and thought deeply.

Virginia: How would such a formula be proved?

Galileo: In modern notation, simply represent the vertices of a triangle by vectors $\mathbf{u} = (a, b)$ and $\mathbf{v} = (c, d)$ in the plane and compute. It helps to use the fact that the area of the triangle is the absolute value of $\frac{1}{2}(ad - bc)$. However, it is still a bit of a mess. We will leave this problem as an exercise.

Simplicio: (To Virginia) That problem belongs to you.

Galileo: A second example is the golden mean (or ratio) ϕ, which the ancient Greeks felt had special, even mystical, significance. This quantity appeared in their art and architecture as well as their mathematics. The ratio of the height to the width of the Parthenon equals this famous number. A pentagram is loaded with ratios equal to ϕ. The golden ratio is defined as the ratio $\phi = \frac{1}{x}$, where x is the point in a line segment

$[0,1]$ such that $\frac{x}{1} = \frac{1-x}{x}$. In other words, the point x is chosen so that the ratio of the whole segment to the longer subsegment equals the ratio of the longer segment to the shorter. When this proportion is solved for x, the answer is $x = \frac{-1\pm\sqrt{5}}{2}$. Since lengths should be positive quantities, we are only interested in the answer $x = \frac{-1+\sqrt{5}}{2}$. An easy computation shows that $\phi = \frac{1}{x} = \frac{1+\sqrt{5}}{2} = \approx 1.61803\ldots$. Thus, the Greeks had a natural interest in computing the quantity $\sqrt{5}$.

Virginia: If I remember correctly, this number can be approximated by computing the ratios of the terms in the Fibonacci sequence $1, 1, 2, 3, 5, 8, \ldots$.

Galileo: Very good.

Simplicio: Is that why we have note cards of dimension 3×5 and 5×8?

Virginia: You do the math.

Exercise Set 4.4.

1. Compute the golden mean to 8 decimal places.
2. Compute the area of a triangle, whose sides have lengths $1, 1$, and 1.
3. Compute the area of a triangle, whose sides have lengths a, a, and a.
4. Compute the area of a triangle, whose sides have lengths $1, 2$, and 3.
5. Compute the area of a triangle, whose sides have lengths $1, 2$, and 4. Why do you have an OOPS?
6. Prove the Archimedes/Heron formula for the area of a triangle, whose sides have lengths a, b, c.

4.5 Rigor

Figure 4.6: Kurt Gödel (1906-1978)

> The development of mathematics towards greater precision has led, as is well known, to the formalization of large tracts of it, so that one can prove any theorem using nothing but a few mechanical rules.-Kurt Gödel

Simplicio: OK, what's next?

Galileo: A solid understanding of Geometry is built on a foundation of mathematical rigor. I insist you are comfortable with logical arguments.

Simplicio: I knew this discussion was going to deteriorate. Here it comes.

Galileo: Before you can understand the strengths and weaknesses of a mathematical technique, you need to have an understanding of when it works and when it fails. A bit of logic and mathematical formalism will aid in the understanding of when you can trust a method. Key examples can be used to point out when you should be suspicious. The first requirement in formal mathematics is that you must understand the difference between an axiom, a definition, and a theorem.

Simplicio: Groan.

Galileo: Unfortunately, the beauty of numerical analysis is that the subject is ruled by Murphy's Law. Namely, "What can go wrong, will go wrong." A technique that works

well for one application may fail for another. Worse yet, for any given technique, an example can invariably be found, where it provides answers that make no sense. It is important to understand why one method is preferred over another. Definitions and theorems can be used to make these thoughts precise. I now introduce Professor Gödel, who has agreed to help clarify these issues for us. Professor Gödel.

Virginia: I am pleased to meet you sir.

Simplicio: Good day sir. (To Virginia) He looks mean. This meeting could get ugly.

Gödel: I am not sure I am welcome. Maybe I should retreat to my office.

Galileo: Please enlighten these young people about the nature of mathematics.

Gödel: I will try. First, every theorem consists of two parts. The first is the hypothesis, while the second is the conclusion. If the theorem is valid and the hypotheses are true, then we can conclude that the conclusion is also true. Symbolically, every theorem is a conditional sentence of the form: If p, then q. If the theorem is true and we know that the statement p is also true for our particular situation, then we immediately know that q is true as well. This bit of logic is called *modus ponens*.

Galileo: Let me note that our friends in statistics are also quite fond of conditional sentences. The theorem of the Presbyterian minister Thomas Bayes (1702-1761) is central to any discussion of conditional probability. Thus, people other than myself require you to understand the structure of language. In any case, what is the hypothesis of the Pythagorean Theorem?

Virginia: Actually, we have two hypotheses. The first hypothesis is that the geometric object we are dealing with is a triangle. The second is that this triangle is of a special type. Namely, one of its three angles is 90 degrees.

Galileo: Correct. Now what is the conclusion?

Virginia: The relationship between the length of the hypotenuse and the lengths of the other two sides of the triangle. Namely, the equation $c^2 = a^2 + b^2$.

Galileo: Correct again.

Simplicio: Why are you boring us with these discussions? I know the formula $c^2 = a^2 + b^2$ has been established. But if I know the formula, then isn't that good

enough? What else matters?

Gödel: How can this guy be so obtuse? Children are evil. (Gödel departs)

Simplicio: This wizened little guy is mean.

Virginia: Maybe he was a pediatrician and had you as a patient.

Galileo: How about a bit less disrespect and a bit more discussion?

Gödel: (Gödel returns) Has anyone seen a small black valise? It contained important work.

Galileo: What if the triangle is not a right triangle? In particular, what if the triangle is acute or obtuse? You need to know when it is appropriate to apply the formula.

Virginia: Obviously, the formula does not apply for all triangles.

Galileo: Correct again. If the hypothesis is not satisfied, then the theorem does not apply and you cannot pretend the conclusion holds.

Simplicio: What do you do then?

Gödel: This discussion is outrageous. Plato understood these issues 2500 years ago. These young people should have mastered logic and rigor when they studied Euclid. We should not be having these discussions.

Galileo: Patience good sir. However, my experience has been that people in applications tend to be sloppy in these matters. I find it is better to discuss them up front. Later, when the setting is more abstract, a discussion of rigor might get lost in the mud. We might as well address the issue now while we are in the familiar setting of geometry. You will be well served if you make the effort to clarify these questions of rigor and logic now. Don't worry, we will revisit these issues.

Gödel: Let's just reduce the discussion to the essentials.

1. A theorem is a statement of the form: "If p, then q."

2. The converse of the theorem "If p, then q." is the statement "If q, then p."

3. The contrapositive (modus tollens) of the theorem "If p, then q." is a statement of the form "If $\sim q$, then $\sim p$."

4. If a statement "If p, then q" and its converse "If q, then p" are both true, then p and q are considered equivalent. In this setting, the statements p and q are either both true or both false.

While politicians and preachers would like you to believe that a theorem and its converse are equivalent, nothing could be further from the truth.

Simplicio: What are those little squiggles "$\sim$" doing in this discussion?

Virginia: Obviously, the symbol $\sim p$ denotes the negation of p. In other words, if p is true, then $\sim p$ is false and vice versa.

Simplicio: How about an example?

Gödel: Consider the statement: "If you are Franklin Delano Roosevelt, then you are famous."

Simplicio: I would rather consider the statement: "If you are Emmitt Smith, then you are famous."

Virginia: Who is Emmitt Smith? Is he famous?

Galileo: I think we are off topic here. In any case, let us assume the statement is true.

Gödel: The converse of MY version of the statement is: "If you are famous, then you are Franklin Delano Roosevelt." Do you think this converse is also true?

Simplicio: No. Barbara Bush is famous and she is not even a male much less a president. In particular, the two statements are not equivalent.

Virginia: On the other hand, the contrapositive of this statement is: "If you are not famous, then you are not Franklin Delano Roosevelt." Note that this statement is indeed equivalent to the original statement.

Galileo: Correct again.

Simplicio: So why should I care?

Gödel: I am done.

Galileo: Good sir. Before you depart, could you give us a quick summary of what these young people need to know.

Gödel: All these truths are encapsulated in Table 4.2.

p	q	$p \wedge q$	$p \vee q$	$p \rightarrow q$
T	T	T	T	T
T	F	F	T	F
F	T	F	T	T
F	F	F	F	T

Table 4.2: The Truth Table for "And," "Or," and "If."

Simplicio: I don't understand all those symbols.

Virginia: Obviously, T = True and F = False.

Simplicio: I figured that out. Also, while I assume the symbol $p \rightarrow q$ represents the conditional statement "If p, then q." What do the symbols $\wedge$ and $\vee$ represent?

Gödel: The symbol $\wedge$ means "And," while $\vee$ means "Or."

Virginia: Ok, I understand that if p and q are both true, then we should define $p \wedge q$ to be true. However, if you are ordering a meal at a restaurant and the choice is "tea or coffee," then you surely don't get both.

Gödel: Don't confuse the "exclusive or" with the "inclusive or." In a restaurant, you will get tea or coffee, but not both. In Logic we are more generous and will give you both.

Simplicio: I guess that's why all the math restaurants have gone out of business.

Gödel: The concept of a theorem is the most important idea to take away from Table 4.2. In particular, if a theorem $p \rightarrow q$ is true and the hypothesis p is true, then the conclusion q is also true. This logic is exactly what use when we apply a general theorem to a specific instance.

Virginia: And if we don't satisfy the hypothesis, then we may be disappointed when q turns out to be false.

Galileo: Correct.

Gödel: In Table 4.3 we observe that the 3^{rd} column represents a statement and the 4^{th} column represents its converse. Note that these two columns are not the same.

Virginia: However, the 3^{rd} and 7^{th} columns *are* the same.

Galileo: Correct again.

Gödel: I must be gone. (Gödel picks up his valise and departs.)

p	q	$p \to q$	$q \to p$	$\sim p$	$\sim q$	$\sim q \to \sim p$
T	T	T	T	F	F	T
T	F	F	T	F	T	F
F	T	T	F	T	F	T
F	F	T	T	T	T	T

Table 4.3: The Truth Table for the Contrapositive

.

Galileo: Very good. Your observation is at the heart of a proof by contradiction. In other words, we will assume that the statement q is false and then will show that the statement p is also false. In summary, an understanding of definitions, theorems, converses, and contrapositives is about all the logic you will need to know.

Virginia: If I remember my Geometry correctly, we also considered lemmas, propositions, and corollaries.

Galileo: These three words all represent different names for for small theorems. A lemma is interesting only because it can be used to help prove a more important theorem. Sometimes they are called helping theorems because they help organize the proof of an important theorem. A proposition is a small (but usually useful) theorem, which is more of a stepping stone than a reservoir containing a big concept. A corollary will usually represent an easy consequence of an important theorem. For example, the Mean Value Theorem has several important corollaries that we will use more often than the theorem itself.

Virginia: So when we are studying for an exam, we study the theorems first, the corollaries second, and the propositions last.

Simplicio: Do we get to forget the lemmas?

Virginia: For you, the answer is probably yes. For the rest of us, a lemma helps us organize and remember the proof. What do you have to say about axioms and definitions?

Galileo: Axioms are something you assume true. For example, in algebra we assume that equals added to equals are equal.

Virginia: So, if $a = b$ and $c = d$, then $a + c = b + d$.

Galileo: While definitions are written in the same "If p, then q." format we use for theorems, their purpose is to define a new concept.

Simplicio: An example please!

Galileo: How about the definition of a right triangle?

Definition 4.5.1. *If a triangle has the property that one of its angles is a right angle, then it is a right triangle.*

Note that while this definition is written as a statement of the form "If p, then q," it is understood that the p and q are equivalent.

Virginia: In other words, there are no converses for definitions. If the triangle doesn't have a 90 degree angle, it cannot be a right triangle.

Galileo: Looks like you understand the hierarchy. I would only add that you pay special attention to theorems with names such as the Pythagorean Theorem, Taylor's Theorem, the Mean Value Theorem, and the Intermediate Value Theorem. We will think of a theorem as an item in a bookkeeper's ledger. Whenever you need to know if something is true, you simply check the list of theorems in the ledger. If you find one that you think might be relevant, all you have to do is check the hypotheses. If they are satisfied, you get the conclusion for free. In other words, the hard work has already been done. Now, you have to admit that this logic and rigor is easy. All you have to know is four logic rules and the difference between an axion, a definition, and a theorem.

Simplicio: I should have gone to church this morning.

Galileo: Remember, math is easy, it's life that's uncertain.

Simplicio: Let's move on before I become rigor-mortified.

Galileo: We end with the definition of the inverse of a statement. I will leave it for you to show the inverse of a statement is equivalent to the converse.

Definition 4.5.2. *The inverse of the statement "If p, then q." is the statement "If $\sim p$, then $\sim q$."*

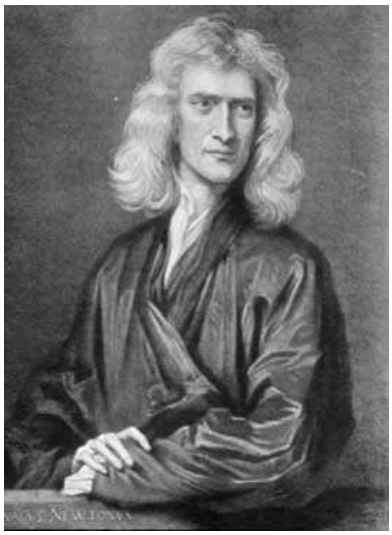
Exercise Set 4.5.

1. Use a truth table to show the inverse is equivalent to the converse.

2. Use a truth table to show the statement "If p, then $\sim q$." is equivalent to the statement "$(\sim p) \vee (\sim q)$."

Part III

Day 3. Methods for Finding Roots



Isaac Newton (1642-1727)

> Truth is ever to be found in the simplicity, and not in the multiplicity and confusion of things.-Isaac Newton

Simplicio: What are the topics for today's lesson?

Galileo: The first topic will be an algorithm for computing the cube root of a number. This technique is a natural an extension of the Archimedes/Heron algorithm for computing the square root of a number. As before, this technique is easy to understand, always works, and converges quickly. For an engineer this is the best of all possible worlds. To illustrate how the algorithm works, we will compute a number of examples such as $\sqrt[3]{2}$.

Simplicio: Wait a minute. I am a bit confused here. The other day you talked about the root of a function $f(x)$. Today you are talking about the root of a positive number K. Do I detect double talk here?

Galileo: You have made a good observation. However, this confusion can be quickly explained away because the quantity $r = \sqrt{K}$ is a root of the function $f(x) = x^2 - K$.

Simplicio: Oh, I see all you have to do is substitute $r = \sqrt{K}$ into the function $f(x)$ and get $f(r) = f(\sqrt{K}) = (\sqrt{K})^2 - K = K - K = 0$. I now understand that point. What is next?

Galileo: After the cube root algorithm, we introduce a similar algorithm for comput-

ing n^{th} roots.

Simplicio: While I can understand why someone might be interested in computing a cube root, why in heaven's name would I care about n^{th} roots?

Galileo: What about music? Recall that a piano has 12 keys for each octave. Each key is represents a different frequency. The frequency represented by C in one octave is twice the frequency for C in the previous octave. The 12^{th} root of 2 is the key. Also, the formula for the n^{th} root algorithm motivates the formula for the method of Newton/Raphson. As it turns out, the square root, cube root, and n^{th} root methods are all special cases of Newton/Raphson.

Simplicio: Why would we bother with the special cases then?

Galileo: Now you are thinking like a mathematician. If you have a general method, why not keep it simple and discard the special cases? However, from a pedagogical point of view, we like to discuss the easy cases first. Building on the experience we have gained from the easy cases, the general cases should be more accessible. We could begin our discussion with the method of Newton/Raphson. However, simple examples exist, which demonstrate that this method doesn't always work. Our square root method doesn't have this problem.

Simplicio: Now you have me worried.

Galileo: Mathematicians always worry. However, after showing you how to compute square roots, cube roots, and n^{th} roots, we present Cardano's formula for computing the roots of a cubic polynomial. This nifty formula requires that you are able to compute square roots and cube roots.

Simplicio: That sounds fine.

Galileo: The next set of topics will be focused on different root finding techniques. In particular, we will present the Newton/Raphson, secant, and bisection methods.

Simplicio: Techniques are good. I am sure I will enjoy it.

Galileo: After we discuss these three algorithms, the story turns ugly. We first show that Newton/Raphson fails in a fundamental way. Sometimes the algorithm produces a sequence, which diverges to infinity. Sometimes the sequence converges

to an unexpected answer. Occasionally, the sequence simply oscillates.

Simplicio: This is not the news I wanted to hear.

Galileo: Unfortunately, the evil Mr. Murphy is lurking behind every clever algorithm. He will pounce when you least expect it. In addition to we will mention a famous example of James Wilkinson, which shows that the roots of a 20 degree polynomial can lead to dangerous instabilities. In other words, you are insane if you model a real-world problem with a high degree polynomial.

Simplicio: OK, OK.

Galileo: The next discussion will focus on the successes we can salvage from our collection of disasters. In an effort to understand and rectify these issues, we turn to mathematics.

Simplicio: Does this mean theory?

Galileo: When you hit the square root button on your calculator, you would like to get the correct answer, wouldn't you?

Simplicio: I have no argument with correct answers.

Galileo: Actually, you are making too much of a big deal about mathematical rigor. We did all the heavy lifting yesterday when we defined and discussed convergence. We will show the method of Archimedes/Heron "always works." The words bounded and increasing will reappear.

Virginia: I look forward to these insights.

Virginia: What's next?

Galileo: The next goal is to demonstrate mathematically why one method might be preferred over another.

Simplicio: What does the word "preferred" mean in this context?

Galileo: If it takes 5 iterations to compute the square root of a number with one method and 30 iterations with another, which would you prefer?

Simplicio: Hmmm.

Galileo: Surprisingly, the Mean Value Theorem and Taylor's Theorem will drive this discussion. We are interested in the problem of when one sequence converges faster

than another.

Simplicio: Wait a minute. What does it mean for one sequence to converge faster than another?

Galileo: Now you are thinking like a mathematician. The first type of convergence is called *first order* or *linear*. The second is called *second order* or *quadratic*. The Mean Value Theorem is the tool for showing a sequence converges linearly. Taylor's Theorem is used to show Newton/Raphson (usually) converges quadratically. As you will see, quadratic convergence is preferred.

Virginia: So Newton/Raphson is preferred when it works!

Galileo: Correct. If one is not careful, Murphy will get you.

Virginia: What is next?

Galileo: The process of understanding the method of Newton/Raphson leads to the amazingly general Contraction Mapping Theorem. Once the terms *contraction* and *fixed point* have been defined, this theorem is easy to state, easy to prove, and even easier to implement. The method always works. Better yet, a multitude of applications are connected with this theorem including the solution of linear equations, non-linear equations, the solution of differential equations, and the creation of fractal patterns. This technique represents the best of all possible mathematical worlds.

Virginia: Great.

Galileo: We will finish the day with a discussion of Aitken's method. The goal of this technique is to speed up the rate of convergence from linear to quadratic. While it works well in some cases, it is not as useful as one might hope.

Simplicio: What? You are going to waste our time by showing us methods that don't work?

Galileo: While Aitken has his place in the world of numerical methods, his technique does little to speed up the bisection method. This is just one example. The sad truth is that the highway of numerical techniques is littered with good ideas that failed to perform as hoped.

Virginia: Let me summarize today's agenda:

1. the square root technique of Archimedes/Heron,
2. general root finding techniques,
3. failure of general methods,
4. success of general methods,
5. analysis of convergence rates,
6. generalization of Newton/Raphson to the Contraction Mapping Theorem, and
7. Aitken's Method to improve the convergence rate.

Galileo: You got it.

Simplicio: The program makes sense to me.

Chapter 5

The Computation of n^{th} Roots

5.1 Cube Roots

Galileo: Since we now understand how to compute square roots, we now turn to the problem of computing cube roots. Our strategy will be to imitate the approach described for square roots. This time we will again assume that the quantity x is a reasonably close approximation of $\sqrt[3]{K}$ and now search for the quantity Δx such that $(x + \Delta x)^3 = K$. While the picture is more difficult to draw than for the 2-dimensional case, it can be visualized by simply replacing the square by a cube as we have attempted in Figure 5.1.

Again, if Δx is small, then Δx^2 and Δx^3 are even smaller, so we find

$$\begin{aligned} K &= (x + \Delta x)^3 \\ &= x^3 + 3x^2\Delta x + 3x\Delta x^2 + \Delta x^3 \\ &\approx x^3 + 3x^2\Delta x. \end{aligned}$$

Thus, if we let $\Delta x = \frac{K - x^3}{3x^2}$ and replace x by x_n and $x + \Delta x$ by x_{n+1}, we have the following cube root algorithm:

$$\begin{aligned} x_0 &= 1, \\ x_{n+1} &= x_n - \tfrac{x_n^3 - K}{3x_n^2}, n \geq 0. \end{aligned}$$

Simplicio: This discussion is quite familiar.

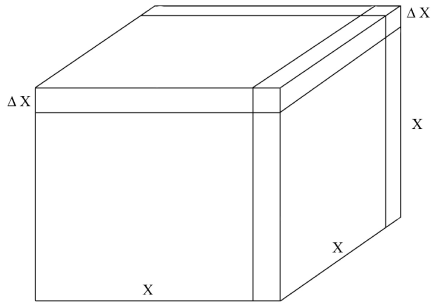


Figure 5.1: The Geometry Underneath the Cube Root Algorithm

Example 5.1.1. *Galileo: OK, it is time to work an example. In Table 5.1 we display the first six approximations of* $\sqrt[3]{2}$.

x_0	1.000000000000000
x_1	1.333333333333333
x_2	1.263888888888889
x_3	1.259933493449977
x_4	1.259921050017770
x_5	1.259921049894873
x_6	1.259921049894873

Table 5.1: Six Estimates of $\sqrt[3]{2}$

Simplicio: This set of computations is amazing! Once again, the 5^{th} *and* 6^{th} *terms are identical out to* 15 *decimal places.*

Galileo: What else do you notice?

Virginia: After the initial guess, the terms are decreasing.

Galileo: In Figure 5.2 we once again display the locations of these estimates on the real number line. As you have noticed, the third estimate is less than the second.

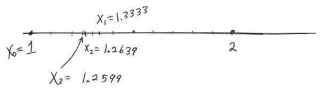


Figure 5.2: The First Three Estimates of $\sqrt[3]{2}$

Galileo: Very good. Now let's make a few remarks about the algorithm. Since the formula for x_{n+1} can also be written as

$$x_{n+1} = \frac{2x_n + \frac{K}{x_n^2}}{3} = \frac{2}{3}x_n + \frac{1}{3}\frac{K}{x_n^2},$$

it becomes apparent that x_{n+1} is the weighted average of x_n and $\frac{K}{x_n^2}$, where the first weight is $\frac{2}{3}$ and the second weight is $\frac{1}{3}$.

Archimedes: While I get annoyed when others try to take credit for my ideas, I am a bit embarrassed that you are assigning this method to me. We didn't even think about cube roots in those days.

Galileo: While you are correct, you must admit the concept is the same. While this generalization to the computation of cube roots may seem like an easy generalization of the method of Archimedes/Heron, the time gap is in terms of millennia.

Simplicio: Probably nobody cared.

Galileo: You may be right. Even today, square roots are used much more often than cube roots. In any case, the concept that bridged the gap was an improved understanding of algebra and the binomial theorem.

Exercise Set 5.1.

1. Let $K = 5$ and $x_0 = 1$. Compute the first five iterations of the cube root algorithm to estimate $\sqrt[3]{5}$. What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

2. Let $K = 10$ and $x_0 = 1$. Compute the first five iterations of the cube root algorithm to estimate $\sqrt[3]{10}$. What is the difference between your estimate and the exact answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

3. Let $K = 1000$ and $x_0 = 1$. Compute the first five iterations of the cube root algorithm to estimate $\sqrt[3]{1000}$. How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

4. Let $K = 1,000,000$ and $x_0 = 1$. Compute the first five iterations of the cube root algorithm to estimate $\sqrt[3]{1,000,000}$. How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? Compare the number of iterations with your answer for $\sqrt[2]{1,000,000}$. Which algorithm takes more iterations? (Make your computations with 10 digits of accuracy.)

5. Let $K = 10^9$ and $x_0 = 1$. Compute the first five iterations of the cube root algorithm to estimate $\sqrt[3]{10^9}$. What do you notice? How close is the last estimate to the correct answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? (Make your computations with 10 digits of accuracy.)

6. Let $K = 0$ and $x_0 = 1$. Compute the first five iterations of the cube root algorithm to estimate $\sqrt[3]{0}$. How close is the last estimate to the correct answer? How many iterations does it take before the difference between x_n and the exact answer is less than 0.000001? Compare the number of iterations require for this problem and when you approximated $\sqrt{0}$. (Make your computations with 10 digits of accuracy.)

5.2 n^{th} Roots

Galileo: We now show how to generalize the method of computing cube roots to a method that can be used to compute the n^{th} root of a number.

Simplicio: Why would we care about n^{th} roots?

Galileo: What about music? Let's ask Pythagoras.

Pythagoras: Long ago I observed that two blacksmith's striking different anvils at the same time can produce resonating frequencies when one is twice the size of the other. With string instruments two strings produce resonating sounds when one is twice (or three times) the length of another and under the same tension.

Simplicio: How do you get the tensions to be the same?

Pythagoras: If you place the fret at the midpoint, the frequency is doubled.

Galileo: While we are at it, let me comment that a major concern of Fourier series is the problem of approximating functions $f(x) : [-\pi, \pi] \to \Re$ by linear combinations of functions of the form $1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots, \cos(nx), \sin(nx)$. Note that the frequency of $\cos(2x)$ is twice that of $\cos(x)$ and the frequency of $\cos(3x)$ is triple that of $\cos(x)$. We will return to this topic.

Simplicio: Interesting.

Galileo: Since my father was a musician, I find this subject of particular interest and would like to make a couple of additional remarks. Every piano has 12 notes from one octave to the next. As you progress up the scale, the frequency changes by the factor $\sqrt[12]{2}$. In the key of C, you begin with middle C as the first note, D is the second note, E is the third, F is the fourth, and G is the fifth. Thus, if you strike the fourth white key to the right of middle C, you have the perfect fifth. The frequency of middle C is 252 Hertz so the frequency of the perfect fifth is $252 * (\sqrt[12]{2})^7$.

Simplicio: What a strange way to tune an instrument? Why not simply tune the piano so the frequencies are equally spaced? That method would seem more reasonable to me.

Pythagoras: As I just remarked, if we were to use your strategy, then the frequency

of C (or any other note) in one octave would not be twice the frequency of C in the previous. Thus, our notes would not be harmonious. On the other hand, if the frequencies are spaced multiplicatively, then harmony is preserved.

Simplicio: I have another question. If the note G is called the perfect fifth, then why isn't it computed as $252 * (\sqrt[12]{2})^5$?

Galileo: The modern piano has black keys as well as white keys. These black keys are tuned as half notes (also known as semitones). The perfect fifth is seven half steps above middle C.

Pythagoras: And note that the quantity $(\sqrt[12]{2})^7 \approx \frac{3}{2}$.

Simplicio: Interesting.

Galileo: People frequently remark that music and mathematics go together. Well, there it is.

Now let's get back to the mathematical issue of computing the n^{th} root of a number K by following the strategy used for computing cube roots. To that end, suppose we have a number x which is a reasonably close approximation of $\sqrt[n]{K}$. We now would like want to approximate the quantity Δx with the property that $(x + \Delta x)^n = K$.

Again, if Δx is small, then for any integer $k > 1$, the power Δx^k is even smaller. For example, if $\Delta x = 0.1$, then $\Delta x^2 = 0.01$ and $\Delta x^3 = 0.001$. Thus, by the binomial theorem we find that

$$\begin{aligned} K &= (x + \Delta x)^n \\ &= x^n + nx^{n-1}\Delta x + \frac{n(n-1)}{2!}x^{n-2}\Delta x^2 + \frac{n(n-1)(n-2)}{3!}x^{n-3}\Delta x^3 + \cdots + \Delta x^n \\ &\approx x^n + nx^{n-1}\Delta x. \end{aligned}$$

Thus, a good choice for the approximate Δx is to set $\Delta x = \frac{K - x^n}{nx^{n-1}}$. If we set $x_k = x$ and $x_{k+1} = x + \Delta x$, then we have the following recursive algorithm for any $K > 0$:

$$\begin{aligned} x_0 &= 1, \\ x_{k+1} &= x_k - \frac{x_k^n - K}{nx_k^{n-1}}. \end{aligned}$$

Simplicio: Given the previous discussions on square roots and cube roots, the technique is quite understandable.

Galileo: Again, note that we have taken a difficult problem, non-linear in the variable Δx, and made it linear in that variable.

Virginia: Is that so the problem is easier?

Galileo: Correct. Note also that we can again write x_{k+1} as the weighted sum of x_k and $\frac{K}{x_k^{n-1}}$. In particular,

$$x_{k+1} = \frac{n-1}{n} x_k + \frac{1}{n} \frac{K}{x_k^{n-1}},$$

where the two weights are $w_0 = \frac{n-1}{n}$ and $w_1 = \frac{1}{n}$.

Simplicio: OK, this discussion is getting all too familiar. How about an example?

Example 5.2.1. *Galileo:*

We have presented the first six approximations for $\sqrt[5]{2}$ in Table 5.2.

x_0	1.000000000000000
x_1	1.200000000000000
x_2	1.152901234567901
x_3	1.148728886527325
x_4	1.148698356619959
x_5	1.148698354997035
x_6	1.148698354997035

Table 5.2: Six Estimates of $\sqrt[5]{2}$

Simplicio: These computations are getting boring. I can see that the questions and answers are the same as for square roots and cube roots.

Example 5.2.2. *Galileo: We have presented the first six approximations for* $\sqrt[12]{2}$ *in Table 5.3.*

x_0	1.000000000000000
x_1	1.083333333333333
x_2	1.062153572038919
x_3	1.059500262653840
x_4	1.059463101529905
x_5	1.059463094359296
x_6	1.059463094359295

Table 5.3: Six Estimates of $\sqrt[12]{2}$

Simplicio: Finally something happened! At least we have a difference in the 15^{th} *digit for the* 5^{th} *and* 6^{th} *estimates.*

Galileo: This algorithm is worthy.

Exercise Set 5.2.

1. Compute $\sqrt[5]{2}$ using $x_0 = 1$ to initialize the algorithm. How many iterations does it take before the error is less than 0.000001? (Make your computations with 10 digits of accuracy.)

2. Compute $\sqrt[7]{2}$ using $x_0 = 1$ to initialize the algorithm. How many iterations does it take before the error is less than 0.000001? (Make your computations with 10 digits of accuracy.)

3. Compute the first five iterations of the n^{th} root algorithm to estimate $\sqrt[12]{2}$ using $x_0 = 1$ to initialize the method. How many iterations does it take before the error is less than 0.000001? (Make your computations with 10 digits of accuracy.)

4. Compute the first five iterations of the n^{th} root algorithm to estimate $\sqrt[20]{2}$ using $x_0 = 1$ to initialize the method. How many iterations does it take before the

error is less than 0.000001? Compare the number of iterations required with the previous three problems. (Make your computations with 10 digits of accuracy.)

5. Compute the first five iterations of the n^{th} root algorithm to estimate $\sqrt[12]{0}$ using $x_0 = 1$ to initialize the method. How many iterations does it take before the error is less than 0.0001? (Make your computations with 10 digits of accuracy.)

Chapter 6

Cardano's Method for Cubic Polynomials

Girolamo Cardano (1501-1576)

> I wrote it out five times, may it last the same number of millennia.- Girolamo Cardano

Galileo: Since we now understand how to compute square roots, cube roots, and n^{th} roots, we now turn to the problem of computing roots of cubic polynomials. First, let us remind you that the solutions of the quadratic equation $Ax^2 + Bx + C = 0$ are given by $r = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$.

Simplicio: Sure, I remember that formula. I learned it many years ago.

Galileo: Well then, can you solve the general cubic equation $Ax^3+Bx^2+Cx+D=0$?

Simplicio: I must admit I have forgotten that formula.

Galileo: Actually, the development of these formulas has a long and sometimes bitter history.

While it may be true that the Babylonians were the first to solve quadratic equations sometime around 400 B.C.E., this statement is a bit of an oversimplification since the Babylonians had no notion of "equation." What they did develop was an algorithmic approach to solving problems which, in our terminology, would give rise to a quadratic equation. The method is essentially the technique of "completing the square." Of course, the ancient Greek mathematicians knew how to solve the quadratic formula by ruler and compass.

Omar Khayyam (1048 - 1122)

Algebras are geometric facts which are proved.-Omar Khayyam

Nearly 1500 years later, we find the first success at solving a cubic equation. While trying to solve the problem of finding a right triangle with the property that the hypotenuse equals the sum of one leg plus the altitude of the hypotenuse, the Persian mathematician and poet, Omar Khayyam (1048 - 1131), found a positive root to the cubic equation $x^3 + 200x = 20x^2 + 2000$. The mathematics world would have to wait another 400 years for a solution to the general cubic equation and the solution would

not come easily. The Italian mathematician Scipione del Ferro (1465-1526) designed algebraic solutions to cubic equations of the form $x^3 + mx = n$.

Simplicio: Did del Ferro publish his work?

Virginia: He made the mistake of showing his ideas to his student Antonio Fior.

Simplicio: How so?

Virginia: Didn't he compete in a challenge, where each contestant gave the other thirty problems to solve?

Figure 6.1: Niccolo Fontana (1499-1557), aka Tartaglia, the Stutterer

When the cube and the things together
Are equal to some discrete number,
Find two other numbers differing in this one.
Then you will keep this as a habit
That their product shall always be equal
Exactly to the cube of a third of the things.
The remainder then as a general rule
Of their cube roots subtracted
Will be equal to your principal thing.-Niccolo Fontana

Galileo: Correct. The other contestant was another Italian mathematician, Niccolo Fontana (1499-1557), known as Tartaglia, the stutterer.

Simplicio: Why was he called the stutterer?

Galileo: When he was a teenager, the French invaded his home town. In the process, a soldier bashed the young fellow in the head causing such severe and permanent injuries he found it difficult to speak.

Simplicio: So what contribution did Tartaglia make to the problem of solving cubics?

Galileo: Tartaglia's methods were more general and were able to solve cubics of the form $x^3 + mx^2 = n$. Fior's methods cold not handle this case and Tartaglia won the challenge. This challenge between Fior and Tartaglia sparked the interest of yet another Italian mathematician, Girolamo Cardano (1501-1576).

Simplicio: So who was Cardano?

Galileo: Cardano was an unusually cantankerous fellow, who was schooled in the field of medicine. However, because of his reputation as a difficult man he was not admitted to the College of Physicians in Milan. This rejection forced him to establish a small medical practice of his own. Cardano's practice, however, could not pay his gambling bills, so when a mathematics lecturing position became available at the Piatti Foundation in Milan, he took it. After hearing of Tartaglia's success with a solution to the cubic equation, Cardano attempted, without success, to learn Tartaglia's methods. Cardano first contacted Tartaglia through an intermediary to request that his method be included in Cardano's soon-to-be published book. Tartaglia declined Cardano's request stating that he intended to publish the method himself. Cardano then persuaded Tartaglia to explain his method.

Tartaglia did not just simply tell Cardano his results. Instead, he wrote them in a poem, so that if it were to fall into the wrong hands, they would still be safe. Furthermore, he insisted that Cardano would not publish the results. Cardano, with the help of Tartaglia's method, was able to find proofs for all cases of the cubic. He even solved the quartic equation. Some years later, Tartaglia still had not published his results. Cardano then learned that del Ferro, not Tartaglia, had been the first to solve the

cubic. Cardano used this new information to justify publishing Tartaglia's method. While Cardano gave Tartaglia full recognition, Tartaglia never forgave Cardano.

Virginia: I can understand why. The formulas are known as Cardano's formulas. Poor old Tartaglia is never mentioned.

Galileo: There are many bitter stories like this one in academics. The profession seems to attract people who have a tendency to involve themselves in this type of politics.

Simplicio: I think my decision to go into business may have been wise.

Galileo: As we noted the general cubic equation can be reduced to an equation of the form, where the quadratic term equals zero. Thus, we can assume that the cubic has the form:

$$p(x) = x^3 + px + q = 0.$$

For a cubic equation of this form, Cardano's Formula 6.2 shows that one root can be written in the form:

$$r = \frac{1}{\sqrt[3]{2}}\sqrt[3]{-q + \sqrt{q^2 + \frac{4}{27}p^3}} + \frac{1}{\sqrt[3]{2}}\sqrt[3]{-q - \sqrt{q^2 + \frac{4}{27}p^3}}.$$

Figure 6.2: Cardano's Formula

Virginia: I like this formula because it shows the roots of a cubic equation can be written in terms of square roots and cube roots.

Simplicio: I agree that Cardano and his friends have produced an amazing formula.

Galileo: Not so fast. Note that care must be exercised when we actually apply the formula. A problem arises because the square root always generates two answers and the cube root function always generates three answers. (Of course, the square root and cube root of zero is zero, so that number is an exception.) Thus, this expression for r could generate as many as 12 different "answers." However, this problem will be avoided if we assume p and q are real numbers and the expression $q^2 + \frac{4}{27}p^3$ is positive. In this setting, we can make the convention that we choose the positive

square root $\sqrt{q^2 + \frac{4}{27}p^3}$ in both parts of the formula for r. Since $-q + \sqrt{q^2 + \frac{4}{27}p^3} > 0$ and $-q - \sqrt{q^2 + \frac{4}{27}p^3} < 0$, we can always find a unique real cube root of each. If we follow this convention and thus avoid choosing complex numbers, then r will be a root.

Virginia: What if $q^2 + \frac{4}{27}p^3$ is negative?

Galileo: We then have to get distracted by the subject of complex numbers. Since we have many more topics to discuss, let us move on.

Virginia: Are there similar formulas for polynomials of all degrees?

Galileo: Unfortunately, the answer to that question is no. While the general quartic equation can also be solved using only square roots and cube roots, the Norwegian mathematician Niels Henrik Abel (1802-1829) and the French mathematician Everiste Galois (1811-1832) showed that no such formula exists for the equation $x^5 + x + 1 = 0$. Of course, we should not forget that Gauss proved the Fundamental Theorem of Algebra around 1800. In fact, he produced five different proofs. The beauty of this theorem is that it states that every polynomial $p_n(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_1x + a_0$, where each a_k is a complex number, has the property that it can be factored as a product of linear factors in its roots. In other words, roots $r_1, r_2, \ldots, r_n$ can be found so that $p_n(x) = (x-r_1)(x-r_2)\ldots(x-r_n)$. If we count multiplicities, we see that every polynomial of degree $n \geq 1$ has exactly n real roots. Unfortunately, the bad news is that the work of Abel and Galois shows that we will be unable to find a tidy little formula for these roots.

Simplicio: I notice that these two fellows Abel and Galois both died at an early age.

Galileo: While Abel died of tuberculosis, Galois was shot and killed in a duel over politics or a woman. It seems that he had a penchant for getting into trouble. A year before his death, he made threats against King Louis-Phillipe while at a dinner with 200 Republicans. While making his speech, he may have been holding a dagger in his hand.

Virginia: Is it not true that trouble seems to have followed you as well.

Galileo: At least I left my daggers at home.

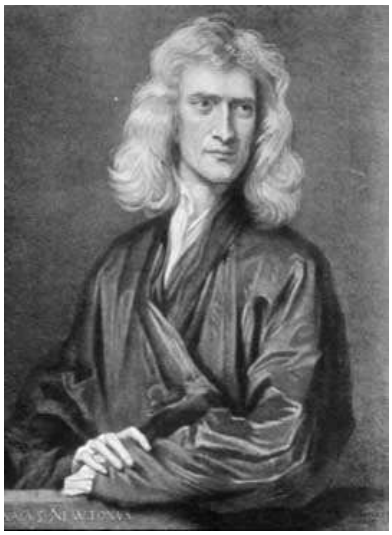
Simplicio: Again, I think my decision to avoid a career in academics may have been wise.

Exercise Set 6.1.

1. Compute a root of the equation $x^3 + x + 1 = 0$.
2. Find a root for Omar Khayyam's equation $x^3 + 200x = 20x^2 + 2000$.
3. Show that the quantity r given by the Cardano Formula 6.2 actually produces a root for the equation $x^3 + px + q = 0$. (Hint: Substitute $x = r$ into $p(x)$.)
4. Compute a root of the equation $x^3 + x^2 + 1 = 0$.
5. Find a formula for a root of the equation $x^3 + Ax^2 + Bx + C = 0$. (Suggestion: Surf the internet to see what others have done.)
6. Show the equation $x^3 + x + 1 = 0$ has exactly one real root.

Chapter 7

Algorithms for Finding Roots



Isaac Newton (1642-1727)

> If I have been able to see further, it was only because I stood on the shoulders of giants.-Isaac Newton

Galileo: We now introduce the English mathematician Isaac Newton (1642-1727), who is one of the giants in physics and mathematics. His treatise, *Principia*, is probably the most important science book ever written because it created mathematical models that explained the motion of the projectiles, planets, pendulums, fluids, and the tides. These models are based on fundamental principles concerning the nature of force, including gravitational and centripetal. His Second Law of Motion, $F = ma$ and his

inverse square law for gravitation are probably his most famous. The mathematical foundation for this work was geometry, geometry, geometry.

Simplicio: Wait a minute. What about Calculus?

Galileo: If you actually open this magnificent book, you will notice an abundance of triangles, parallelograms, and ellipses. You will find no derivatives $\frac{dy}{dx}$. Old Is aac was too smart to justify his methods on mathematics that was not quite ready for prime time. Of course, the spirit of Calculus was present everywhere.

Simplicio: Sounds like a lot of math theory to me. Did he include any data to support his theory?

Galileo: In fact, he did. Remember that the idea that the orbits of the planets might be elliptical comes from Kepler. The basis for his ideas was the data set acquired by Tycho Brahe (1546-1601). Newton actually included other astronomical data in his "Principia."

Tell us about yourself, Sir Isaac.

Newton: While I was interested in a variety of different subjects including chemistry and theology, my main interest was in physics and mathematics. In physics, I made fundamental contributions to dynamics, statics, optics, hydrodynamics, hydrostatics, and of course I discovered Calculus.

Virginia: I thought Gottfried Wilhelm von Leibniz (1646-1716) also invented Calculus.

Newton: Yes, you might have heard about that controversy. However, as the president of the Royal Society, I appointed an "impartial" committee to decide whether Leibniz or myself was the sole inventor. The official report of this illustrious committee concluded that I deserve full credit for the Calculus as we know it. Of course, I used the Calculus to explain the motion of falling bodies, Kepler's three laws of planetary motion, as well as the tides.

Galileo: But who wrote the report?

Newton: Well, I did.

Galileo: Enough of that. Let us mention, however, that Joseph Raphson (1648-1715) was a contemporary of yours, but used the same method to approximate roots of an

equation. Raphson, however, was one of the few people who you allowed to see your mathematical papers.

Newton: He took a clear position in favor of my claims over those of Leibniz. I appreciated his support.

(Newton leaves.)

Virginia: I am not certain that I would like to converse with that Mr. Newton again. He is a most unpleasant fellow.

Galileo: A great mind may possess a small personality. How about if we forgot all that politics and refocus our energies on his method. Since has been such a cad about the efforts of others, I think we should give Raphson equal credit?

7.1 The Method of Newton/Raphson

Galileo: Professor Newton, could you explain the ideas behind your method?

Newton: Certainly. Let us begin this section with the definition of the term root.

Definition 7.1.1. *If X is an interval and $f(x) : X \to \Re$ is a function, then a point $r \in X$ is called a root of $f(x)$ if $f(r) = 0$.*

Newton: The fundamental principle underlying the method is to "linearize the problem" by approximating a non-linear function by a straight line. Thus, easiest starting point is to find the root of the function $f(x) = m(x - x_0) + b$.

Simplicio: Even I can do that. All you have to do is solve the equation $0 = m(r - x_0) + b$. As long as $m \neq 0$, the root $r = x_0 - \frac{b}{m}$.

Newton: My method is not much more difficult. Since the first derivative of a function is the slope of the line that "best approximates" the curve $y = f(x)$ at a given point $(x_0, f(x_0))$, we begin the process by drawing a tangent line to the curve at this point. Since the tangent line to the curve $y = f(x)$ at a point x_0 is given by $y = f(x_0) + f'(x_0)(x - x_0)$, and the root of this linear equation is found when $y = 0$, the x-intercept is found by solving the equation $0 = f(x_0) + f'(x_0)(x - x_0)$, for x.

When we do this, we find that $x = x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. If x_n represents the approximation at the n^{th} iteration, then $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.

The Newton/Raphson Algorithm:

$$\begin{aligned} x_0 &= \text{an initial guess.} \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \text{ for all } n \geq 0. \end{aligned}$$

The recursive part of the algorithm can be thought of as a generalization of the cube root algorithm $x_{n+1} = x_n - \frac{x_n^3 - K}{3x_n^2}$, where the denominator of the fractional expression is also the derivative of the numerator.

Simplicio: Actually, I am quite comfortable with this algorithm.

Example 7.1.1. *Galileo: We now include a practice problem. If we would like to approximate the value of* $\sqrt{2}$*, then we can let* $x_0 = 1$ *and begin computing using the recursive formula stated in the algorithm. Notice that the first step is to think up a function* $f(x)$ *which has the property that* $r = \sqrt{2}$ *is a root.*

Virginia: How about the function $f(x) = x^2 - K$*?*

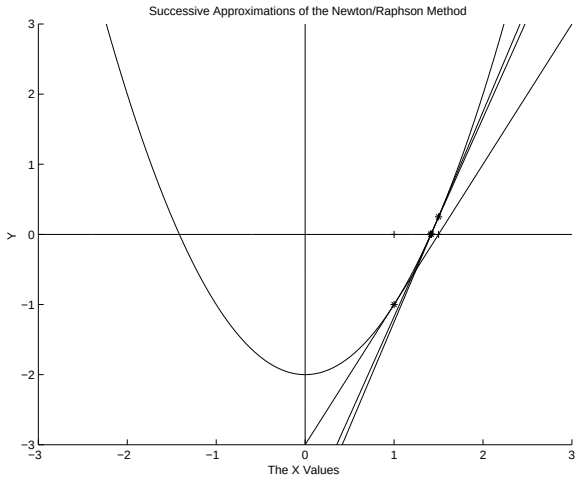


Figure 7.1: Five Steps Newton/Raphson Estimates for $f(x) = x^2 - 2$

Galileo: The approximations provided by the first five steps of the method are displayed in Figure 7.1. Note that x_2 is between the root $r = \sqrt{K}$ and x_1, x_3 is between $r = \sqrt{K}$ and x_2, and x_4 is between $r = \sqrt{K}$ and x_3. This pattern continues indicating that there is a strong probability that the sequence of $x-$intercepts for the tangent lines will converge to the root.

Virginia: Is the concavity of the curve important?

Galileo: In fact, it is. But we will discuss that thought in more detail at a later time.

Example 7.1.2. *Galileo: A second example is the polynomial $p(x) = x^5 + x + 1$. This example is of particular interest because our friends Abel and Galois showed we have no option except numerical computation of the roots.*

Here is the algorithm.

$$\begin{aligned} \textit{Step 0. } x_0 &= 1.0 \\ \textit{Step 1. } x_1 &= x_0 - \frac{x_0^5 + x_0 + 1}{5x_0^4 + 1} \\ \textit{Step 2. } x_2 &= x_1 - \frac{x_1^5 + x_1 + 1}{5x_1^4 + 1} \\ \textit{Step n. } x_{n+1} &= x_n - \frac{x_n^5 + x_n + 1}{5x_n^4 + 1} \end{aligned}$$

The first seven estimates of the real root are listed in Table 7.1 when the algorithm is initialized with $x_0 = 1$.

Simplicio: What a great algorithm! While not quite as good as the square root and cube root methods, this technique is still in my comfort zone.

Galileo: The method of Newton/Raphson is popular.

Virginia: I can see why.

Simplicio: I do have one quick question. If this method includes the square root and cube root techniques as special cases, why not skip them? It certainly would have been more efficient to simply discuss the Newton/Raphson Algorithm at the beginning.

x_0	1.000000000000000
x_1	0.500000000000000
x_2	-0.666666666666667
x_3	-0.768115942028985
x_4	-0.755162523060901
x_5	-0.754877799264274
x_6	-0.754877666246722
x_7	-0.754877666246693

Table 7.1: Seven Estimates of a Root of $p(x) = x^5 + x + 1$

Galileo: While we could have, there is a difference between presenting mathematics in its most perfect final form and presenting concepts to someone unfamiliar with the subject. In my experience, the human brain works inductively from particular cases to more general ones. Mathematics is a process, which has been unfolding for several thousand years. The pedagogic rule we will follow is to proceed from the particular to the abstract.

Simplicio: I actually agree with this approach. Simple is good.

Galileo: We will soon discuss examples, where the method of Newton/Raphson fails. These examples will encourage us to search for algorithms, which "always work." The square root and cube root algorithms do in fact enjoy this comforting property.

Exercise Set 7.1.

1. Set up the Newton/Raphson algorithm to compute $\sqrt[5]{2}$. Test the method by using $x_0 = 2$ to initialize the method and compute 6 iterations.

2. Use the method of Newton/Raphson to compute a root of the polynomial $p_3(x) = x^3 + x + 1$ with error less than 10^{-5}. Initialize the method with $x_0 = 1.0$.

3. Use the method of Newton/Raphson to compute a root of the the polynomial $p_3(x) = x^3 + x^2 + 1$ with error less than 10^{-5}. Initialize the method with $x_0 = 1.0$.

4. Use the method of Newton/Raphson to compute a root of the polynomial $p_5(x) = (x-1)(x-2)(x-3)(x-4)(x-5)$ with error less than 10^{-5}. Initialize the method with $x_0 = 5.10$.

5. Use the method of Newton/Raphson to compute a solution of Omar Khayyam's equation $x^3 + 200x = 20x^2 + 2000$ with error less than 10^{-5}. Initialize the method with $x_0 = 1.0$. Compare your answer with the one produced by Cardano's Formula 6.2.

6. Use the method of Newton/Raphson to compute a root of the function $f(x) = x\cos(x)$ with error less than 10^{-5}. Initialize the method with $x_0 = 10$. Be sure to make your computations using radians rather than degrees.

7. Use the method of Newton/Raphson to compute a root of the function $f(x) = x\ e^x$ with error less than 10^{-5}. Initialize the method with $x_0 = 1.00$ and $x_0 = -2.00$.

8. Use the Newton/Raphson method to approximate a root of the polynomial $p_7(x) = x^7 + x + 1$ with error less than 10^{-5}. Initialize the method with $x_0 = 1.0$.

9. Use the method of Newton/Raphson to approximate a solution of the equation $\sin(x) = e^x$ with error less than 10^{-5}. Initialize with $x_0 = 0$ and $x_0 = 5$. What do you notice?

10. Use the method of Newton/Raphson to approximate a solution of the equation $e^x = 3x^2$ with error less than 10^{-5}. Initialize with $x_0 = 0$ and $x_0 = 5$. What do you notice?

11. Use the method of Newton/Raphson to approximate a solution of the equation $log_e(x) = -\cos(x)$ with error less than 10^{-5}. Initialize the method with $x_0 = 0.5$. If the initialization is changed to $x_0 = 2.0$, then what happens?

12. Let $p_2(x) = (x-1000)^2$ and $q_2(x) = x^2 - 1000000$. Note that $x = 1000$ is a root for both $p_2(x)$ and $q_2(x)$. Use the method of Newton/Raphson to approximate

this root for both polynomials. Initialize the method with $x_0 = 1001$. Compare the number of iterations required to achieve an error of less than 10^{-5}. What do you notice? What is different about the roots of the two polynomials?

7.2 The Secant Method

Galileo: We now turn to a variant of Newton/Raphson known as the secant method, where the first derivative is approximated numerically as the slope of the line through the two previous approximations produced by the algorithm. This modification is important in applications, where the first derivative is difficult to compute using the usual rules of differential Calculus. Instead of having the term $f'(x)$ in the denominator of the second term, the approximation $\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$ is used.

Thus, the $(n+1)^{st}$ term becomes:

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\ &\approx x_n - \frac{f(x_n)}{\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}} \\ &= x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}. \end{aligned}$$

Since we require two values to initialize the algorithm, the secant method can implemented as:

The Secant Algorithm:

$$\begin{aligned} \text{Step 0. } x_0, x_1 &= \text{initial estimates} \\ \text{Step n. } x_{n+1} &= x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}. \end{aligned}$$

Simplicio: OK, I see that the secant method has the advantage that you don't have to compute the first derivative. How about an example?

Example 7.2.1. *Galileo: While this example is a bit embarrassing becasue the first derivative is easy to compute, why not begin by applying the secant method to recompute our old friend* $\sqrt{2}$*? For this computation, we choose* $f(x) = x^2 - 2$*. If we initialize*

the method with the points $x_0 = 1$ *and* $x_1 = 2$*, the first secant line approximation is given by the equation* $y = -1 + \frac{2+1}{2-1}(x-1) = 3x - 4$.

Simplicio: I see that this secant line intersects the $x-$*axis at* $x_2 = 1.33333333333333$.

Galileo: Correct. A graph of the function $y = f(x) = x^2 - 2$ *and the first approximating secant line are graphed in Figure 7.2.*

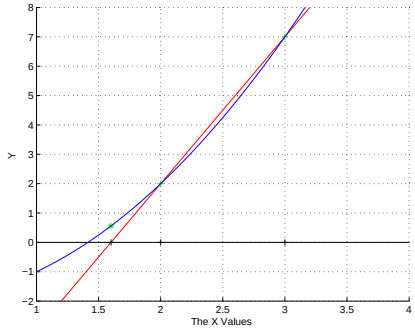


Figure 7.2: The First Secant Approximation for $f(x) = x^2 - 2$

Galileo: In addition, we display the first eight data points generated by the algorithm in Table 7.2.

x_0	1.00000000000000
x_1	2.00000000000000
x_2	1.33333333333333
x_3	1.40000000000000
x_4	1.41463414634146
x_5	1.41421143847487
x_6	1.41421356205732
x_7	1.41421356237310
x_8	1.41421356237310

Table 7.2: Eight Secant Method Estimates of $\sqrt{2}$ Initialized by $x_0 = 1, x_1 = 2$

Simplicio: While the algorithm seems to converge quickly, it does appear to be a shade

slower than the method of Archimedes/Heron.

Galileo: Good observation. As it turns out, the convergence rate for the secant method is generally slower than the convergence rate for Newton/Raphson. We will make that statement more precise on another occasion.

Virginia: It doesn't look like the sequence of approximations is decreasing.

Galileo: Another good observation. However, if we change the initialization to $x_0 = 1$ *and* $x_1 = 2$*, then the algorithm behaves the way you might expect. We have listed this data in Table 7.3.*

x_0	3.00000000000000
x_1	2.00000000000000
x_2	1.60000000000000
x_3	1.44444444444444
x_4	1.41605839416058
x_5	1.41423305925716
x_6	1.41421357508149
x_7	1.41421356237318
x_8	1.41421356237309

Table 7.3: Eight Secant Method Estimates of $\sqrt{2}$ Initialized by $x_0 = 3, x_1 = 2$

Virginia: Now the sequence is decreasing.

Simplicio: Does that always happen?

Galileo: Stick around and you will see.

Virginia: Are there any disadvantages to this technique?

Galileo: The first problem is that you need two starting points instead of one.

Simplicio: Why should that matter?

Galileo: If they aren't chosen close to the answer, the estimates may fail to converge to the desired answer. Later, we will give an example illustrating this issue.

Simplicio: Are there any other issues?

Galileo: You also have to be careful not to divide by zero. This problem is a real and dangerous possibility with the secant method whenever two successive approximations, $f(x_n)$ and $f(x_{n-1})$ are approximately equal because their difference $f(x_n) - f(x_{n-1})$ is close to zero and is in the denominator. In fact, if we had computed a few more terms with our approximations of $\sqrt{2}$, we would have had an explosion caused by a division by zero.

Simplicio: I think I can program around that issue.

Exercise Set 7.2.

1. If $K = 2, f(x) = x^2 - K, x_0 = 1$, and $x_1 = K = 2$, then use the secant method to compute x_{11} and x_{12}. What happens?

2. If $K = 5, f(x) = x^2 - K, x_0 = 1$, and $x_1 = K = 5$, then use the secant method to compute the root with an accuracy of $\frac{1}{10{,}000}$. How many iterations are required? Compare the estimates generated by the secant method with those generated by the Newton/Raphson method when $x_0 = 1$. Which is faster: the secant method or Newton/Raphson?

3. If $K = 1{,}000{,}000, f(x) = x^2 - K, x_0 = 1$, and $x_1 = K = 1{,}000{,}000$, then use the secant method to compute the root with an accuracy of $\frac{1}{10{,}000}$. How many iterations are required? Compare the estimates generated by the secant method with those generated by the Newton/Raphson method when $x_0 = 1$.

4. If $K = 2, x_0 = 1, x_1 = K = 2$, and $f(x) = x^3 - K$, then how many iterations will be required for the secant method to estimate a root of $f(x)$ to an accuracy of $\frac{1}{10{,}000}$. Compare the number of iterations required for the secant method and the number required by the Newton/Raphson method when $x_0 = 1$.

5. Use the secant method to compute a root of the polynomial $p(x) = x^3 + x + 1$ with error less than 10^{-5}. Initialize the method with $x_0 = 0.0$ and $x_1 = 1.0$. Compare the number of iterations required for the secant method and the number required by the Newton/Raphson method when $x_0 = 1$.

6. Use the secant method to compute a root of the polynomial $p(x) = x^5+x+1$ with error less than 10^{-5}. Initialize the method with $x_0 = 0.0$ and $x_1 = 1.0$. Compare the number of iterations required for the secant method and the number required by the Newton/Raphson method when $x_0 = 1$.

7. Use the secant method to compute a root of the of the polynomial $p(x) = (x-1)(x-2)(x-3)(x-4)(x-5)$ with error less than 10^{-5}. Initialize the method with $x_0 = 0.5$ and $x_1 = 1.5$. Compare the number of iterations required for the secant method and the number required by the Newton/Raphson method when $x_0 = 0.5$

8. Use the secant method to compute a root of the Omar Khayyam's equation $x^3 + 200x = 20x^2 + 2000$ with error less than 10^{-5}. Initialize the method with $x_0 = 0.0$ and $x_1 = 1.0$. Compare the number of iterations required for the secant method and the number required by the Newton/Raphson method when $x_0 = 1$.

9. Müller's Method: Determine a recursive formula that uses three successive points to determine the next approximation to a root r for a function $y = f(x)$. In other words, given three points x_0, x_1, x_2, find a parabola $p_2(x) = A(x - x_2)^2 + B(x - x_2) + C$ with the property that $p_2(x_0) = f(x_0), p_2(x_1) = f(x_1)$, and $p_2(x_2) = f(x_2)$. After computing the constants A, B, and C, then use the quadratic formula to compute an approximate root x_3. Note further that since the quadratic formula provides two roots, the choice with the largest denominator is preferred.
(Answer: $A = \frac{(x_1-x_2)[f(x_0)-f(x_2)]-(x_0-x_2)[f(x_1)-f(x_2)}{(x_0-x_2)(x_1-x_2)(x_0-x_1)}$,
$B = \frac{(x_0-x_2)^2[f(x_1)-f(x_2)]-(x_1-x_2)^2[f(x_0)-f(x_2)]}{(x_0-x_2)(x_1-x_2)(x_0-x_1)}$, and $C = f(x_2)$.)

Simplicio: But wait a minute. The functions in these exercises all have first derivatives that are easy to compute. Wouldn't we simply use Newton/Raphson?
Galileo: To illustrate a situation, where you might want to choose the secant method consider the polynomial $p_{20}(x) = (x-1)(x-2)\ldots(x-20)$. Note that the roots of

$p_{20}(x)$ are the integers $r = 1, 2, \ldots, 20$. While the value of $p_{20}(x)$ can be computed for any value of x, the first derivative requires you to either expand the function as a 20 degree polynomial or compute 20 product rules. Take your pick. Better yet, implement the secant method for finding a root for $p_{20}(x)$ and then test the method for two initial input points x_0 and x_1, where x_0 and x_1, are chosen near the root $r = 1$ and near the root $r = 20$. Compare your results for two different sets of inputs.

Simplicio: I get the concept, but what about computing $p_{20}(x)$ when $x = 21$? By my calculation, I get 20!, which is a very large number. In fact, it turns out to be equal to about 2.4329×10^{18}.

Galileo: You are very perceptive. We will see shortly that the computation of the roots of this polynomial lead to a fundamentally unstable problem. In fact, this problem offers a view into exactly the type of problem applications people must either avoid or enter into at great risk.

7.3 The Bisection Method

Galileo: The bisection method is probably the most basic method for finding a root of a continuous function. This method is a straightforward application of the Intermediate Value Theorem 10.2 for the case when $y = 0$. We now give the exact statement of the theorem.

Theorem 7.3.1 (Intermediate Value Theorem). *If $f(x) : [a, b] \to \Re$ is continuous at each $x \in [a, b]$ and $f(a) < y_0 < f(b)$ (or $f(a) > y_0 > f(b)$), then there is a point $z_0 \in [a, b]$ such that $f(z_0) = y_0$.*

Simplicio: This theorem is much too abstract. Bring it down to earth.

Galileo: The Intermediate Value Theorem states something quite natural about the way we perceive the world around us. For example, I contend that at some point in your life you were exactly 4 feet tall.

Simplicio: No problem. Since I was less than 2 feet tall when I was born and am now over 5 feet, at some moment in time I must have been exactly 4 feet tall.

Galileo: While our friends in philosophy and physics might have objections, that is the answer I was looking for. Your reasoning is encapsulated by the Intermediate Value Theorem, where the function $f(x)$ represents your height at time x.

Simplicio: How about another example?

Galileo: If the temperature is less than 50 degrees in the morning and more than 80 degrees in the afternoon, then at some moment during the day, the temperature must have been exactly 70 degrees. For this example the function $f(x)$ represents the temperature at time x.

Virginia: But why is this theorem called the Intermediate Value Theorem?

Galileo: In the examples just mentioned, the temperature 70 degrees is intermediate between 50 and 80 and the height of 4 feet is intermediate between 2 feet and 5 feet. Assuming temperature and height vary continuously with time, the Intermediate Value Theorem will guarantee that there is some instant in time when these values are attained exactly.

Simplicio: But what if I was a midget and never got to be 4 feet tall?

Virginia: If you don't satisfy the hypotheses, the theorem does not apply.

Galileo: We will apply the theorem when $f(x)$ is a continuous function on an interval $[a, b]$ and $f(a)$ and $f(b)$ have opposite signs. (i.e. Either $f(a) > 0$ and $f(b) < 0$ or $f(a) < 0$ and $f(b) > 0$). In this setting the value $y = 0$ is *intermediate* between $f(a)$ and $f(b)$ so the function $f(x)$ has a root between a and b. If we let $a_0 = a$, $b_0 = b$, and $m_0 = \frac{a_0+b_0}{2}$, then we have two cases. If $f(a_0)$ and $f(m_0)$ have opposite signs, then define $a_1 = a_0$ and $b_1 = m_0$. If not, then define $a_1 = m_0$ and $b_1 = b_0$. Repeating this process, let $m_1 = \frac{a_1+b_1}{2}$. If $f(a_1)$ and $f(m_1)$ have opposite signs, then define $a_2 = a_1$ and $b_2 = m_1$. If not, then define $a_2 = m_1$ and $b_2 = b_1$.

Inductively, if a_{k-1} and b_{k-1} have been found, then define $m_{k-1} = \frac{a_{k-1}+b_{k-1}}{2}$. If $f(a_{k-1})$ and $f(m_{k-1})$ have opposite signs, then define $a_k = a_{k-1}$ and $b_k = m_{k-1}$. If not, then define $a_k = m_{k-1}$ and $b_k = b_{k-1}$.

Note that a root will lie in the interval $[a_k, b_k]$ and the length of the interval is $\frac{b-a}{2^k}$. Thus, the value $m_k = \frac{a_k+b_k}{2}$ will approximate the root with an error no more than

$\frac{b-a}{2^{k+1}}$. In fact, for any given function $f(x)$ the convergence rate only depends on the length of the interval $[a, b]$. Thus, this estimate of the convergence rate is the same for every function.

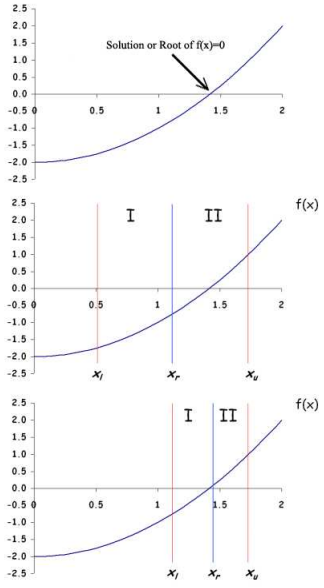


Figure 7.3: The Bisection Method for the function $f(x) = x^2 - 2$

Galileo: In general, the technique can be stated as the

Bisection Algorithm:

1. Let $f(x)$ be a continuous real-valued function on a closed bounded interval $[a, b]$, which has the property that $f(a)$ and $f(b)$ have opposite signs.

2. Let $m = \frac{a+b}{2}$.

3. If $f(a)$ and $f(m)$ have opposite signs, then set $b = m$.

4. If $f(a)$ and $f(m)$ do not have opposite signs, then set $a = m$.

5. Continue this process (i.e. repeat steps 2-4) until the required accuracy has been achieved.

Simplicio: This method seems to be quite understandable.

Galileo: If a function $f(x)$ crosses the $x-$axis at some point in an interval $[a, b]$ and $f(a)$ and $f(b)$ have opposite signs, then this method has the virtue that it "always works." While the method may always work, its downside is that the convergence rate is slower than the method of Newton/Raphson.

Simplicio: How about an example?

Example 7.3.1. *Galileo: Let's revisit our old friend* $f(x) = x^2 - 2$, *where the method is initialized with* $a = 1$ *and* $b = 2$. *The results of the bisection algorithm's first eight estimates are listed in Table 7.4.*

x_0	1.000000000000000
x_1	1.500000000000000
x_2	1.250000000000000
x_3	1.375000000000000
x_4	1.437500000000000
x_5	1.406250000000000
x_6	1.421875000000000
x_7	1.414062500000000
x_8	1.417968750000000

Table 7.4: Eight Estimates of a Root of the $\sqrt{2}$

Simplicio: You are right. The convergence rate of this method is glacial in comparison with either the Newton/Raphson or secant method. With theses other methods we are almost perfect after eight steps. Since $\sqrt{2} = 1.414213562373095$, *we have achieved only two digits of accuracy with the bisection method. Why would anyone use it?*

Galileo: The method is important because it always works and because it can be used in combination with other less stable methods such as Newton/Raphson. In particular, the bisection method can sometimes be iterated enough times to guarantee convergence. We will discuss this issue again in more detail. The combination of two such methods

results in a hybrid, which is sometimes better than each used separately.

Virginia: What can you say about the error?
Galileo: Since the midpoint m is half way between the points a and b, note that the error is cut in half at each iteration. Thus, the initial error is $b-a$ and the first error is $\frac{b-a}{2}$. The general formula for the error can be summarized in the following proposition.

Proposition 7.3.2 (Bisection Error Formula). *If $f(x)$ is a continuous real-valued function defined on the interval $[a,b]$ and $f(a)$ and $f(b)$ have opposite signs, then the error E_n at the n^{th} iteration satisfies the inequality $|E_n| < \frac{b-a}{2^n}$.*

Proof. Since a root of the function lies in the interval $[a,b]$ which has length $b-a$, the error E_0 satisfies $|E_0| < b-a$. Similarly, since a root of the function lies in either the interval $[a, \frac{a+b}{2}]$ or $[\frac{a+b}{2}, b]$ and both these closed intervals have length $\frac{b-a}{2}$, the error $|E_1| < \frac{b-a}{2^1}$. Since the length of the interval containing the root is halved at each iteration of the process, $|E_n| < \frac{b-a}{2^n}$. □

Example 7.3.2. *Galileo: How many iterations are required for the bisection to guarantee 14 digits of accuracy when computing $\sqrt{2}$ on the interval $[1,2]$?*
Virginia: Simply find an integer n with the property that $\frac{1}{2^n} < \frac{5}{10^{15}}$. When we take logs of both side of this expression, we find that this inequality will be satisfied if $n > 15log(10)/log(2) - log(5)/log(2) \approx 47.5$. Thus, if we choose $n = 48$, we will achieve the required accuracy.

Simplicio: That's worse than I thought it would be.
Galileo: In summary, while the method of Newton/Raphson may converge faster than the bisection method, the bisection method has the advantage that it "works" as long as the function $f(x)$ is continuous and satisfies the initial condition that $f(a)$ and $f(b)$ have opposite signs.
Simplicio: Something bothers me about the error formula $|E_n| \leq \frac{b-a}{2^n}$. While it contains the initial endpoints a and b, it seems to be the same for every function.

Galileo: Yes, your observation is correct. While it is reliable, its convergence rate is the same for all functions.

Virginia: I would like to back up and ask a question about the Intermediate Value Theorem. While I have an intuitive idea what the word continuous means, I am not sure I could define what it means for a function to be continuous. Could you give me more precision here?

Galileo: While we won't discuss that topic today, No worries. We will address all these issues in detail when we discuss the theory underlying Calculus. We will even provide a proof.

Simplicio: I can't wait.

Exercise Set 7.3.

1. If $K = 2, 5, 20,000$, $a = 1, b = K$, and $f(x) = x^2 - K, x^3 - K$ or $x^5 - K$, then how many iterations will be required for the bisection method to estimate a root of $f(x)$ to an accuracy of $\frac{1}{10,000}$? Compare the number of iterations with that needed by the Newton/Raphson method. Which do you prefer?

2. Using the bisection method how many iterations will be needed to approximate the real root of the function $f(x) = x^3 + x + 1$ if $a = -1, b = 0$, and the error is required to be less than 0.000001? Compare your answer with the answer you get when the method of Newton/Raphson is used with $x_0 = 0$ as the initial guess.

3. If the bisection method is used to compute a root of the function $f(x) = x^2 + 1$, then what goes wrong? Why does the bisection method fail when we were promised that it "always works."

4. If the bisection method is used to compute a root of the function $f(x) = xe^{-x^2}$ initialized by the points $a = -2$ and $b = 3$, then does the method work? How many iterations will be required to estimate the root of $f(x)$ to an accuracy of $\frac{1}{10,000}$.

Chapter 8

Problems With Root Finding

> If anything can go wrong, it will.-Murphy
>
> Nothing is ever as simple as it seems.-Murphy
>
> Nature always sides with the hidden flaw.-Murphy

Galileo: We now devote a few minutes to a discussion of examples that require us to be careful when computing roots.
Simplicio: Why discuss failure? Everything seems to be going well at the moment.
Galileo: Nothing is ever as simple as it seems.

8.1 Failure of Newton/Raphson

Galileo: Let us begin by reviewing the Newton/Raphson problems I assigned?
Simplicio: Everything went well. I had no problems. I even seemed to get all the right answers.
Galileo: How about if we take a more careful look at the method? What if we begin by computing the square root of K, where we initialize the method with a value of $x_0 = 0$?
Simplicio: Since the method of Archimedes/Heron is given by the recursive formula $x_{n+1} = \frac{x_n + \frac{K}{x_n}}{2} = x_n - \frac{x_n^2 - K}{2x_n}$, a division by zero occurs. Obviously, this event will not

be well-received in the mathematics community.

Galileo: Since the general formulation of Newton/Raphson is given by the equation $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, the strategy will be to avoid roots of the first derivative $f'(x)$.

Simplicio: Since the probability of making such a choice is about zero, we should not worry too much about that case. Right?

Galileo: While this avoidance task is easy for functions like $f(x) = x^2 - K$ when $K > 0$ and $x_0 = 1$, it can actually happen in simple settings. For example, consider the function $f(x) = x^2 + 1$. While this polynomial has real coefficients, its two roots are the complex numbers $r = \pm i$, where $i = \sqrt{-1}$. If the method of Newton/Raphson is initialized with $x_0 = 1$, then note that $x_1 = 0$, which leads to a division by zero in the computation of x_2. Thus, the value for x_2 can't even be computed. However, even if we choose another number, say $x_0 = 2$, so that division by zero never occurs, each recursively computed x_{n+1} will always be a real number. Thus, the method has no chance to converge to either $r_1 = i = \sqrt{-1}$ or $r_2 = -i = -\sqrt{-1}$.

Simplicio: Suddenly complex number have raised their ugly head, a worrisome situation.

Galileo: On the contrary.

Simplicio: You mean the method of Newton/Raphson can be used if the numbers are complex? Your motivation and graphs only seemed to apply to real-valued functions.

Galileo: Not a problem. The key is that you can compute the first derivative. The rules for derivatives are exactly the same as those you learned for real variable Calculus. The only difference is that you change the letter x to the letter $z = a + bi$. For the function we just considered, we let $f(z) = z^2 - K$. The derivative turns out to be $f'(z) = 2z$ and the recursive step in the algorithm becomes $z_{n+1} = z_n - \frac{f(z)}{f'(z)} = z_n - \frac{z_n^2 - K}{2z_n}$. An amazing feature of this example is that if the initial guess z_0 is chosen to be any complex number other than one of those on the real line, then the method in fact works. Work the first problem in the set of exercises listed below and you should begin to appreciate these remarks.

Simplicio: Interesting. What is the next example you have in mind?

Galileo: While dividing by zero is an obvious problem, we might also worry about functions with large derivatives near a root. For example, consider the function $f(x) = x^{\frac{1}{3}}$. Note that $f(0) = 0$ so $x = 0$ is a root. If we apply Newton/Raphson to this function, we find that the recursive relation becomes

$$x_{n+1} = x_n - \frac{x_n^{\frac{1}{3}}}{\frac{1}{3}x_n^{\frac{-2}{3}}} = x_n - 3x_n = -2x_n.$$

Thus, unless your initial guess $x_0 = 0$, you will have problems.

Simplicio: Is that it?

Galileo: As you might guess, the situation gets worse.

Let us now consider the differentiable function $f(x) = x \cdot e^{-x^2}$, which is graphed in Figure 8.1. This function illustrates a fundamental problem with the method of Newton/Raphson. While the function $f(x)$ has a unique root at $x = 0$ and has a graph which is almost a straight line near zero, a poor initial guess can lead to a sequence of points that converge to infinity.

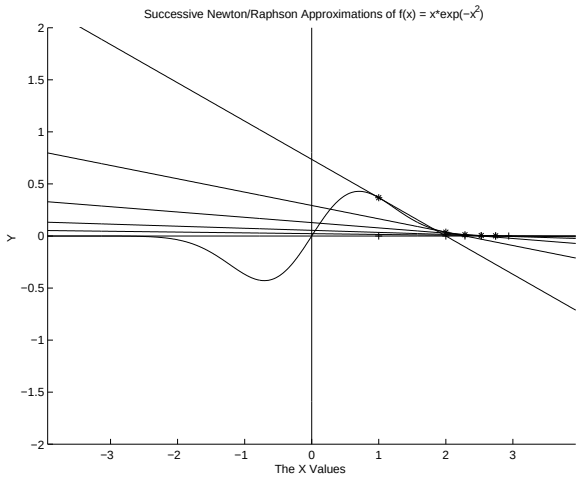


Figure 8.1: Failure of Newton/Raphson for the function $f(x) = x \cdot e^{-x^2}$.

Simplicio: How does that happen?

Galileo: Since the derivative is $f'(x) = (1 - 2x^2)e^{-x^2}$, $f(x)$ has critical points at $x = \pm\frac{\sqrt{2}}{2}$, which are the locations of the minimum and maximum. Thus, if the initial

guess x_0 for the Newton/Raphson method is chosen to the right of the location of the maximum, then it is clear that the subsequent terms in the sequence each be further to the right than the previous. In other words, $x_0 \leq x_1 \leq x_2, etc.$ We can actually show that the sequence converges to infinity. Similarly, if the initial guess x_0 is chosen to be to the left of the location of the minimum, then the resulting sequence will converge to negative infinity. On the other hand, if the initial point is chosen close to zero, then Newton/Raphson converges without a problem. Thus, the method works in some situations and not in others. One of our tasks will be to establish conditions which will guarantee convergence.

Virginia: Looks like we have a theorem to look forward to.

Galileo: Correct.

Simplicio: Groan. These examples make me worry that the method of Newton/Raphson is not as perfect as I had hoped.

Galileo: Just another example, where Murphy's Law applies to numerical methods. However, our next discussion will focus on the success of the method. As you will see, a number of very smart people have thought about these issues for a very long period of time.

Simplicio: Could you summarize the problems with Newtion/Raphson?

Galileo: Sure, the previous examples indicate the types of trouble we can expect to encounter with Newton/Raphson. These potential problems can be summarized as:

Example 8.1.1. *(Division by Zero) The derivative $f'(x_n) = 0$ for some integer n. If $f(x) = x^2 - 2$ and Newton/Raphson is initialized with $x_0 = 0$, then $f'(x_0) = f'(0) = 0$ so x_1 cannot be computed.*

If Newton/Raphson is initialized with any other real number $x_0 < 0$, then the sequence x_n converges to $-\sqrt{2}$. If Newton/Raphson is initialized with any other real number $x_0 > 0$, then the sequence x_n converges to $\sqrt{2}$.

Example 8.1.2. *(Unexpected Answer) The initial guess x_0 was not chosen sufficiently close to the root $x = r$ and the Newton/Raphson sequence converges to an unexpected answer.*

If $f(x) = \sin(x)$ and Newton/Raphson is initialized with $x_0 = \frac{\pi}{2} + 0.001$, then the sequence converges to a root r. However, the root r is far to the right of the initial guess.

Example 8.1.3. *(No Answer) The function $f(x)$ fails to have a real root.*
If $f(x) = x^2 + 1$ and Newton/Raphson is initialized with any real number x_0, then the sequence x_n simply bounces around and never has any hope of converging.

Example 8.1.4. *(First Derivative Problem) The first derivative $f'(x)$ does not exist at the root and the Newton/Raphson sequence diverges.*
If $f(x) = x^{\frac{1}{3}}$ and Newton/Raphson is initialized with any real number $x_0 \neq 0$, then $x_{n+1} = -2x_n$ and the sequence diverges to ∞.

Example 8.1.5. *(Poor Initialization) The initial guess x_0 was not chosen sufficiently close to the root $x = r$ and the Newton/Raphson sequence oscillates.*
If $f(x) = xe^{-x^2}$ and Newton/Raphson is initialized with $x_0 = 0.5$, then the sequence x_n oscillates between ± 0.5.

Example 8.1.6. *(Poor Initialization) The initial guess x_0 was not chosen sufficiently close to the root $x = r$ and the Newton/Raphson sequence diverges to infinity.*
If $f(x) = xe^{-x^2}$ and Newton/Raphson is initialized with $x_0 = 1$, then the sequence x_n diverges to $+\infty$.

Simplicio: So if I am computing my Newton/Raphson Algorithm for a particular function and it hasn't converged in 200 iterations, then I need to take a second look at the problem to make sure the method has a chance of working.
Galileo: Correct. And remember, the type of problem most likely to occur is the one depicted in Figure 8.1. In higher dimensional vector spaces, this problem is so common it is labeled "The Curse of Dimensionality."

Exercise Set 8.1.

1. Use the method of Newton/Raphson to compute a root of the polynomial $p(x) = x^2 + 1$. Begin by Initializing the method with $x_0 = 1$ and compute a thousand

terms. What do you observe? Can you decide whether or not the resulting sequence diverges to infinity or is bounded? Initialize the method a second time with the complex point $x_0 = 1 + i$, where $i = \sqrt{-1}$. What do you notice about this sequence of iterates?

2. Use the method of Newton/Raphson to compute a root of the function $f(x) = x^{\frac{1}{3}}$. Note that $x = 0$ is a root of $f(x)$. Initialize Newton/Raphson with values of $x_0 = 0.1, 0.2, \ldots, 1$. What do you notice? How about if we initialize with $x_0 = 0.01$ or $x_0 = 0.001$?

3. If $f(x) = x \cdot e^{-x^2}$, then implement Newton/Raphson with the values $x_0 = 0.25, x_0 = 0.50$, and $x_0 = 0.75$. What do you observe with these three examples? Find the largest real number L such that if $x_0 \in (-L, L)$, then the Newton/Raphson sequence $\{x_n\}_{n=1}^{\infty}$ converges to the root 0.

4. If the secant method is used to compute a root of the function $f(x) = xe^{-x^2}$ with $x_0 = 1/2$ and $x_1 = 1$, then does the method work? How many iterations will be required to estimate a root of $f(x)$ to an accuracy of $\frac{1}{10,000}$. Compare the number of iterations required by the Newton/Raphson method when $x_0 = 1/2$ or $x_0 = 1$.

5. Use the method of Newton/Raphson to compute a root of the function $f(x) = \sin(x)$. Note that $x = 0$ is a root of $f(x)$. Initialize the method with values of $x_0 = \frac{\pi}{2} + 0.1$ and $x_0 = \frac{\pi}{2} + 0.001$. Does the method converge to a root? If so, find it.

8.2 Newton/Raphson and Double Roots

Circles to square and cubes to double would give a man excessive trouble.-
Matthew Prior(1664-1721)

Galileo: We would now like to mention some examples, which reflect on the the rate

of convergence for the method of Newton/Raphson. As it turns out, different choices of functions $f(x)$ may produce different rates of convergence. In some of the exercises we assigned the convergence took 6 iterations to achieve as much convergence as you could want, while others took more than 30.

Simplicio: Yes, I remember that computing the square root of 5 worked great, while the square root of zero took much longer. I wondered about that.

Example 8.2.1. *Galileo: Consider the example, where* $p_2(x) = f(x) = x^2 - 1000^2 = x^2 - 1,000,000$. *Note that the roots are* $r_1 = 1000$ *and* $r_2 = -1000$. *The algorithm for Newton/Raphson is given by the recursive expression* $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 1000^2}{2x_n} = \frac{1}{2}x_n + \frac{500,000}{x_n}$. *We have the output from this algorithm summarized in Table 8.1, where the initialization was chosen to be* $x_0 = 1001$.

x_0	1001.00000000000
x_1	1000.00049950050
x_2	1000.00000000012
x_3	1000.00000000000

Table 8.1: Three Estimates of $\sqrt{1,000,000}$

Simplicio: Since the method converges in three steps, there is no problem.

Galileo: Correct.

Example 8.2.2. *Galileo: Now let's compute a second example that looks almost the same. If* $q_2(x) = f(x) = (x - 1000)^2$, *then the roots are* $r_1 = 1000$ *and* $r_2 = 1000$. *(We have a double root!) The algorithm for Newton/Raphson is given by the recursive expression* $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{(x_n-1000)^2}{2(x_n-1000)} = x_n - \frac{x_n-1000}{2} = \frac{1}{2}x_n + 500$. *The computations from this algorithm are displayed in Table 8.2, where we again have initialized with the value* $x_0 = 1001$.

Simplicio: Hey, this algorithm is as bad as the bisection method. The error simply drops by 50% *for each iteration. Not good.*

x_0	1001.00000000000
x_1	1000.50000000000
x_2	1000.25000000000
x_3	1000.12500000000
x_4	1000.06250000000
x_5	1000.03125000000
x_6	1000.01562500000
x_7	1000.00781250000
x_8	1000.00390625000
x_9	1000.00195312500
x_{10}	1000.00097656250
x_{11}	1000.00048828125
x_{12}	1000.00024414063
x_{13}	1000.00012207031
x_{14}	1000.00006103516

Table 8.2: Three Estimates of $\sqrt{1,000,000}$

Galileo: In both these examples, we see that the sequence of numbers $\{x_n\}_{n=1}^{\infty}$ is converging to the number 1000. In the first example we have a sequence that produces 11 digits of accuracy after only three iterations. In the second example, the algorithm has produced ly 4 digits of accuracy after 15 iterations. It is getting there, but even after 30 iterations, we have $x_{30} = 1.00000000000093$, which still isn't quite there.

Simplicio: What seems to be the problem?

Galileo: While the first example has distinct roots that are far apart, the second has the double root $r_1 = r_2 = 1000$. Double roots slow down the convergence rate from quadratic to linear.

Simplicio: What are these quadratic and linear convergence rates?

Galileo: The sequence $x_n = \frac{1}{2^n}$ converges linearly to zero. The sequence $x_n = \frac{1}{2^{2^n}}$ converges quadratically to zero. These examples typify the different convergence rates. Make a few calculations and you will see the difference. You do the math.

Virginia: Our initial guess $x_0 = 1001$ is reasonably close to the final answers. What if we had made a poor initial guess?

Galileo: If we use $x_0 = 1$ as our initial guess, then the method of Newton/Raphson produces $x_{10} = 1296.191592707$ for $p_2(x)$ and $x_{10} = 999.024414063$ for the root of $q_2(x)$. However, after 14 iterations, the method produces $x_{14} = 1000.000000000$ for $p_2(x)$ and $x_{14} = 999.939025879$ for $q_2(x)$. Thus, our convergence is complete for the root of $p_2(x)$, but still has an error of more than 0.939 for the root of $q_2(x)$. Thus, while the linearly convergent sequence converges better for the first ten terms, the quadratically convergent sequence quickly overtakes it once it gets close. The Mean Value Theorem will provide our main tool for showing linear convergence. Taylor's Theorem will provide our main tool for showing quadratic convergence.

Simplicio: I am not quite sure what is going on here.

Galileo: Don't worry. We will return to this topic.

Exercise Set 8.2.

1. If $p_2(x) = f(x) = x^2 - 10^8$ and $x_0 = 10,001$, then how many iterations of Newton/Raphson are required to achieve an accuracy of 10 decimal places?

2. If $q_2(x) = f(x) = (x - 10000)^2$ and $x_0 = 10,001$, then how many iterations of Newton/Raphson are required to achieve an accuracy of 10 decimal places? Compare your answer with your answer to problem 1.

3. If $f(x) = (x + 3)^2$ and $x_0 = 1$, then compute the first 30 iterations of the Newton/Raphson algorithm. Format your output in a column. How does the convergence rate of the last five computations compare with the first 25?

4. Compute 15 iterations in the Archimedes/Heron/Newton/Raphson algorithm to approximate the square root of $K = 1,000,000$. Initialize the algorithm with $x_0 = 1$. Format your output in a column. How does the convergence rate of the last five computations compare with the first 10?

8.3 Instabilities With Root Finding

James Hardy Wilkinson (1919-1986)

Mother Nature is a bitch.-Murphy

Galileo: Before moving on to the topic of the theory of convergent sequences, let us take a closer look at the problem of computing the roots of the polynomials. First, to give you an idea of where the problems lie, let us look at the graph of the polynomials $p_4(x) = (x-1)(x-2)(x-3)(x-4)$ and $p_5(x) = (x-1)(x-2)(x-3)(x-4)(x-5)$.

These polynomials are of particular importance because the roots are simple (i.e. not double roots) and equally spaced. However, also note that the graphs are almost flat between the roots. Thus, a small change of one of the coefficients can lead to a large change in the placement of the roots.

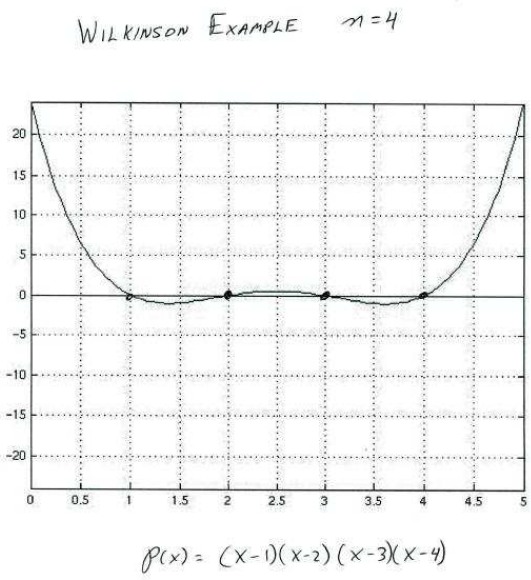


Figure 8.2: The Graph of the polynomial $y = p_4(x)$

The British mathematician, James Wilkinson (1919-1986), noticed that the roots of the polynomial $p_{20}(x) = (x-1)(x-2)\dots(x-20)$ have even more bazaar instabilities. First, he noticed that if this polynomial is multiplied out, then the coefficient of the $19^{th}-$degree term is -210.

Simplicio: That calculation is easy because that coefficient is simply the sum of the integers $-1, -2, \dots, -20$. I know how to use the formula for the arithmetic sum to compute this quantity.

Example 8.3.1. *Galileo: Wilkinson also noticed that if this coefficient of* x^{19} *is*

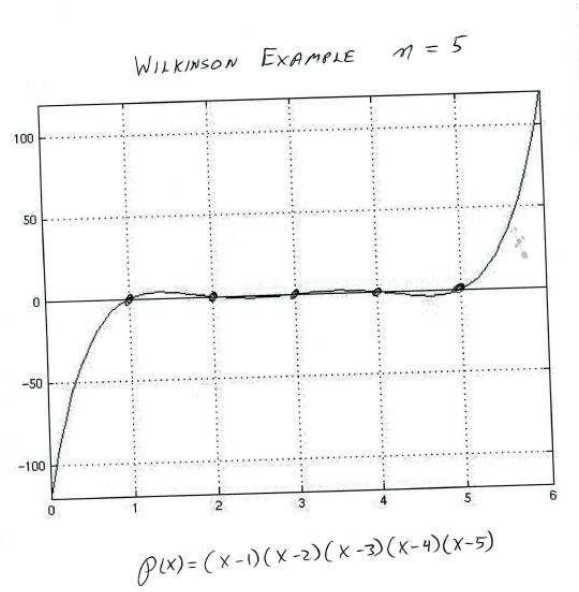


Figure 8.3: The Graph of the polynomial $y = p_5(x)$

changed by $2^{-23} \approx 10^{-7}$, *then the roots become*

$$
\begin{aligned}
1.0,\\
2.0,\\
3.0,\\
4.0,\\
5.0,\\
6.0,\\
7.0,\\
8.0,\\
8.9,\\
10.1 \pm 0.6i,\\
11.8 \pm 1.7i,\\
14.0 \pm 2.5i,\\
16.7 \pm 2.8i,\\
19.5 \pm 1.9i,\\
20.8.
\end{aligned}
$$

In particular, a very small change in one coefficient can lead to a large change in the values of the roots. Worse yet, half of the roots are complex.

Simplicio: That example is amazing!! Not only did the last root change by 0.8, but ten of the roots suddenly became imaginary. It makes one worry about finding the roots of any function.

Galileo: I couldn't agree more. The rule is: Small changes in the coefficients may lead to large changes in the values of the roots. This type of problem occurs when the function is very "flat" near the root. Try graphing the function locally near $r = 20$.

Simplicio: Has anyone ever tried to build something using these high-degree polynomials?

Galileo: Indeed, a group of my engineering colleagues tried to use 16 and 32 degree polynomials in a mathematical model designed to control the motion of an arm of one of their robots. Their efforts were a disaster. One of their students was almost killed.

Simplicio: So avoiding an unstable mathematical method could save lives.

Galileo: If you model a phenomenon with an unstable method, you are asking for trouble. As always, the mantra for numerical analysis remains the same: "The name of the game is control."

Exercise Set 8.3.

1. Note that the polynomial of degree 9 with roots $1, 2, 3, 4, 5, 6, 7, 8, 9$ can be expanded into the form $p_9(x) = x^9 - 45 * x^8 + 870 * x^7 - 9450 * x^6 + 63273 * x^5 - 269325 * x^4 + 723680 * x^3 - 1172700 * x^2 + 1026576 * x - 362880$. Using available software, compute the roots of the polynomials $q_9(x), r_9(x)$, and $s_9(x)$ listed below.

   (a) $q_9(x) = x^9 - (45 + \frac{1}{10^5}) * x^8 + 870 * x^7 - 9450 * x^6 + 63273 * x^5 - 269325 * x^4 + 723680 * x^3 - 1172700 * x^2 + 1026576 * x - 362880$,

   (b) $r_9(x) = x^9 - (45 + \frac{1}{10^4}) * x^8 + 870 * x^7 - 9450 * x^6 + 63273 * x^5 - 269325 * x^4 + 723680 * x^3 - 1172700 * x^2 + 1026576 * x - 362880$, and

(c) $s_9(x) = x^9 - (45 + \frac{1}{10^3}) * x^8 + 870 * x^7 - 9450 * x^6 + 63273 * x^5 - 269325 * x^4 + 723680 * x^3 - 1172700 * x^2 + 1026576 * x - 362880.$

How many real and how many imaginary roots do each of these polynomials have? What is the distance between corresponding roots of $p_9(x)$ and $q_9(x)$, $p_9(x)$ and $r_9(x)$, and $p_9(x)$ and $s_9(x)$?

Part IV

Day 4. Advanced Calculus

Chapter 9

Limits

Augustin Louis Cauchy (1789-1857)

> Men pass away, but their deeds abide.-Augustin Louis Cauchy [His last words?]

9.1 Sequences

> Calculus has its limits.-unknown

Galileo: We now introduce Augustin Louis Cauchy (1789-1857) for an explanation of the theory underlying limits. His text "Cours d'analyse" (written in 1821) was an important step towards bringing rigor to Calculus. Professor Cauchy grew up during

the French Revolution so he knows how to bring order out of chaos.

Virginia: If I count correctly, Newton's Principia was written in 1689 so it took more than 100 years to bring rigor to Calculus.

Galileo: Actually, this issue has been around since Plato recorded the paradoxes of Zeno of Elea (490-450 B.C.E.) in his dialogue Parmenides.

Simplicio: As far as I am concerned, infinity has nothing to do with the real world. Why don't we just focus on algorithms. Something useful an employer would appreciate.

Virginia: Your goal is to earn a blue collar wage?

Galileo: Before we begin, we let us take a minute and have a brief quiz to make sure you will follow each nuance of the discussion.

Quiz:

1. What is a conditional sentence?

2. What is the purpose of a definition?

3. What is the difference between a definition and a theorem?

If you can't answer these question, then there is no point continuing.

Simplicio: But we just covered these issues?

Galileo: I am never quite sure what you retain. Professor Cauchy, where should we begin?

Cauchy: Let us begin by admitting we have a problem. Namely. some sequences converge and some do not. The issue is simple. We must get the language straight. Namely, we must make some carefully worded definitions that set the ground rules for what we want. Let us begin with two examples, which encapsulate the issues.

Example 9.1.1. *First, the alternating sequence of points defined by* $\{x_n\}_{n=1}^{\infty} = \{(-1)^n\}_{n=1}^{\infty} = -1, 1, -1, 1, -1, \ldots,$ *causes trouble because it seems to converge to two points at the same time, namely* $+1$ *and* -1*. However, if you are going to allow a sequence to converge to two numbers, then why not three? Why not four? Now the*

situation is out of control so we decided that we wanted a sequence to converge to only one number.

Example 9.1.2. *Second, while some people might want the first sequence o converge to both* $+1$ *and* -1*, I don't think anyone would allow a sequence to converge to infinity. Thus, the sequences* $\{x_n\}_{n=1}^{\infty} = \{n\}_{n=1}^{\infty} = 1, 2, 3, 4, 5, \ldots, n, \ldots$ *and* $\{y_n\}_{n=1}^{\infty} = \{n^2\}_{n=1}^{\infty} = 1^2, 2^2, 3^2, 4^2, 5^2, \ldots, n^2, \ldots$ *march off to infinity. The theory and applications work much better if we simply rule them out. For example, looking ahead, we would like to have a theorem which states that the limit of the sum equals the sum of the limits. However, if we had that theorem, we might try to compute the limit of the sequence*

$$\lim_{n\to\infty}\{z_n\}_{n=1}^{\infty} = \lim_{n\to\infty}\{n - n^2\}_{n=1}^{\infty} = \lim_{n\to\infty}\{n\}_{n=1}^{\infty} - \lim_{n\to\infty}\{n^2\}_{n=1}^{\infty} = \infty - \infty = ???$$

Thus, we don't want to deal with unbounded limits-at least not at this time.

Simplicio: How about something more positive?

Cauchy: No matter what your attitude, the following three sequences should converge.

Example 9.1.3. *The sequence* $\{\frac{1}{n}\}_{n=1}^{\infty} = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots,$ *should converge to zero.*

Example 9.1.4. *The sequence* $\{\frac{(-1)^n}{n}\}_{n=1}^{\infty} = -1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \ldots,$ *should also converge to zero.*

Example 9.1.5. *The sequence* $\{\frac{n-1}{n}\}_{n=1}^{\infty} = 0, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots,$ *should converge to one.*

Cauchy: To rectify the situation with the first two examples, we first need to decide what the word convergence means.

Simplicio: You mean you we get to make up the rules?

Cauchy: You are in control. But remember, once you have made a choice, you have to stick with it. You don't get to change the rules.

Virginia: But how do you make up a rule to test for something that goes on forever?

Cauchy: First, we are given a sequence of numbers $\{x_n\}_{n=1}^{\infty}$.

Second, we have an idea of what number the sequence is supposed to converge to. Since that number is going to be the LIMIT of a sequence, we will denote it by the letter L.

Third, we need to devise a test (or criterion) to decide whether or not the sequence converges to the number L.

Simplicio: What's wrong with the rule that the sequence simply stabilizes. Namely, a sequence converges when $a_k = a_{k+1} = a_{k+2} = a_{k+3} = \ldots$. That idea worked fine when we computed square roots.

Cauchy: Unfortunately, that idea only worked because of the finite precision of your calculator or computer. The successive terms just look equal. There are even examples of sequences that have the property that successive terms are equal, while the sequence converges to ∞.

Simplicio: Like what?

Cauchy: Consider the sequence $x_n = \sum_{k=1}^{n} \frac{1}{k}$. Compute x_n when
$n = 100,000,000,000,000$ and when $n = 100,000,000,000,001$ and then check to see if they are the same.

Simplicio: But who would be dumb enough to ever compute that many terms of the sequence.

Virginia: We are not talking about computing yet. We are simply trying to get the language straight.

Galileo: I can think of a number of situations, where you might want to compute even more terms.

Cauchy: In any case, there are valid mathematical and engineering reasons to proceed with a bit of caution right at the beginning.

Galileo: Proceed.

Cauchy: The tricky part about the definition of a limit is the test (or criterion). This test is given in terms of a conditional sentence.

Galileo: Remember: "If p, then q."?

Virginia: I do.

Cauchy: This conditional sentence can be thought of as a challenge, where I begin by giving you a distance and then you are expected to show that almost all of the terms of the sequence are within this given distance of the limit L. Historically, this distance has been denoted by the Greek letter ϵ. Since distances are always positive, we insist that $\epsilon > 0$. In other words, eventually all the terms of the sequence are within a distance of ϵ from L.

Galileo: Mr. Simplicio, let me ask you one last time: Are you clear about the difference between a definition and a theorem?

Simplicio: I know, I know. I was listening.

Cauchy: We have two different ways of measuring distance at our disposal. The first is the open interval. The second is the absolute value function. These two different techniques are equivalent. In other words, it doesn't matter which you choose, the results will be the same.

Simplicio: Why not just give us the easiest one?

Cauchy: The open interval definition is easier to visualize, while the absolute value is usually easier to compute. The advantage of the absolute value function is that you are often able to condense multiple cases in a mathematical argument into a single case. Thus, the arguments are shorter.

Galileo: And sometimes it provides a more conceptual framework because you can think in terms of distances from the limit L.

Cauchy: We begin by defining the terms interval, open interval, and closed interval. We also let the symbol $\Re$ denote the set of real numbers.

Definition 9.1.1. *A subset X of $\Re$ is called an interval if there are points a and b in $\Re$ such that one of the following four cases is true:*

1. $X = (a, b) = \{x \in \Re : a < x < b\}$,
2. $X = (a, b] = \{x \in \Re : a < x \leq b\}$,
3. $X = [a, b) = \{x \in \Re : a \leq x < b\}$,

4. $X = [a, b] = \{x \in \Re : a \leq x \leq b\}$.

If $a, b \in \Re$, then an *open* interval has the form $(a, b), (a, \infty), (-\infty, b)$ or $(-\infty, \infty)$ and a *closed* interval has the form $[a, b], [a, \infty), (-\infty, b]$, or $(-\infty, \infty)$. In particular, the set $\Re$ is considered both an open and closed interval. While the empty set is considered an interval, it will seldom be of interest. In fact, in the definition of limit, we will want to rule it out by assuming our open intervals U are non-empty.

Simplicio: These ideas are easy so far. If someone gives you two points a and b, then an interval defined by a and b will be all the points between a and b and possibly one or both endpoints.

Cauchy: Maybe now is a good time to give a formal definition of the absolute value function.

Definition 9.1.2 (The Absolute Value Function). *If $x \in \Re$, then the absolute value of x is defined by the rule*

$$|x| = \begin{cases} x \text{ if } x \geq 0 \\ -x \text{ if } x < 0 \end{cases}.$$

Cauchy: This function is intimately connected with finding the distance between two points. The properties of the absolute function are summarized in the following proposition.

Proposition 9.1.3. *If $x, y \in \Re$, then*

1. $|x| \geq 0$,
2. $|x| = 0$ *if and only if* $x = 0$,
3. $|x + y| \leq |x| + |y|$, *and*
4. $||x| - |y|| \leq |x - y|$.

Proof. The proofs of these items are straightforward. □

Cauchy: While we are at it, why don't we define the distance between two real numbers?

Definition 9.1.4 (Distance). *If $x, y \in \Re$, then the distance between x and y is defined to be $dist(x, y) = |x - y|$.*

Cauchy: The properties of the distance function are summarized in the following proposition.

Proposition 9.1.5. *If $x, y, z \in \Re$, then*

1. *$dist(x, y) \geq 0$, and $dist(x, y) = 0$ if and only if $x = y$ (positive definite),*
2. *$dist(x, y) = dist(y, x)$ (symmetry), and*
3. *$dist(x, y) \leq dist(x, z) + dist(z, y)$ (triangle inequality).*

Virginia: So, am I to understand that whenever I see the absolute value function, I should think length. Also, whenever I see the absolute value of the difference of two numbers, I should think distance.

Cauchy: Absolutely. Note also that while these propositions are important, we have not labeled them as theorems. We will save that designation for the big boys like the Mean Value Theorem and Fundamental Theorem of Calculus. We now offer three equivalent definitions for a sequence to converge to a number L. The first definition is conceptual. If you don't like it, ignore it. We won't use it often.

Definition 9.1.6 (Convergence of Sequence 1). *A sequence of real numbers $\{x_n\}_{n=1}^{\infty}$ is said to converge to a number L if for any non-empty open interval U of the form $U = (L - \epsilon, L + \epsilon)$, then all but a finite number of terms of the sequence lie in U.*

Simplicio: I am not sure I understand that definition at all.

Cauchy: In other words, for any open interval U containing L, there is an integer N with the property that if $n \geq N$, then $x_n \in U$. If you draw a picture with the first five

terms of the sequence x_1, x_2, x_3, x_4, x_5 outside the interval, but $x_6, x_7, x_8, \ldots$ all inside the interval U, then you have the idea. Let's go back to one our successful examples.
Simplicio: I find the use of that symbol ϵ annoying.
Cauchy: The use of the letter ϵ has been around for a long time and probably won't change any time soon. While any other letter or symbol could be used, this letter is indelibly etched in mathematical culture. If it helps, think of it as a tolerance or precision forced on you by your employer. For example, if you are expected to build some structure within a certain precision, then the amount of error you are allowed is ϵ. If you prefer, you can use any symbol you want. However, we will follow our cultural traditions. Sorry.

Example 9.1.6. *We would now like to show the sequence* $\{\frac{1}{n}\}_{n=1}^{\infty} = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots,$ *converges to the limit* $L = 0$. *The procedure is as follows. If I give you an open interval* $U = (-\frac{1}{10}, \frac{1}{10})$, *your job is to find an integer* N, *which has the property that whenever* $n \geq N$, *then* $x_n = \frac{1}{n} \in U$.
Virginia: Obviously, if $N = 11$, *then* $x_{11} = \frac{1}{11}, x_{12} = \frac{1}{12}, x_{13} = \frac{1}{13}, x_{14} = \frac{1}{14}, \ldots$ *all lie in* U. *Since all but* 10 *terms in the sequence lie in* U, *we are done.*
Cauchy: Very good. Now, how about a smaller interval? Say, $U = (-\frac{1}{100}, \frac{1}{100})$.
Virginia: Obviously, if $N = 101$, *then* $x_{101} = \frac{1}{101}, x_{102} = \frac{1}{102}, x_{103} = \frac{1}{103}, x_{104} = \frac{1}{104}, \ldots$ *all lie in* U. *Since all but* 100 *terms in the sequence lie in* U, *we are done.*
Cauchy: Very good again. Now let's try the case when $\epsilon = 0$.
Simplicio: Even I can see that if $\epsilon = 0$, *then the interval* $U = [0, 0]$ *is simply a single point and we will never have any terms of the sequence in* U.
Virginia: Now I see why made this annoying distinction between open and closed intervals. Obviously, we only want open intervals for these types of problems. While the interval $U = [0, 0]$ *is closed, it is not open. Thus, we don't have to worry this case.*
Cauchy: Excellent. Now let's try the case when $\epsilon = -1$.
Simplicio: But, if $\epsilon = -1$, *then the interval* $U = (+1, -1)$ *is the empty set and we will never have any terms of any sequence in* U.

Virginia: And we now see why the definition only expects us to consider NONEMPTY open intervals.

Galileo: I think we are getting somewhere.

Cauchy: Of course, you realize that you haven't satisfied the definition at all. These first few choices of U were just for practice. The real test comes when we choose $U = (-\epsilon, \epsilon)$, where $\epsilon > 0$.

Galileo: However, before we do that, let's follow the example of George Polya and think in terms of his four steps to solving a problem. Do you know what they are?

Virginia: I know:

1. *understanding the problem,*
2. *devise a plan,*
3. *carry out the plan, and*
4. *look back and review what was done.*

Cauchy: She is good. How do you recruit such good students Professor Galileo? OK, so do you understand the problem?

Simplicio: I am not sure.

Galileo: So, now may be a good time to devise a general plan of attack.

Cauchy: When using the definition to prove a sequence converges to a particular number L, the plan of attack is always the same and can be broken down into three steps:

1. *The Challenge,*
2. *The Choice, and*
3. *The Check.*

In the first example, we were considering the sequence $x_n = \frac{1}{n}$ and I Challenged you with the interval $U = (-\frac{1}{10}, \frac{1}{10})$.

Virginia: We then noticed that if we Choose the integer $N = 11$, then it might be a good candidate to separate the terms that are members of U and those that are not. We then had to Check that if $n \geq N = 11$, then the term $x_n = \frac{1}{n}$ is a member of U.

Simplicio: Even I can see that when you gave us the interval $U = (-\frac{1}{100}, \frac{1}{100})$, the process was exactly the same. The same three steps work.

Cauchy: OK, now I want you to consider the fourth step in Professor Polya's plan. Namely, let's review what we have done and generalize the process. As you will see, the first step is ALWAYS the same:

Step 1. The Challenge: Let $\epsilon > 0$ be given.

If you miss that step on an exam problem, your professor will classify you as a slow learner. As you can see in our practice problems, the positive quantity ϵ defines the endpoints of our open interval $U = (-\epsilon, \epsilon)$. This quantity has to be positive because if it equals zero, the interval is not open and if it equals a negative number, the interval U is empty. We are only interested in nonempty open intervals. OK, what do you do next?

Virginia: Now it is time to choose the integer N. Obviously, for this problem,

$$\textit{Step 2. The Choice: Choose } N > \frac{1}{\epsilon}.$$

Simplicio: How did you know to do that?

Cauchy: In general, making an intelligent choice for N is almost always the hardest of the three steps.

Virginia: But, for this problem, we simply work backwards from what we want. Namely, since we would like $\frac{1}{n} < \epsilon$, then we assume what we want and solve for n. In this case, this step is easy because if we multiply the above expression by n and divide by ϵ, then we get $n > \frac{1}{\epsilon}$.

Cauchy: To complete the process, we must now Check that your Choice works.

Virginia: For this problem, this last step is easy because all we have to do is reverse the process from Step 2.

$$\textit{Step 3. The Check: If } n \geq N, \textit{ then we must show } x_n = \frac{1}{n} \in U = (-\epsilon, \epsilon).$$

For if $n \geq N > \frac{1}{\epsilon}$, then $0 < \frac{1}{n} \leq \frac{1}{N} < \epsilon$. Thus, $x_n = \frac{1}{n}$ lies in $U = (-\epsilon, \epsilon)$ and we are done.

Cauchy: Excellent. Professor Galileo, you should be proud.

Galileo: I am.

Simplicio: How did you figure that out?

Cauchy: Did you notice that we used a conditional sentence in step three? Namely, we only needed to check that $x_n = \frac{1}{n}$ is in U for "large" n. Namely, those larger than N. In fact, in the definition of convergence, that's what we meant by the phrase "all but a finite number of terms of the sequence lie in U.."

Example 9.1.7. *Cauchy: In this next example we will show the sequence $\{\frac{(-1)^n}{n}\}_{n=1}^{\infty} = -1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \ldots$, converges to the limit $L = 0$. The procedure is the same as before. If $\epsilon = \frac{1}{10}$, then could you outline the process?*

Virginia: Step 1. The Challenge:

We begin with the challenge: Let $\frac{1}{10}$ be given. Again, this quantity defines the open interval $U = (-\frac{1}{10}, \frac{1}{10})$.

Step 2. The Choice:

We also choose N as before. Namely, we choose $N = 11$.

Step 3. The Check:

We must now check that whenever $n \geq N$, then $x_n = \frac{(-1)^n}{n} \in U$. However, $x_{11} = -\frac{1}{11}, x_{12} = \frac{1}{12}, x_{13} = -\frac{1}{13}, x_{14} = \frac{1}{14}, \ldots$ all lie in $U = (-\frac{1}{10}, \frac{1}{10})$. In general, if $n \geq N = 11$, then $-\frac{1}{10} < \frac{(-1)^n}{n} < \frac{1}{10}$.

Cauchy: Very good. Now, how about a smaller interval? Say, $U = (-\frac{1}{100}, \frac{1}{100})$.

Virginia: Obviously, if $N = 101$, then the discussion we just gave guides you through the three steps.

Simplicio: Even I am beginning to get it.

Cauchy: Very good again. As before, these first two choices of U were just for practice. Now let's attack the general case, where I give you the following

Step 1. The Challenge: Let $\epsilon > 0$ be given.

How do you show all but a finite number of the terms of the sequence $x_n = \frac{(-1)^n}{n}$ *lie in* $U = (-\epsilon, \epsilon)$. *Note that I just did* 33% *of the problem for you!*

Virginia: Obviously, we can make the same choice as before.

$$\textit{Step 2. The Choice: Choose } N > \frac{1}{\epsilon}.$$

We now have to show that this choice works by giving the following short proof.

$$\textit{Step 3. The Check: If } n \geq N, \textit{ then we must show } x_n = \frac{(-1)^n}{n} \in U = (-\epsilon, \epsilon).$$

Proof. For if $n \geq N > \frac{1}{\epsilon}$, then $-\epsilon < -\frac{1}{N} \leq -\frac{1}{n} \leq \frac{(-1)^n}{n}$ and $\frac{(-1)^n}{n} \leq \frac{1}{n} \leq \frac{1}{N} < \epsilon$. Thus, $x_n = \frac{(-1)^n}{n}$ lies in $U = (-\epsilon, \epsilon)$ and we are done. □

Cauchy: Professor Galileo, where do you find such excellent students?

Galileo: I am a lucky man.

Simplicio: I think I am beginning to figure it out. The open interval U needs to surround the limit L so it traps terms of the sequence coming from both sides.

Virginia: That's why the interval is nonempty and open.

Cauchy: In the spirit of Polya's looking back, I would like to comment on the phrase "all but a finite number of terms of the sequence lie in U.," which appears in the definition of a convergent sequence. While this phrase makes sense, it is a bit of a mouthful and it is not expressed mathematically.

Virginia: But isn't that why we went to the trouble to find the integer N with the property that if $n \geq N$, then $x_n \in U$.

Cauchy: Exactly. Note also that the phrase "if $n \geq N$, then $x_n \in U$" is a conditional statement. Thus, when we check a sequence converges, the Check will always be a test phrased as a conditional sentence.

Virginia: Now we understand why we discussed conditional sentences when we reviewed logic and rigor.

Simplicio: I didn't say anything. Why are you looking at me?

Cauchy: For practical problems we have two standard choices for ϵ. To ensure that our sequence is within single precision accuracy of the limit, we would choose $\epsilon = \frac{1}{10^7}$.

To ensure that our sequence is within double precision accuracy of the limit, we would choose $\epsilon = \frac{1}{10^{14}}$. Thus, for single precision accuracy, we let $U = (L - \frac{1}{10^7}, L + \frac{1}{10^7})$. For double precision accuracy, we let $U = (L - \frac{1}{10^{14}}, L + \frac{1}{10^{14}})$. Of course, ϵ can represent any positive number. Conceptually, ϵ measures the distance from the center of the interval to the two endpoints of U. I think you can now see why we insist ϵ MUST always be positive. If it were negative, the set U would represent the empty set. Also, since it represents a distance, it must be positive.

From an engineering point of view this definition can be thought of as an employer/employee challenge, where the employer gives the employee the specs (or tolerance for error) on the project and the employee is expected to search until he/she can guarantee that all the remaining terms of the sequence are within that specification. The number ϵ represents the tolerance forced by the employer on the employee. For example, if I wanted to build a house with 2500 square feet and I gave you a tolerance of 10 square feet, I would be upset if I ended up with only 2450 square feet.

We would now like to give a second definition of convergence.

Simplicio: You have got to be kidding. One definition was bad enough, but now I have to deal with another one?

Cauchy: The idea behind the first definition is to get the language as simple and natural as possible. The only difference between the first and second is the observation that an open interval $U = (L - \epsilon, L + \epsilon)$ is equal to the set of all numbers $x \in \Re$ such that $|x - L| < \epsilon$. For the sake of completeness, we formalize this bit of information in the next proposition.

Proposition 9.1.7. *If $L, \epsilon, x \in$ Re, then x is a member of the set $U = (L - \epsilon, L + \epsilon)$ if and only if $|x - L| < \epsilon$.*

Simplicio: Am I correct in noting in this proposition that if $\epsilon \leq 0$, then the set U is the empty.

Cauchy: True, but we aren't interested in negative values for ϵ. The second definition of convergence can be given as:

Definition 9.1.8 (Convergence of Sequence 2). *A sequence of real numbers* $\{x_n\}_{n=1}^{\infty}$ *is said to converge to a number* $L \in \Re$ *if for every* $\epsilon > 0$ *there is an integer* N *with the property that if* $n \geq N$, *then* $|x_n - L| < \epsilon$.

Proposition 9.1.9. *Definition 1 for convergence is equivalent to Definition 2.*

Proof. By the previous proposition, we know x is a member of the set $U = (L-\epsilon, L+\epsilon)$ if and only if $|x - L| < \epsilon$. Thus, we are done. □

Cauchy: While this last definition may be a bit less transparent, the test for convergence has changed from open interval to distance. In other words, the test requires the distance between x_n and the limit L is less than ϵ for all but a finite number of the terms of the sequence. Since we now have the idea of distance, we see that the sequence $\{x_n\}_{n=1}^{\infty}$ converges to L if for any positive distance ϵ, we can find an integer N with the property that if $n \geq N$, then the distance between x_n and L is less than ϵ. If the limit of a sequence $\{x_n\}_{n=1}^{\infty}$ equals L, then we will write $\lim_{n\to\infty}\{x_n\} = L$.
Simplicio: So, let's see if I can phrase the definition in engineering terms. First, the inputs are:

1. a sequence $\{x_n\}_{n=1}^{\infty}$,
2. a number L, and
3. a tolerance $\epsilon > 0$.

Second, if the test for convergence is successful, the output is an integer N, which has the property that if $n \geq N$, then $|x_n - L| < \epsilon$. Moreover, if your employer has insisted your precision is within $\epsilon = \frac{1}{10^{14}}$, then you might as well have used my definition that a sequence converges when you can find an integer N with the property that if $n \geq N$, then $x_n = x_{n+1} = x_{n+2} = x_{n+3} = \ldots$.
Galileo: I think he's got it!
Cauchy: As with the first definition, each argument can be broken down into three steps.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice of N:

The second step in the limit definition is to choose an integer N that "work's" If you have no idea how to choose this integer, you might leave this step blank until after you have made a few preliminary mathematical calculations. These calculations are usually begin by assuming what you want to be true and working backwards until you uncover an expression for n in terms of ϵ.

Step 3. The Check that N "works":

The third step in the process is to check that your Choice of N has the property: If $n \geq N$, then $|x_n - L| < \epsilon$.

Another tip: When first learning about a new type of mathematical argument, it is often a good idea to write down what you are expected to do. For limits, a helpful starting point is to write the sentence: We MUST show: If whenever $n \geq N$, then $|x_n - L| < \epsilon$.

Galileo: OK, let's go through this process to prove that $\lim\{\frac{1}{n}\} = 0$. I think you will agree that the limit should equal zero.

Example 9.1.8. *Cauchy: Using the definition of limit, show that* $\lim_{n\to\infty}\{\frac{1}{n}\} = 0$.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice of N:

Since we want $|x_n - 0| = |\frac{1}{n}| < \epsilon$, we can multiply both side of the inequality by n and observe that we require $n > \frac{1}{\epsilon}$. Thus, our Choice for N is any integer larger than $\frac{1}{\epsilon}$.

Step 3. The Check that N works:

Let us begin this step by writing down what we are expected to do. Namely, we MUST show: If $n \geq N$, then $|x_n - L| = |\frac{1}{n} - 0| = \frac{1}{n} < \epsilon$.

Since we only have to test integers $n \geq N$, we know that $n \geq N > \frac{1}{\epsilon}$, we know $n > \frac{1}{\epsilon}$. By dividing both sides of the inequality by n and multiplying both sides by ϵ, we see that $\frac{1}{n} < \epsilon$. Thus, $|x_n - 0| = \frac{1}{n} < \epsilon$ and we are done.

Simplicio: That argument was the same as for the first Definition.

Galileo: I think you have got it. Let's move on to the next example.

Example 9.1.9. *Cauchy: Using the definition of limit, prove that* $\lim_{n\to\infty}\{\frac{1}{n^2}\} = 0$.

Galileo: How about if you present the argument this time?

Simplicio: To begin the discussion I simply write:

Step 1. The Challenge:

Let $\epsilon > 0$ *be given.*

Is that correct?

Galileo: Correct, you are 33% *of the way to the goal. Moreover, you have absolutely no excuse for getting this step wrong. It is the same for every problem of this type.*

Simplicio: But I have no idea how to choose N.

Galileo: No worries. Simply make the same choice we made for the first problem and see what happens.

Simplicio: OK, I will simply repeat your choice. Not having to think is good.

Step 2. The Choice of N*:*

Choose $N > \frac{1}{\epsilon}$.

Step 3. The Check:

We MUST show: If $n \geq N$, *then* $|x_n - L| = |\frac{1}{n^2} - 0| = \frac{1}{n^2} < \epsilon$. *If* $n \geq N > \frac{1}{\epsilon}$, *then* $n > \frac{1}{\epsilon}$. *When we divide by* n *and multiply by* ϵ, *we find that* $\frac{1}{n} < \epsilon$ *as before. Since* $1 \leq n, n \leq n^2$. *Thus,* $|x_n - L| = \frac{1}{n^2} < \frac{1}{n} < \epsilon$.

Cauchy: Note that this last sequence converges to zero much more quickly than the sequence $\lim_{n\to\infty}\{\frac{1}{n}\}$. The difference in the rate of convergence will be discussed again when we compare the bisection and Newton/Raphson methods.

Simplicio: I don't see any reason for this new definition. How about an example that illustrates the benefits of this second definition?

Example 9.1.10. *Cauchy: OK, how about if we prove the* $\lim_{n\to\infty}\{\frac{2n-3}{5n+1}\}$ *exists.*

Virginia: Since we aren't told what the limit should equal, we have a problem even getting started. Maybe we should add an extra "Step" to the process, where we make

an educated guess for L.. In this example, it isn't too difficult to figure out that $L = \frac{2}{5}$.

Simplicio: How so?

Virginia: If we divide both numerator and denominator by the integer n, then we see that $\frac{2n-3}{5n+1} = \frac{2-\frac{3}{n}}{5+\frac{1}{n}}$. *Thus, if n is large, then the numerator is close to 2 and the denominator is close to 5. Thus, the limit L should equal* $\frac{2}{5}$.

Step 0. The Candidate for L:

Let $L = \frac{2}{5}$.

Step 1. The Challenge:

Let $\epsilon > 0$ *be given.*

Step 2. The Choice for N:

Since I have no idea how to choose N, I will simply assume what I am trying to prove and set $|\frac{2n-3}{5n+1} - \frac{2}{5}| < \epsilon$.

Simplicio: Wait a minute. Even I know that that can't assume what you are trying to prove.

Virginia: The idea is that we will be able to make an "educated guess" for a value of N that might work. In other words, if we are clever, we will be able to reverse the steps. All we are going to do is solve this inequality for n in the following steps:

1. $|\frac{2n-3}{5n+1} - \frac{2}{5}| < \epsilon$.
2. $|\frac{5(2n-3)-2(5n+1)}{5(5n+1)}| < \epsilon$.
3. $|\frac{-15-2}{5(5n+1)}| < \epsilon$.
4. $|\frac{-17}{5(5n+1)}| < \epsilon$.
5. $\frac{17}{5(5n+1)} < \epsilon$.
6. $\frac{17}{\epsilon} < 25n + 5$.
7. $\frac{17}{\epsilon} - 5 < 25n$.
8. $\frac{\frac{17}{\epsilon}-5}{25} < n$.

Now choose N to be any integer so that $N > \frac{\frac{17}{\epsilon}-5}{25} > \frac{17}{25\epsilon}$.

Note that if $N > \frac{17}{25\epsilon}$, then $\frac{17}{25N} < \epsilon$.

Thus, to find the integer N all you need to do is:

1. *Write down the absolute value of the difference between the limit L (in this case $L = \frac{2}{5}$) and the formula for x_n $(= \frac{2n-3}{5n+1})$,*

2. *Determine a common denominator $(= 5(5n+1))$,*

3. *Simplify the numerator $(= 17)$, and*

4. *Solve for n.*

Step 3. Check N works:

If $n > N$, then

$$\begin{aligned} \left| \frac{2n-3}{5n+1} - \frac{2}{5} \right| &= \frac{|5(2n-3) - 2(5n+1)|}{5(5n+1)} \\ &= \frac{|-15-2|}{5(5n+1)} \\ &= \frac{|-17|}{5(5n+1)} \\ &= \frac{17}{5(5n+1)} < \frac{17}{25N} < \epsilon. \end{aligned}$$

Galileo: For this example, Definition 2 has a technical advantage over Definition 1 because the absolute value function takes care of different cases that you would have had to separate. Thus, the argument is cleaner. OK, Mr. Simplicio. How about if you try the next example. It is going to reappear many times before these gathering are finished.

Example 9.1.11. *Using the definition of limit, prove that* $\lim_{n\to\infty}\{\frac{1}{2^n}\} = 0$.

Simplicio:

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Working backwards again, how about if we choose N so that $\frac{1}{2^N} < \epsilon$? If we solve this

inequality, we see that $\frac{1}{\epsilon} < 2^N$. *Taking logarithms of both side of the inequality, we see that* $-\log(\epsilon) = \log(\frac{1}{\epsilon}) < \log(2^N) = N\log(2)$. *Thus, we should choose* $N > -\frac{\log(\epsilon)}{\log(2)}$.

Step 3. The Check:

To complete the problem, simply reverse the steps. In other words, if $n \geq N$, *then* $n \geq N > -\frac{\log(\epsilon)}{\log(2)}$ *so that* $n\log(2) > -\log(\epsilon)$. *Thus,* $\log(2^n) > \log(\frac{1}{\epsilon}), 2^n > \frac{1}{\epsilon}$, *and* $\epsilon > \frac{1}{2^n}$.

Cauchy: I think he has got it!

Galileo: While not all limit problems can be solved in such a straightforward fashion, at least we have a method for these. In the spirit of Professor Polya, we should look back at what we have done and generalize the method. The next proposition does exactly that.

Proposition 9.1.10. *If* $x \in \Re$ *and* $|x| < 1$, *then* $\lim_{n\to\infty} x^n = 0$.

Proof. Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Working backwards again, how about if we choose N so that $|x|^N < \epsilon$? If we take logarithms of both side of this inequality, we see that $Nlog(|x|) < log(\epsilon)$. Since $|x| < 1, log(|x|) < 0$. Thus, when we divide both sides of the inequality by $log(|x|)$, the sign of the inequality reverses and we find that $N > \frac{\log(\epsilon)}{\log(|x|)}$.

Step 3. The Check:

To complete the problem, simply reverse the steps. In other words, show that if $n \geq N$, then $|x|^n < \epsilon$. □

Simplicio: I really like that proof,

Virginia: Really?

Simplicio: But, why is it important?

Galileo: As you will soon see, we can use this fact to show that the square root method of Archimedes/Heron always converges. For this application, $x = \frac{1}{2}$, which tells you that the error drops by 50% for each iteration of the algorithm. For the cube

root algorithm, $x = \frac{2}{3}$, which means that the error drops by 33% for each iteration. This fact will also appear in the error formula for the Contraction Mapping Theorem.

Cauchy: Once again following the dictums of Professor Polya, we should review what we have done and think bigger. At the beginning of our conversation about convergence, we began by defining the absolute value function and a distance metric. Distance is a very general concept and works in all dimensions.

Virginia: Pythagoras provides us with distance formulas for vectors in the plane and three space.

Cauchy: Better yet, Pythagoras provides us with distance formulas in infinite dimensional spaces.

Simplicio: I bet those formulas are really complicated.

Galileo: Actually, no. The formula for $\Re^n$ generalizes in a completely natural way.

Definition 9.1.11. *If $f(x), g(x) : [a, b] \to \Re$ are continuous functions, the distance between $f(x)$ and $g(x)$ is defined by*

$$d(f(x), g(x)) = \sqrt{\int_a^b (f(x) - g(x))^2 \, dx}.$$

If you think of the points $x \in [a, b]$ as coordinates, then this formula is exactly the Pythagorean Theorem. Moreover, it satisfies the same symmetry and triangle inequality properties that the absolute value function does. Thus, we can now talk about limits of functions.

Simplicio: OK, but why would we want to? How could that formula be useful?

Galileo: Since a multitude of applications are based on frequency and since frequencies can be modeled by the trigonometric functions $\cos(nx)$ and $\sin(nx)$ defined on the interval $[-\pi, \pi]$, we confront these problems everywhere. The heat equation and the wave equation are just the beginning.

Cauchy: True, but we are going to need to be more general than that. As it turns out,

Exercise Set 9.1.

1. Using either definition of limit, prove that $\lim_{n\to\infty}\{\frac{1}{n^3}\} = 0$.

2. Using the definition of limit, prove that $\lim_{n\to\infty}\{\frac{1}{n^4}\} = 0$.

3. Assume you have a sequence defined by the following rules:
   $x_0 = 2$.
   $x_{n+1} = \frac{x_n - \frac{1}{x_n}}{2}$.
   After the first fifty terms are computed, are you close to convergence yet? What can you conclude after the first million terms are computed? Do they seem to be bounded? Is the sequence increasing?

4. Using the definition of limit, prove the following limit exists: $\lim_{n\to\infty}\{\frac{3n-7}{2n+5}\}$.

5. Using the definition of limit, prove the following limit exists: $\lim_{n\to\infty}\{\frac{2n+5}{3n-7}\}$.

6. Prove: If $\lim_{n\to\infty} x_n = L$, then $\lim_{n\to\infty} |x_n| = |L|$. (Hint: This fact is easier to prove if you select the right fact from the right proposition. Otherwise, you have to consider a number of special cases.)

7. Find a sequence $\{x_n\}_{n=1}^{\infty}$ with the property that the statement $\lim_{n\to\infty} |x_n| = |L|$ is true, but the statement $\lim_{n\to\infty} x_n = L$ is false. (Remark: In other words, the converse to the previous problem may not be true.)

8. Using the definition of limit, prove that $\lim_{n\to\infty}\{\frac{1}{4^n}\} = 0$.

9.2 The Geometric Series

Galileo: Before we move on to more theoretical issues, we should discuss the Geometric series. This special case has played an important role in mathematics since Archimedes used it to compute the area under a parabola.
Virginia: But isn't that a Calculus issue?
Galileo: If that Roman soldier hadn't run the old man through with a spear, we would have had integration several thousand years ago. Archimedes was an amazingly

productive individual. When you read his proof of the volume of a sphere, all you can do is wonder at his imagination and energy. In any case, we now turn from the problem of computing the limit of a sequence to computing the sum of an infinite series.

Simplicio: What is difference between a sequence and a series?

Galileo: The sum of an infinite series is a special case of a limit of sequence. Thus, any fact we prove about the limit of a sequence immediately translates into a fact about series. However, before we do that, let's compute the sum of a finite series. This formula should be familiar.

Proposition 9.2.1 (Sum Formula for the Finite Geometric Series). *If $x \in \Re$ and $x \neq 1$ and $S_n = \sum_{k=0}^{n} x^k$, then $S_n = \frac{1-x^{n+1}}{1-x}$.*

Proof. If $S_n = \sum_{k=0}^{n} x^k$, then $xS_n = \sum_{k=0}^{n} x^{k+1}$. If we subtract these two equations, then only two terms remain on the right hand side. Thus, $(1-x)S_n = 1 - x^{n+1}$ and the result follows by dividing both sides of the equation by $1 - x$. □

Simplicio: That proof was too easy.

Example 9.2.1. *Galileo: How about the special case when $x = \frac{1}{4}$? Archimedes needed this case when he computed the area under a parabola.*

Virginia: But that is easy. By the formula, we can see that

$$S_n = 1 + \frac{1}{4^1} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots + \frac{1}{4^n} = \frac{1 - \frac{1}{4^{n+1}}}{1 - \frac{1}{4}} = \frac{4 - \frac{1}{4^n}}{3}.$$

Galileo: So what number is this sum close to?

Virginia: If n is large, then $\frac{1}{4^n}$ is small, which implies $S_n \approx \frac{4}{3}$.

Galileo: So, can you find a parabola with area $\frac{4}{3}$ under the curve?

Galileo: This example leads to the question: How do you sum an infinite series? When we computed in the proposition, note that we added up the first n terms of the sequence, which we denoted by S_n.

Virginia: We then observed the limit of this sequence of sums converges to $\frac{4}{3}$.

Galileo: We not make two definitions to formalize the ideas in this example.

Definition 9.2.2. *If $\sum_{k=0}^{\infty} x_k$ is an infinite series, then the sum $S_n = \sum_{k=0}^{n} x_k$ is called the n^{th} partial sum.*

Definition 9.2.3. *An infinite series $\sum_{k=0}^{\infty} x_k$ is said to converge to a number S, if the limit of the n^{th} partial sums converges to S. More precisely, $S = \sum_{k=0}^{\infty} x_k$ if and only if $\lim_{n\to\infty} S_n = S$, where $S_n = \sum_{k=0}^{n} x_k$.*

Galileo: In other words, the infinite sum S equals the limit of the sequence of partial sums. We are now in a position to compute the infinite version of the Finite Geometric series.

Proposition 9.2.4 (Sum Formula for the Infinite Geometric Series). *If $x \in \Re$ and $|x| < 1$ and $S_n = \sum_{k=0}^{n} x^k$, then $\sum_{k=0}^{\infty} x^k = \lim_{n\to\infty} S_n = \frac{1}{1-x}$.*

Proof. Step 1. The Challenge:

Let $\epsilon > 0$ be given. Step 2. The Choice:

Since $S_n = \sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}$, we only need to find an integer n with the property that

$$|\frac{1}{1-x} - \frac{1-x^{n+1}}{1-x}| < \epsilon.$$

Since $|\frac{1}{1-x} - \frac{1-x^{n+1}}{1-x}| = |\frac{x^{n+1}}{1-x}|$, we only need to show that $|\frac{x^{n+1}}{1-x}| < \epsilon$.

Working backwards, we see that

$$\begin{aligned} |x|^{n+1} &< (1-x)\epsilon \\ (n+1)log(|x|) &< log\{(1-x)\epsilon\} \\ n+1 &> \frac{log\{(1-x)\epsilon\}}{log(|x|)} \\ n &> \frac{log\{(1-x)\epsilon\}}{log(|x|)} - 1 \end{aligned}$$

Thus, we choose N to be any integer with the property that $N > \frac{log\{(1-x)\epsilon\}}{log(|x|)} - 1$.

Step 3. The Check:

To check that N works, simply assume $n \geq N$ and reverse the above inequalities. □

Simplicio: I noticed you reversed inequalities in the middle of the argument, where you chose N.

Galileo: Good observation. Since we assumed that $|x| < 1$, the quantity $log(|x|)$ is negative. Thus, we must reverse the inequality.

Simplicio: Does the argument work better if $x > 1$?

Galileo: Unfortunately, the proposition is false if $x > 1$.

Virginia: Which *log* function did you use? Natural or base 10?

Galileo: Choose your weapons. Either, in fact, any logarithm will work just fine.

Exercise Set 9.2.

1. Sum the finite series $S_n = 1 + 2 + 2^2 + \cdots + 2^n$.
2. Sum the terms in the finite sequence $S_n = 1 + 3 + 3^2 + \cdots + 3^n$.
3. Sum the terms in the infinite sequence $S = 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n} \ldots$.
4. Sum the terms in the infinite sequence $S = 1 + \frac{1}{3} + \frac{1}{3^2} + \cdots + \frac{1}{3^n} \ldots$.
5. Sum the terms in the infinite sequence $S = 1 - \frac{1}{2} + \frac{1}{2^2} - \cdots + (-1)^n \frac{1}{2^n} + \ldots$.
6. Sum the terms in the infinite sequence $S = 1 - \frac{1}{3} + \frac{1}{3^2} - \cdots + (-1)^n \frac{1}{3^n} + \ldots$.

9.3 Limit Theorems For Sequences

Cauchy: We next turn to the idea of making limits a bit easier so we don't always have to grind our way through this three step process of proving limits. For example, if you try to show that $\lim_{n\to\infty} \frac{2n^2+3n+5}{7n^2+1} = \frac{2}{7}$, you will find that annoying technical difficulties arise. Thus, while we still want to have the capability of using the definition to prove a limit, we would also like to have more weapons at our disposal. The point of our discussion will be to make limits and convergence easier.

Simplicio: I like easy.

Cauchy: However, before we start, I would like remark that we are going to be proving theorems and propositions. These proofs require that you understand the logic and

rigor of a mathematical argument. Before we proceed, it is necessary that you can answer the following questions.

1. What is the triangle inequality for the absolute value function?
2. What is the contrapositive of the statement "If p, then q."?
3. What is a proof by contradiction?
4. What is the connection between a proof by contradiction and the contrapositive of a statement?

Do you remember the contrapositive and modus tollens?
Virginia: Yes, I do.
Simplicio: I'm not sure.
Cauchy: Well, there is no point in proceeding until you know. Go back and review these concepts.
Simplicio: I think we should move on before my brain melts.
Virginia: I am ready.
Cauchy: Good. Let us begin. While you should have already seen these ideas in your previous study of Calculus, you may not have seen the proofs. The facts we will establish are:

1. The limit of the sum is the sum of the limit.
2. The limit of the product is the product of the limit.
3. The limit of the quotient is the quotient of the limit.
4. The uniqueness of limits.
5. Several squeezing propositions.

The proofs of the first three facts will all have the same 3 step structure that we just employed for our examples. For the sum, product, and quotient proofs, we will

use the absolute value function extensively. For the uniqueness and squeezing facts we will use a proof by contradiction strategy. Let's now state and prove the first proposition.

Proposition 9.3.1 (Limit Facts for Sequences). *Let $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ be sequences in $\Re$. If $\lim_{n\to\infty}\{x_n\} = L$ and $\lim_{n\to\infty}\{y_n\} = M$, then*

1. $\lim_{n\to\infty}\{x_n + y_n\} = \lim_{n\to\infty}\{x_n\} + \lim_{n\to\infty}\{y_n\} = L + M$,
   (i.e. The limit of the sum equals the sum of the limits or LS = SL.)

2. $\lim_{n\to\infty}\{x_n * y_n\} = \lim_{n\to\infty}\{x_n\} * \lim_{n\to\infty}\{y_n\} = L * M$,
   (i.e. The limit of the product equals the product of the limits or LP = PL.)

3. *If $M \neq 0$, then* $lim_{n\to\infty}\{\frac{x_n}{y_n}\} = \frac{\lim_{n\to\infty}\{x_n\}}{\lim_{n\to\infty}\{y_n\}} = \frac{L}{M}$.
   (i.e. The limit of the quotient equals the quotient of the limits or LQ = QL.)

Proof. 1. Let us begin by proving $\lim_{n\to\infty}\{x_n + y_n\} = L + M$.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Since we are assuming that $\lim_{n\to\infty}\{x_n\} = L$, we can find an integer N_1 with the property that if $n \geq N_1$, then $|x_n - L| < \frac{\epsilon}{2}$.

Since we are assuming that $\lim_{n\to\infty}\{y_n\} = M$, we can find an integer N_2 with the property that if $n \geq N_2$, then $|y_n - L| < \frac{\epsilon}{2}$.

Since we want both of these statements to be true, we choose N to be any integer larger than both N_1 and N_2. The best choice is $N = max\{N_1, N_2\}$.

Step 3. The Check:

If $n \geq N$, then by the triangle inequality

$$\begin{aligned} |x_n + y_n - (L + M)| &= |(x_n - L) + (y_n - M)| \\ &\leq |x_n - L| + |y_n - M| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon. \end{aligned}$$

2. Next let us prove $\lim_{n\to\infty}\{x_n * y_n\} = L * M$.

While the proof of this proposition is often considered more difficult than LS = SL, the approach is the same. The main difference is that we are confronted by the distributive law.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Since we are assuming that $\lim_{n\to\infty}\{x_n\} = L$, we can find an integer N_1 with the property that if $n \geq N_1$, then $|x_n - L| < \epsilon_1$.

Since we assume that $\lim_{n\to\infty}\{y_n\} = M$, we can find an integer N_2 with the property that if $n \geq N_2$, then $|y_n - L| < \epsilon_2$.

We again choose $N = max\{N_1, N_2\}$.

After we make a couple of computations, we will figure out reasonable choices for ϵ_1 and ϵ_2. For LS = SL, it was easy to see that ϵ_1 and ϵ_2 should both be chosen equal to $\frac{\epsilon}{2}$.

Step 3. The Check:

If $n \geq N$ and we have been smart enough to choose ϵ_2 so small that $|x_n|\epsilon_2 < \frac{\epsilon}{2}$ and $\epsilon_1|M| < \frac{\epsilon}{3}$, then by the distributive law and the triangle inequality we see that

$$\begin{aligned}
|x_n * y_n - (L * M)| &= |x_n * y_n - x_nM + x_nM - LM| \\
&= |x_n(y_n - M) + (x_n - L)M| \\
&\leq |x_n||(y_n - M)| + |(x_n - L)||M| \\
&\leq |x_n|\epsilon_2 + \epsilon_1|M| \\
&< |x_n|\epsilon_2 + \epsilon_1|M| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{3} \\
&< \epsilon.
\end{aligned}$$

Virginia: While I am not sure about ϵ_2, I can see that we should choose $\epsilon_1 = \frac{1}{3|M|}$, then $\epsilon_1|M| < \frac{\epsilon}{3}$.

Cauchy: But if $M = 0$, then you are dividing by zero. Bad idea.

Virginia: You are correct. I guess I had better choose $\epsilon_1 = \frac{\epsilon}{3|M|+1}$ so the denominator can never equal zero AND the choice of ϵ_1 will still have the property that $\epsilon_1|M| < \frac{\epsilon}{3}$.

Cauchy: Yes, you have now covered all the cases.

Simplicio: But what about choosing ϵ_2 so that $|x_n|\epsilon_2 < \frac{\epsilon}{2}$? I don't see that choice at all.

Cauchy: We can begin addressing that question by observing that if we choose $\epsilon_1 < \frac{1}{2}$, then we will know that $|x_n| < |L| + \frac{1}{2}$ for all $n \geq N_1$.

Virginia: In other words, if we had chosen $\epsilon_2 = \frac{\epsilon}{3|L|+1}$, then we can guarantee that $|x_n|\epsilon_2 < (|L| + \frac{1}{2}) * \frac{\epsilon}{3|L|+1} < \frac{\epsilon}{2}$. Thus, to complete the argument, we only need to choose $\epsilon_1 = Min\{\frac{1}{2}, \frac{\epsilon}{3|M|+1}\}$.

Cauchy: Correct.

3. Next let us prove the quotient rule: If $M \neq 0$, then $\lim_{n\to\infty}\{\frac{x_n}{y_n}\} = \frac{L}{M}$.

Since the strategy for proof of LQ = QL is similar to LP = PL, we will leave the proof as an exercise. However, since we have just proved that the limit of the product equals the product of the limit, note that we only need to prove the special case: $\lim_{n\to\infty}\{\frac{1}{y_n}\} = \frac{1}{M}$. □

Simplicio: Thanks. I have had enough anyway. How about an example?

Example 9.3.1. *Cauchy: Suppose you are asked to show* $\lim_{n\to\infty}\{\frac{2n^2+3n+5}{7n^2+1}\} = \frac{2}{7}$. *If you try to use the definition, you will find the process annoying. However, with the Basic Limit Facts, we simply make the following computations:*

$$\begin{aligned}\lim_{n\to\infty}\{\frac{2n^2+3n+5}{7n^2+1}\} &= \frac{\lim_{n\to\infty}\{2+\frac{3}{n}+\frac{5}{n^2}\}}{\lim_{n\to\infty}\{7+\frac{1}{n^2}\}} && (LQ = QL)\\ &= \frac{2+0+0}{7+0} && (LS = SL)\\ &= \frac{2}{7}.\end{aligned}$$

Cauchy: The next corollary shows that we can "pull" a constant across the limit sign.

Corollary 9.3.2. *If K is a real number and $\{x_n\}_{n=1}^{\infty}$ is a sequence of numbers such that $\lim_{n\to\infty} x_n = L$, then $\lim_{n\to\infty} Kx_n = K\lim_{n\to\infty} x_n = KL$.*

Proof. This result follows immediately from the limit of the product equals the product of the limits because we can define $y_n = K$ for all n. Since the limit of the constant sequence $K, K, \ldots, K, \ldots$ is K, we are done. □

Cauchy: We now give a second proof of the sum formula for the Geometric series.

Simplicio: A second proof?

Galileo: The result is useful and Repetition is a great teacher. You will see this formula again.

Proposition 9.3.3 (Sum Formula for the Infinite Geometric Series). *If* $x \in \Re$, $|x| < 1$, *and* $S_n = \sum_{k=0}^{n} x^k$, *then* $\sum_{k=0}^{\infty} x^k = \lim_{n\to\infty} S_n = \frac{1}{1-x}$.

Proof. Since we are assuming that $|x| < 1$, we know $\lim_{n\to\infty} x^n = 0$. By the limit of the sum equals the sum of the limits and the previous corollary we can see that

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \frac{1-x^{n+1}}{1-x} = \frac{1}{1-x} \lim_{n\to\infty} (1-x^{n+1}) = \frac{1}{1-x} - \frac{1}{1-x} \lim_{n\to\infty} x^{n+1} = \frac{1}{1-x}.$$

□

Cauchy: We now prove uniqueness for limits.

Simplicio: Uniqueness? I have been patient until now, but this theory stuff is killing me.

Cauchy: While you may not think uniqueness is important, engineers really do want to know when there is only one answer. In some sense, the sequence $x_n = (-1)^n$ has both -1 and $+1$ as it limits. Rather than deal with this ambiguity, the mathematics community has voted to say the sequence does not converge. While these facts may seem obvious, they require proof.

Simplicio: But every test problem I ever did only had one answer. (To Virginia) Did you ever bubble in more than one answer?

Virginia: No, but few of my tests were multiple guess.

Cauchy: OK, but quadratic polynomials usually have two roots. A multitude of computational problems have more than one answer. Life is easier when we have uniqueness.

Simplicio: One wife, one mother-in-law?

Proposition 9.3.4 (Uniqueness Theorem for Limits of Sequences). *Let* $\{x_n\}_{n=1}^{\infty}$ *be a sequence of numbers in* $\Re$. *If* $\lim_{n\to\infty}\{x_n\} = L_1$ *and* $\lim_{n\to\infty}\{x_n\} = L_2$, *then* $L_1 = L_2$.

Proof. Cauchy: By way of contradiction, we will assume the proposition is false. In other words, we will assume $L_1 \neq L_2$. If you make a smart choice of ϵ–namely $\epsilon = \frac{1}{2}dist(L_1, L_2) = \frac{1}{2}|L_1 - L_2|$, then you will find that all but a finite number of the terms of the sequence must lie in both of the intervals $(L_1-\epsilon, L_1+\epsilon)$ and $(L_2-\epsilon, L_2+\epsilon)$. However, by the choice of ϵ, there are no points in both of these intervals. Thus, we have a contradiction. Now, that wasn't so bad was it? □

Simplicio: Short is good. It was OK.

Cauchy: Now it is time to squeeze.

Simplicio: And I must ask again. What are these facts good for?

Cauchy: A basic rule for applications is that inequalities are more important than equalities. As physicist Werner Heisenberg (1901-1976) pointed out, measurements are not exact and we are thus forced to settle for approximate answers. Under these circumstances, we are comfortable if we can control a sequence by squeezing it between two constants. Many of the algorithms we will be using can be controlled this way.

Simplicio: How about an example.

Cauchy: While root finding method of Newton/Raphson and the Contraction Mapping Theorem are the first settings where we will need these ideas, we will also need tools of estimation everywhere in Fourier series. Squeezing helps.

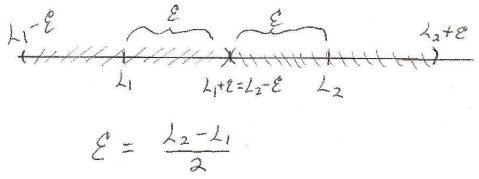


Figure 9.1: The Uniqueness of Limits

Proposition 9.3.5 (The Squeezing Theorem for Sequences). *Let* $\{x_n\}_{n=1}^{\infty}$, $\{y_n\}_{n=1}^{\infty}$, *and* $\{z_n\}_{n=1}^{\infty}$ *be sequences in* $\Re$, *where* $x_n \leq y_n \leq z_n$.

1. *Fact 1. If* $\lim_{n\to\infty}\{x_n\} = L$ *and* $\lim_{n\to\infty}\{z_n\} = M$, *then* $L \leq M$.

2. *Fact 2. If the sequence* $\{y_n\}_{n=1}^{\infty}$ *converges and* $y_n \leq M$ *for all* n, *then* $\lim_{n\to\infty}\{y_n\} \leq M$.

3. *Fact 3. If* $\lim_{n\to\infty}\{x_n\} = L = \lim_{n\to\infty}\{z_n\}$, *then the sequence* $\{y_n\}_{n=1}^{\infty}$ *converges and* $\lim_{n\to\infty}\{y_n\} = L$.

Proof. Proof of Fact 1.

The proof of the first squeezing fact, is again by contradiction. Thus, we begin by assuming that $L > M$. The next step is to let $\epsilon = \frac{1}{2}dist(L, M) = \frac{1}{2}|L - M|$. Since $L > M$, we have the situation that all but a finite number of the terms of the sequence $\{x_n\}_{n=1}^{\infty}$ lie in the interval $(L - \epsilon, L + \epsilon)$ and all but a finite number of the terms of the sequence $\{y_n\}_{n=1}^{\infty}$ are in the interval $(M - \epsilon, M + \epsilon)$. Since these two intervals are disjoint and $L > M$, we have now created the problem that all $y_n < x_n$ for all but a finite number of the integers n. Thus, we have a contradiction to our assumption that $x_n \leq y_n$ for ALL n.

Proof of Fact 2.

This fact follows immediately from Fact 1 because the constant M can be thought of as a sequence where $z_n = M$, for all n.

Proof of Fact 3.

Since we are not assuming that the sequence $\{y_n\}_{n=1}^{\infty}$ converges to any number, this fact doesn't immediately follow from Facts 1 or 2. However, we can go back to basics.

Step 0. The Candidate:

The only possibility is that the sequence $\{y_n\}_{n=1}^{\infty}$ should converge to M.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

The integer N will be the maximum of the integers N_1 and N_2, where

1. If $n \geq N_1$, then $x_n \in (M - \epsilon, M + \epsilon)$.

2. If $n \geq N_2$, then $z_n \in (M - \epsilon, M + \epsilon)$.

Step 3. The Check:

Thus, if $n \geq N$, then both x_n and z_n lie in the interval $(M - \epsilon, M + \epsilon)$. Since we are assuming $x_n \leq y_n \leq z_n$, $y_n \in (M - \epsilon, M + \epsilon)$.

□

Exercise Set 9.3.

1. Using limit facts, prove that $\lim_{n\to\infty}\{\frac{1}{n^3}\} = 0$.

2. Using limit facts, prove that $\lim_{n\to\infty}\{\frac{1}{n^4}\} = 0$.

3. Using limit facts, prove that $\lim_{n\to\infty}\{\frac{3n-7}{2n+5}\} = \frac{3}{2}$.

4. Using limit facts, prove that $\lim_{n\to\infty}\{\frac{2n+5}{3n-7}\} = \frac{2}{3}$.

5. Using limit facts, prove that $\lim_{n\to\infty}\{\frac{2n^2+7}{3n^2-5}\} = \frac{2}{3}$.

6. Using limit facts, prove that $\lim_{n\to\infty}\{\frac{2n^3+5}{3n^3-7}\} = \frac{2}{3}$.

9.4 Every Bounded Increasing Sequence Converges

> Numbers are the free creation of the human mind.-Julius Wilhelm Richard Dedekind (1831-1916)

Galileo: We now turn to the problem of showing that every bounded increasing sequence converges.

Simplicio: I hate to be predictable, but why should I care?

Galileo: The short answer is that if we can show an algorithm produces a sequence of numbers which is both bounded and increasing, then the method will "work." For an engineer, it is important that the method produce accurate answers reliably.

Virginia: The long answer?

Galileo: The long answer is that it took well over 2000 years to figure out how to fill in the holes in the real numbers. Since checking all the details of this construction is really really boring, we are only going to present the flavor of the ideas. This topic is probably the most theoretical we will encounter in this tutorial. If you do not remember our discussion of rigor and logic, it might be a good time to review definitions, contrapositives, and proof by contradiction,

Simplicio: I believe in the real numbers. Maybe I will take a short nap.

Galileo: The following two examples should set the stage for the main theorem.

Example 9.4.1. *The sequence $x_k = k^2$ is increasing, but not bounded.*

Example 9.4.2. *The sequence $x_k = (-1)^k$ is bounded, but not increasing.*

Simplicio: And?

Galileo: As we have already remarked, an engineer wants to have confidence in his answers. In other words, if he hits the square root button on his calculator, he would like to know the answer is correct. The beauty of the Archimedes/Heron square root method is that it always produces a bounded decreasing sequence. The beauty of the bisection method is that it produces a sequence of closed intervals, where the left endpoints are increasing and the right endpoints are decreasing. Thus, the answer is always "trapped." Thus, if we can show that every bounded increasing sequence converges, then we will have shown that these two methods "always work."

Galileo: We now turn to a fascinating little problem that has caused 2000 years of consternation. Namely, how do we "fill in" the "holes" in the real line so we can be sure the irrational numbers such as $\sqrt{2}, e, \pi$, and e^π are well-defined.

Simplicio: Wait a minute. What does the word "well-defined" mean?

Galileo: Julius Wilhelm Richard Dedekind (1831-1916) went to great lengths to get arithmetic right. With his idea of a "cut" he showed that the associative, commutative, and distributive laws for addition and multiplication can not only be extended from the positive and negative integers Z to the rational numbers $Q = \{\frac{p}{q} : p, q \in Z \text{ and } q > 0\}$, but can also be extended to the real numbers $\Re$. A large part of this problem is the exact definition of a real number.

Definition 9.4.1. *A non-empty subset S of Q is a called a cut if the following conditions hold:*

1. *The set S is not equal to Q.*
2. *If whenever $p \in S$ and $q < p$, then $q \in S$.*
3. *The set S contains no largest rational number.*

Virginia: Thus, the number $\sqrt{2}$ can be represented by the set
$S = \{\frac{p}{q} : (\frac{p}{q})^2 < 2 \text{ or } \frac{p}{q} < 0\}$. In general, a real number can be represented by a "connected" open interval of rational numbers! And the real numbers $\Re$ is the collection of all such connected open intervals.

Galileo: Correct.
Simplicio: But I thought a real number was a point? Now you tell me it is a set.
Galileo: No worries. You can go back to thinking a real number is a point. While this construction represents an important milestone in establishing the rigor of arithmetic, I agree that it can only be described as tedious. The details are guaranteed to put even the sleep deprived into a sound slumber.
Simplicio: I am a man of faith. Let's move on.

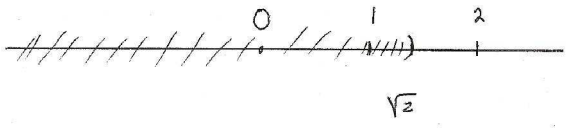


Figure 9.2: A Dedekind Cut Representing $\sqrt{2}$

Galileo: The Least Upper Bound Principle is a consequence of Dedekind's construction. The importance of this principle is that it "fills in" all the "holes" in the real number line.

Virginia: When you use the word consequence, I suspect you mean that this Principle is really a theorem which must be proved from other more basic assumptions

Galileo: Correct again. While the Least Upper Bound Principle is a theorem, which can be proved from the properties of Dedekind's construction, we will not go there. In the interests of time, we will assume it is true.

Virginia: Like an axiom, a postulate, or a definition?

Galileo: Yes.

Simplicio: As I said, let's move on.

Galileo: Before we can state this important principle, we must define what it means for a set to have an upper bound.

Definition 9.4.2 (Bounded Above). *A non-empty set $S \subset \Re$ is bounded above if there is a number $M \in \Re$ with the property that $x \leq M$ for all $x \in S$. The number M is called an upper bound for the set S.*

Galileo: We now define the least upper bound (lub) of a set of real numbers.

Definition 9.4.3 (Least Upper Bound). *If a real number L is an upper bound for a non-empty set $S \subset \Re$, then L is called the least upper bound (lub) of S if for any upper bound M of the set S, it is always true that $L \leq M$.*

We now state the Least Upper Bound Principle.

Principle 9.4.4 (The Least Upper Bound Principle). *If a non empty set $S \in \Re$ is bounded above, then it has a least upper bound.*

Simplicio: I failed to get that principle at all. I need an example.

Galileo: If we consider the sequence $x_n = (-1)^n$, we notice that the terms oscillate between $+1$ and -1. While the sequence has a multitude of upper bounds such as $2, 47$, and 1001, the number $+1$ is not only an upper bound but, in fact, the least

upper bound. On the other hand, if we consider the sequence $x_n = \frac{n}{n+1}$, we again notice that the sequence has a multitude of upper bounds including $2, 47,$ and 1001. Again, the least upper bound of the sequence is $+1$.

Simplicio: Why did you give us two examples with the same answer?

Galileo: To point out that in the first example the least upper bound is equal to one of the terms of the sequence, while the least upper bound in the second case never equals any term in the sequence. If the least upper bound was always one of the terms in the sequence, it never would have been invented. In fact, if the least upper bound was always a rational number, it never would have been invented. In other words, the Least Upper Bound Principle fills in the "holes" in the real number system vacated by numbers such as $\sqrt{2}, \sqrt[3]{2}, e, \pi$.

Simplicio: Let's move on.

Galileo: Certainly. We begin with two important concepts associated with sequences: increasing and bounded. These two ideas will provide a test for when a sequence converges. The definitions of these terms are now presented. We begin with the definition of an upper bound for a sequence.

Definition 9.4.5. *A sequence $\{x_k\}_{k=1}^{\infty}$ is bounded above if there is a number $M \in \Re$ such that $x_k \leq M$ for all integers $k \geq 1$.*

Definition 9.4.6. *A sequence $\{x_k\}_{k=1}^{\infty}$ is increasing if $x_k \leq x_{k+1}$ for all $k \geq 1$.*

Theorem 9.4.7 (Every Bounded Increasing Sequence Converges). *If a sequence $\{x_n\}_{n=1}^{\infty}$ is both bounded above and increasing, then there is a number L such that $\lim_{n\to\infty}\{x_n\} = L$. In particular, if M is any upper bound, then $x_n \leq L \leq M$ for all n.*

Proof. The reason we mention the least upper bound principle is to identify the limit L.

Step 0. The Candidate:

Set L equal to the least upper bound of the set of points consisting of all the terms of the sequence $\{x_n\}_{n=1}^{\infty}$. In particular, $L = lub\{x_n : n = 1, 2, 3, \ldots, n, \ldots\}$. We must

now show that $\lim_{n\to\infty}\{x_n\} = L$.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose N so that $x_N > L - \epsilon$.

Simplicio: How do we know we can find such an N?

Galileo: Good question. Once again, the only viable proof for the existence of such an integer N is by contradiction. To this end, we assume that no such integer N exists. But, if we make this assumption, then $x_N \leq L - \epsilon$ for ALL integers N. Thus, $L - \epsilon$ is also an upper bound for the sequence. Since $L < L - \epsilon$, we would have a contradiction of the assumption that L is the least (or smallest) upper bound.

Step 3. The Check:

We must now show that if $n \geq N$, then $x_n \in (L - \epsilon, L + \epsilon)$. Since $n \geq N$, and we are assuming the sequence is increasing, we know that $X_N \leq x_{N+1} \leq x_{N+2} \leq \ldots, x_n$. Thus, $L - \epsilon < x_N \leq x_n$.

Since we are assuming that L is an upper bound for the sequence, $x_n \leq L < L + \epsilon$.

Thus, $x_n \in (L - \epsilon, L + \epsilon)$ and the sequence converges to L. □

Galileo: Now that proof wasn't so bad, was it?

Simplicio: This proof seems to have the same four steps as the others.

Galileo: An equivalent formulation of this theorem (and the one that we will need) can be stated in terms of bounded decreasing sequences.

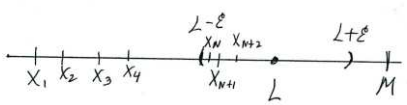


Figure 9.3: Every Bounded Increasing Sequence Converges

Definition 9.4.8. *A sequence $\{x_k\}_{k=1}^{\infty}$ is said to be bounded below if there is a number M such that $x_k \geq M$ for all integers $k \geq 1$.*

Definition 9.4.9. *A sequence $\{x_k\}_{k=1}^{\infty}$ is said to be decreasing if $x_k \geq x_{k+1}$ for all integers $k \geq 1$.*

Theorem 9.4.10. *If a sequence $\{x_n\}_{n=1}^{\infty}$ is both bounded below and decreasing, then there is a number L such that $\lim_{n\to\infty}\{x_n\} = L$.*

Galileo: For bounded decreasing sequences, we will see that the sequence will actually converge to the greatest lower bound.
Simplicio: I have a question. In a real-world problem, you don't know the answer so you can't begin to test if some number L is a limit. If you did, you wouldn't do all this checking. Why waste your time when a client wants the results yesterday.
Galileo: You have a good point. All we have done so far is set the context. We will return to your question when we discuss Cauchy sequences. His sequences are the ones engineers care about.
Simplicio: Cauchy again?

Exercise Set 9.4.

1. Compute the least upper bound of the sequence $\{\frac{(-1)^n}{n}\}_{n=1}^{\infty}$. Compute the greatest lower bound. Does the sequence converge to the least upper bound?

2. Compute the least upper bound of the sequence $\{(-1)^n \frac{n-1}{n}\}_{n=1}^{\infty}$. Compute the greatest lower bound. Does the sequence converge to the least upper bound?

3. Prove: If a sequence $\{x_n\}_{n=1}^{\infty}$ is both bounded below and decreasing, then there is a number L such that $\lim_{n\to\infty}\{x_n\} = L$.

9.5 Cauchy Sequences

Galileo: We now recall our friend Cauchy to provide a brief introduction to a criterion that guarantees a sequence converges.

Simplicio: I dread the thought of more theory.

Cauchy: The reason for defining this new concept is that we would like to be certain a sequence converges even when we have no idea what the limit will be. If we know the answer, then why waste time computing limits!! Since the limit is missing, the setting is more like the situations engineers face with real-world problems. Namely, they don't know the answer before they start. However, it will turn out that while we don't know the limit exactly, it can be contained somewhere in a small interval.

Galileo: Actually, Mr. Simplicio has already encountered these ideas in Calculus when he was introduced to the ratio and n^{th} root tests.

Simplicio: I liked the ratio test. It was easy because all you had to do was compute $r = \lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}$. If $r < 1$, then the series $\sum_{n=0}^{\infty} a_n$ converges. If $r > 1$, then the series diverges.

Galileo: Very good.

Simplicio: Actually, that is the only technique I remember on that subject.

Galileo: The only problem is that several cards were dealt from the bottom of the deck.

Simplicio: How so?

Galileo: The technique didn't actually give you the answer.

Simplicio: You are correct. The answer to those problems was simply "convergent" or "divergent."

Virginia: But wait a minute. If you think about the proofs of the ratio test, you are dominating the given series by a Geometric series. That information ought to help.

Simplicio: I do my best to avoid proofs and here she comes.

Virginia: If we assume the series $\sum_{n=0}^{\infty} a_n$ has the property $\frac{|a_n|}{|a_{n-1}|} \leq r$ for all integers

$n = 0, 1, 2, \ldots, n, \ldots$, then $|a_n| \leq |a_{n-1}|r$ for all n. Thus,

$$\begin{aligned}
|a_0| &\leq |a_0|r^0. \\
|a_1| &\leq |a_0|r^1. \\
|a_2| &\leq |a_1|r \leq |a_0|r^2. \\
|a_3| &\leq |a_2|r \leq |a_0|r^3. \\
|a_4| &\leq |a_3|r \leq |a_0|r^4. \\
&\vdots \\
|a_n| &\leq |a_{n-1}|r \leq |a_0|r^n.
\end{aligned}$$

Adding these quantities, we see by the sum formula for the Geometric series that

$$|\sum_{n=0}^{\infty} a_n| \leq \sum_{n=0}^{\infty} |a_n| \leq |a_0| \sum_{n=0}^{\infty} r^n = |a_0| \frac{1}{1-r}.$$

We can always estimate the error by comparing the tails of series

$$|E_n| = |\sum_{k=0}^{\infty} a_k - \sum_{k=0}^{n} a_k| = |\sum_{k=n+1}^{\infty} a_k| \leq \sum_{k=n+1}^{\infty} |a_k| \leq |a_0| \sum_{k=n+1}^{\infty} r^k = |a_0| \frac{r^{n+1}}{1-r},$$

Since $\lim_{n\to\infty} |a_0| \frac{r^{n+1}}{1-r} = 0$, we have convergence.

Galileo: Very good! However, it isn't immediately clear that the symbol $\sum_{k=0}^{\infty} a_k$ actually represents a real number.

Simplicio: But isn't that obvious?

Galileo: Show me the sum.

Virginia: If you think about it, the only general condition we have that guarantees a sequence converges is that it is bounded and increasing.

Galileo: Correct. The reason for Cauchy sequences is to guarantee convergence. Once we have completed this task, the ratio test will guarantee that the symbol $\sum_{k=0}^{\infty} a_k$ makes sense. By the way, Cauchy is involved whenever we are apply any comparison test. In particular, the root test and the integral test are involved.

Simplicio: OK, enough of these old tests, how about this Contraction Mapping Theorem?

Galileo: The strategy is the same with the Contraction Mapping Theorem, Namely, you use an iterated function computation $x_{n+1} = T(x_n)$ to create an infinite sequence $\{x_n\}_{n=0}^{\infty}$ of points. Since $T(x)$ is a contraction with contraction factor $M < 1$, we can use the same Geometric series argument Virginia just mentioned to show that $|x_n - x_N| \leq \frac{M^n}{1-M}|x_0 - x_1|$ for all $n \geq N$. This inequality will be sufficient to show that the sequence is Cauchy. Later we will see we have the same issues with Fourier series. While it is easy to show the series $\sum_{n=0}^{\infty} \frac{1}{n^3}\cos(nx)$ converges for all $x \in \Re$, it is not so easy to figure out a tidy little formula for the function it represents.

Simplicio: So where do we begin?

Galileo: We begin with the definition, which poses the following challenge: If given a sequence $\{x_n\}_{n=1}^{\infty}$ and a tolerance $\epsilon > 0$, then find an integer N so that whenever $n \geq N$, the point x_n will lie in the interval $(X_N - \epsilon, X_N + \epsilon)$. In particular, all but a finite number of the terms in the sequence will lie in the interval $(X_N - \epsilon, X_N + \epsilon)$. As we did with the second definition for convergence, we will use the absolute value function and distance in the definition of Cauchy Sequence.

Definition 9.5.1 (Cauchy Sequence). *A sequence $\{x_n\}_{n=1}^{\infty}$ is called Cauchy, if for every $\epsilon > 0$, there is an integer N with the property that if $n \geq N$, then $|x_n - x_N| < \epsilon$.*

Cauchy: Note that this definition is exactly the same as the definition of limit except there is no mention of the limit L. Consider the following examples.

Example 9.5.1. *The sequence $x_n = \frac{(-1)^n}{n}$ is Cauchy.*

The argument this statement is true is the same as we encountered for convergent sequences.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose $N > \frac{2}{\epsilon}$.

Step 3. The Check:

If $n \geq N$, then $|\frac{(-1)^n}{n} - \frac{(-1)^N}{N}| \leq |\frac{(-1)^n}{n}| + |-\frac{(-1)^N}{N}| \leq \frac{1}{N} + \frac{1}{N} \leq \frac{2}{N} < \epsilon$.

Simplicio: That argument is certainly within my comfort zone.

Example 9.5.2. *The sequence* $x_n = \frac{(-1)^n}{n^2}$ *is Cauchy.*

Step 1. The Challenge:

Let $\epsilon > 0$ *be given.*

Step 2. The Choice:

Choose $N > \sqrt{\frac{2}{\epsilon}}$. *Thus,* $N^2 > \frac{2}{\epsilon}$.

Step 3. The Check:

If $n \geq N$, *then* $|\frac{(-1)^n}{n^2} - \frac{(-1)^N}{N^2}| \leq \frac{1}{N^2} + \frac{1}{N^2} \leq \frac{2}{N^2} < \epsilon$.

Simplicio: So it looks like we need to choose the integer N a bit larger than before.

Cauchy: I knew you would like this topic.

Galileo: The beauty of the situation is that convergent sequences are Cauchy and vice versa. Our first theorem is the observation that if a sequence is convergent, then it must also be Cauchy. Note that the format of the proof exactly parallels the proofs of the previous limit theorems. Note also, that the triangle inequality is evident.

Theorem 9.5.2 (Convergent Sequences are Cauchy). *If a sequence of real numbers* $\{x_k\}_{k=1}^{\infty}$ *is convergent, then it is a Cauchy sequence. In particular, if there is a number* L *so that* $lim_{n\to\infty} x_n = L$, *then* $\{x_k\}_{k=1}^{\infty}$ *is Cauchy.*

Proof. Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose N so that if $n \geq N$, then $|x_n - L| < \frac{\epsilon}{2}$.

Step 3. The Check:

We must show that if $\epsilon > 0$ is given, then we can always find an integer N such that whenever $n \geq N$, then $|x_n - x_N| < \epsilon$.

However, since the sequence converges to some limit L, we know by the definition of limit that there is an integer N such that if $n \geq N$, then $|x_n - L| < \epsilon/2$.

Thus, $|x_n - x_N| = |x_n - L + L - x_N| \leq |x_n - L| + |L - x_N| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$.

Thus, the sequence is Cauchy. □

Cauchy: We now prove the converse of the previous theorem, which shows that every Cauchy sequence converges to some number.

Simplicio: But I thought you said we couldn't find the number.

Cauchy: As you might have guessed, the answer comes to us as a least upper bound or a greatest lower bound of a set of numbers. While it is a bit theoretical, we do have it trapped in an arbitrarily small closed bounded interval.

Theorem 9.5.3 (Cauchy Sequences Converge). *If a sequence of real numbers* $\{x_n\}_{n=1}^{\infty}$ *is Cauchy, then there is a unique* L *such that* $lim_{n\to\infty}\{x_n\} = L$.

Proof. We will find two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ such that:

1. $a_{n-1} \leq a_n \leq b_n \leq b_{n-1}$ for all integers n,
2. $b_n - a_n \leq \frac{2}{n}$ for all integers n, and
3. for each integer n there is an integer N_n with the property that if $k \geq N_n$, then $x_k \in [a_n, b_n]$.

The essence of the argument is to simply set ϵ equal to smaller and smaller numbers and then apply the definition of Cauchy sequence. While any sequence of numbers which converges to zero will do, we simply let $\epsilon = \frac{1}{n}$ for larger and larger values of n.

Case $n = 1$. Let $\epsilon = 1$.

Find an integer N_1 such that if $k \geq N_1$, then $|x_k - x_{N_1}| < 1$. Let $a_1 = x_{N_1} - 1$ and $b_1 = x_{N_1} + 1$. Note that $b_1 - a_1 = \frac{2}{1}$ $a_1 \leq x_k \leq b_1$ for all $k \geq N_1$.

Case $n = 2$. Let $\epsilon = \frac{1}{2}$.

Find an integer $N_2 > N_1$ such that if $k \geq N_2$, then $|x_k - x_{N_2}| < \frac{1}{2}$. Let $a_2 = max\{a_1, x_{N_2} - \frac{1}{2}\}$ and $b_2 = min\{b_1, x_{N_2} + \frac{1}{2}\}$. Note that $b_2 - a_2 \leq \frac{2}{2}$ and $a_1 \leq a_2 \leq x_k \leq b_2 \leq b_1$ for all $k \geq N_2$.

Case $n = 3$. Let $\epsilon = \frac{1}{3}$.

Find an integer $N_3 > N_2$ such that if $k \geq N_3$, then $|x_k - x_{N_3}| < \frac{1}{3}$.

Let $a_3 = max\{a_2, x_{N_3} - \frac{1}{3}\}$ and $b_3 = min\{b_2, x_{N_3} + \frac{1}{3}\}$. Note that $b_3 - a_3 \leq \frac{2}{3}$ and $a_1 \leq a_2 \leq a_3 \leq x_k \leq b_3|leb_2 \leq b_1$ for all $k \geq N_3$.

Case $n = n$. Let $\epsilon = \frac{1}{n}$.

Continuing inductively, find an integer $N_n > N_{n-1}$ such that if $k \geq N_n$, then $|x_k - x_{N_n}| < \frac{1}{n}$.

Let $a_n = max\{a_{n-1}, x_{N_n} - \frac{1}{n}\}$ and $b_n = min\{b_{n-1}, x_{N_n} + \frac{1}{n}\}$. Note that $b_n - a_n \leq \frac{2}{n}$ and $a_1 \leq a_2 \leq a_3 \leq \cdots \leq a_n \leq x_k \leq b_n \leq \cdots \leq b_3 \leq b_2 \leq b_1$ for all $k \geq N_n$.

Since the sequence $\{a_n\}_{n=1}^{\infty}$ is bounded and increasing, it converges to some number L. Since the sequence $\{b_n\}_{n=1}^{\infty}$ is bounded and decreasing, it also converges. Since $b_n - a_n \leq \frac{2}{n}$ for all integers n, the sequences must converge to the same number L. Note that $a_n \leq L \leq b_n$ for all n.

We now have to prove that the sequence $\{x_n\}_{n=1}^{\infty}$ converges to L.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose N large enough that $\frac{2}{N} < \epsilon$ and N large enough so that whenever $n \geq N$, then $a_N \leq x_n \leq b_N$ in the above construction. In particular, we know $b_N - a_N \leq \frac{2}{N} < \epsilon$.

Step 3. The Check:

If $n \geq N$, then $x_n \in [a_N, b_N]$. Since $L \in [a_N, b_N]$, $|x_n - L| \leq b_N - a_N \leq \frac{2}{N} < \epsilon$.

Thus, $\{x_n\}_{n=1}^{\infty}$ must converge to L. □

Galileo: In the spirit of Professor Polya, let's think about the key components contained in this proof.

1. Construct a nested sequence of closed bounded intervals $\{[a_n, b_n]\}_{n=1}^{\infty}$.

2. Note that since $a_n \leq a_{n+1} \leq b_{n+1} \leq b_n$ for all n, both $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ converge.

3. If $\lim_{n\to\infty}(b_n - a_n) = 0$, then both sequences converge to the same number. In other words, there is a number L so that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$.

4. Any sequence which is frequently in each of these intervals has a subsequence which converges to L. In other words, if $\{x_n\}_{n=1}^{\infty}$ is a sequence with the property

that there are integers $n_1 < n_2 < \cdots < n_k < n_{k+1} < \ldots$ such that $x_{n_1} \in [a_1, b_1], x_{n_2} \in [a_2, b_2], x_{n_3} \in [a_3, b_3], etc.$ then $\lim_{k\to\infty} x_{n_k} = L$.

5. Any sequence squeezed by these intervals also converges to L. In other words, if $\{x_n\}_{n=1}^{\infty}$ is a sequence with the property that for every integer n there is an integer N_n such that whenever $k \geq N_n$, then $x_k \in [a_n, b_n]$, then $\lim_{k\to\infty} x_k = L$.

The first three items in this construction can be encapsulated in a proposition.

Proposition 9.5.4. *If $\{[a_n, b_n]\}_{n=1}^{\infty}$ is a nested sequence of closed bounded intervals with the property that $\lim_{n\to\infty}(b_n - a_n) = 0$, then there is a unique point L which is contained in every interval $[a_n, b_n]$. Moreover, $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$.*

We will see this construction again when we discuss compactness. We will need this property when was show integrals of reasonable functions exist.

Exercise Set 9.5.

1. Show the sequence $x_n = \frac{(-1)^n}{n^3}$ is Cauchy.
2. If $|x| < 1$ and $S_n = \sum_{k=0}^{n} x^k$, then show the sequence S_n is Cauchy.
3. If $|x| < 1$ and $S_n = \sum_{k=0}^{n} (-x)^k$, then show the sequence S_n is Cauchy.
4. If $S_n = \sum_{k=0}^{n} (-1)^k \frac{1}{k!}$, then show the sequence S_n is Cauchy. (Hint: Think ratio test.)
5. If $S_n = \sum_{k=0}^{n} (-1)^k \frac{1}{k^k}$, then show the sequence S_n is Cauchy. (Hint: Think n^{th} root test.)

9.6 Series

Galileo: Let us return to the topic of series by reminding you of what it means for a series to converge. The idea is to bring precision to the addition of an infinite number of terms.

Simplicio: Where are we going to use these ideas?

Galileo: Approximation theory is all about infinite sums. Taylor series and Fourier series are probably the most notable. We just want to make sure they make sense.

9.6.1 Series Facts

Virginia: As you mentioned earlier, we divide this definition into two pieces. The first part is the definition of partial sum.

Definition 9.6.1. *If $\sum_{k=0}^{\infty} x_k$ is an infinite series, then the sum $S_n = \sum_{k=0}^{n} x_k$ is called the n^{th} partial sum.*

Virginia: We now can define the sum of an infinite series to be the limit of the sequence of partial sums. Thus, the study of series simply reduces to the study of a special type of sequence.

Definition 9.6.2. *An infinite series $\sum_{k=0}^{\infty} x_k$ is said to converge to a number S, if the limit of the n^{th} partial sums converges to S. More precisely, $S = \sum_{k=0}^{\infty} x_k$ if and only if $\lim_{n\to\infty} S_n = S$, where $S_n = \sum_{k=0}^{n} x_k$.*

Galileo: Correct.

Virginia: Actually, if series are a subset of sequences, life should be a bit easier because you don't have to prove theorems twice. For example, we immediately have the Sum Theorem for Infinite Series.

Theorem 9.6.3 (The Sum Theorem for Infinite Series). *If $S = \sum_{k=0}^{\infty} x_k$ and $T = \sum_{k=0}^{\infty} y_k$, then $\sum_{k=0}^{\infty}(x_k + y_k) = \sum_{k=0}^{\infty} x_k + \sum_{k=0}^{\infty} y_k = S + T$.*

Proof. Since the limit of the sum equals the sum of the limits for sequences 9.3.1,

$$S + T = \lim_{n\to\infty} S_n + \lim_{n\to\infty} T_n = \lim_{n\to\infty} (S_n + T_n) = \sum_{k=0}^{\infty}(x_k + y_k).$$

□

Simplicio: We also can pull constants across the summation.

Theorem 9.6.4 (The Distributive Law for Series). *If $S = \sum_{k=0}^{\infty} x_k$ and C is a real numer, then $\sum_{k=0}^{\infty} Cx_k = C\sum_{k=0}^{\infty} x_k = CS$.*

Proof. This theorem follows immediately from the fact that we can pull constants across limits of sequences. 9.3.1. Namely,

$$\sum_{k=0}^{\infty} Cx_k = \lim_{n\to\infty}(\sum_{k=0}^{n} Cx_k) = \lim_{n\to\infty}(C\sum_{k=0}^{n} x_k) = C\lim_{n\to\infty}(\sum_{k=0}^{n} x_k) = C\sum_{k=0}^{\infty} x_k = CS.$$

□

Galileo: Very good observation.
Virginia: Don't forget uniqueness and squeezing.

Theorem 9.6.5 (Uniqueness for Infinite Series). *If $S_1 = \sum_{k=0}^{\infty} x_k$ and $S_2 = \sum_{k=0}^{\infty} x_k$, then $S_1 = S_2$.*

Proof. This theorem follows immediately from the Uniqueness Theorem for Sequences 9.3.4. □

Theorem 9.6.6 (The Squeezing Theorem for Series). *If $S = \sum_{k=0}^{\infty} x_k$, $T = \sum_{k=0}^{\infty} y_k$, and $x_k \le y_k$ for all $k = 0, 1, 2, \ldots, \infty$, then $S = \sum_{k=0}^{\infty} x_k \le \sum_{k=0}^{\infty} y_k = T$.*

Proof. If $S_n = \sum_{k=0}^{n} x_k$ and $T_n = \sum_{k=0}^{n} y_k$, then the assumption $x_k \le y_k$ implies that $S_n \le T_n$ for all n.

Thus, by the Squeezing Theorem for Sequences 9.3.5

$$S = \lim_{n\to\infty} S_n \le \lim_{n\to\infty} T_n = T.$$

□

Simplicio: How about an example?

Example 9.6.1. *Galileo: How about if we compute $\sum_{k=0}^{\infty}(2\frac{1}{3^k} + 7\frac{1}{5^k})$?*

Virginia: How about if we decompose the sum into:

$$\begin{aligned}\sum_{k=0}^{\infty}(2\frac{1}{3^k}+7\frac{1}{5^k}) &= \sum_{k=0}^{\infty}2\frac{1}{3^k}+\sum_{k=0}^{\infty}7\frac{1}{5^k}\\ &= 2\sum_{k=0}^{\infty}\frac{1}{3^k}+7\sum_{k=0}^{\infty}\frac{1}{5^k}\\ &= 2\frac{1}{1-\frac{1}{3}}+7\frac{1}{1-\frac{1}{5}} = 2\frac{3}{2}+7\frac{5}{4} = 3+\frac{35}{4} = \frac{47}{4}.\end{aligned}$$

Simplicio: That was easy. How about an example to illustrate the Squeezing Theorem for series?

Example 9.6.2. *Galileo: How about if we show the series* $\sum_{k=0}^{\infty}\frac{k}{k+1}\frac{1}{3^k}$ *converges?*

Virginia: Easy. All we have to do is notice that $\frac{k}{k+1}\frac{1}{3^k} \leq \frac{1}{3^k}$ *for all* $k = 0, 1, 2, \ldots$. *Since* $S_n = \sum_{k=0}^{n}\frac{k}{k+1}\frac{1}{3^k} \leq \sum_{k=0}^{n}\frac{1}{3^k} \leq \frac{1}{1-\frac{1}{3}} = \frac{3}{2}$, *the sequence of partial sums* $\{S_n\}_{n=0}^{\infty}$ *is bounded.*

Since each term $\frac{k}{k+1}\frac{1}{3^k}$ *is positive, the sequence* $\{S_n\}_{n=0}^{\infty}$ *is also increasing.*

Thus, the sequence $\{S_n\}_{n=0}^{\infty}$ *converges.*

9.6.2 Euler's Constant

Galileo: We now turn to the important constant e discovered by the Swiss mathematician and astronomer Leonhard Euler (1707-1783). Professor Euler was probably the most prolific mathematician of all time. He was amazingly productive. Any complete collection of his books is an incredible nuisance to the librarian in charge of finding shelf space.

Example 9.6.3. *We begin with a definition of the constant that bears his name.*

Definition 9.6.7 (Euler's Constant). $e = \sum_{k=0}^{\infty}\frac{1}{k!}$.

Simplicio: Even I remember that $e = 2.71828182845905$.

Virginia: How do you remember all those numbers?

Simplicio: Andrew Jackson (1767-1845) was elected president of the United States in 1828.

Galileo: But, does the infinite sum make any sense?

Theorem 9.6.8. *There is a constant e such that $e = \sum_{k=0}^{\infty} \frac{1}{k!}$.*

Proof. Virginia: Since $e = \sum_{k=0}^{\infty} \frac{1}{k!}. = \lim_{n\to\infty} S_n$, where $S_n = \sum_{k=0}^{n} \frac{1}{k!}$, all we have to do is show the sequence of partial sums $\{S_n\}_{n=1}^{\infty}$ is bounded and increasing.
Simplicio: But $S_{n+1} = S_n + \frac{1}{(n+1)!}$ so the sequence is increasing.
Virginia: Since $\frac{1}{k!} \leq \frac{1}{2^k}$ for all $k = 0, 1, 2, \ldots$, $\sum_{k=0}^{n} \frac{1}{k!} \leq 1 + \sum_{k=0}^{n} \frac{1}{2^k} \leq 1 + 2 = 3$, for all $n = 0, 1, 2, \ldots$, Since the sequence of partial sums $S_n = \sum_{k=0}^{n} \frac{1}{k!}$ is bounded and increasing, there is a real number e with the property that $e = \lim_{n\to\infty} S_n$. □

9.6.3 Convergence Tests for Series

Galileo: In the spirit of Professor Polya, let's take a second look at the argument that the number e is well defined. What do you observe about the series?
Virginia: Since the terms of the series are positive, the sequence of partial sums is increasing.
Simplicio: But that is obvious. The only hard part of the argument is to show these partial sums are bounded.
Galileo: You have just generalized our example into a theorem.

Theorem 9.6.9. *If $\sum_{k=0}^{\infty} a_k$ is a series with the property that $a_k \geq 0$ for all $k = 0, 1, \ldots,$ and the partial sums $S_n = \sum_{k=0}^{n} a_k$ are bounded, then the series converges. In particular, if $S_n \leq M$ for all n, then $S = \sum_{k=0}^{\infty} a_k \leq M$.*

Proof. Simplicio: Even I can see that this theorem is an obvious consequence of the fact the sequence of partial sums S_n is bounded and increasing. Thus, the series $\sum_{k=0}^{\infty} a_k$ converges. □

Galileo: Very good. Note that whenever we have identified a series $\sum_{k=0}^{\infty} a_k$ as convergent, we have observed that $\lim_{k\to\infty} a_k = 0$. Let's encapsulate this observation into a theorem.

Theorem 9.6.10. *If the series $\sum_{k=0}^{\infty} a_k$ is convergent, then $\lim_{n\to\infty} a_n = 0$.*

Proof. Virginia: But this fact is easy to prove. All we have to notice is that

$$\begin{aligned}\lim_{n\to\infty} a_n &= \lim_{n\to\infty}(\sum_{k=0}^{n} a_k - \sum_{k=0}^{n-1} a_k) \\ &= \lim_{n\to\infty}(S_n - S_{n-1}) = \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} = S - S = 0.\end{aligned}$$

□

Example 9.6.4. *Galileo: Before we move on, let's consider an example, which shows how this theorem can be applied. Consider the series* $\sum_{k=0}^{\infty}(-1)^k = 1 + (-1) + 1 + (-1) + \cdots +$ *What do you think this series should be?*

Simplicio: Since we can group the sum as

$$1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \cdots = 1 + (0) + (0) + (0) + \cdots = 1,$$

it looks to me like the series should equal 1.

Virginia: Since we can group the sum as

$$(1 + -1) + (1 + -1) + (1 + -1) + (1 + -1) + \cdots = 0 + (0) + (0) + (0) + \cdots = 0,$$

it looks to me like the series should equal 0.

Galileo: Mathematicians decided a while back that certain expressions of symbols should be classified as nonsense. Since the contrapositive of Theorem 9.6.10 states that if the sequence $\{a_k\}_{k=0}^{\infty}$ *does anything other than converge to zero, then the series* $\sum_{k=0}^{\infty} a_k$ *diverges.*

Virginia: In other words, the series is nonsense.

Galileo: Correct.

Simplicio: Wait a minute! I have a better theorem:

If $\lim_{n\to\infty} a_n = 0$, then the series $\sum_{k=0}^{\infty} a_k$ converges.

I am sure it is true.

Galileo: Whenever a mathematician proves a theorem, he/she immediately asks the question: Is the converse? Are you making a conjecture that the converse of Theorem 9.6.10 is true?

Simplicio: I guess so.

Example 9.6.5. *Galileo: How about if we sum the famous Harmonic Series given by the formula* $\sum_{k=1}^{\infty} \frac{1}{k}$*? If we sum the first few billion terms, the series seems to converge. In particular, consider the data in Table 9.1.*

N	Harmonic Sum
10	2.92896825396825
100	5.1873775176396
1,000	7.48547086055034
10,000	9.78760603604435
100,000	12.09014612986334
1,000,000	14.39272672286499
10,000,000	16.69531136585727
100,000,000	18.99789641385255
1,000,000,000	21.30048150234855
10,000,000,000	22.06477826202586
100,000,000,000	22.06477826202586

Table 9.1: The Sum of the Harmonic Series $\sum_{k=1}^{N} \frac{1}{k}$

Simplicio: Looks to me like we have convergence. The last two computations are identical.

Galileo: Sadly, while it looks like the series converges to a number a bit larger than 22.064778, our optimism is unjustified. Consider the following proposition.

Proposition 9.6.11 (The Harmonic Series Diverges). *If* $N = 2^n$*, then* $\sum_{k=1}^{N} \frac{1}{k} > \frac{n}{2}$*. Thus, the series* $\sum_{k=1}^{\infty} \frac{1}{k}$ *diverges.*

Proof. If $n = 1$, then $N = 2^1$ and $\sum_{k=1}^{N} \frac{1}{k} = 1 + \frac{1}{2} > \frac{1}{2}$.

If $n = 2$, then $N = 2^2$ and $\sum_{k=1}^{N} \frac{1}{k} = (1 + \frac{1}{2}) + (\frac{1}{3} + \frac{1}{4}) > \frac{1}{2} + \frac{1}{2} = 2\frac{1}{2}$.

If $n = 3$, then $N = 2^3$ and $\sum_{k=1}^{N} \frac{1}{k} = (1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}) + (\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}) > 2\frac{1}{2} + \frac{1}{2} = 3\frac{1}{2}$.

And so it goes. □

Simplicio: OOPS! I was wrong.

Galileo: No worries. Not all of our first thoughts are correct.

Simplicio: So if $n = 40$, then $N = 2^{40} \approx 1.0995 * 10^{12}$. Thus, the sum of the finite harmonic series is $\sum_{k=1}^{N} \frac{1}{k} > \frac{n}{2} = \frac{40}{2} = 20$. Hey, that's about right! Looks like my conjecture is out the window.

Virginia: I don't quite understand this example yet. What will happen if you just keep adding more numbers of the form $\frac{1}{k}$?

Galileo: If you are above the precision of the computer, you will simply be adding real numbers of the form $\frac{1}{k}$. Since k is "large," $\frac{1}{k} = 0$. In other words, the activity won't be very productive.

Example 9.6.6. *Now let's modify the definition of Euler's constant and ask the question: Does the series* $\sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$ *converge?*

Simplicio: Can we compare this series to the geometric series $\sum_{k=0}^{\infty} \frac{(-1)^k}{2^k}$*?*

Galileo: There are good ideas and bad ideas. Your idea does not work so well. Virginia?

Virginia: While the sequence of partial sums fail to be increasing for these series, they are still Cauchy. In particular, we can check this fact by following the steps in the usual program.

Step 1. The Challenge
Let $\epsilon > 0$ *be given.*

Step 2. The Choice
Choose N *so that* $\frac{1}{2^N} < \epsilon$.

Step 3. The Check

If $n \geq N$, then the difference

$$\begin{aligned}
|S_n - S_N| &= |\sum_{k=0}^{n} \frac{(-1)^k}{k!} - \sum_{k=0}^{N} \frac{(-1)^k}{k!}| \\
&= |\sum_{k=N+1}^{n} \frac{(-1)^k}{k!}| \\
&\leq \sum_{k=N+1}^{n} |\frac{(-1)^k}{k!}| \\
&= \sum_{k=N+1}^{n} \frac{1}{k!} \\
&\leq \sum_{k=N+1}^{n} \frac{1}{2^k} < \frac{1}{2^{N+1}} \frac{1}{1-\frac{1}{2}} = \frac{1}{2^{N+1}} 2 = \frac{1}{2^N} < \epsilon
\end{aligned}$$

Thus, we have shown the sequence of partial sums is Cauchy.

Galileo: Very good.

Galileo: In the spirit of Professor Polya, let's take a second look at this example and make a number of observations about this example.

1. The positive and negative signs don't make a difference.

2. The comparison with a known series, the geometric series, does make a difference.

We now generalize these examples and observations into a theorem.

Simplicio: Why didn't you compare the series $\sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$ with the series for Euler's constant $e = \sum_{k=0}^{\infty} \frac{1}{k!}$?

Galileo: Good observation. While we could have done that, I thought you would be more comfortable with the familiar geometric series. However, your observation is useful because it leads to a general theorem.

Theorem 9.6.12 (Absolute Convergence). *If the series $\sum_{k=0}^{\infty} |a_k|$ converges, then the series $\sum_{k=0}^{\infty} a_k$ converges. In particular, if $\sum_{k=0}^{\infty} |a_k| < \infty$, then $\sum_{k=0}^{\infty} a_k$ converges.*

Proof. Virginia: Using the previous example as a guide, we need only show the sequence of partial sums $\{S_n = \sum_{k=0}^{n} a_k\}_{n=0}^{\infty}$ is Cauchy.

Step 1. The Challenge

Let $\epsilon > 0$ be given.

Step 2. The Choice

Since we are assuming the series $\sum_{k=0}^{\infty} |a_k|$ converges, the partial sums $T_n = \sum_{k=0}^{n} |a_k|$ are Cauchy 9.5.2. Thus, we can find an integer N with the property that if $n \geq N$, then $|T_n - T_N| < \epsilon$.

Step 3. The Check

If $n \geq N$, then the difference

$$\begin{aligned} |S_n - S_N| &= |\sum_{k=0}^{n} a_k - \sum_{k=0}^{N} a_k| \\ &= |\sum_{k=N+1}^{n} a_k| \\ &\leq \sum_{k=N+1}^{n} |a_k| \\ &= |T_n - T_N| < \epsilon. \end{aligned}$$

Since the sequence of partial sums $\{S_n\}_{k=0}^{\infty}$ is Cauchy, we know by Theorem 9.5.3 that it converges. □

Simplicio: How about an example?

Example 9.6.7. *Galileo: How about the series* $\sum_{k=1}^{\infty} \frac{k-1}{k} \frac{(-1)^k}{k!}$*?*

Simplicio: Now that we have the Absolute Convergence Theorem 9.6.12 all we have to do is show the series $\sum_{k=1}^{\infty} |\frac{k-1}{k} \frac{(-1)^k}{k!}|$ *is bounded.*

However, by the Squeezing Theorem for Series 9.6.6 we simply note that

$$\sum_{k=1}^{\infty} |\frac{k-1}{k} \frac{(-1)^k}{k!}| = \sum_{k=1}^{\infty} \frac{k-1}{k} \frac{1}{k!} \leq \sum_{k=1}^{\infty} \frac{1}{k!} = e - 1 < \infty.$$

Thus, the series $\sum_{k=1}^{\infty} \frac{k-1}{k} \frac{(-1)^k}{k!}$ *converges.*

Simplicio: Actually, we showed more. Namely, we showed the series $\sum_{k=1}^{\infty} \frac{k-1}{k} \frac{1}{k!}$ *also converges.*

Galileo: Once again, we can encapsulate this special case as a new theorem called the Comparison Test 9.6.13.

Theorem 9.6.13 (Comparison Test). . *If a_k and b_k are real numbers for $k = 0, 1, 2, \ldots$, such that*

1. *$b_k \geq 0$ for $k = 0, 1, 2, \ldots$,*
2. *$|a_k| \leq b_k$ for $k = 0, 1, \ldots$, and*
3. *$\sum_{k=0}^{\infty} b_k \leq M < +\infty$,*

then the series $\sum_{k=0}^{\infty} a_k$ converges.

Proof. Simplicio: But even an engineer can now prove this theorem. By the the Squeezing Theorem 9.6.6 $\sum_{k=0}^{\infty} |a_k| \leq \sum_{k=0}^{\infty} b_k \leq M < +\infty$. By the Absolute Convergence Theorem 9.6.12, the series $\sum_{k=0}^{\infty} a_k$ converges.

□

Galileo: Very good. Note that the Absolute Convergence Theorem 9.6.12 and Comparison Test 9.6.13 inspire the following definition.

Definition 9.6.14 (Absolute Convergence). *If a series of real numbers $\sum_{k=0}^{\infty} a_k$ has the property that $\sum_{k=0}^{\infty} |a_k|$ converges, then the series converges Absolutely.*

As it turns out, whenever you successfully apply a comparison test, you will be able to declare your series converges absolutely. Most of your favorite tests will be comparison tests. What I have found through the ages is that students have a great preference for the Ratio Test 9.6.15. It is easy to understand and easy to apply. In fact, it is easy to prove because all you have to do is compare a given series with the appropriately chosen Geometric Series.

Corollary 9.6.15 (Ratio Test). *If $0 \leq r < 1$ and $|a_{k+1}| \leq r|a_k|$ for $k = 0, 1, 2, \ldots$, then the series $\sum_{k=0}^{\infty} a_k$ converges. In particular, $|\sum_{k=0}^{\infty} a_k| \leq \sum_{k=0}^{\infty} |a_k| \leq \frac{|a_0|}{1-r} < \infty$.*

If $r > 1$, $a_0 \neq 0$, and $|a_{k+1}| \geq r|a_k|$ for $k = 0, 1, 2, \ldots$, then the series $\sum_{k=0}^{\infty} a_k$ diverges.

Proof. Since

1. $|a_0| = |a_0| = r^0|a_0|$
2. $|a_1| \leq r|a_0| = r^1|a_0|$
3. $|a_2| \leq r|a_1| \leq r^2|a_0|$
4. $|a_3| \leq r|a_2| \leq r^3|a_0|$
5. $|a_4| \leq r|a_3| \leq r^4|a_0|$
6. $\vdots$
7. $|a_k| \leq r|a_{k-1}| \leq r^k|a_0|$

and $\sum_{k=0}^{\infty} |a_0|r^k = \frac{|a_0|}{1-r} < \infty$, the series $\sum_{k=0}^{\infty} a_k$ converges by the Comparison Test 9.6.13.

If $r > 1$ and $|a_{k+1}| \geq r|a_k|$ for $k = 0, 1, 2, \ldots$, then

1. $|a_0| = |a_0| = r^0|a_0|$
2. $|a_1| \geq r|a_0| = r^1|a_0|$
3. $|a_2| \geq r|a_1| \geq r^2|a_0|$
4. $|a_3| \geq r|a_2| \geq r^3|a_0|$
5. $|a_4| \geq r|a_3| \geq r^4|a_0|$
6. $\vdots$
7. $|a_k| \geq r|a_{k-1}| \geq r^k|a_0|$.

Thus, $\lim_{k\to\infty} |a_k| = +\infty$. By Theorem 9.6.10, the series $\sum_{k=0}^{\infty} a_k$ diverges. □

Virginia: So the Ratio Test begins and ends with the Geometric Series?

Galileo: Correct.

Simplicio: How about an example?

Example 9.6.8. *Galileo: How about if we show the series* $\sum_{k=0}^{\infty} \frac{k-1}{k}\frac{1}{2^k}$ *is convergent? Simplicio: No problem. All we have to do is observe that* $\frac{k-1}{k}\frac{1}{2^k} < \frac{1}{2^k}$ *for all* $k > 0$ *so that* $\sum_{k=0}^{\infty} \frac{k-1}{k}\frac{1}{2^k} \leq \sum_{k=0}^{\infty} \frac{1}{2^k} = 2 < \infty$.

Galileo: Very good.

Example 9.6.9. *Virginia: What if we modify the previous problem so it reads: Show the series* $\sum_{k=0}^{\infty} k\frac{1}{2^k}$ *is convergent?*
Galileo: I like this question because it forces us to rethink our choice for r. We also have a problem making the comparison work for the first few terms.
Virginia: How about if we choose the ratio r somewhere between 0 *and* 1? *Say,* $r = \frac{2}{3}$?
Simplicio: I see that we have a problem with the first few terms. $a_k = k\frac{1}{2^k} < \frac{2}{3}(k-1)\frac{1}{2^{k-1}} = a_{k-1}$. *For example, if we compute the fraction* $\frac{a_k}{a_{k-1}} = \frac{k\frac{1}{2^k}}{(k-1)\frac{1}{2^{k-1}}}$, *then we find that*

1. *If* $k = 1$, *then* $\frac{a_k}{a_{k-1}} = \frac{a_1}{a_0} = \frac{1\frac{1}{2^1}}{0\frac{1}{2^0}} = +\infty$.
2. *If* $k = 2$, *then* $\frac{a_k}{a_{k-1}} = \frac{a_2}{a_1} = \frac{2\frac{1}{2^2}}{1\frac{1}{2^1}} = 1$.
3. *If* $k = 3$, *then* $\frac{a_k}{a_{k-1}} = \frac{a_3}{a_2} = \frac{3\frac{1}{2^3}}{2\frac{1}{2^2}} = \frac{3}{4}$.
4. *If* $k = 4$, *then* $\frac{a_k}{a_{k-1}} = \frac{a_4}{a_3} = \frac{4\frac{1}{2^4}}{3\frac{1}{2^3}} = \frac{2}{3}$.

Virginia: But obviously, if $k \geq 4$, *then* $0 \leq \frac{a_k}{a_{k-1}} \leq \frac{2}{3}$. *Thus, after the first four terms of the series, our sum is dominated by the series*

$$\sum_{k=4}^{\infty}(\frac{2}{3})^k = (\frac{2}{3})^4\sum_{k=0}^{\infty}(\frac{2}{3})^k = (\frac{2}{3})^4\frac{1}{1-\frac{2}{3}} = \frac{16}{81} * \frac{3}{2} = \frac{24}{81}.$$

Thus, the series converges. Note that we shifted the indices in the summation by 4.

Galileo: Once again, in the spirit of Professor Polya let's convert this example into a theorem.

Theorem 9.6.16 (Ratio Test 2). *If a series* $\sum_{k=0}^{\infty} a_k$ *has the property that* $\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = L < 1$, *then the series converges. Moreover, if r is any real number*

strictly between L and 1 (i.e. $0 \leq L < r < 1$), then there is a constant $K > 0$ so the series is dominated by the series $K \sum_{k=0}^{\infty} r^k = \frac{K}{1-r}$. If $\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = L > 1$, then the series $\sum_{k=0}^{\infty} a_k$ diverges.

Proof. Virginia: Since the open interval $(-r, r)$ contains the limit L, all we have to do is find an integer $N > 0$ with the property that if $n \geq N$,then $\frac{|a_{n+1}|}{|a_n|} \in (-r, r)$.

The argument now repeats the exact same pattern discussed in the first Ratio Test 9.6.15. The only difference is that we begin our comparisons farther out in the series.

1. $|a_{N+0}| = |a_N| = r^0|a_N|$
2. $|a_{N+1}| \leq r|a_N| = r^1|a_N|$
3. $|a_{N+2}| \leq r|a_{N+1}| \leq r^2|a_N|$
4. $|a_{N+3}| \leq r|a_{N+2}| \leq r^3|a_N|$
5. $|a_{N+4}| \leq r|a_{N+3}| \leq r^4|a_N|$
6. $\vdots$
7. $|a_{N+k}| \leq r|a_{N+k-1}| \leq r^k|a_N|$ or (substituting $n = N + k$)
   $|a_n| \leq r|a_{n-1}| \leq r^{n-N}|a_N|$.

Thus,

$$\sum_{k=0}^{\infty} |a_{N+k}| \leq \sum_{k=0}^{\infty} r^k|a_N| = |a_N| \sum_{k=0}^{\infty} r^k = |a_N| \frac{1}{1-r}.$$

Simplicio: So the secret constant K is equal to $|a_N|$?

Galileo: Almost, but don't forget the terms a_k before a_N. If they are larger than a_N, then K wil have to be increased so that the inequality $|a_k| \leq r^k K$ holds for all $k = 0, 1, 2, \ldots$ While the constant K might have to be adjusted, it is the "tail" of the series (i.e. the terms out "near" ∞) that determine convergence. □

Simplicio: How about if we compute one easy example to show how to apply this second Ratio Test?

Example 9.6.10. *Galileo: Moments ago we showed the series* $\sum_{k=0}^{\infty} \frac{k}{2^k}$ *converges. Using Ratio Test 2, all we have to do is compute the limit*

$$L = \lim_{k\to\infty} \frac{a_{k+1}}{a_k} = \lim_{k\to\infty} \frac{\frac{(k+1)}{2^{k+1}}}{\frac{k}{2^k}} = \lim_{k\to\infty} \frac{1}{2}\frac{(k+1)}{k} = \frac{1}{2}\lim_{k\to\infty} \frac{(k+1)}{k} = \frac{1}{2} < 1.$$

Since $L < 1$, *the series converges.*

Example 9.6.11. *Galileo: How about if we compute one more example illustrating how the Geometric Series can be used to show convergence? Namely, let's show that the series* $\sum_{k=1}^{\infty} \frac{1}{k^k}$ *converges.*
Virginia: This problem is easy because $\frac{1}{k^k} \leq \frac{1}{2^k}$ *for all* $k \geq 2$.
Galileo: True, but I don't want to do it that way. Instead, I want to compute the k^{th} *root of* $\frac{1}{k^k}$ *and notice that* $\sqrt[k]{\frac{1}{k^k}} = \frac{1}{k} \leq \frac{1}{2}$ *for all* $k \geq 2$. *Thus, computing the* k^{th} *power of both sides of this inequality we see that* $\frac{1}{k^k} \leq \frac{1}{2^k}$ *for* $k \geq 2$. *Thus,* $\sum_{k=2}^{\infty} \frac{1}{k^k} \leq \sum_{k=2}^{\infty} \frac{1}{2^k} = \frac{1}{2^2}\sum_{k=0}^{\infty} \frac{1}{2^k} = \frac{1}{2^2}\frac{1}{1-\frac{1}{2}} = \frac{1}{2}$.

Galileo: Now let's take a second look at this process and generalize it into a theorem.
Virginia: Professor Polya again?

Theorem 9.6.17 (n^{th} **Root Test).** *If a series* $\sum_{k=0}^{\infty} a_k$ *has the property that* $\sqrt[k]{|a_k|} \leq r < 1$ *for all* $k = 0, 1, 2, \ldots$, *then the series* $S = \sum_{k=0}^{\infty} a_k$ *converges. Moreover,* $|S| \leq \frac{1}{1-r} < \infty$.

Proof. Since $\sqrt[k]{|a_k|} \leq r < 1$ for all $k = 0, 1, 2, \ldots$, $|a_k| \leq r^k$ for all $k = 0, 1, 2, \ldots$. Thus,

1. $|a_0| \leq r^0$,
2. $|a_1| \leq r^1$,
3. $|a_2| \leq r^2$,
4. $|a_3| \leq r^3$,
5. $\vdots$

6. $|a_k| \leq r^k$.

Thus, $\sum_{k=0}^{\infty} |a_k| \leq \sum_{k=0}^{\infty} r^k = \frac{1}{1-r} < +\infty$. □

Simplicio: So, the secret to life is to compare with the Geometric Series!

Galileo: Not so fast.

Virginia: Actually, I am a bit worried. It seems to me that we have neglected a special case in Theorem 9.6.16 when the limit $L = \lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = 1$. I noticed that the series $\sum_{k=1}^{\infty} \frac{1}{k}$ has the property that $L = \lim_{k\to\infty} \frac{k}{k+1} = 1$. Yet, it diverges. Can we conclude that a series always diverges when this limit $L = 1$?

Example 9.6.12. *Galileo: The standard student mistake is to apply the Ratio Test to every series problem. For example, let's consider the series* $\sum_{k=1}^{\infty} \frac{1}{k^2}$. *Using Fourier Series we can show that* $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$.

Virginia: I will interested to learn why that is true.

Galileo: However, if we apply the Ratio Test, we see that

$$L = \lim_{k\to\infty} \frac{\frac{1}{(k+1)^2}}{\frac{1}{k^2}} = \lim_{k\to\infty} \frac{k^2}{(k+1)^2} = (\lim_{k\to\infty} \frac{k}{k+1})^2 = 1.$$

Simplicio: What does that tell us?

Virginia: Since we have observed there are both divergent and convergent series with the property that the limit $L = 1$, *the Ratio Test provides no useful information in this setting.*

Galileo: To be blunt, the Ratio Test cannot be applied.

Simplicio: So we need more techniques?

Galileo: Unfortunately, the answer to your question is yes.

Simplicio: So, math is not so easy after all.

Example 9.6.13. *Galileo: Let's now turn to a slightly more delicate series*

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + (-1)^{k+1}\frac{1}{k} + \ldots$$

converges. From a geometric point of view, this series must converge because for n equal to an even integer we see that

$$S_n = \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} = (1 - \frac{1}{2}) + (\frac{1}{3} - \frac{1}{4}) + \cdots + (\frac{1}{n-1} - \frac{1}{n}).$$

1. *the difference $1 - \frac{1}{2}$ equals the length of the interval $[\frac{1}{2}, 1]$,*

2. *the difference $\frac{1}{3} - \frac{1}{4}$ equals the length of the interval $[\frac{1}{4}, \frac{1}{3}]$,*

3. *the difference $\frac{1}{5} - \frac{1}{6}$ equals the length of the interval $[\frac{1}{6}, \frac{1}{5}]$,*

4. *the difference $\frac{1}{n-1} - \frac{1}{n}$ equals the length of the interval $[\frac{1}{n-1}, \frac{1}{n}]$,*

Since these intervals are pairwise disjoint, the partial sum is (at least for n even)

$$S_n = \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} = (1 - \frac{1}{2}) + (\frac{1}{3} - \frac{1}{4}) + \cdots + (\frac{1}{n-1} - \frac{1}{n}) \leq 1.$$

Thus, the sequence of partial sum $\{S_{2n}\}_{n=1}^{\infty}$ is bounded and increasing and thus converges to some number S. While the difference between S_n and S_{n+1} is $S_{n+1} - S_n = \frac{1}{n+1}$ and thus small, you have to be careful about about the difference $S_n - S_N$ because it is possible that the sum of many small differences could accumulate into a large one. The argument is a bit cleaner if we simply show the sequence is Cauchy.

Virginia: I can finish the argument.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose N to be an even number with the property that $N > \frac{2}{\epsilon}$.

Step 3. The Check:

If $n \geq N$, then (again since the intervals $[\frac{1}{k+2}, \frac{1}{k+1}]$ are disjoint)

$$\begin{aligned}
|S_n - S_N| &= |\sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} - \sum_{k=1}^{N} \frac{(-1)^{k+1}}{k}| \\
&= |\sum_{k=N+1}^{n} \frac{(-1)^{k+1}}{k}| \\
&= (\frac{1}{N+1} - \frac{1}{N+2}) + (\frac{1}{N+3} - \frac{1}{N+4}) + (\frac{1}{N+5} - \frac{1}{N+6}) \cdots \\
&\quad + (\frac{1}{n-1} - \frac{1}{n}) \\
&= (\frac{1}{N+1} - \frac{1}{N+2}) + (\frac{1}{N+2} - \frac{1}{N+4}) + (\frac{1}{N+4} - \frac{1}{N+6}) \cdots \\
&\quad + (\frac{1}{n-2} - \frac{1}{n}) \\
&= \frac{1}{N+1} - \frac{1}{n} \ < \ \frac{1}{N} < \ \epsilon.
\end{aligned}$$

Simplicio: What if the integer n is odd?

Virginia: No worries. You simply get an extra copy of the fraction $\frac{1}{n}$ hanging out on the end. That is why we chose $N > \frac{2}{\epsilon}$.

Simplicio: Is there any way to add up the terms of this series?

Galileo: Actually, we will see that ideas from Taylor Series can be used to show that $ln(2) = \ log_e(2) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$.

Galileo: In the spirit of Professor Polya, we would now like to generalize this example into a theorem.

Theorem 9.6.18 (Alternating Series Test). *If $\{a_k\}_{k=0}^{\infty}$ is a sequence of real numbers with the property that $a_k \geq a_{k+1} \geq 0$ and $\lim_{k\to\infty} a_k = 0$, then $\sum_{k=0}^{\infty} a_k(-1)^k$ converges to a number less than a_0.*

Proof. Virginia: I would like to work this problem. Following the outline provided by the example we just discussed, all we have to do is show the sequence of partial sums is Cauchy.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose N to be an even number with the property that $a_N < \frac{\epsilon}{2}$.

Step 3. The Check:

If n is an even integer and $n \geq N$, then

$$\begin{aligned}
|S_n - S_{N-1}| &= |\sum_{k=0}^{n} a_k - \sum_{k=0}^{N-1} a_k| \\
&= |\sum_{k=N}^{n} a_k| \\
&= (a_N - a_{N+1}) + (a_{N+2} - a_{N+3}) + (a_{N+4} - a_{N+5}) + \cdots + (a_n - a_{n-1}) \\
&< (a_N - a_{N+1}) + (a_{N+1} - a_{N+3}) + (a_{N+3} - a_{N+5}) + \cdots + (a_n - a_{n-1}) \\
&= a_N - a_{n-1} < a_N < \epsilon.
\end{aligned}$$

Simplicio: In other words, if you increase a_{N+2} to a_{N+1}, a_{N+4} to a_{N+3}, a_{N+6} to a_{N+5}, etc., then the sum $\sum_{k=N}^{n} a_k$ collapses to $a_N - a_{n-1}$.

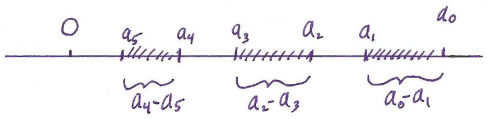


Figure 9.4: The Proof of the Alternating Series Test

Virginia: Just like the picture in Figure 9.4.

Simplicio: What if n is an odd integer?

Virginia: If n is an odd, then we have one more term to deal with. Namely, the sum $|\sum_{k=N}^{n} a_k| \leq a_N - a_{n-1} + a_{n+1} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$.

□

Simplicio: So, are we done yet?

Galileo: The fact that the series $\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$ converges, while the series $\sum_{k=1}^{\infty} \frac{1}{k}$ diverges leads to the concept of conditional convergence, which we now define.

Definition 9.6.19. *If the series $\sum_{k=0}^{\infty} a_k$ converges, while the series $\sum_{k=0}^{\infty} a_k$ diverges, then the series is called conditionally convergent.*

Example 9.6.14. *Galileo: Obviously the series $\sum_{k=0}^{\infty}(-1)^{k+1}\frac{1}{k}$ is conditionally convergent.*

Simplicio: What about our Geometric Series and our comparison tests?
Galileo: Think about it. Whenever you apply a comparison test to show a series converges, you ALWAYS prove absolute convergence. If a series converges absolutely, it NEVER converges conditionally.

9.6.4 Power Series

Galileo: We now turn to the topic of Power Series. While Isaac Newton considered every function to be a polynomial (finite or infinite) and while Power Series have a life of their own, we are not going to spend an excessive amount of time on this topic, Instead, our goal is to use this topic as a bridge between convergence tests for series and Taylor Series.
Simplicio: So, what is a Power Series?

Definition 9.6.20. *A Power Series is a series of the form $\sum_{k=0}^{\infty} a_k x^k$.*

Simplicio: So a Power Series is a finite or infinite polynomial.
Galileo: The next theorem is an immediate consequence of the Ratio Test 29.6.16.

Theorem 9.6.21. *If a series $\sum_{k=0}^{\infty} a_k$ has the property that $L = \lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|}$, then the Power Series $\sum_{k=0}^{\infty} a_k x^k$ converges for all $|x| < \frac{1}{L}$.*

Proof. If $|x| < \frac{1}{L}$, then

$$\lim_{k\to\infty} \frac{|a_{k+1}x^{k+1}|}{|a_k x^k|} = \lim_{k\to\infty} |x|\frac{|a_{k+1}|}{|a_k|} = |x| \lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = |x|L < \frac{1}{L}L < 1.$$

Thus, the series converges by the Ratio Test 2 9.6.16. □

Simplicio: How about some examples?

Example 9.6.15. *Galileo: You have five friends:*

1. $\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \ldots,$ *for* $|x| < 1,$
2. $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots,$ *for* $x \in \Re$
3. $\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^5}{5!} + \ldots,$ *for* $x \in \Re$
4. $\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots,$ *for* $x \in \Re$ *and*
5. $\ln(1-x) = log_e(1-x) = -\sum_{k=0}^{\infty} \frac{x^{k+1}}{k+1} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots,$ *for* $x \in [-1, 1).$

Simplicio: Where did these formulas come from? How am I going to be able to remember them?
Galileo: While we will wait until our discussion of Taylor Series to justify these series, they should be in your comfort zone.

1. The first equation is our old friend the Geometric Series.
2. The second is the exponential function, where you need only remember the $k!$ in the denominator of the fraction $\frac{x^k}{k!}$.
3. The third is the cosine function, which is almost the same as the exponential except for the alternating sign. If you remember that the function $\cos(x)$ is an *even* function (i.e. $f(x) = f(-x)$, for all $x \in \Re$), then only the terms x^k with even exponents will appear.
4. The fourth is the sine function, which is almost the same as the cosine. If you remember that the function $\sin(x)$ is an *odd* function (i.e. $f(x) = -f(-x)$, for all $x \in \Re$), then only the terms x^k with odd exponents will appear.
5. The function $log_e(x)$ is the integral of the Geometric Series.

Simplicio: I see the first example is our old friend the Geometric Series. The others examples look familiar from my study of Calculus. Where did those formulas come from?

Virginia: These formulas are all special cases of the Taylor Series formula:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k,$$

where x_0 is some point in $\Re$. Of course, we are assuming that the function $f(x)$ has infinitely many derivatives $f^{(k)}(x)$.

Galileo: Very good.

Simplicio: Could we work out the coefficients for one of these friends?

Example 9.6.16. *Galileo: If $f(x) = e^x$, then recall that $f'(x) = e^x$ for all $x \in \Re$. Thus, all the higher derivatives $f^{(k)}(x) = e^x$ for all $x \in \Re$. If we let $x_0 = 0$, then $f^{(k)}(0) = e^0 = 1$ for all $k = 0, 1, 2, \ldots$. Thus, the Taylor series is*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

Simplicio: That computation wasn't so bad.

Galileo: The justification for the equal sign takes more work, but we are going to dodge that issue for the moment. Let's drive the remainder of our discussion by asking three key questions.

1. Where and why does the series converge?

2. Can the series be differentiated term by term?

3. Can the series be integrated term by term?

For Power Series, the key to convergence is a comparison with a Geometric Series and the associated radius of convergence. In Examples 1 and 5, each series has a radius of convergence of $R = 1$. Examples 2, 3, and 4 each series has a radius of convergence of $R = +\infty$. These radii can be computed using Theorem 9.6.21.

We now present the formal (and slightly more general) definition of radius of convergence.

Definition 9.6.22. *If $x_0 \in \Re$, then radius of convergence of the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ is*

$$R = lub\{r \in \Re : \text{ if } |x-x_0| < r, \text{ then } \sum_{k=0}^{\infty} |a_k(x-x_0)^k| < \infty\}.$$

Galileo: For all the examples we will consider, $R = \frac{1}{L}$, where $L = \lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|}$. The *interval* of convergence is the set of all points $x\Re$ with the property that the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges.

Simplicio: But is this set necessarily an interval? Couldn't it be disconnected?

Galileo: No, by the Ratio Test/Geometric Series we know that if the series $\sum_{k=0}^{\infty} a_k r^k$ converges and $|x-x_0| < r$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges. Thus, the set of convergence points is always an interval of the form (x_0-R, x_0+R) plus either one or both endpoints x_0-R or x_0-R. Note that the interval of convergence for the function $\ln(x)$ is $[-1,1)$.

Simplicio: Why did you make the definition more general to include powers of $x-x_0$?

Galileo: When we discuss the rate of convergence of the Newton/Raphson algorithm, we will let $x_0 = r$, where $x = r$ is a root of the given function $f(x)$. As you will see, this slight change will appear in other applications as well.

Example 9.6.17. *Galileo: By substituting $y = 1-x$ in $\ln(1-x)$ we generate a second representation for $\ln(x)$ centered at $x_0 = 1$. In particular,*

$$\ln(x) = log_e(x) = -\sum_{k=0}^{\infty} \frac{(1-x)^{k+1}}{k+1} = \sum_{k=0}^{\infty} \frac{(x-1)^{k+1}}{k+1} \text{ for } x \in [0,2).$$

Note that the interval of convergence has shifted to the interval $[0,2)$.

Virginia: If we substitute $x = 0$ in the formula for $\ln(x)$ we get

$$\ln(2) = \sum_{k=0}^{\infty} \frac{(0-1)^{k+1}}{k+1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots.$$

Simplicio: Very interesting.

Galileo: But, we do need theorems and proofs to justify these formulas.

Virginia: Speaking of formulas, I noticed that if we compute the derivative of the series for e^x we simply get back e^x. Similarly, if we compute the derivatives of the terms of series for $\cos(x)$ and $\sin(x)$ we get the appropriate series for the derivatives. Is that always true?

Galileo: In fact, yes. As long as you stay inside the interval of convergence, everything is fine.

Simplicio: I noticed that if we integrate the Geometric Series, we produce the series for the log function.

$$\ln(1-x) = -\int_0^x \frac{1}{1-t}\, dt = -\sum_{k=0}^{\infty} \int_0^x t^k\, dt = -\sum_{k=0}^{\infty} \frac{x^{k+1}}{k+1}, .$$

What about integration?

Galileo: Again, as long as you stay inside the interval of convergence, you can integrate a series term by term. The next theorem summarizes these remarks.

Theorem 9.6.23 (Differentiation of Power Series). *If* $f(x) = \sum_{k=0}^{\infty} a_k x^k$ *for all* $x \in (-R, R)$, *then* $f'(x) = \sum_{k=1}^{\infty} k a_k x^{k-1}$ *for all* $x \in (-R, R)$.

Galileo: We have a similar result for integration.

Theorem 9.6.24 (Integration of Power Series). *If* $f(t) = \sum_{k=0}^{\infty} a_k t^k$ *for all* $t \in (-R, R)$ *and* $x \in (-R, R)$, *then* $F(x) = \int_0^x f(t)\, dt = \sum_{k=0}^{\infty} a_k \frac{x^{k+1}}{k+1}$.

Galileo: In every infinite sum of the form $f(x) = \sum_{k=0}^{\infty} a_k x^k$ the equal sign always means that for a fixed value of x, the sequence of partial sums $S_n = S_n(x) = \sum_{k=0}^{n} a_k x^k$ forms a Cauchy sequence. (The Comparison Test 9.6.13 guarantees our sequence of partial sums $\{S_n = S_n(x)\}_{n=0}^{\infty}$ will always be Cauchy.) Since a Cauchy sequence always converges to some quantity, there is no problem denoting the limit by the function $f(x) = \lim_{n\to\infty} S_n(x)$. A consequence of these last two theorems 9.6.23 9.6.24 is that a function of the form $f(x) = \sum_{k=0}^{\infty} a_k x^k$ can be differentiated and integrated with impunity.

Virginia: It all fits together.

9.6.5 Trigonometric/Fourier Series

Galileo: We now turn our discussion to Trigonometric Series of the form

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} \{a_k \cos(kx) + b_k \sin(kx)\}, \text{ for } x \in [-\pi, \pi].$$

Simplicio: Groan. More math?

Galileo: Maybe so, but a multitude of engineering and real-world applications are connected with functions of this type. In particular, any application associated with waves, vibrations, or periodic behavior can (and probably should) be modeled by functions of this form. Sound, light, radio waves, ocean waves, and planetary motion are only the beginning. Physicists love these functions. For the moment, however, let's limit our discussion to a few key questions.

1. Where and why does the series converge?
2. Can the series be differentiated term by term?
3. Can the series be integrated term by term?
4. How do we compute the coefficients a_k and b_k?
5. How can we use these series to compute certain infinite sums?

Simplicio: Sounds familiar.

Galileo: Before we get started though, let's make a couple of remarks about the big picture. First, we are now in the position of looking at the collection of all integrable functions on the interval $[-\pi, \pi]$. Since the sum of two integrable functions is integrable and the product of a scalar (i.e a real number) and an integrable function is integrable, it is easy to show that the collection of all integrable functions on $[-\pi, \pi]$ forms a vector space.

Simplicio: I am not sure I remember the definition of a vector space.

Virginia: A vector space is simply a collection of points with two operations: addition and scale multiplication. These two operations obey the usual associative, commu-

tative, and distributive laws of Algebra. The additive operation also has an identity and inverses.

Simplicio: But when I took Linear Algebra, our points were in the plane or three space. I never thought of $\cos(x)$ and $\sin(x)$ as vectors.

Galileo: Hermann Grassmann (1809-1877), Giuseppe Peano (1858-1932), and David Hilbert (1862-1943) changed the equation, In particular, they made the axioms of a vector space general enough to include functions as vectors?

Simplicio: So, what do I need to know?

Galileo: While we will give a more complete discussion of Linear Algebra in a day or so, the key idea hear is the notion of writing a vector as a linear combination of vectors residing in a given basis.

Simplicio: An example please.

Galileo: Since you like the plane let's start with the vectors

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Given a vector $\mathbf{v} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$, we can write $\mathbf{v} = 2\mathbf{e}_1 + 3\mathbf{e}_2$. Thus, we have written the vector $\mathbf{v}$ as a linear combination of the vectors in the basis $\mathbf{B} = \{\mathbf{e}_1, \mathbf{e}_2\}$.

Simplicio: No issue here.

Galileo: The polynomial $p_2(x) = 3x^2 + 5x + 7$ is a linear combination of vectors in the basis $\mathbf{B} = \{1, x, x^2\}$.

Simplicio: So, you are thinking of the functions $1, x, x^2$ and $p_2(x)$ as vectors?

Galileo: You can add them; you can multiply them by a constant; the associative, commutative, and distributive laws apply. Now consider the function $T_1(x) = 2 + 3\cos(x) + 5\sin(x)$.

Virginia: This time we have the function $T_1(x)$ written as a linear combination of vectors in the basis $\mathbf{B} = \{1, \cos(x), \sin(x)\}$.

Simplicio: I am not sure I like this discussion.

Galileo: As a software engineer, you do write your subroutines to be as general as possible. Don't you?

Simplicio: Sure. It is expected.

Galileo: Then you should appreciate the economy of having one concept cover such a broad collection of examples. Now let's think about the infinite. If a particular function $f(x)$ happens to have derivatives of all orders, the Taylor Series expansion shows that the function can be written as a linear combination of members from the basis

$$\mathbf{B}_P = \{1, x, x^2, x^3, \ldots, x^n, \ldots\}.$$

Simplicio: Except that we now have the small problem that the sum is infinite.

Virginia: Fortunately, through our understanding of the convergence of series, we know what the sum of an infinite numer of numbers means.

Galileo: The goal now is to change our representation form the basis $\mathbf{B}_P$ to a new basis

$$\mathbf{B}_T = \{1, \cos(x), \cos(2x), \cos(3x), \ldots, \sin(x), \sin(2x), \sin(3x), \ldots\}.$$

Simplicio: How about a couple of examples to get started?

Example 9.6.18. *Galileo: Here are a couple of series, where we have represented the polynomial functions $\frac{x}{2}$ and x^2 in terms of sines and cosines. Note that this strategy is the opposite of the strategy invoked for Taylor Series.*

1. $\frac{x}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} \sin(kx)$, *for* $x \in (-\pi, \pi)$,
2. $x^2 = \frac{\pi^2}{3} - 4 \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^2} \cos(kx)$, *for* $x \in [-\pi, \pi]$,
3. $|x| = \frac{\pi}{2} - \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \cos((2k-1)x)$, *for* $x \in [-\pi, \pi]$,

Simplicio: I hope there is a formula for computing the coefficients for these series.

Galileo: No worries. While we will eventually give you a tidy little formula, let's focus on the convergence, differentiation, and integration issues first. What do you notice?

Virginia: I notice with these examples that you gave an interval of convergence.

Galileo: Since the functions $\cos(x)$ and $\sin(x)$ are 2π periodic and x represents an angle (in radians of course), the interval of convergence will almost invariably be chosen to as $[-\pi, \pi]$ or $[0, 2\pi]$.

Simplicio: I notice that the function $\frac{x}{2}$ is odd and is written as a linear combination of the odd functions $\sin(kx)$, for $k = 1, 2, 3, \ldots$ A also that the functions x^2 and $|x|$ are even and can be written as a linear combination of the even functions $\cos(kx)$, for $k = 1, 2, 3, \ldots$

Galileo: In fact, you have noted a completely general property about Trigonometric functions.

Simplicio: I also noticed that we won't have to compute the radius of convergence for this type of series.

Galileo: Correct.

Virginia: What about convergence?

Galileo: With Trigonometric Series, convergence is a delicate issue. There is good news and bad news.

Simplicio: I vote to hear the good news first.

Galileo: OK, let's begin by looking at examples 2 and 3 above. What do you notice about the series

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^2} \quad \text{and} \quad \sum_{k=1}^{\infty} a_{2k-1} = \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2}?$$

Virginia: They both converge absolutely.

Galileo: Correct. So what does that tell you about the series

$$\sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^2} \cos(kx) \quad \text{and} \quad \frac{\pi^2}{3} - 4 \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^2} \cos(kx)?$$

Simplicio: Since $|cos(kx)| \leq 1$ for any k and all x, they both converge absolutely by the Comparison Test 9.6.13. In particular,

$$\sum_{k=1}^{\infty} |(-1)^{k+1} \frac{1}{k^2} \cos(kx)| \leq \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} < \infty.$$

Galileo: Correct. So, could someone please state the next theorem?

Virginia: I can.

Theorem 9.6.25. *If*

$$\sum_{k=1}^{\infty}(|a_k| + |b_k|) < +\infty,$$

then the series

$$\sum_{k=1}^{\infty}\{a_k \cos(kx) + b_k \sin(kx)\}$$

converges absolutely for all $x \in [-\pi, \pi]$.

Proof. Galileo: So, how about a proof?

Virginia: Easy.

If $x \in [-\pi, \pi]$, then

$$\sum_{k=1}^{\infty}\{|a_k \cos(kx) + b_k \sin(kx)|\} \leq \sum_{k=1}^{\infty}\{|a_k| + |b_k|\} < +\infty.$$

□

Galileo: Watching the human mind extrapolate general theorems from a few special cases is a wonderful thing. How about some more good news?

Simplicio: Good news is good.

Galileo: If we let $x = \pi$ in the equation

$$x^2 = \frac{\pi^2}{3} - 4\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k^2}\cos(kx),$$

then we see that

$$\pi^2 = \frac{\pi^2}{3} - 4\sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k^2}\cos(k\pi) = \frac{\pi^2}{3} - 4\sum_{k=1}^{\infty}(-1)^{2k+1}\frac{1}{k^2}.$$

Thus,

$$\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}.$$

Simplicio: Magic!!

Virginia: Its even a good way to compute π.

Galileo: Better than Archimedes' method for computing π..

Simplicio: With all this good news, what's the problem with these Trig Series?

Galileo: How about if we go back to equation 1? If $x = \pi$, then

$$\frac{\pi}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} \sin(k\pi) = 0 + 0 + 0 + \cdots = 0.$$

Simplicio: OOPS!

Virginia: Now I understand why you didn't include $\pm\pi$ in the interval of convergence.

Galileo: The news gets worse. What can you say about the convergence of the series

$$\sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} \sin(kx)?$$

Simplicio: Nothing.

Galileo: That's right.

Virginia: None of the Convergence Tests work. The Ratio, Root, and Comparison Tests can't be applied because the series $\sum_{k=1}^{\infty} \frac{1}{k}$ diverges.

Simplicio: What about the Alternating Series Test?

Virginia: Unfortunately, the sign of k^{th} term $a_k = (-1)^{k+1} \frac{1}{k} \sin(kx)$ of the sequence alternates so irregularly (almost randomly) that no pattern emerges. Thus, there is no hope for the Alternating Series Test.

Galileo: In fact, the argument that this series converges for $x \in (-\pi, \pi)$ is quite tricky.

Simplicio: I don't know if I can stand any more of this good news.

Galileo: The prrof will be left for another day.

Simplicio: Sounds like good news to me.

Galileo: Quickly now. I am running out of time. Lets finish with an observation about differentiation and integration. Note that if we differentiate Equation 2, we arrive at Equation 1.

Simplicio: And if we integrate Equation 1, we get Equation 2. What's the big deal? This technique worked fine for Taylor.

Galileo: Equation 2 has excellent convergence properties. Equation 1 has poor convergence properties. Every time you differentiate a function of the form

$$t_k(x) = a_k \cos(kx) + b_k \sin(kx),$$

you find that

$$t_k'(x) = -ka_k \sin(kx) + kb_k \cos(kx).$$

Every time you integrate a function of the form

$$t_k(x) = a_k \cos(kx) + b_k \sin(kx),$$

you find that

$$\int t_k(x)\ dx = \frac{1}{k} a_k \sin(kx) - \frac{1}{k} b_k \cos(kx) + C.$$

The factor k produced by differentiation retards convergence. The factor $\frac{1}{k}$ produced by integration improves convergence. The bottom line is that integration is good while differentiation is dangerous.

Virginia: Wait a minute! I see a problem if the constant $C \neq 0$. For example, if $C = \frac{1}{2}$, then if we integrate a second time then we will have that unhappy series for $\frac{x}{2}$ appearing. I anticipate the formulas becoming more complicated and the convergence getting worse.

Example 9.6.19. *Galileo: In fact, you are correct. While Mathematicians lust for tidy little formulas, Mother Nature does not always cooperate. Here are a couple more examples:*

1. $\frac{\pi^2 x - x^3}{12} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^3} \sin(kx), \quad x \in [-\pi, \pi]$
2. $\frac{x^3 - 3\pi x^2 + 2\pi^2 x}{12} = \sum_{k=1}^{\infty} \frac{1}{k^3} \sin(kx), \quad x \in [0, 2\pi]$

Simplicio: How about a quick hint at an application before we leave?

Galileo: The First Harmonic (or Fundamental Overtone) of the series is the term $a_1 \cos(x) + b_1 \sin(x)$. The Second Harmonic is given by $a_2 \cos(2x) + b_2 \sin(2x)$. These two harmonics are important in speech recognition, filtering, and a host of other applications. In signal compression (e.g. JPEG), radio, and television the key idea is to filter out the frequency terms $a_k \cos(kx) + b_k \sin(kx)$, where k is large.

Simplicio: How do you do that?

Galileo: If you compute the Fourier Transform (i.e. compute the a_k and b_k terms), delete the high frequency components, and then compute the inverse Fourier Transform, then this new signal is the filtered version of the old.

Simplicio: By the way, you promised to give us the formulas for the Fourier Transform.

Galileo: OK, here are the formulas for the coefficients.

Theorem 9.6.26 (Fourier Coefficients). *If $f(x) : [-\pi, \pi] \to \Re$ is continuous, then*

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx)\ dx \text{ for } k = 0, 1, 2, 3, \ldots$$
$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx)\ dx \text{ for } k = 1, 2, 3, \ldots$$

Simplicio: But where did these formulas come from?

Galileo: Pythagoras.

Simplicio: Pythagoras? Surely, you are joking, Professor Galileo. What did Pythagoras know about Trigonometric Series?

Galileo: We will explain. First, consider the following proposition, which effectively states that that the functions $\cos(kx)$ and $\sin(kx)$ are orthogonal. This proposition will get us half way to Pythagoras.

Proposition 9.6.27 (Orthogonality of Cos(x) and Sin(x)). *If m and n are positive integers, then*

1. $\int_{-\pi}^{\pi} \cos(mx)\ dx = 0.$
2. $\int_{-\pi}^{\pi} \sin(nx)\ dx = 0.$
3. $\int_{-\pi}^{\pi} \cos(mx)\sin(nx)\ dx = 0.$
4. *If $m \neq n$, then* $\int_{-\pi}^{\pi} \cos(mx)\cos(nx)\ dx = 0.$
5. *If $m \neq n$, then* $\int_{-\pi}^{\pi} \sin(mx)\sin(nx)\ dx = 0.$

Proof. Galileo: What about proofs?

1. $\int_{-\pi}^{\pi} \cos(mx)\ dx = 0.$

Simplicio: This integral is zero because when we draw the graph of the function $y = \cos(x)$, it is obvious that the area under the curve is zero on both of the intervals $[-\pi, 0]$ and $[0, \pi]$. If m is a positive integer, then the function $\cos(mx)$ is the same as $\cos(x)$ except that it goes up and down m times.

Virginia: You can also apply the Fundamental Theorem of Calculus 11.7.3 to observe that $\int_{-\pi}^{\pi} \cos(mx)\, dx = \frac{\sin(mx)}{m}|_{x=-\pi}^{\pi} = 0 - 0 = 0$.

Simplicio: The Fundamental Theorem of Calculus works too.

2. $\int_{-\pi}^{\pi} \sin(nx)\, dx = 0$.

Simplicio: This integral is zero because when we draw the graph of the function $y = \sin(nx)$ is an *odd* function on $[-\pi, \pi]$.

Virginia: The Fundamental Theorem of Calculus also works.

3. $\int_{-\pi}^{\pi} \cos(mx)\sin(nx)\, dx = 0$.

Simplicio: Since the function $y = \cos(mx)$ is *even* and the function $y = \sin(nx)$ is odd, the product is odd. Thus, integral is zero.

4. If $m \neq n$, then $\int_{-\pi}^{\pi} \cos(mx)\cos(nx)\, dx = 0$.

Simplicio: I don't see how to prove this fact.

Virginia: Neither do I.

Galileo: A little trigonometry goes a long way here. Recall your sum formulas for $\cos(x)$ and observe.

1. $\cos(A - B) = \cos(A)\cos(B) + \sin(A)\sin(B)$
2. $\cos(A + B) = \cos(A)\cos(B) - \sin(A)\sin(B)$
3. $\cos(A - B) + \cos(A + B) = 2\cos(A)\cos(B)$

Note that the third equation is the sum of the first two. Thus,

$$\cos(A)\cos(B) = \frac{1}{2}\{\cos(A - B) + \cos(A + B)\}.$$

Virginia: I see how to finish the argument. All we have to do is let $A = mx$ and

$b = nx$ and substitute into the integral. Thus,

$$\begin{aligned}\int_{-\pi}^{\pi} \cos(mx)\cos(nx)\ dx &= \int_{-\pi}^{\pi} \frac{1}{2}\{\cos(mx-nx)+\cos(mx+nx)\}\ dx \\ &= \int_{-\pi}^{\pi} \frac{1}{2}\{\cos((m-n)x)+\cos((m+n)x)\}\ dx \\ &= \frac{1}{2}\int_{-\pi}^{\pi} \cos((m-n)x)\ dx + \frac{1}{2}\int_{-\pi}^{\pi} \cos((m+n)x)\ dx \\ &= 0+0=0.\end{aligned}$$

Simplicio: Looks like we used Fact 1 twice to get the last two zeros.

5. If $m \neq n$, then $\int_{-\pi}^{\pi} \sin(mx)\sin(nx)\ dx = 0$.

Simplicio: Once again, I don't see how to prove this fact.

Virginia: I think I do. All we have to do is subtract the equations we had before. In particular,

1. $\cos(A-B) = \cos(A)\cos(B)+\sin(A)\sin(B)$
2. $\cos(A+B) = \cos(A)\cos(B)-\sin(A)\sin(B)$
3. $\cos(A-B)-\cos(A+B) = 2\sin(A)\sin(B)$.

Note that the third equation is equation 2 subtracted from eqation 1. Thus,

$$\sin(A)\sin(B) = \frac{1}{2}\{\cos(A-B)-\cos(A+B)\}.$$

The rest of the argument is the same as before because

$$\begin{aligned}\int_{-\pi}^{\pi} \sin(mx)\sin(nx)\ dx &= \int_{-\pi}^{\pi} \frac{1}{2}\{\cos((m-n)x)-\cos((m+n)x)\}\ dx \\ &= \frac{1}{2}\int_{-\pi}^{\pi} \cos((m-n)x)\ dx - \frac{1}{2}\int_{-\pi}^{\pi} \cos((m+n)x)\ dx \\ &= 0+0=0.\end{aligned}$$

□

Simplicio: While that proposition was a bit long, it really was quite understandable because it only require you know basic facts from Trigonometry and Calculus.

Galileo: The next proposition provides us with the lengths of the basis vectors $1, \cos(nx), \sin(nx)$.

Proposition 9.6.28 (Fourier Equal Lengths Formulas for Cos(x) and Sin(x)). *If n is a positive integer, then*

1. $\int_{-\pi}^{\pi} 1 \, dx = 2\pi$,
2. $\int_{-\pi}^{\pi} \cos^2(nx) \, dx = \pi$,
3. $\int_{-\pi}^{\pi} \sin^2(nx) \, dx = \pi$.

Proof. Simplicio: What Trig fact do we need this time?
Galileo: While the first integral is easy, the other two rely on the half angle formulas relating the square of the functions $\cos(x)$ and $\sin(x)$ and $\cos(2x)$. In particular,

1. $\cos^2(x) = \frac{1+ \cos(2x)}{2}$ and
2. $\sin^2(x) = \frac{1- \cos(2x)}{2}$.

Virginia: Thus,

$$\int_{-\pi}^{\pi} \cos^2(nx) \, dx = \int_{-\pi}^{\pi} \frac{1+ \cos(2x)}{2} \, dx = \int_{-\pi}^{\pi} \frac{1}{2} \, dx + \int_{-\pi}^{\pi} \frac{\cos(2x)}{2} \, dx = \pi + 0 = \pi.$$

Simplicio: And,

$$\int_{-\pi}^{\pi} \sin^2(nx) \, dx = \int_{-\pi}^{\pi} \frac{1- \cos(2x)}{2} \, dx = \int_{-\pi}^{\pi} \frac{1}{2} \, dx - \int_{-\pi}^{\pi} \frac{\cos(2x)}{2} \, dx = \pi + 0 = \pi.$$

□

Galileo: Now that we have discussed the Orthogonality and Equal Lengths Propositions 9.6.27, 9.6.28, we are ready to prove the Fourier Coefficients Formula 9.6.26.

Proof. Galileo: While the general proof of the Fourier Coefficients Theorem 9.6.26 is difficult and requires a deep understanding of integration theory, we are now ready to prove it for the finite dimensinal case. To keep the subscripts and notation out of the discussion, let's consider the special case when $f(x) = \frac{a_0}{2} + a_1 \mathbb{Q}s(x) + a_2 \cos(2x) +$

$b_1 \sin(x) + b_2 \sin(2x)$. How about if we show you how to compute the formula for the coefficient a_2?

Simplicio: Simple is good.

Galileo: Step 1. Multiply both sides of the equation by the function $\cos(2x)$.

When we do this, we find that

$$\begin{aligned} f(x)\cos(2x) =& (\frac{a_0}{2} + a_1\cos(x) + a_2\cos(2x) + b_1\sin(x) + b_2\sin(2x)\ )\ \cos(2x) \\ =& \frac{a_0}{2}\cos(2x) + a_1\cos(x)\cos(2x) + a_2\cos(2x)\cos(2x) \\ &+\ b_1\sin(x)\cos(2x) + b_2\sin(2x)\cos(2x). \end{aligned}$$

Step 2. Integrate both sides of the equation.

When we do this, we find by the Orthogonality Property (Proposition 9.6.27) and the Equal Lengths Property (Proposition 9.6.28)

$$\begin{aligned} \int_{-\pi}^{\pi} f(x)\cos(2x)\ dx =& \int_{-\pi}^{\pi} \frac{a_0}{2}\cos(2x)\ dx + \int_{-\pi}^{\pi} a_1\cos(x)\cos(2x)\ dx \\ &+\ \int_{-\pi}^{\pi} a_2\cos(2x)\cos(2x)\ dx + \int_{-\pi}^{\pi} b_1\sin(x)\cos(2x)\ dx \\ &+\ \int_{-\pi}^{\pi} b_2\sin(2x)\cos(2x)\ dx \\ =& \ 0 + 0 + \int_{-\pi}^{\pi} a_2\cos(2x)\cos(2x)\ dx + 0 + 0 \\ =& \ a_2 \int_{-\pi}^{\pi} \cos(2x)\cos(2x)\ dx = a_2\pi. \end{aligned}$$

Thus,

$$a_2 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(2x)\ dx.$$

□

Simplicio: How about an example?

Example 9.6.20. *Galileo: If $f(x) = 1$ for $x \in [-\pi, \pi]$, then $a_0 = 2$ and $a_k = b_k = 0$ for all $k = 1, 2, \ldots$.*

Simplicio: That example was too easy. How about a more challenging one?

Example 9.6.21. *Galileo: If*

$$f(x) = \begin{cases} 1, & x \in [-\pi, 0] \\ -1, & x \in [0, \pi] \end{cases}$$

Virginia: Since the function $f(x)$ is odd, we know that $a_k = 0$ for all $k = 0, 1, 2, 3, \ldots$

Simplicio: On the other hand, since $f(x)$ is odd, the function $f(x)\sin(kx)$ is even. Thus,

$$\begin{aligned} b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)\ dx &= \frac{2}{\pi}\int_0^{\pi} sin(kx)\ dx \\ &= \frac{2}{\pi}\frac{-\cos(kx)}{k}\Big|_{x=0}^{\pi} \\ &= -\frac{2}{\pi}(\cos(k\pi) - 1) \\ &= -\frac{2}{\pi}\frac{(-1)^k - 1}{k}. \end{aligned}$$

In particular,

$$b_k = \begin{cases} \frac{4}{k\pi} & \text{if } k = 1, 3, 5, \ldots \\ 0 & \text{if } k = 2, 4, 6, \ldots \end{cases}$$

and

$$f(x) = \frac{4}{\pi}(\sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \ldots)$$

Example 9.6.22. *Galileo: If $f(x) = x$ for $x \in [-\pi, \pi]$, then $f(x)$ is an odd function. Thus, the function $f(x)\cos(kx) = x\cos(kx)$ is an odd function for all k, which implies $a_k = 0$, for all $k = 0, 1, 2, \ldots$. Since the function $x\sin(kx)$ is the product of two functions so you have to integrate by parts. While not a bad exercise for you, the antiderivative is*

$$\int x\sin(kx)\ dx = -x\frac{\cos(kx)}{k} + \frac{\sin(kx)}{k^2}.$$

Since the function $f(x)\sin(kx) = x\sin(kx)$ is the product of two odd functions, it

is even. Thus,

$$\begin{aligned} b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)\ dx &= \frac{2}{\pi}\int_0^{\pi} xsin(kx)\ dx \\ &= \frac{2}{\pi}(-x\frac{\cos(kx)}{k} + \frac{\sin(kx)}{k^2})|_{x=0}^{\pi} \\ &= \frac{2}{\pi} - \pi\frac{\cos(k\pi)}{k} - -0\frac{\cos(0)}{k} \\ &= (-1)^{k+1}\frac{2}{k}. \end{aligned}$$

Simplicio: Actually, your answer agrees with the formula you posted at the beginning of the discussion.

Exercise Set 9.6.

Exercises on Convergence of Series

1. Compute: $\sum_{k=0}^{\infty}(3\frac{1}{5^k} + 2\frac{1}{7^k})$.
2. Show the series $\sum_{k=1}^{\infty}\frac{k+1}{k}\frac{1}{5^k}$ converges.
3. Show the series $\sum_{k=1}^{\infty}\frac{(-1)^k}{k^k}$ converges.
4. Show the series $\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}$ converges.
5. Show: If the series $\sum_{k=1}^{\infty} a_k$ diverges and $a_k \geq 0$, then $\sum_{k=1}^{\infty}\frac{a_k}{1+a_k}$ diverges.

Exercises on Power/Taylor Series

1. Determine the interval of convergence of the series $\sum_{k=0}^{\infty}\frac{x^k}{k!}$.
2. Determine the interval of convergence of the series $\sum_{k=0}^{\infty} k^2x^k$.
3. Determine the interval of convergence of the series $\sum_{k=0}^{\infty}\frac{2^k}{k!}(x-3)^k$.
4. Determine the interval of convergence of the series $\sum_{k=0}^{\infty}\frac{k^3}{5^k}(x-7)^k$.
5. Determine the interval of convergence of the series $\sum_{k=0}^{\infty}\frac{k^3}{5^k}(x-7)^k$.

Exercises on Trigonometric/Fourier Series

1. Use the Fourier Coefficient Theorem to show:

$$\frac{x}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} \sin(kx), \text{ for } x \in (-\pi, \pi),$$

2. Use the Fourier Coefficient Theorem to show:

$$x^2 = \frac{\pi^2}{3} - 4 \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^2} \cos(kx), \text{ for } x \in [-\pi, \pi].$$

3. Use the Fourier Coefficient Theorem to show:

$$|x| = \frac{\pi}{2} - \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \cos((2k-1)x), \text{ for } x \in [-\pi, \pi].$$

4. Show:

$$\sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k^2} = \frac{\pi^2}{12}.$$

5. Show:

$$\sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} = \frac{\pi^2}{8}.$$

9.7 Limits of Functions

Galileo: We turn now to the topic of the limit of a function. I am sure you studied this topic in your Calculus courses.

Simplicio: It has been a long time since I took Calculus. Much knowledge has since evaporated. So where are we headed?

Galileo: The first theorem we will discuss is the Mean Value Theorem, which contains the idea that a function cannot grow faster than the maximum of its first derivative. The second key theorem is Taylor's Theorem, which basically states that a smooth function can be approximated by a polynomial.

Simplicio: If we are interested in sequences and data, why should we have to discuss functions?

Galileo: For the Archimedes/Heron algorithm, an understanding of the function $T(x) = x - \frac{x^2 - K}{2x}$ becomes central. Since an easy calculation shows that $|T'(x)| \leq \frac{1}{2}$ for

all $x \geq \sqrt{K}$, we will be able to conclude that the difference between the n^{th} approximation x_n and the answer $\sqrt{K}$ drops by 50% for each iteration. Such a convergence rate is known as linear (or first order) convergence. These ideas are completely general and apply to a wide range of problems including cube roots and beyond.

Simplicio: A 50% improvement at each iterations sounds good.

Galileo: As you will see, we are actually doing better than 50%. Taylor's Theorem will be the key to understanding why this algorithm converges so rapidly. In fact, of all the theorems you visited in Calculus, Taylor's Theorem is probably the most important for numerical computations. This theorem allows us to compute first and second derivatives numerically. Thus, many differential equations and partial differential equations can be solved numerically including heat transfer, fluid flow, airfoil design, electromagnetism, and weather modeling. The basic techniques of signal and image processing also involve these methods. In other words, the applications are everywhere.

Simplicio: I like these applications.

Galileo: Unfortunately, before we can even think about modeling a real-world problem, we have to develop the requisite language. Since the Intermediate Value Theorem, the Mean Value Theorem, and Taylor's Theorem have hypotheses where functions are assumed continuous or differentiable, we begin our discussion with the definition of the limit of a function. We begin our discussion with the definition of a limit of a function.

Definition 9.7.1 (Limit of a Function). *If X is an interval and $f(x) : X \to \Re$, then $lim_{x \to a} f(x) = L$, if for every $\epsilon > 0$, there is a $\delta > 0$ with the property that if $x \in X, |x - a| < \delta$, and $x \neq a$, then $|f(x) - L| < \epsilon$.*

Simplicio: Brutal. For sequences we had one Greek letter, now we are doubly blessed. I am confused.

Galileo: True, but the real problem is that the definition is backwards. While the function $f(x)$ assigns a point x in the domain to a point $f(x)$ in the range, the

tolerance $\epsilon > 0$ is associated with a distance in the range of $f(x)$, while the $\delta > 0$ measures a distance in the domain of $f(x)$. The ϵ appears first, while the δ is second.
Virginia: Hey, this definition is not so bad. In fact, it is almost the same as the definition for the limit of a sequence. The ϵ functions exactly as it did before, while the integer N is replaced by the quantity δ.
Galileo: In other words, a given an accuracy between $f(x)$ and L can be assured if a given precision between x and a is required.
Simplicio: OK, but why do you have that little condition that $x \neq a$?
Galileo: Because Calculus is the study of being close. For example, if we compute the derivative of the function $f(x) = x^2$ at the point $x = 2$, then we must investigate the values of the difference quotient $DQ(x) = \frac{x^2-4}{x-2}$ close to (but not at) the number 2. If we are careless and substitute $x = 2$ into this function, we get $DQ(x) = \frac{x^2-4}{x-2} = \frac{0}{0}$. Since division by zero is always evil, we must avoid that "bad" point $x = 2$. How about if we use the definition to show that $lim_{x\to 2} DQ(x) = 4$?
Virginia: We simply follow the same "Challenge, Choice, and Check" process we did for sequences.

Example 9.7.1. *Using the DEFINITION of limit show:* $lim_{x\to 2} \frac{x^2-4}{x-2} = 4$.

Step 1. The Challenge:

Let $\epsilon > 0$ *be given.*

Step 2. The Choice of δ :

While I am not exactly sure how to choose δ*, I will make the guess that* $\delta = \epsilon$*. If we are wrong, we will make adjustments and do it again.*

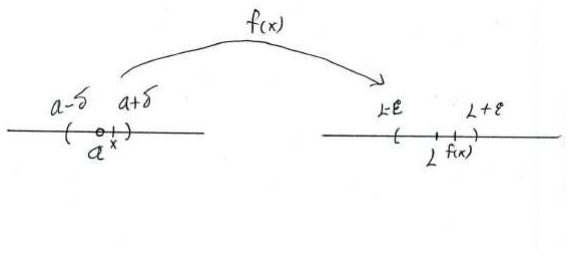


Figure 9.5: The Definition of a Limit

Step 3. The Check that δ works:

If we can show the absolute value of the difference between $DQ(x) = \frac{x^2-4}{x-2}$ and 4 is less than ϵ, then we are done. However, if we assume that $x \neq 2$ and $|x-2| < \delta = \epsilon$, then we see that

$$|\frac{x^2-4}{x-2} - 4| = |\frac{(x-2)(x+2)}{x-2} - 4| = |(x+2)-4| = |x-2| < \delta = \epsilon.$$

Thus, we are done.

Galileo: Very good.

Simplicio: How about another example?

Example 9.7.2. *Using the DEFINITION of limit show: $lim_{x\to 2}(3x+5) = 11$.*

Virginia: I bet you can do it.

Simplicio: OK, I'll give it a try.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice of δ :

Since I have no clue how to choose δ, I will simply follow your lead and let $\delta = \epsilon$.

Step 3. (The Check that δ works)

Again, following your lead, I will compute the absolute value of the difference between $3x+5$ and 11. We find that $|3x+5-11| = |3x-6| = 3|x-2| < 3\delta < 3\epsilon$.

Simplicio: OOPS. Now I am stuck.

Virginia: But think about it. If you had simply been a bit smarter and had chosen $\delta = \frac{\epsilon}{3}$, you would have been fine. With this choice we now see that if $|x-2| < \delta$, then $|3x+5-11| = |3x-6| = 3|x-2| < 3\delta = 3\frac{\epsilon}{3} = \epsilon$. Now you are done.

Simplicio: Actually, that wasn't so bad.

Galileo: Note that there is a general strategy here. Namely, choose $\delta = \frac{\epsilon}{slope}$.

Simplicio: Sounds good, but what if the slope equals zero?

Virginia: And what if the slope is negative?

Galileo: OK, choose $\delta = \frac{\epsilon}{|slope|+1}$.

Virginia: Much better. Now we know that δ can never be negative or zero.

Simplicio: However, I do have just one more question. When I took Calculus, we always described limits by saying that if a sequence of points $x_1, x_2, \ldots, x_n, \ldots$ gets close to a point a, then the sequence of points $f(x_1), f(x_2), \ldots, f(x_n), \ldots$ gets close to the limit L.

Galileo: Good question. In fact, your idea turns out to be equivalent to the definition I just gave you. A more careful statement of the definition of limits in terms of sequences is given in the following theorem.

Theorem 9.7.2 (The Sequence Definition for Limit of a Function). *If X is an interval, $f(x) : X \to \Re$, and $\lim_{x\to a} f(x) = L$, then for any sequence $\{x_n\}_{n=1}^{\infty}$ with the property that $x_n \in X, \lim_{n\to\infty} x_n = a$, and $x_n \neq a$ for all n, then $lim_{n\to\infty} f(x_n) = L$.*

Proof. The proof follows the same format as our other proofs that sequences converge.. Begin by assuming we have a sequence $\{x_n\}_{n=1}^{\infty}$ with the property that $\lim_{n\to a} x_n = a$ and $x_n \neq a$ for all n.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice of N :

Since we don't have a formula for the function $f(x)$, we are forced to use our hypotheses to find N. However, since we are assuming that $\lim_{x\to a} f(x) = L$, we know there is a $\delta > 0$ with the property that if $|x - a| < \delta$ and $x \neq a$, then $|f(x) - L| < \epsilon$. Since $\delta > 0$ and since $\lim_{n\to a} x_n = a$, we can find an integer N with the property that $|x_n - a| < \delta$. This integer N is our choice.

Step 3. The Check that N works:

Since $|x_n - a| < \delta$ and $x_n \neq a$, we know immediately that $|f(x_n) - L| < \epsilon$. □

Galileo: Now that wasn't so bad was it?

Simplicio: I guess the proof was similar to the others. But why would you bring up this tangential topic?

Galileo: It may be tangential, but from a pedagogical point of view, sequences are probably a bit easier to visualize than functions.

Virginia: But, are sequences good enough?

Galileo: Actually, the converse of the above theorem is also true so we have actually formulated an equivalent definition of limits that only involves sequences.

Virginia: Should we prove it?

Galileo: While similar to the proof that every Cauchy sequence converges, the proof is by contradiction and we have other topics to cover. I will leave it as an exercise.

Exercise Set 9.7.

1. Using the definition of limit show: $lim_{x\to 3}\frac{x^2-9}{x-3} = 6$.
2. Using the definition of limit show: $lim_{x\to a}(mx+b) = ma+b$.
3. Prove that the two Definitions of Limit are equivalent.

9.8 Limit Facts for Functions

Galileo: Just as we assembled basic facts for limits of sequences, we now mention similar facts for limits of functions. The same sum, product, and quotient rules hold for functions as hold for sequences. Note that the spirit of the proofs is the same.

Theorem 9.8.1 (Basic Limit Facts for Functions). *If X is an interval, $a \in X$, and $f(x), g(x) : X \to \Re$ are functions with the property that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$, then:*

1. *Fact 1.* $\lim_{x\to a}(f(x)+g(x)) = L+M$,
   (The limit of the sum equals the sum of the limits or LS = SL.)
2. *Fact 2.* $\lim_{x\to a}(f(x)*g(x)) = L*M$, *and*
   (The limit of the product equals the product of the limits or LP = PL.)
3. *Fact 3. If $M \neq 0$, then* $\lim_{x\to a}(\frac{f(x)}{g(x)}) = \frac{L}{M}$.
   (The limit of the quotient equals the quotient of the limits or LQ = QL.)

Proof. Fact 1. The limit of the sum equals the sum of the limits.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Actually, we need to make two choices.

Choice 1: Since $\lim_{x\to a} f(x) = L$, we know that there is a quantity $\delta_1 > 0$ with the property that if $x \neq a$ and $|x - a| < \delta_1$, then $|f(x) - L| < \frac{\epsilon}{2}$.

Choice 2: Since $\lim_{x\to a} g(x) = M$, we know that there is a quantity $\delta_2 > 0$ with the property that if $x \neq a$ and $|x - a| < \delta_2$, then $|g(x) - M| < \frac{\epsilon}{2}$.

Since we want both of the statements $|f(x) - L| < \frac{\epsilon}{2}$ and $|g(x) - M| < \frac{\epsilon}{2}$ to be true, we choose δ to be the smaller of the two numbers δ_1 and δ_2.

Step 3. The Check:

Thus, if $x \neq a$ and $|x - a| < \delta$, then

$$\begin{aligned} |f(x) + g(x) - (L + M)| &\leq |(f(x) - L) + (g(x) - M)| \\ &\leq |f(x) - L| + |g(x) - M| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

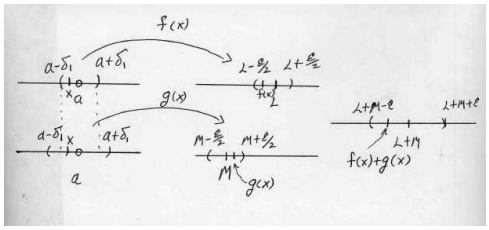


Figure 9.6: The Limit of the Sum Equals the Sum of the Limits

Fact 2. The limit of the product equals the product of the limits.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Actually, we need to make three choices.

Choice 1: Since $\lim_{x\to a} f(x) = L$, we know that there is a quantity $\delta_1 > 0$ with the property that if $x \neq a$ and $|x - a| < \delta_1$, then $|f(x) - L| < \frac{\epsilon}{3|M|+1}$.

Choice 2: Since $\lim_{x\to a} f(x) = L$, we know that there is a quantity $\delta_2 > 0$ with the property that if $x \neq a$ and $|x - a| < \delta_2$, then $|f(x) - L| < \frac{1}{2}$.

Choice 3: Since $\lim_{x\to a} g(x) = M$, we know that there is a quantity $\delta_3 > 0$ with the property that if $x \neq a$ and $|x - a| < \delta_3$, then $|g(x) - M| < \frac{\epsilon}{3|L|+1}$.

Since we want all three of the statements $|f(x) - L| < \frac{\epsilon}{3|M|+1}$, $|f(x) - L| < \frac{1}{2}$, and $|g(x) - M| < \frac{\epsilon}{3|L|+1}$ to be true, we choose δ to be the minimum of the three numbers δ_1, δ_2 and δ_3.

Step 3. The Check:

Thus, if $x \neq a$ and $|x - a| < \delta$, then we know by the choices for δ_1 and δ_2 that

$$\begin{aligned}
|f(x) * g(x) - L * M| &= |f(x) * g(x) - f(x) * M + f(x) * M - L * M| \\
&\leq |f(x) * g(x) - f(x) * M| + |f(x) * M - L * M| \\
&\leq |f(x)|\ |g(x) - M| + |f(x) - L|\ |M| \\
&\leq |f(x)|\frac{\epsilon}{3|L|+1} + \frac{\epsilon}{3|M|+1}M < \\
&\leq |f(x)|\frac{\epsilon}{3|L|+1} + \frac{\epsilon}{3}.
\end{aligned}$$

Since $x \neq a$ and $|x - a| < \delta_2$, we know by the second choice that

$$|\ |f(x)| - |L|\ | \leq |\ f(x) - L\ | < \frac{1}{2},$$

which implies

$$|\ f(x)\ | \leq |L| + \frac{1}{2}.$$

Thus,

$$(|L| + \frac{1}{2})\frac{\epsilon}{3|L|+1} < \frac{2}{3}\epsilon$$

and

$$|f(x) * g(x) - L * M| < |f(x)|\frac{\epsilon}{3|L|+1} + \frac{\epsilon}{3} < (|L| + \frac{1}{2})\frac{\epsilon}{3|L|+1} + \frac{\epsilon}{3} < \frac{2}{3}\epsilon + \frac{\epsilon}{3} < \epsilon.$$

Thus, the proof is complete.

Fact 3. The limit of the quotient equals the quotient of the limits.

This proof is left as an exercise. □

Simplicio: But wait a minute, I don't quite see why we know

$$(|L| + \frac{1}{2})\frac{\epsilon}{3|L|+1} < \frac{2}{3}\epsilon.$$

Galileo: Whenever you are expected to show one fraction is less than another, simply assume the relation holds, cross multiply, and simplify. More than likely, you can figure it out.

We now turn to a special case of the theorem that the limit of the product is the product of the limits when one of the functions is a constant. We single out this case because it is one of the details that needs to be checked when we show the collection of continuous functions forms a vector space. In particular, if $f(x) : X \to \Re$ is a function which is continuous at each $x \in X$ and $K \in \Re$, then the function $Kf(x)$ is also continuous.

Corollary 9.8.2 (Pulling Constants Across Limit Signs). *If X is an interval, $a \in X$, K is a real number, and $f(x) : X \to \Re$ is a function with the property that $\lim_{x\to a} f(x) = L$, then $\lim_{x\to a}(K * f(x)) = K * \lim_{x\to a} f(x) = K * L$.*

Proof. This fact follows immediately from the limit of the product equals the product of the limit. (i.e. Fact 2, above.) You only have to set $g(x) = K$, for all $x \in X$. □

Exercise Set 9.8.

1. Using your limit facts, show: $\lim_{x\to a}(mx + b) = ma + b$.

2. Using your limit facts, show: $\lim_{x\to a} x^2 = a^2$.

3. Using your limit facts, show: $\lim_{x\to a} x^3 = a^3$.

4. Using your limit facts, show: $\lim_{x\to 2} x\frac{x^2-4}{x-2} = 8$.

5. Using your limit facts, show: $\lim_{x\to 3} x\frac{x^2-9}{x-3} = 18$.

6. To complete the proof that the limit of the product equals the product of the limit, show: If $L > 0$, then $\frac{L+\frac{1}{2}}{3L+1} < \frac{2}{3}$.

7. Prove: The limit of the quotient equals the quotient of the limits.

Chapter 10

Connectedness and Compactness

Galileo: A solid understanding of Calculus is a must. While we will review the big named theorems, we do expect you to be able to compute derivatives and sketch graphs. In particular, you should know the product rule, the quotient rule, and the chain rule.

Simplicio: I have forgotten the chain rule. Remind me.

Galileo: Go look it up.

Simplicio: I sold my book.

Galileo: Sorry, I don't have time to reteach all of Calculus.

Virginia: What about those word problems? I found them difficult.

Galileo: Any skills you learned solving extrema (e. g. max/min) problems should help. Root finding and data fitting are techniques connected to real applications. Real applications invariably involve transforming words into symbols.

Simplicio: Actually, while I also found some of those problems to be hard, I enjoyed connecting the techniques to something in the real world.

Galileo: For Isaac Newton, Calculus was always connected to velocity, acceleration, force, mass, and volume. Unfortunately, while these applications are the real reason to study Calculus, we are now going to take a major detour and discuss the theory. You should recall that the grandfather of all the theorems in Calculus is the Fundamental Theorem of Calculus, which not only states that the two big ideas of Calculus are

related, but that they are actually inverse operations of one another. While we will prove this theorem along the way, our main goals are to prove the Intermediate Value Theorem, the Mean Value Theorem, and Taylor's Theorem.

Simplicio: And why do we care about these wondrous theorems?

Galileo: The Intermediate Value Theorem is exactly the type of information we need to guarantee the existence of a root for a continuous function. This theorem assures us that the bisection algorithm always works.

Simplicio: And the Mean Value Theorem?

Galileo: The Mean Value Theorem provides a tool for showing certain methods converge linearly.

Simplicio: Linearly convergence?

Galileo: While the sequence $\{\frac{1}{n}\}_{n=1}^{\infty}$ converges to zero, the rate is glacial. If you want 6 digits of accuracy, you have to compute more than a million terms. On the other hand, the sequence $\{\frac{1}{2^n}\}_{n=1}^{\infty}$ converges much faster.

Simplicio: Looks to me like you only need 20 terms this time.

Galileo: Very good. In fact, the error drops by 50% for each new term. The Mean Value Theorem helps us to uncover when this preferred convergence rate will occur. In particular, under reasonable conditions, the method of Newton/Raphson converges linearly. This theorem also sets the stage for the algorithm associated with the Contraction Mapping Theorem

Simplicio: And Taylor's Theorem?

Galileo: Consider the sequence $\{\frac{1}{2^{2^n}}\}_{n=1}^{\infty}$. How many terms do you have to compute before you have 6 digits of accuracy this time?

Simplicio: Looks like you only need to compute 5 terms this time.

Galileo: Excellent! You should have been a computer scientist. OK, now think about it. If you only have a paper and pencil, which sequence would you rather compute. I think the answer is obvious. In any case, as long as the function $f(x)$ doesn't have multiple roots, the Newton/Raphson algorithm usually provides quadratic convergence. Later, we will show how Taylor's Theorem provides a technique for computing

derivatives numerically. Thus, they can be used to solve differential equations and partial differential equations. These derivatives are also used extensively in signal processing and image processing applications. You can find employment in these areas.

10.1 Continuous Functions

Galileo: When we discussed the bisection method, we mentioned that the Intermediate Value Theorem can be used to show that the method always works. Since continuity of the function $f(x)$ is not only a key hypothesis for this theorem, but also for the Fundamental Theorem of Calculus, the Mean Value Theorem, and Taylor's Theorem, it is now time to nail the Jello to the wall. Before we can give careful proofs of these theorems, we need to prove a number of other theorems along the way including the Extremum Theorem and the Intermediate Value Theorem for Integrals. Every one of these theorems requires the assumption that the function $f(x)$ is continuous. In fact, whenever we integrate a function, we will assume it is continuous to make sure the integral exists. The bottom line: continuity is an omnipresent assumption that insures good things will happen.

Simplicio: I guess theory awaits us.

Galileo: We now turn to the task of giving a careful definition of what it means for a function $f(x) : X \to \Re$ to be continuous at a point a in an interval X. As we have already mentioned, this idea is quite natural. Time is probably the best example of a continuous phenomenon. At least, we would like to think time changes continuously. A multitude of physical quantities are measured as functions of time in a continuous way. Examples include: the distance a projectile has traveled, the distance from the earth to the sun, your age, your height, and your weight.

Virginia: How does nature connect with mathematics?

Galileo: Since we think of time as a linear progression, we can think of time as a copy of the real numbers. Since we are giving ourselves the Least Upper Bound

Principle, we have no holes or jumps in the real numbers. The Intermediate Value Theorem states that a function which is continuous at every point in an interval actually preserves this property.

Virginia: In other words, the analogy is that time corresponds to the real numbers and measurements dependent on time correspond to continuous functions.

Galileo: Deep in our hearts we believe atoms move through space in a continuous fashion.

Simplicio: I bet your colleagues in Quantum Mechanics would have something to say about this.

Galileo: No doubt. But we don't have time for such a diversion.

Virginia: Let's get back to the mathematics.

Galileo: As you will soon notice, a continuous function will be one whose limits are EASY to compute. Namely, limits are computed by simple substituting. We now give the precise formulation of the definition

Definition 10.1.1. *If $a \in X$, where $X \subset \Re$ is an interval, $f(x) : X \to \Re$ is a function, and $\lim_{x\to a} f(x) = f(a)$, then $f(x)$ is continuous at $x = a$.*

Simplicio: How about a few examples?

Galileo: Moments ago, we showed that $\lim_{x\to a} mx+b = ma+b$. This exercise showed that the function $f(x) = mx + b$ is continuous at the point $x = a$. Thus, straight lines are always continuous. In fact, all your old friends including polynomials $p_n(x)$, trigonometric functions (e.g. $\cos(x)$ and $\sin(x)$), and exponential functions (such as e^x) are continuous at every point $x \in \Re$.. Any sum, product, or quotient of these functions will also be continuous. While functions like $f(x) = \frac{1}{x}$ and $\tan(x) = \frac{\sin(x)}{\cos(x)}$ are continuous at most points, they both shoot off to ∞ at points where the denominator equals zero. For example, the function $f(x) = \frac{1}{x}$ heads off to infinity at $x = 0$ and thus is not continuous at this point. However, they have the enjoyable property that they are continuous at every point where the denominator is different from zero. During our discussions, we will frequently need to assume that the functions under consideration are continuous

Virginia: How about an example of a function, which is not continuous?

Example 10.1.1. *Galileo: Consider the Heaviside function*

$$H(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}.$$

Note that while it is continuous at every point except $x = 0$, there is no point x with the property that $H(x) = \frac{1}{2}$. Thus, the function $H(x)$ tears apart the real numbers into two sets. The first set is all the negative numbers, which gets mapped to zero. The second is the set of all the non-negative numbers, which gets sent to 1. Thus, nothing gets mapped to $\frac{1}{2}$. This example will become important when we discuss the Intermediate Value Theorem 10.2.

The purpose of the next theorem is to formalize the fact that the sum, product, and quotient of two continuous functions is continuous.

Theorem 10.1.2 (Sum, Product, and Quotient of Continuous Functions). *If $a \in X$, where X is an interval, and $f(x), g(x) : X \to \Re$ are both continuous at the point $x = a$, then*

1. *the function $(f + g)(x) = f(x) + g(x)$ is continuous at $x = a$.*
2. *the function $(f * g)(x) = f(x) * g(x)$ is continuous at $x = a$.*
3. *if $g(a) \neq 0$, then the function $(\frac{f}{g})(x) = \frac{f(x)}{g(x)}$ is continuous at $x = a$.*

Proof. If $f(x)$ and $g(x)$ are both continuous at $x = a$, then $\lim_{x\to a} f(x) = f(a)$ and $\lim_{x\to a} g(x) = g(a)$.

From the Basic Limit Facts for Function 9.8.1, we now make three observations:

1. $\lim_{x\to a} f(x) + g(x) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x) = f(a) + g(a)$.
2. $\lim_{x\to a} f(x) * g(x) = \lim_{x\to a} f(x) * \lim_{x\to a} g(x) = f(a) * g(a)$.
3. If $g(a) \neq 0$, then $\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)} = \frac{f(a)}{g(a)}$.

□

Simplicio: OK, those three proofs are easy, but what can I do with them?

Galileo: Since $f(x) = x$ is continuous, we now know that $f(x) * f(x) = x^2$ and $f(x) * f(x) * f(x) = x^3$ are also continuous. In general, we now know that any polynomial $p_n(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ is continuous at every point. Even more generally, we know that if $p_n(x)$ and $q_m(x)$ are two polynomials, then the rational function $r(x) = \frac{p_n(x)}{q_m(x)}$ is continuous at any point $x = a$, where $q_m(a) \neq 0$. While we won't take the time to show it now, the trigonometric functions $\cos(x)$ and $\sin(x)$ also turn out to be continuous. Thus, functions like $f(x) = 2x + 3\cos(x) + x^2\sin(x)$ will be continuous.

Virginia: Wait a minute! I just noticed that the functions $\cos(\pi x), \sin(\pi x), \cos(2\pi x), \sin(2\pi x)$ are not covered by our Sums, Products and Quotients Theorem. In other words, how do I know these functions are continuous?

Galileo: You caught me. I forgot to mention that the composition of two continuous functions is continuous. Since $g(y) = \cos(y)$ and $f(x) = 2\pi x$ are continuous at every point, then the next proposition justifies the claim that the function $h(x) = g(f(x)) = \cos(2\pi x)$ is continuous at every point x.

Proposition 10.1.3 (The Composition of Continuous Functions is Continuous). *Let X, Y be intervals in $\Re$. Let $f(x) : X \to Y$ and $g(y) : Y \to \Re$ be functions. If $f(x)$ is continuous at a point $a \in X$ and $g(y)$ is continuous at the point $f(a)$ in Y, then the composition $g(f(x))$ is continuous at $x = a$.*

Proof. Galileo: We can prove this proposition right from the definition. As usual, the proof is backwards. Namely, we begin with the function $g(y)$ and then with the function $f(x)$. The only idea is that we have to choose two "$\delta's$." We first choose δ_1 for the function $g(y)$ and then (depending on the size of δ_1) we choose δ.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Our job is to find a $\delta > 0$ with the property that if $x \in (a - \delta, a + \delta)$, then $g(f(x) \in (g(fa)) - \epsilon, g(fa)) + \epsilon)$.

Step 2. The Choice:

Since $g(y)$ is continuous at $y = f(a)$ and $\epsilon > 0$, choose $\delta_1 > 0$ with the property that if $y \in (f(a) - \delta_1, f(a) + \delta_1)$, then $g(y) \in (g(f(a)) - \epsilon, g(f(a)) + \epsilon)$.

Since $f(x)$ is continuous at $x = a$ and $\delta_1 > 0$, choose $\delta > 0$ with the property that if $x \in (a - \delta, a + \delta)$, then $f(x) \in (f(a) - \delta_1, f(a) + \delta_1)$.

Step 3. The Check:

If $x \in (a - \delta, a + \delta)$, then $f(x) \in (f(a) - \delta_1, f(a) + \delta_1)$.

Since $f(x) \in (f(a) - \delta_1, f(a) + \delta_1)$, $g(f(x) \in (g(f(a)) - \epsilon, g(f(a)) + \epsilon)$.

□

Simplicio: Not so bad.

Exercise Set 10.1.

1. Discuss why the function $f(x) = \sin(x^2 + 1)$ is continuous.

2. Discuss where the function $f(x) = \frac{x^2+1}{x-9}$ is continuous. Justify your answer.

3. Show the function $f(x) = |x|$ is continuous.

4. Explain why the function $f(x) = \frac{2x+5}{7x+11}$ is continuous at $x = 3$.

5. Evaluate the limit $\lim_{x\to 3} \frac{2x+5}{7x+11}$.

6. Show the function $f(x) = \frac{1}{1-x}$ is continuous. Where does it fail to be continuous?

7. Explain why the function $\tan(x) = \frac{\sin(x)}{\cos(x)}$ is continuous at most points. Where does it fail to be continuous?

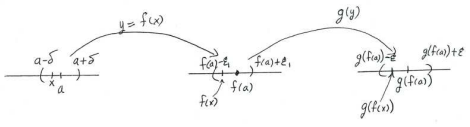


Figure 10.1: The Composition of Continuous Functions Is Continuous

8. Explain why the function $f(x) = \sin(x^2 + 3)$ is continuous. (You may assume the function $\sin(x)$ is continuous.

9. Prove: If $T(x) : [a, b] \to \Re$ is a function with the property that $|T(x_1) - T(x_2)| \leq M \ |x_1 - x_2|$ for all $x_1, x_2 \in [a, b]$, then show that $T(x)$ is continuous at each $x \in [a, b]$.

10.2 Intermediate Values and Connectedness

Galileo: We now return to the Intermediate Value Theorem, which we already mentioned when we presented the bisection method.

Simplicio: Remind my why I should care about this theorem?

Galileo: The Intermediate Value Theorem is exactly what is needed to guarantee the bisection method always works. The first mathematician/philosopher to attempt placing these ideas on a firm mathematical foundation was Bernard Bolzano (1781-1848). His goal was to make the idea of an infinitesimal precise. While he published a proof in 1817, he achieved little recognition for his efforts until after his death. In fact, he had a rough time since he lost his teaching position at the University of Prague for his pacifist views. He was even put under house arrest and forbidden to publish.

Virginia: I think you could identify with the plight of this fellow.

Galileo: Indeed I do. While unaware of Bolzano's ideas, Augustin Cauchy (1789-1857) published many of these results in 1821. We now state and prove a technical proposition, which will help us prove the theorem. Intuitively, this proposition states that if a function $f(x)$ maps a point x_0 to a value above y_0, then a whole open interval of points must also be mapped above y_0. A similar statement can be made if $f(x)$ maps a point x_0 to a location below y_0,

Proposition 10.2.1. *Let $f(x) : (a, b) \to \Re$ be a function, which is continuous at a point $x_0 \in (a, b)$.*

1. *If $f(x_0) > y_0$, then there is a $\delta > 0$ with the property that $f(x) > y_0$ for all $x \in (x_0 - \delta, x_0 + \delta)$.*

2. *If $f(x_0) < y_0$, then there is a $\delta > 0$ with the property that $f(x) < y_0$ for all $x \in (x_0 - \delta, x_0 + \delta)$.*

Proof. 1. If $f(x_0) > y_0$, then let $\epsilon = f(x_0) - y_0 > 0$. Since $f(x)$ is continuous at $x = x_0$, there is a $\delta > 0$ with the property that if $x \in (x_0 - \delta, x_0 + \delta)$, then $f(x) \in (f(x_0) - \epsilon, f(x_0) + \epsilon)$. Thus, $f(x) > f(x_0) - \epsilon = y_0$ for all $x \in (x_0 - \delta, x_0 + \delta)$.

2. If $f(x_0) < y_0$, then let $\epsilon = y_0 - f(x_0) > 0$. Since $f(x)$ is continuous at $x = x_0$, there is a $\delta > 0$ with the property that if $x \in (x_0 - \delta, x_0 + \delta)$, then $f(x) \in (f(x_0) - \epsilon, f(x_0) + \epsilon)$. Thus, $f(x) < f(x_0) + \epsilon = y_0$ for all $x \in (x_0 - \delta, x_0 + \delta)$.

□

Simplicio: I didn't like that proposition. I hope I never see it again.

Galileo: Unfortunately, we will see it again when we discuss extrema and compactness. This proposition contains useful connections between continuous functions and open intervals.

Virginia: Open intervals aren't so hard.

Galileo: Let us now state and prove the Intermediate Value Theorem. If we use our example to illustrate the theorem, we should let the function $f(x)$ be your height at time x. This function will be a continuous function of time. Since you were less than 2 feet tall when you were born, $f(0) < 2$. If b denotes your current age, $f(b) > 5$. Since $y_0 = 4$ is intermediate between 2 and 5, the theorem guarantees that there will be a time z_0 with the property $f(z_0) = 4$. Now, for the theorem itself.

Theorem 10.2.2 (Intermediate Value Theorem). *If $f(x) : [a, b] \to \Re$ is continuous at each $x \in [a, b]$ and $f(a) < y_0 < f(b)$ (or $f(a) > y_0 > f(b)$), then there is a point $z_0 \in [a, b]$ such that $f(z_0) = y_0$.*

Proof. The proof rests on the Law of Trichotomy, the Least Upper Bound Principle, and the previous proposition.

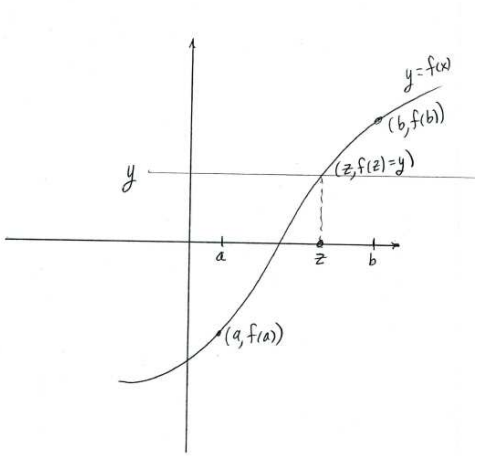


Figure 10.2: The Intermediate Value Theorem

Simplicio: What the heck is the Law of Trichotomy?

Galileo: The prefix "Tri" indicates three possibilities. The Law of Trichotomy is a fancy way of saying that if someone gives you two real numbers x and y, then one of the following three possibilities must hold: $x > y, x < y$, or $x = y$.

Simplicio: That Law is obvious.

Galileo: Well OK, but it can be proved from basic principles. In any case, our strategy is going to be to find a number z_0 with the property that if $f(a) < y_0 < f(b)$, then there is a number $z_0 \in [a, b]$ such that the statements $f(z_0) > y_0$ and $f(z_0) < y_0$ are both false.

Virginia: So, by the Law of Trichotomy, there is no other possibility except that $f(z_0) = y_0$.

Galileo: Correct.

Virginia: But how do we find z_0?

Galileo: The point z_0 will be defined as the least upper bound of all those points x in $[a, b]$, such that $f(x)$ is "below" the line $y = y_0$. To formalize this statement, we define this set by the rule $S = \{x \in [a, b] : f(x) \leq y_0\}$. A detail that needs to be checked is that this set is non empty.

Virginia: Since $f(a) < y_0$, we immediately know that $a \in S$.

Galileo: Correct. Now we simply identify z_0 as the least upper bound of S.

Virginia: And show the two other cases $f(z_0) > y_0$ and $f(z_0) < y_0$ are both false.

Galileo: Correct.

Case 1. Suppose the statement $f(z_0) > y_0$ is true.

By the previous proposition we can find a $\delta > 0$ so that if $x \in (z_0 - \delta, z_0 + \delta)$, then $f(x) > y_0$. Thus, if $x \in (z_0 - \delta, b]$, then x is NOT in the set S and the number $z_1 = z_0 - \delta$ must be an upper bound for S. Since $z_1 = z_0 - \delta < \delta$, we have a contradiction to the assumption that z_0 is the smallest upper bound. This contradiction forces us to abandon the supposition that $f(z_0) > y_0$ is true.

Case 2. Suppose the statement $f(z_0) < y_0$ is true.

Again, by the previous proposition we can find a $\delta > 0$ so that if $x \in (z_0 - \delta, z_0 + \delta)$, then $f(x) < y_0$. Thus, if $x \in (z_0 - \delta, z_0 + \delta)$, then $x \in S$. In particular, the point $x = \frac{z_0 + \delta}{2}$ is NOT in the set S. Thus, we have a contradiction to the assumption that z_0 is an upper bound of S.

□

Galileo: Notice that the idea underlying this proof is that the problem of "breaks" or "jumps" in the curve $y = f(x)$ is thrown back to the problem of no "holes" in the real number line. Actually, what we are saying is that if X is an interval and the image set $Y = f(X)$ is defined by $Y = f(X) = \{y \in \Re : y = f(x) \text{ for some } x \in X\}$, then Y is an interval. In other words, the continuous image of a connected set is connected.

Virginia: The Least Upper Bound Principle is what makes it all work.

Galileo: Before we leave this subject, let's follow Professor Polya's dictum that we should look back at what we have accomplished. First, let me comment that the idea of connectedness is a completely general concept, which is valid in any dimension. In our setting, the point y_0 separates the real line into the two open intervals $V_1 = (-\infty, y_0)$ and $V_2 = (y_0, \infty)$. The proposition shows that the two sets $S_1 = \{x \in [a, b] : f(x) \in V_1\}$ and $S_2 = \{x \in [a, b] : f(x) \in V_2\}$ are unions of open intervals. Such sets are called open. Since the sets V_1 and V_2 are disjoint, the sets S_1 and S_2 are disjoint. Thus, we have separated the interval $[a, b]$ into the union two non-empty disjoint open

sets. The point z_0 we found shows this is impossible.

Virginia: Why do we need the assumption that the function is continuous?

Galileo: Recall the Heaviside example

$$H(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases},$$

where there is no point x with the property that $H(x) = \frac{1}{2}$. Thus, the intermediate value $\frac{1}{2}$ is never attained.

Virginia: Where might we see these ideas again?

Galileo: In Complex Variables you will immediately be confronted by the Jordan Curve Theorem, which says that any simple closed curve C separates the plane into two open sets, an "inside" and an "outside." Thus, the set $\Re^2 - C$ is not connected.

Simplicio: That stuff sounds way too theoretical to be useful.

Galileo: Not only is Complex Variables a beautiful subject, but it is used everywhere in engineering and physics applications.

Exercise Set 10.2.

1. Show that the function $f(x) = x^5 + x + 1$ has a root in the interval $[-1, 0]$.

2. Show that the function $f(x) = x - e^x$ has a root in the interval $[0, 1]$.

3. Prove the following theorem: If $f(x) : [0, 1] \to [0, 1]$ is a function that is continuous at each $x \in [0, 1]$, then there is a point $z \in [0, 1]$ with the property that $f(z) = z$. (Hint: Apply the Intermediate Value Theorem to the function $h(x) = x - f(x)$.)

10.3 Extreme Values and Compactness

Galileo: We now turn to the Extremum Theorem for continuous functions. This theorem states that a continuous function $f(x) : [a, b] \to \Re$ always attains its maximum.

In other words, there is a point $z_0 \in [a,b]$ with the property that $f(z_0) \geq f(x)$ for all $x \in [a,b]$.

Simplicio: So, if I toss a ball into the air and catch it a few moments later, then at some instant z_0 in time, the ball will be at its highest. Seems obvious to me.

Galileo: Not so fast. What about the function $f(x) = \frac{1}{x}$ defined on the interval $(0.1]$. While the function is continuous, the graph becomes arbitrarily high as x gets close to zero.

Simplicio: In other words, the ball just keeps on going up.

Galileo: Correct.

Virginia: How do we keep that from happening?

Galileo: Our friend the Least Upper Bound Principle will once again save us. Note that the theorem states that not only is the function $f(x)$ bounded above, but that there is a particular point (or instant in time) z_0 which is the highest point on the curve.

Theorem 10.3.1 (Extremum Theorem). *If $f(x) : [a,b] \to \Re$ is continuous at each point $x \in [a,b]$, then there is a point $z_0 \in [a,b]$ with the property that $f(z_0) \geq f(x)$ for all $x \in [a,b]$. Similarly, there is a point $z_1 \in [a,b]$ with the property that $f(z_1) \leq f(x)$ for all $x \in [a,b]$.*

Proof. This theorem is proved in two steps.

Our first step is to show the function $f(x)$ must be bounded. In other words, there is a constant M with the property that $f(x) \leq M$ for all $x \in [a,b]$. In particular, $f(x)$ cannot be unbounded the way the function $f(x) = \frac{1}{x}$ is.

The second step in the proof is to guarantee that there is a point $z_0 \in [a,b]$ with the property that $f(z_0) = L$, where $L = lub(f([a,b])) = lub\{y \in \Re : y = f(x) \text{ for some } x \in [a,b]\}$. By the definition of L, $L \geq f(x)$ for all $x \in [a,b]$. If $f(z_0) = L$, then $f(z_0) \geq f(x)$ for all $x \in [a,b]$.

Step 1. There is a constant M th the property that $f(x) \leq M$ for all $x \in [a,b]$.

Suppose this statement is false. If false, then for each integer n the set $S_n = \{x \in [a,b] : f(x) \geq n\}$ is nonempty. Note that each S_n is nonempty and that $S_{n+1} \subset S_n$ for

all n. If $b_n = lub(S_n)$, then $a \leq b_{n+1} \leq b_n \leq b$, for all n. Thus, the sequence $\{b_n\}_{n=1}^{\infty}$ is a decreasing sequence, which is bounded below by the number a. Hence the sequence converges to some number $z_0 \in [a, b]$. Note that $z_0 \leq b_n$ for all n.

Choose an integer $n > f(z_0)$. Since the function $f(x)$ is continuous at $x = z_0$, we know by Proposition 10.2.1 there is a $\delta > 0$ with the property that if $x \in (z_0 - \delta, z_0 + \delta)$, then $f(x) < n$. Since no point x can be in both S_n and the interval $(z_0 - \delta, z_0 + \delta)$, the number $z_0 - \delta$ is an upper bound for the set S_n. Since $z_0 - \delta < z_0 \leq b_n$, Thus, the number $z_0 - \delta$ is an upper bound for the set S_n, which is smaller than its least upper bound b_n.

This contradiction shows that there is a constant M with the property that $f(x) \leq M$ for all $x \in [a, b]$.

Step 2. If $L = lub(f([a, b]))$, then there is a point $z_0 \in [a, b]$ such that $f(z_0) = L$.

Suppose this statement is false. If false, then define the function $g(x) = \frac{1}{L - f(x)}$. Since $f(x)$ is continuous for all $x \in [a, b]$ and $f(x) \neq L$ for all $x \in [a, b]$, we know by Theorem 10.1.2 that the quotient $g(x) = \frac{1}{L - f(x)}$ is also continuous. By Step 1, we know there is a constant $M > 0$ with the property $|\frac{1}{L - f(x)}| = |g(x)| \leq M$ for all $x \in [a, b]$.

Since $L - f(x) > 0$ for all $x \in [a, b]$, $\frac{1}{L - f(x)} \leq M$ for all $x \in [a, b]$.

Thus, $\frac{1}{M} \leq L - f(x)$ for all $x \in [a, b]$ or $f(x) \leq L - \frac{1}{M}$ for all $x \in [a, b]$. Thus, $L - \frac{1}{M}$ is an upper bound for the set $\{y \in \Re : y = f(x) \text{ for some } x \in [a, b]\}$, which is smaller than L.

Thus, we have a contradiction to the assumption that L is the least upper bound for the set $f([a, b])$. Thus, there is a point $z_0 \in [a, b]$ with the property that $f(z_0) = L \geq f(x)$ for all $x \in [a, b]$. □

Galileo: In the spirit of Professor Polya let us think about what we have accomplished. Note that we have just considered two big ideas: connectedness and compactness.

Simplicio: So?

Galileo: So the continuous image of a closed bounded interval is a closed bounded interval. Thus, continuous functions preserve this type of interval. Note also that our

proofs of both the Intermediate Value Theorem and the Extremum Theorem employ Proposition 10.2.1. What is the key idea embedded in this Proposition?

Virginia: It seems to start with an open interval in the range of the function and then work backwards to the domain.

Simplicio: The resulting set in the domain turns out to be the union of a bunch of open intervals.

Galileo: Exactly. If we introduce a bit of notation, we can clarify the concept. In particular, if we define the open interval in the range of the function by the rule $V = \{y \in \Re : y > y_0\}$, then we showed that the inverse image set $U = f^{-1}(V) = \{x \in (a, b) : f(x) > y_0\}$ is the union of open intervals back in the domain. Better yet, if we combine the two parts of Proposition 10.2.1 we have shown that the inverse image of an open set is open.

Simplicio: So why is this idea a big deal?

Galileo: First, it throws all the problems back to an open interval in the real line $\Re$. Thus, once we understand the real numbers, we are ready to go.

Simplicio: I have understood the real numbers for a long time.

Galileo: Maybe so, but it wasn't until Cantor and Dedekind came along that people felt the Jello was nailed to the wall. Two thousand years is a long time. While students think that complex numbers are weird, the real difficulties lie in the real numbers, where Dedekind showed the associative, commutative, and distributive laws can be extended from the rational numbers to this bigger set of numbers.

Simplicio: Is that all?

Galileo: A second reason to think in terms of open intervals is that these ideas generalize to all dimensions. In particular, the generalization of an open interval is an open disk in the plane and an open ball in three space. An open set is the union these simple building blocks.

Simplicio: So.

Galileo: If we define a continuous function to be one with the property that $U = f^{-1}(V)$ is an open set whenever V is open then we can show that the properties of

compactness and connectedness are both preserved by continuous functions.

Simplicio: But that means we have to go through all that theory again. More proofs!

Galileo: But this time the proofs are more conceptual and much easier because we don't have all those ϵs, δs, and limits. This branch of mathematics is called Topology.

Virginia: Why don't we do it?

Galileo: We could, but it would be a distraction from our main mission.

Simplicio: If this approach is easier, why didn't we skip all the limit stuff and just do Topology?

Galileo: We could have, but you would have found the discussions weird and abstract. You would have constantly been asking where this stuff came from.

Virginia: It is interesting that one little proposition could lead to a whole new view on a subject.

Galileo: Topology provides a wonderfully elegant framework for these ideas.

Exercise Set 10.3.

1. Identify the extreme values of the function $f(x) = x^2 - 1$ on the interval $[-1, 1]$.

2. Identify the extreme values of the function $f(x) = x^2 - 5x + 6$ on the interval $[2, 3]$.

3. Identify the extreme values of the function $f(x) = x^3 - 9x + 1$ on the interval $[-4, 4]$.

Chapter 11

Mean Value Theorems

11.1 Differentiation

Galileo: While you have seen the definition of derivative and the different rules for computing the sum, product, and quotient of differentiable functions, we now provide a quick review.

Simplicio: It has been years since I took Calculus. A review would be appreciated.

Galileo: We will need the assumption of differentiability as an assumption in many of our theorems. We will also need to compute derivatives when we use the error formulas to determine an upper bound on the error.

Simplicio: But aren't continuous functions good enough? Every continuous function is differentiable. I am sure that is true.

Galileo: Sorry, but you are mistaken once again.

Virginia: Don't you remember that the function $f(x) = |x|$ is continuous at every point but has a sharp corner at $x = 0$?

Simplicio: OK, OK.

Galileo: Since we have felt the impact of Murphy's fist when we discussed the failures of Newton/Raphson, our goal now is to get the language exactly right. As a polite reminder we begin with the familiar definition for a function $f(x)$ to be differentiable.

Definition 11.1.1. *If X is an interval, $f(x) : X \to \Re$, and the limit $lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$*

exists, then $f(x)$ is said to be differentiable at the point $x \in X$. The derivative is defined by $f'(x) = lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$.

Galileo: If $y = f(x)$, we will sometimes write $f'(x) = \frac{dy}{dx}$. Just as we remarked for continuous functions, the assumption of differentiability will occur in most of our theorems including the Fundamental Theorem of Calculus, the Mean Value Theorem, Taylor's Theorem, and the Lagrange Error Formula for polynomial interpolation.

Example 11.1.1. *If $x \in \Re$, then recall the following derivatives.*

1. *If $f(x) = \cos(x)$, then $f'(x) = -\sin(x)$.*
2. *If $f(x) = \sin(x)$, then $f'(x) = \cos(x)$.*
3. *If $f(x) = e^x$, then $f'(x) = e^x$.*
4. *If $x > 0$ and $f(x) = \log_e(x)$, then $f'(x) = \frac{1}{x}$.*

Simplicio: No problem, I think I remember seeing all those rules.
Galileo: What about the derivative of $h(x) = e^{x^2}$?
Simplicio: Hmmm. Not sure.
Virginia: That derivative follows from the chain rule, where you compute the derivative of the composition of two functions as the derivative of the outside holding the inside fixed and then multiply by the derivative of the inside. For this example, you simply think of the function $h(x)$ as the composition of the two functions $f(x) = x^2$ and $g(y) = e^y$. Since $h(x) = e^{x^2} = g(f(x))$, $h'(x) = g'(f(x))f'(x) = e^{x^2}2x$.
Galileo: Very good. The important computational facts about the sum, product, quotient, and composition of two differentiable functions are summarized in the following theorem.

Theorem 11.1.2 (Differentiation Rules). *If X is an interval and $f(x), g(x) : X \to Y \subset \Re$ are both differentiable at the point $x \in X$ and $h(y) : Y \to Z \subset \Re$ is differentiable at the point $y = g(x)$, then*

1. $(f+g)'(x) = f'(x) + g'(x)$,
   (The derivative of the sum equals the sum of the derivatives.)

2. $(f*g)'(x) = f(x)*g'(x) + f'(x)*g(x)$,
   (The Product Rule.)

3. *if* $g(x) \neq 0$, *then* $(\frac{f}{g})'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{(g(x))^2}$, *and*
   (The Quotient Rule.)

4. $h(g(x))' = h'(g(x))g'(x)$.
   (The Chain Rule.)

Proof. Galileo: You should be familiar with these formulas so we will skip the proofs. □

Simplicio: Not a problem.

Galileo: Just as we commented for continuous functions, we see by the first derivative rule that the sum of two differentiable functions is differentiable. By the second derivative rule, we see that constants can pulled across derivative signs.

Simplicio: What?

Virginia: In other words, $\frac{dKf(x)}{dx} = K\frac{df(x)}{dx}$.

Simplicio: Why would I care?

Virginia: Because you now know that the collection of all differentiable functions on an interval $[a,b]$ forms a vector space.

Galileo: Correct.

Simplicio: Why is this important?

Galileo: The general rule is that the more smoothness you have in your data, the the easier it is to find accurate approximations.

Simplicio: Smoothness?

Galileo: The more derivatives a function $f(x) : [a,b] \to \Re$ has, the smoother it is. Let us make the following inductive definition for the n^{th} derivative as the derivative of the $(n-1)^{st}$ derivative.

Definition 11.1.3. *If $f(x) : [a,b] \to \Re$, then the n^{th} derivative of $y = f(x)$ is defined as $\frac{d^n y}{dx^n} = f^{(n)}(x) = \frac{df^{(n-1)}(x)}{dx}$, where $f^{(0)}(x) = f(x)$, for all $x \in [a,b]$.*

Simplicio: So, if $y = f(x) = \sin(x)$, then $\frac{dy}{dx} = f^{(1)}(x) = f'(x) = \cos(x)$ and $\frac{d^2y}{dx^2} = f^{(2)}(x) = f''(x) = -\sin(x)$.

Galileo: Correct. In other words, not only is $f^{(0)}(x) = f(x)$, but also $f^{(1)}(x) = f'(x)$ and $f^{(2)}(x) = f'(f^{(1)}(x)) = f''(x)$, etc. The purpose of the next definition is to grade a function by the number of derivatives it has. The more derivatives $f(x)$ has, the smoother it is. The smoother it is, the easier it is to find accurate approximations.

Definition 11.1.4. *The symbol $C^0[a,b]$ denotes the collection of all functions on the interval $[a,b]$ with the property that $f(x)$ is continuous at each $x \in [a,b]$.*

Definition 11.1.5. *The symbol $C^n[a,b]$ denotes the collection of all functions on the interval $[a,b]$ with the property that $f(x), f'(x), f''(x), \ldots, f^{(n)}(x)$ are all continuous at each $x \in [a,b]$.*

The larger the integer n, the smoother the functions in the collection.

The next proposition shows that if $f(x) \in C^1[a,b]$, then $f(x) \in C^0[a,b]$.

Proposition 11.1.6. *If $f(x) : [a,b] \to \Re$ is differentiable at a point $x = z \in [a,b]$, then $f(x)$ is continuous at $x = z$.*

Proof. We must show that $\lim_{x\to z} f(x) = f(z)$.

Since the statement $\lim_{x\to z} f(x) = f(z)$ is equivalent to $\lim_{x\to z}(f(x) - f(z)\ ) = 0$, we need only prove this last equality.

We know by the limit of the product equals the product of the limits that

$$\begin{aligned}\lim_{x\to z}(f(x) - f(z)\ ) &= \lim_{x\to z} \frac{f(x) - f(z)}{x - z}(x - z) \\ &= \lim_{x\to z} \frac{f(x) - f(z)}{x - z} \lim_{x\to z}(x - z) \\ &= f'(x) * 0 = 0.\end{aligned}$$

Thus, $\lim_{x\to z} f(x) = f(z)$ and $f(x)$ is continuous at $x = z$. □

Exercise Set 11.1.

1. If $f(x) = \sin(\frac{x}{2})$, then compute $f'(x)$.

2. If $f(x) = e^{x^2}$, then compute $f'(x)$.

3. If $f(x) = e^{\frac{x}{2}}$, then compute $f'(x)$.

11.2 Rolle's Theorem

Michel Rolle (1652-1719)

Galileo: Let us begin by introducing the ideas of Michel Rolle (1652-1719), a French mathematician, who lived during the rein of King Louis XIV. While we will not give a formal proof of this theorem, an easy physics application can be used to help visualize where it comes from. In particular, if the variable x represents time and $f(x)$ represents the height of a ball thrown into the air, then the theorem states that if the ball leaves your hand at 4 feet above the ground at time $x = a$ and is caught at this same height at a second time $x = b$, then there will be some time z when the instantaneous velocity is zero. as it turns out, that time is at the exact moment when the ball achieves its greatest height.

Simplicio: But what about a bungee jumper, who jumps off a bridge at time $x = a$ and returns to the same height a few seconds later at time $x = b$?

Galileo: You are optimistic to think that the bungee jumper will return to his initial height. However, if he does, then we can visualize the point z as the moment in time when a bungee jumper is at the bottom of his fall. Both situations are covered in his theorem.

Theorem 11.2.1 (Rolle). *If $f(x) : [a,b] \to \Re$, where $f(x), f'(x)$ are continuous, and $f(a) = f(b)$, then there is a point $z \in (a,b)$ such that $f'(z) = 0$.*

Proof. Galileo: To ease your pain we will skip the difficult part of the proof. You might be surprised to learn that the difficulties lie in showing that the function actually attains a highest (and lowest) value at some point z. However, if we can find a point $z \in (a,b)$ with the property that $f(z) \geq f(x)$ for all $x \in [a,b]$, then all we have to do is compute the difference quotient on each side. The difference quotient will be positive on the left and negative on the right. Thus, the derivative at the top of the mountain must be zero.

A more quantitative argument can be given by simply noticing when the numerator and denominator of the difference quotient are positive and negative. Since $f(z) \geq f(x)$ for all $x \in [a,b]$, the numerator of the difference quotient $f(z+h) - f(z)$ is negative. If the point $z+h$ is to the left of z, then the quantity h must also be negative. Thus the fraction $\frac{f(z+h)-f(z)}{h}$ must be positive.

Similarly, if the point $z+h$ is to the right of z, then the quantity h must be positive. Thus, the difference quotient $\frac{f(z+h)-f(z)}{h}$ equals a positive number divided by a negative number and thus negative. Thus, $f'(z)$ is the limit of both a sequence of positive numbers and a sequence of negative numbers. Thus, $f'(z) = 0$. □

Galileo: An application of Rolle's Theorem is in the area of roof repair. For example, when you are in need of a hammer and call to your assistant to get one to you right away, what is the fastest method?

Simplicio: The answer is simple. You simply throw it at him.

Galileo: Very good. However, fewer injuries will occur if the highest point of the trajectory occurs where you are standing on the roof. If the velocity is zero, then you

can simply pluck the hammer out of the air.

Simplicio: I think I am beginning to see that locations where a function has zero velocity might be useful.

Galileo: Others have made this observation before you. The next definition makes this idea official.

Definition 11.2.2. *If $X \subset \Re$, and $f(x) : X \to \Re$ is differentiable at each point in X, then a point $c \in X$ is a critical point of $f(x)$ if $f'(c) = 0$. The value $y = f(c)$ is called a critical value.*

In other words, a critical point is where the curve $y = f(x)$ has a horizontal tangent.

Simplicio: Ah! So the point $x = c$ is nothing but a root of the first derivative. Why do you call it a critical point?

Galileo: Because something important might be happening at that point. For us, the word important means a maximum or minimum value of $f(x)$ occurs at that location. If you remember from Calculus, maxima and minima occur at critical points or endpoints. Finding a root of a function's first derivative $f'(x)$ is a big deal.

Virginia: Aren't we talking about roots tomorrow?

Galileo: Absolutely. However, our immediate need for Rolle's Theorem is that it provides a quick proof of the Mean Value Theorem.

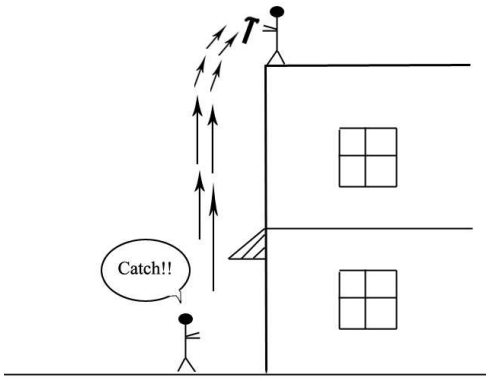


Figure 11.1: An Application of Rolle's Theorem

Exercise Set 11.2.

1. If $f(x) = -x^2 + 3x - 2$, then find a critical point for $f(x)$. What is the critical value? (Graph the function $y = f(x)$.)

2. If $f(x) = x^3 + 2x$, then show that $f(x)$ has exactly one real root. (Graph the function $y = f(x)$.)

3. Compute the critical points and critical values of the function $f(x) = xe^{-x^2}$. (Graph the function $y = f(x)$.)

11.3 The Mean Value Theorem

Galileo: Now we turn to the proof of the Mean Value Theorem.

Simplicio: What is the idea underneath the Mean Value Theorem? How am I going to remember it?

Galileo: Sometimes we refer to this theorem as the "Highway Patrol Theorem."

Simplicio: Why is that?

Galileo: Suppose you decide to visit your grandmother, who lives 80 miles away. Since you have just purchased a new car, you decide to drive. If you get there in one hour, then do you deserve a ticket?

Simplicio: I am not sure. The time does sound a bit short.

Galileo: Hopefully, the local police officer will be taking a lunch break. If not, you might warrant a speeding ticket, which could cost you a serious amount of money.

Simplicio: How so?

Galileo: Since the distance traveled in one hour was 80 miles, the average velocity is 80mph. The Mean Value Theorem guarantees that at some time during the trip your instantaneous velocity will be exactly 80mph. If the maximum speed limit over the duration of the trip is 70mph, then you will need a very bright and energetic lawyer to get you off.

Simplicio: How about if I get a fuzz-buster?

Galileo: Let's turn to the theorem.

Theorem 11.3.1 (Mean Value Theorem). *If $f(x) : [a,b] \to \Re$ has the property that $f(x), f'(x)$ are continuous, then there is a point $z \in (a,b)$ such that $f'(z) = \frac{f(b)-f(a)}{b-a}$.*

Proof. Define the function $F(x) = f(x) - (f(a) + \frac{f(b)-f(a)}{b-a}(x-a))$. Note that $F(a) = f(a) - f(a) = 0$ and $F(b) = f(b) - f(a) - (f(b) - f(a)) = 0$. Since $F'(x) = f'(x) - \frac{f(b)-f(a)}{b-a}$, we can conclude from Rolle's Theorem that there is a point $z \in (a,b)$ such that $F'(z) = f'(z) - \frac{f(b)-f(a)}{b-a} = 0$. Thus, $f'(z) = \frac{f(b)-f(a)}{b-a}$. □

Simplicio: I do not like that proof. How did some one think of that idea?

Galileo: While the proof of the theorem may appear artificial, the basic idea is to reduce the Mean Value Theorem to Rolle's theorem by subtracting the straight line $y = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$ from the function $f(x)$. The next version of the Mean Value Theorem is rewritten into a form similar to Taylor's Theorem, which we will consider shortly.

Theorem 11.3.2 (Mean Value Theorem 2). *If $f(x) : [a,b] \to \Re$ has the property that $f(x), f'(x)$ are continuous, then for every pair of points $x, x_0 \in (a,b)$ there is a point $z \in (a,b)$ such that $f(x) = f(x_0) + f'(z)(x - x_0)$.*

Proof. In the Mean Value Theorem 11.3.1 simply let $x_0 = a, x = b$, and substitute into the expression $f'(z) = \frac{f(b)-f(a)}{b-a}$ to get $f'(z) = \frac{f(x)-f(x_0)}{x-x_0}$. If we multiply both

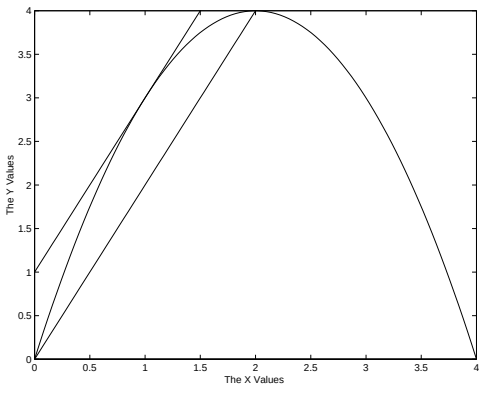


Figure 11.2: The Mean Value Theorem for $f(x) = 4 - (x-2)^2$ on $[0,2]$

sides of the equation by x_0, we see that $f(x) - f(x_0) = f'(z)(x - x_0)$ and the result follows. □

Simplicio: So what is this Mean Value Theorem good for?
Galileo: The next theorem allows us to estimate how much a function expands or contracts.

Corollary 11.3.3 (Corollary to the Mean Value Theorem). *If $f(x) : [a, b] \to \Re$ has the property that $f(x), f'(x)$ are continuous and $M = max\{|f'(x)| : x \in [a, b]\}$, then for every pair of points $x, x_0 \in [a, b]$ we know that $|f(x) - f(x_0)| \leq M|x - x_0|$.*

Proof. By Mean Value Theorem 2 11.3.2 we know that for any two points $x, x_0 \in [a, b]$, there is a point $z \in [a, b]$ so that $f(x) - f(x_0) = f'(z)(x - x_0)$.

Thus, if $M = max\{|f'(x)| : x \in [a, b]\}$, then $|f(x) - f(x_0)| = |f'(z)||x - x_0| \leq M|x - x_0|$. □

Galileo: From an intuitive perspective, the Corollary states that if you drive your rusty old car from your house to a party at your grandmother's house 80 miles away and the jalopy cannot go faster than 45mph, then you had better leave in plenty of time or you will be late.
Simplicio: If you allow only an hour, then you will be assured of being late.
Galileo: There it is, a mathematical fact.

Example 11.3.1. *If $f(x) = \sin(x)$, then we will show that $|\sin(x) - \sin(y)| \leq |x - y|$ for any two real numbers x and y.*

However, since $f'(x) = \cos(x)$ for all $x \in \Re$ and $\cos(x) \leq 1$ for all $x \in \Re$, we know by the Mean Value Theorem 11.3.3 that $|\sin(x) - \sin(y)| \leq |x - y|$ for all $x, y \in \Re$.

How about if you practice on a couple of the following problems?

Exercise Set 11.3.

1. If $f(x) = x^2 - 4, a = 0$, and $b = 1$, then find the point z guaranteed by the Mean Value Theorem 11.3.1. (Graph the function $y = f(x)$.)

2. If $f(x) = x^3 - 4, a = 1$, and $b = 5$, then find the point z guaranteed by the Mean Value Theorem 11.3.1. (Graph the function $y = f(x)$.)

3. If $f(x) = e^x$ and $x, y \in [0, 1]$, then show that $|e^x - e^y| \leq 3|x - y|$. (Graph the function $y = f(x)$.)

4. If $K > 0$ and $T(x) = x - \frac{x^2-K}{2x} = \frac{1}{2}x + \frac{K}{2x}$, then show that $|T(x) - T(y)| \leq \frac{1}{2}|x - y|$ for any two real numbers $x, y \in [\sqrt{K}, \infty)$. (We will see this problem again when we analyze the Archimedes/Heron square root algorithm. Graph the function $y = T'(x)$.)

5. If $K > 0$ and $T(x) = x - \frac{x^3-K}{3x^2} = \frac{2}{3}x + \frac{K}{3x^2}$, then show that $|T(x) - T(y)| \leq \frac{2}{3}|x - y|$ for any two real numbers $x, y \in [\sqrt[3]{K}, \infty)$. (We will see this problem again when we analyze the cube root algorithm. Graph the function $y = T'(x)$.)

6. If $T(x) = \frac{1}{3}\cos(2x) - 3$, then show that $|T(x) - T(y)| \leq \frac{2}{3}|x - y|$ for any two real numbers x and y.

11.4 Uniform Continuity

Galileo: We now turn to the topic of uniform continuity.

Simplicio: Yet a second type of continuity? Isn't one enough?

Galileo: It really isn't a new type of continuity, but rather is involved in the choice of δ when you have been challenged by an ϵ.

Simplicio: I have no idea what you are talking about.

Galileo: Let us begin with a couple of examples.

Example 11.4.1. *If $f(x) : \Re \to \Re$ is defined by the rule $f(x) = 2x, x_0 \in \Re$, and $\epsilon > 0$ is given, then how small must $\delta > 0$ be chosen to guarantee that if $x \in (x_0 - \delta, x_0 + \delta)$, then $f(x) \in (f(x_0) - \epsilon, f(x_0) + \epsilon)$?*

Simplicio: Even I can answer that question. All we have to do is choose $\delta = \frac{\epsilon}{2}$ because to check that this choice works we simply note that $|f(x) - f(x_0)| = |2x - 2x_0| = 2|x - x_0| < 2\frac{\epsilon}{2} = \epsilon$.

Galileo: Very good. Now consider a second example.

Example 11.4.2. *If $f(x) : \Re \to \Re$ is defined by the rule $f(x) = x^2, x_0 \in \Re$, and $\epsilon > 0$ is given, then how small must $\delta > 0$ be chosen to guarantee that if $x \in (x_0 - \delta, x_0 + \delta)$, then $f(x) \in (f(x_0) - \epsilon, f(x_0) + \epsilon)$?*

Simplicio: This question is a bit harder, but let's figure it out. If we assume that $\delta < 1$, then $|x| < |x_0| + 1$. Thus, $|f(x) - f(x_0)| = |x^2 - x_0^2| = |(x - x_0)(x + x_0)| = |x - x_0||x + x_0| < \delta(|x| + |x_0|) < \delta(2|x_0| + 1)$. Thus, if I choose $\delta > 0$ less than the minimum of 1 and $\delta < \frac{\epsilon}{2|x_0|+1}$, then I am done.

Galileo: You are getting good at these computations. I am impressed. OK, what is the difference between the choice of δ in these two examples?

Virginia: In the first example, the choice of δ does not depend on the given point x_0. Namely, $\delta = \frac{\epsilon}{2}$ for any point x_0. In the second example, the choice of δ must be made smaller for larger values of x_0.

Galileo: In other words, in the first example, the choice of δ is independent of the point x_0, while in the second example, the choice of δ depends on x_0. Let's modify the second example and see if you can figure out what the choice should be this time.

Example 11.4.3. *If $x_0 \in [-100, 100]$ and $\epsilon > 0$ are given and $f(x) : [-100, 100] \to \Re$ is defined by the rule $f(x) = x^2$, then how small must $\delta > 0$ be chosen to guarantee that if $x \in (x_0 - \delta, x_0 + \delta)$, then $f(x) \in (f(x_0) - \epsilon, f(x_0) + \epsilon)$?*

Simplicio: This question is easy. If we choose $\delta = \frac{\epsilon}{200}$, then $|f(x) - f(x_0)| = |x^2 - x_0^2| = |(x - x_0)(x + x_0)| = |x - x_0||100 + 100| < \delta(200) < \frac{\epsilon}{200}200 = \epsilon$.

Thus, we are done.

Galileo: Very good. Now, what is the difference between the second and third examples.

Simplicio: Obviously, the only difference is that the interval in the third example is closed and bounded.

Virginia: And you choose $\delta = \frac{\epsilon}{M}$, where $M \geq |f'(x)|$ for all x in the interval.

Galileo: Guess what! You have discovered two new theorems.

Theorem 11.4.1 (Uniform Continuity 1). *If X is an interval in $\Re$ and $f(x) : X \to \Re$ is a differentiable function with the property that $|f'(x)| < M$ for all $x \in X$, then for any $x_0 \in X$ and any $\epsilon > 0$, there is a $\delta > 0$ with the property that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$.*

Proof. Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose $\delta = \frac{\epsilon}{M+1}$.

Step 3. The Check:

If $|x - x_0| < \delta$, then by the Mean Value Theorem 11.3.3

$|f(x) - f(x_0)| \leq M|x - x_0| < M\delta < M\frac{\epsilon}{M+1} = \frac{M}{M+1}\epsilon < \epsilon$.

□

Galileo: The next theorem provides the generality we desire. Note the hypotheses have been changed so that it is no longer necessary to assume that the function is differentiable. However, to make up for this weaker assumption, we must assume that the interval is closed and bounded.

Theorem 11.4.2 (Uniform Continuity 2). *If $f(x) : [a, b] \to \Re$ is a function with the property that $f(x)$ is continuous at each $x \in [a, b]$, then for any $\epsilon > 0$ there is a $\delta > 0$ with the property that if $x_0, x \in [a, b]$ have distance $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$.*

Proof. By way of contradiction, assume that there is no such *delta*.

If this is true, then we have the following cases.

Case $n = 1$.

For $\delta_1 = \frac{1}{1} = 1$ we can find points $y_1, z_1 \in [a, b]$ with the property that $|y_1 - z_1| < \delta_1 = \frac{1}{1}$ and $|f(y_1) - f(z_1)| \geq \epsilon$.

Case $n = 2$.

For $\delta_2 = \frac{1}{2}$ we can find points $y_2, z_2 \in [a, b]$ with the property that $|y_2 - z_2| < \delta_2 = \frac{1}{2}$ and $|f(y_2) - f(z_2)| \geq \epsilon$.

Case $n = 3$.

For $\delta_3 = \frac{1}{3}$ we can find points $y_3, z_3 \in [a, b]$ with the property that $|y_3 - z_3| < \delta_3 = \frac{1}{3}$ and $|f(y_3) - f(z_3)| \geq \epsilon$.

Case $n = n$.

For $\delta_n = \frac{1}{n}$ we can find points $y_n, z_n \in [a, b]$ with the property that $|y_n - z_n| < \delta_n = \frac{1}{n}$ and $|f(y_n) - f(z_n)| \geq \epsilon$.

Since we have assumed the interval $[a, b]$ is closed and bounded, the sequence $\{y_n\}_{n=1}^{\infty}$ has a convergent subsequence. Without loss of generality, we can assume the sequence $\{y_n\}_{n=1}^{\infty}$ converges to some point x_0. Since the function $f(x)$ is continuous at x_0, we can find a $\delta > 0$ with the property that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \frac{\epsilon}{2}$.

Choose an integer N sufficiently large that if $n \geq N$, then $|y_n - x_0| < \frac{\delta}{2}$.

Since $|y_n - x_0| < \frac{\delta}{2} < \delta, |f(y_n) - f(x_0)| < \frac{\epsilon}{2}$.

Since $|z_n - x_0| = |z_n - y_n + y_n - x_0| \leq |z_n - y_n| + |y_n - x_0| < \frac{1}{n} + \frac{\delta}{2} \leq \frac{1}{N} + \frac{\delta}{2} \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta, |f(z_n) - f(x_0)| < \frac{\epsilon}{2}$.

Combining these last two pieces of information, we see that $|f(y_n) - f(z_n)| = |f(y_n) - f(x_0) + f(x_0) - f(z_n)| \leq |f(y_n) - f(x_0)| + |f(x_0) - f(z_n)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$.

Thus, we have a contradiction to our assumption that $|f(y_n) - f(z_n)| \geq \epsilon$ for all integers n.

Thus, the theorem is proved. □

Simplicio: I have the creepy feeling I have seen that argument before.

Galileo: You have. As part of the proof of the Extremum Theorem, we showed that a continuous function on a closed bounded interval is bounded. The argument is the same except for the phrasing. In fact, our theorem on uniform continuity can be used to show a continuous function on a closed bounded interval is bounded. The argument is straightforward.

Simplicio: Well, why didn't you give us this argument before? It would have been more economical.

Galileo: True, but it would have seemed a bit contrived. In any case, repetition is a great teacher.

Simplicio: I have one last question. Why did we go to the trouble to discuss uniform continuity? It seems like a detail.

Galileo: While you are correct that uniform continuity is a detail for an applications person like yourself, it is the key idea in the proof that a continuous function on a closed bounded interval is integrable.

Simplicio: As far as I am concerned, any function can be integrated.

Galileo: The continuous functions on a closed bounded interval form a generally well behaved collection. They possess the extremum and intermediate value properties. As we will see momentarily, they are also integrable. Thus, they form an important subset of the collection of integrable functions. In some sense the collection of continuous functions are a nice subset of the collection of integrable functions. In an effort to isolate the concept of Uniform Continuity and unify the two theorems Theorem 11.4.1 and Theorem 11.4.2, we make the following definition.

Definition 11.4.3. *If X is an interval in $\Re$ and $f(x) : X \to \Re$ is a function with the property that $\epsilon > 0$ there is a $\delta > 0$ with the property that if $x_0, x \in X$ have distance $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$.*

Exercise Set 11.4.

1. If $f(x) = x^3 + 3x$ is defined on the interval $[-2, 2]$ and $\epsilon > 0$, then find a $\delta > 0$ with the property that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$ for all $x, x_0 \in [-2, 2]$.

2. If $f(x) = x^4 + x$ is defined on the interval $[-3, 3]$ and $\epsilon > 0$, then find a $\delta > 0$ with the property that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$ for all $x, x_0 \in [-3, 3]$.

3. If $f(x) = 5|x| + 3|x - 1|$ is defined on the interval $[-2, 2]$ and $\epsilon > 0$, then find a $\delta > 0$ with the property that if $|x - x_0| < \delta$, then $|f(x) - f(x_0)| < \epsilon$ for all $x, x_0 \in [-2, 2]$.

11.5 Integration

Galileo: Since our proofs of both Taylor's Theorem and the Fundamental Theorem of Calculus require the Intermediate Value Theorem for Integrals, I guess we have no choice but to define the integral of a function.

Simplicio: More theory?

Galileo: While you dislike the theory, the definition is in the same spirit as the definitions we gave for limits of sequences and functions. If you have forgotten those details, go back and look at your notes from those discussions.

Virginia: You mean you can phrase the definition in terms of a challenge?

Galileo: Absolutely. First, we have to define the ideas of a partition and a refinement of a partition. These terms will appear in the definition of the integral.

Definition 11.5.1. *A partition of an interval $[a,b]$ is a finite ordered set of points of the form $P = \{a = x_0 < x_1 < x_2 < \cdots < x_n = b\}$.*

Definition 11.5.2. *If P and P' are two partitions of an interval $[a,b]$, then P' is a refinement of P if every member of P' is a member of P.*

Definition 11.5.3. *A bounded function $f(x) : [a,b] \to \Re$ is integrable with integral $\int_a^b f(x)\ dx$ if for every $\epsilon > 0$, there is a partition P with the property that if $P' = \{a = x_0 < x_1 < x_2 < \cdots < x_n = b\}$ is any refinement of P and for any choice of points $x_k^* \in [x_k, x_{k+1}]$, then*

$$|\sum_{k=0}^{n-1} f(x_k^*)(x_{k+1} - x_k) - \int_a^b f(x)\ dx| < \epsilon.$$

Since we have an excess of notation, we will use the notation $S(P) = \sum_{k=0}^{n-1} f(x_k^*)(x_{k+1} - x_k)$ to denote the sums approximating the integral. We will write this sum with the understanding that $x_k^* \in [x_k, x_{k+1}]$. With this notation we can reformulate the definition a bit more succinctly.

Definition 11.5.4. *A bounded function $f(x) : [a,b] \to \Re$ is integrable with integral $\int_a^b f(x)\ dx$ if for every $\epsilon > 0$, there is a partition P with the property that if P' is any*

refinement of P, then

$$|S(P') - \int_a^b f(x)\ dx| < \epsilon.$$

Simplicio: This definition seems unnecessarily complicated.

Virginia: Actually, no. I can already see that it can once again be phrased as a three step process with the usual suspects: Challenge, Choice, and Check. If I challenge you with an $\epsilon > 0$, then you are required to find me a partition P (The Choice) with the property that any "bigger" partition P' has the property that $S(P')$ is within ϵ of the integral $\int_a^b f(x)\ dx$. Once again the ϵ is a measure of our distance from the desired answer. Not complicated at all.

Galileo: The next proposition encapsulates the two most important facts concerning integrals. The first states that the integrable of the sum is the sum of the integrals. The second states that we can pull constants across the integral sign. Recall that derivatives also had these two properties. Together these two properties state that the derivative and integral are linear transformations and thus lie under the purview of Linear Algebra. More about this later.

Proposition 11.5.5 (Linearity Property for Integrals). *If $f(x), g(x) : [a, b] \to \Re$ are integrable and K is a real number, then*

1. $\int_a^b f(x) + g(x)\ dx = \int_a^b f(x)\ dx + \int_a^b g(x)\ dx$.
   (The integral of the sum equals the sum of the integrals.)

2. $\int_a^b Kf(x)\ dx = K \int_a^b f(x)\ dx$.
   (Pulling constants.)

Proof. Fact 1. Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose a partition P with the property that if P' is any refinement of P, then

1. $|S_f(P') - \int_a^b f(x)\ dx| < \frac{\epsilon}{2}$ and

2. $|S_g(P') - \int_a^b g(x)\ dx| < \frac{\epsilon}{2}$,

where $S_f(P')$ and $S_g(P')$ denote the approximating sums associated with $f(x)$ and $g(x)$, respectively. (We assume that the choice of x_k^* is the same for both approximations.)

Step 3. The Check: Since $S_f(P') + S_g(P') = S_{f+g}(P')$,

$$\begin{aligned}|S_{f+g}(P') - (\int_a^b f(x)\ dx + \int_a^b g(x)\ dx)| =& |S_f(P') + S_g(P') - (\int_a^b f(x)\ dx + \int_a^b g(x)\ dx)| \\ =& |S_f(P') - \int_a^b f(x)\ dx + S_g(P') - \int_a^b g(x)\ dx| \\ \leq& |S_f(P') - \int_a^b f(x)\ dx| + |S_g(P') - \int_a^b g(x)\ dx| \\ <& \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.\end{aligned}$$

Proof of Fact 2.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose a partition P with the property that if P' is any refinement of P, then

$|S_f(P') - \int_a^b f(x)\ dx| < \frac{\epsilon}{|K|+1}$.

Step 3. The Check:

If P' is any refinement of P, then $|S_{Kf}(P') - K\int_a^b f(x)\,dx| = |KS_f(P') - K\int_a^b f(x)\,dx| = |K||S_f(P') - \int_a^b f(x)\ dx| < |K|\frac{\epsilon}{|K|+1} < \epsilon$. □

Virginia: Those proofs weren't bad at all. They were almost the same as our limit facts.

Simplicio: But why are they called linearity properties? I don't see any proportions.

Galileo: Do you remember the definition of linear transformation from your studies of Linear Algebra?

Simplicio: I am not sure what you are getting at.

Galileo: If you remember, a transformation $L : U \to V$ from a vector space U to a vector space V is called linear if it satisfies two properties:

1. $L(\mathbf{u}_1 + \mathbf{u}_2) = L(\mathbf{u}_1) + L(\mathbf{u}_2)$ for all $\mathbf{u}_1, \mathbf{u}_2 \in U$ and

2. $L(K\mathbf{u}) = KL(\mathbf{u})$ for all $\mathbf{u} \in U$ and $K \in \Re$.

Of course, the vector space of integrable functions is infinite dimensional.

Simplicio: I have no use for infinite dimensional vector spaces and their transformations.

Galileo: But you will.

Simplicio: Oh.

Galileo: The global strategy will be to approximate infinite dimensional spaces by finite dimensional spaces and linear transformations by matrices. You have heard of a matrix, haven't you? Derivatives, integrals, and Fourier Transformations all operate in the infinite dimensional arena. Fortunately, they all have finite matrix representations. Thus, Linear Algebra will be involved.

Simplicio: OK, OK. An integration example please.

Galileo: Before we present an example, I would like to present two more notations for the lower and upper sums.

Definition 11.5.6. *If $f(x) : [a, b] \to \Re$ is a bounded function, P is a partition of $[a, b]$, and $\underline{z}_k \in [x_k, x_{k+1}]$ has been chosen with the property that $m_k = f(z_k) \leq f(x)$ for all $x \in [x_k, x_{k+1}]$, then define the lower sum on P by $\underline{S}(P) = \sum_{k=0}^{n-1} f(\underline{z}_k)(x_{k+1} - x_k) = \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k)$.*

Definition 11.5.7. *If $f(x) : [a, b] \to \Re$, P is a partition of $[a, b]$, and $\overline{z}_k \in [x_k, x_{k+1}]$ has been chosen with the property that $f(x) \leq f(z_k) = M_k$ for all $x \in [x_k, x_{k+1}]$, then define the upper sum on P by $\overline{S}(P) = \sum_{k=0}^{n-1} f(\overline{z}_k)(x_{k+1} - x_k) = \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k)$.*

Virginia: Actually, I hate to be picky, but I have a complaint about these last two definitions. If we assume the function $f(x)$ is continuous, we know we can find the points $\underline{z}_k$ and $\overline{z}_k$. However, if we don't make this assumption about $f(x)$, we might not be able to find such points. What do we do then?

Galileo: Good point. We would be on safer ground if we defined them more carefully using the concepts greatest lower bound and the least upper bound.

Definition 11.5.8. *If $f(x) : [a,b] \to \Re$ is bounded and $P = \{a = x_0 < x_1 < x_2 < \cdots < x_n = b\}$ is any partition of $[a,b]$, then define the notation $m_k = glb\{f(x) : x \in [x_k, x_{k+1}]$ and $M_k = lub\{f(x) : x \in [x_k, x_{k+1}]$.*

Virginia: I see why you are assuming your functions are bounded. If you had unbounded functions, the quantities m_k and M_k could be infinite.
Galileo: You are correct. We are trying to keep our discussion as simple as possible. Let us begin by making a number of observations.

Proposition 11.5.9. *If $f(x) : [a,b] \to \Re$ is bounded and P is any partition of $[a,b]$, then the lower and upper sums exist and $\underline{S}(P) \leq S(P) \leq \overline{S}(P)$.*

Proof. Since $m_k \leq f(x_k^*) \leq M_k$ for all $x \in [x_k, x_{k+1}]$ and all $k = 0, 1, \ldots, n-1$,

$$\underline{S}(P) = \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k) \leq S(P) = \sum_{k=0}^{n-1} f(x_k^*)(x_{k+1} - x_k)$$
$$\leq \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k) = \overline{S}(P).$$

Thus, we are done. □

Proposition 11.5.10. *Let $f(x) : [a,b] \to \Re$ be bounded. If P and P' are any two partitions of $[a,b]$ where P' is a refinement of P, then $\underline{S}(P) \leq \underline{S}(P') \leq \overline{S}(P') \leq \overline{S}(P)$.*

Proof. Simplicio: Even I can see that this proposition is true.
Galileo: But, you might want to be a bit careful and increase the partition P to P' by adding one point at a time. This technique is called induction. □

Simplicio: Our example please.
Galileo: OOPS! We need to remind you of one more detail. We need the sum formula for the arithmetic series.

Proposition 11.5.11. $\sum_{k=1}^{n} k = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.

Proof. Virginia: I remember the proof.

If we let $S_n = 1 + 2 + \cdots + n$, then

$$\begin{array}{rcccccccc} S_n & = & 1 & + & 2 & + & \ldots & + & n \\ S_n & = & n & + & (n-1) & + & \ldots & + & 1 \\ 2S_n & = & (n+1) & + & (n+1) & + & \ldots & + & (n+1) \end{array}$$

Since the quantity $2S_n$ is written as n sums of the number $n+1$, we see that $2S_n = n(n+1)$. Thus, $S_n = \frac{n(n+1)}{2}$. □

Virginia: Now we should be ready for our example.

Example 11.5.1. *Galileo: How about if we compute the area under the curve $y = f(x) = x$ for $x \in [0,1]$?*

Simplicio: Sure, but I already see the enclosed region is a right triangle with base and height equal to one. The answer equals $\frac{1}{2}$.

Galileo: We shall do as the young lady instructs.

Virginia:

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Begin by choosing an integer n with the property that $n > \frac{1}{\epsilon}$.

Now choose the partition P to be $n+1$ equally spaced points between 0 and 1. In other words, $P = \{0 = x_0 < x_1 < x_2 < \cdots < x_n = 1\}$, where $x_k = \frac{k}{n}$, for $k = 0, 1, 2, \ldots, n$.

Step 3. The Check:

Let P' be any refinement of P with x_k^* any choice of points in the interval $[x_k, x_{k+1}]$. Before we discuss P', let's make a couple of observations about P. Since $x_{k+1} - x_k = \frac{1}{n}$ and $m_k = \frac{k}{n}$, for all $k = 0, 1, \ldots, n-1$,

$$\begin{aligned}\underline{S}(P) &= \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k) \\ &= \sum_{k=0}^{n-1} \frac{k}{n}\frac{1}{n} \\ &= \frac{1}{n^2}\sum_{k=0}^{n-1} k \\ &= \frac{1}{n^2}\frac{(n-1)n}{2} \\ &= \frac{1}{2}\frac{n-1}{n}.\end{aligned}$$

Similarly, since $M_k = \frac{k+1}{n}$, *for all* $k = 0, 1, \ldots, n-1$,

$$\begin{aligned}\overline{S}(P) &= \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k) \\ &= \sum_{k=0}^{n-1} \frac{k+1}{n}\frac{1}{n} \\ &= \frac{1}{n^2}\sum_{k=0}^{n-1}(k+1) \\ &= \frac{1}{n^2}\frac{n(n+1)}{2} \\ &= \frac{1}{2}\frac{n+1}{n}.\end{aligned}$$

Thus,

$$\frac{1}{2}\frac{n-1}{n} = \underline{S}(P) \leq \underline{S}(P') \leq S(P') \leq \overline{S}(P') \leq \overline{S}(P) = \frac{1}{2}\frac{n+1}{n}.$$

Since we have chosen $n > \frac{1}{\epsilon}$ *and* $\frac{1}{2}\frac{n+1}{n} - \frac{1}{2}\frac{n-1}{n} = \frac{2}{2n} = \frac{1}{n}$, *we can see that* $|S(P') - \frac{1}{2}| < \frac{1}{n} < \epsilon$. *Thus,* $\int_0^1 x\, dx = \frac{1}{2}$.

Virginia: Since each estimate of the integral is squeezed between a bit less than $\frac{1}{2}$ and a bit more than $\frac{1}{2}$, I see we have a squeezing type process taking place here.

Simplicio: OK, but I knew before we started that a triangle with height and base equal to one has area equal to $\frac{1}{2}$.

Example 11.5.2. *Galileo: OK, then how do you compute the area under the parabola* $y = f(x) = x^2$, *for* $x \in [0, 1]$?

Simplicio: I would use my antiderivatives from Calculus.

Galileo: But, what if you were Archimedes? He had no antiderivatives.

Simplicio: I would be in trouble.

Galileo: While we won't give his proof, the next proposition provides the key to a proof he would appreciate. Virginia, how about if you lead the way again?

Proposition 11.5.12. $\sum_{k=1}^{n} k^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$.

Proof. Note the following special cases.

If $n = 1$, then $1^2 = \frac{1(1+1)(2+1)}{6}$.

If $n = 2$, then $1^2 + 2^2 = \frac{2(2+1)(2*2+1)}{6}$.

If $n = 3$, then $1^2 + 2^2 + 3^2 = \frac{3(3+1)(2*3+1)}{6}$.

The formal proof is by induction. □

Virginia: Using the definition, we simply go through the same steps as before.

Step 1. The Challenge:

Let $\epsilon > 0$ *be given.*

Step 2. The Choice:

Begin by choosing an integer n *with the property that* $n > \frac{1}{\epsilon}$.

Now choose the partition P *to be* $n + 1$ *equally spaced points between* 0 *and* 1. *In other words,* $P = \{0 = x_0 < x_1 < x_2 < \cdots < x_n = 1\}$, *where* $x_k = \frac{k}{n}$, *for* $k = 0, 1, 2, \ldots, n$.

Step 3. The Check:

Let P' *be any refinement of* P *with* x_k^* *any choice of points in the interval* $[x_k, x_{k+1}]$. *Before we discuss* P', *let's make a couple of observations about* P. *Since* $x_{k+1} - x_k = \frac{1}{n}$ *and* $m_k = (\frac{k}{n})^2$, *for all* $k = 0, 1, \ldots, n - 1$,

$$\begin{aligned}\underline{S}(P) =& \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k)\\ =& \sum_{k=0}^{n-1} (\frac{k}{n})^2 \frac{1}{n}\\ =& \frac{1}{n^3} \sum_{k=0}^{n-1} k^2\\ =& \frac{1}{n^3} \frac{(n-1)n(2n-1)}{6}\\ =& \frac{1}{6} \frac{(n-1)(2n-1)}{n^2}.\end{aligned}$$

Similarly, since $M_k = (\frac{k+1}{n})^2$, *for all* $k = 0, 1, \ldots, n-1$,

$$\begin{aligned}\overline{S}(P) =& \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k)\\ =& \sum_{k=0}^{n-1} (\frac{k+1}{n})^2 \frac{1}{n}\\ =& \frac{1}{n^3} \sum_{k=0}^{n-1} (k+1)^2\\ =& \frac{1}{n^3} \frac{n(n+1)(2n+1)}{6}\\ =& \frac{1}{6} \frac{(n+1)(2n+1)}{n^2}.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{1}{6} \frac{(n-1)(2n-1)}{n^2} =& \underline{S}(P)\\ \leq& \underline{S}(P')\\ \leq& S(P')\\ \leq& \overline{S}(P')\\ \leq& \overline{S}(P)\\ =& \frac{1}{6} \frac{(n+1)(2n+1)}{n^2}.\end{aligned}$$

Since we have chosen $n > \frac{1}{\epsilon}$,

$$\begin{aligned}\overline{S}(P') - \underline{S}(P') \leq & \frac{1}{6}\frac{(n+1)(2n+1)}{6n^2} - \frac{1}{6}\frac{(n-1)(2n-1)}{6n^2} \\ = & \frac{2n^2+3n+1}{6n^2} - \frac{2n^2-3n+1}{6n^2} \\ = & \frac{6n}{6n^2} = \frac{1}{n},\end{aligned}$$

and both $S(P')$ *and* $\frac{2}{6}$ *are trapped between* $\underline{S}(P')$ *and* $\overline{S}(P')$, *these estimates show that* $|S(P') - \frac{2}{6}| < \frac{1}{n} < \epsilon$. *Thus,* $\int_0^1 x^2 \, dx = \frac{1}{3}$.

Simplicio: These examples are not as bad as I would have expected. However, how did you know that $m_k = (\frac{k}{n})^2$ and $M_k = (\frac{k+1}{n})^2$?

Virginia: Since the function $y = f(x) = x^2$ is increasing on the interval $[x, x_{k+1}]$, the lowest point on the curve occurs at the left endpoint x_k. Thus, $m_k = (x_k)^2 = (\frac{k}{n})^2$. Similarly, the value of M_{k+1} is computed at the right endpoint so that $M_k = (x_{k+1})^2 = (\frac{k+1}{n})^2$.

Example 11.5.3. *Galileo: How about if we show that* $\int_0^1 x^3 \, dx = \frac{1}{4}$? *The only fact we need is that* $\sum_{k=1}^n k^3 = (\frac{n(n+1)}{2})^2$.

Simplicio: Holy Mother of Jesus, save me from this maniac. Let's move on. I would rather we do it Isaac Newton's way.

Galileo: So you do appreciate a good theorem when you see one! OK, we will leave it for an exercise.

Galileo: OK, now it is time to move on to inequalities. Note that the next proposition is analogous to the squeezing theorem for sequences. Unfortunately, just as the previous squeeze involved a proof by contradiction, the current proof does as well.

Simplicio: But, why can't we avoid new proofs?

Galileo: Sadly, we did not define the integral in terms of sequences.

Simplicio: I can smell that contrapositive already.

Proposition 11.5.13 (Monotone Property for Integrals). *If* $f(x), g(x) : [a,b] \to \Re$ *are bounded, integrable, and* $f(x) \leq g(x)$ *for all* $x \in [a,b]$, *then* $\int_a^b f(x) \, dx \leq \int_a^b g(x) \, dx$.

Proof. Begin by noting that if P is any partition of $[a,b]$, then our assumption that $f(x) \leq g(x)$ for all $x \in [a,b]$, implies that

$$S_f(P) = \sum_{k=0}^{n-1} f(x_k^*)(x_{k+1} - x_k) \leq \sum_{k=0}^{n-1} g(x_k^*)(x_{k+1} - x_k) = S_g(P).$$

By way of contradiction assume that $\int_a^b f(x)\ dx > \int_a^b g(x)\ dx$. We will show this assumption leads to the absurdity that the number $S_f(P)$ is strictly less than itself.

Step 1. The Choice of epsilon:

Let $\epsilon = \frac{\int_a^b f(x)\ dx - \int_a^b g(x)\ dx}{2} > 0$.

(Since we are proving the contrapositive, we get to choose ϵ to be any number we want. The smart choice is half the distance between the integrals $\int_a^b f(x)\ dx$ and $\int_a^b g(x)\ dx$.)

With this choice of ϵ, we know $2\epsilon = \int_a^b f(x)\ dx - \int_a^b g(x)\ dx$. If we write $2\epsilon = \epsilon + \epsilon$ and move one integral to the other side of the equation, then

$$\int_a^b g(x)\ dx + \epsilon = \int_a^b f(x)\ dx - \epsilon.$$

Step 2. The Choice of the partition P :

(We now get to choose the partition based on the choice of ϵ.)

Choose a partition P with the property that if P' is any refinement of P, then $|S_f(P') - \int_a^b f(x)\ dx| < \epsilon$ and $|S_g(P') - \int_a^b g(x)\ dx| < \epsilon$.

Step 3. The Contradiction:

(We now show that the number $S_f(P')$ is less than itself.)

Since $|S_f(P') - \int_a^b f(x)\ dx| < \epsilon$, $\int_a^b f(x)\ dx - \epsilon < S_f(P')$.

Since $|S_g(P') - \int_a^b g(x)\ dx| < \epsilon$, $S_g(P') < \int_a^b g(x)\ dx + \epsilon$.

Since $S_f(P') \leq S_g(P')$, we see that

$$S_f(P') \leq S_g(P') < \int_a^b g(x)\ dx + \epsilon = \int_a^b f(x)\ dx - \epsilon < S_f(P').$$

Thus, $S_f(P') < S_f(P')$, a contradiction since no number can be less than itself.

Don't let all the notation confuse you. The proof is easier than it looks. Draw a picture.

□

Simplicio: That proposition seems obvious to me. I don't see why it was necessary to prove it.

Galileo: The next corollary will provide the starting point in our proof of the Mean Value Theorem for Integrals.

Corollary 11.5.14 (Integral Bounds). *If $f(x) : [a, b] \to \Re$ is bounded, integrable, and $m \leq f(x) \leq M$ for all $x \in [a, b]$, then $m(b-a) \leq \int_a^b f(x)\ dx \leq M(b-a)$.*

Proof. This corollary follows immediately from the previous proposition.

First, set $g(x) = M$ for all $x \in [a, b]$. Thus, $\int_a^b f(x)\ dx \leq \int_a^b M\ dx = M(b-a)$.

Second, set $g(x) = m$ for all $x \in [a, b]$. Thus, $m(b-a) = \int_a^b m\ dx \leq \int_a^b f(x)\ dx$.

□

Simplicio: I have a quick question. Why all this generality in the definition of the integral? In other words, as soon as you decided to compute, you immediately chose your partition to have equally spaced points. Why not always limit your partitions to equally spaced points?

Galileo: Excellent question! We have partitions with variable length intervals for both practical and theoretical reasons. A practical reason is that the integral can be estimated more efficiently and accurately if we have shorter intervals where the function $y = f(x)$ is changing rapidly and longer intervals where the function is changing more slowly. If the function happens to be differentiable, then the computations will be improved if the lengths of intervals are chosen relatively small in regions where the derivative is large and relatively long in regions where the derivative is close to zero. This process can be automated.

Example 11.5.4. *For example, our friends in statistics are always having to approximate integrals like $\int_{-10}^{10} e^{-x^2}\ dx$. Since the function $f(x) = e^{-x^2}$ and its first derivative are virtually zero on the intervals $[-10, -5]$ and $[5, 10]$, our partition $P = \{-10 = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = 10\}$ can be chosen so that $x_1 = -5$. and $x_{n-1} = 5$. The intermediate points can be clustered in the interval $[-5, 5]$.*

Virginia: And the theoretical reasons?

Galileo: If we use the definition of integral we just gave, it is easy to prove the rule $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$ for $c \in [a, b]$. We simply add the point c to an arbitrary partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ to create a refinement $P' = \{a = x_0 < x_1 < \cdots < x_k < c < x_{k+1} < \cdots < x_n = b\}$. The proof of this fact is a bit of a nuisance if we had only considered equally spaced partitions. We will prove this fact momentarily.

Simplicio: Is that all?

Galileo: As you will see during this discourse, many techniques have difficulty when making approximations near the boundary. The Runge and Gibbs example stand out as examples of this type. Some of these problems can be alleviated when we choose our partition so that most of the points are clustered out near the boundary of the interval. For numerical integration, Gauss Quadrature provides an elegant way to make this choice.

Simplicio: Any other thoughts?

Galileo: While most of our discussions will be restricted to the 1-dimensional setting, most real applications take place in $2, 3$, or even higher dimensional spaces. While upper sums and lower sums may not be well-defined in these settings, the expression $S(P) = \sum_{k=0}^{n-1} f(x_k^*)(x_{k+1} - x_k)$ makes sense as long as the value $f(x_k^*)$ lies in a real vector space and the quantity $(x_{k+1} - x_k)$ is a real number. The other issue is convergence for the partial sums. However, if we define the metric on the range of the function so convergence implies convergence on each coordinate, then we are back to dimension one. Pythagoras does that for us. He is our man. This heavy-handed discussion implies that when we integrate a function of the form $r(t) = (x(t), y(t))$, we simply integrate the two coordinates separately.

Simplicio: Hmmm.

Virginia: I also have a question. When you computed the examples, you immediately turned to the lower and upper sums. If you have equally spaced points, then the lower and upper sums are sequences so you simply could have defined P_n to be the

partition of the interval $[a, b]$ with n equally spaced points. The integral can now be defined simply as the limit of the sequence $\lim_{n\to} \underline{S}(P_n) = \lim_{n\to} \overline{S}(P_n)$. While these two limits might not be equal, I doubt that happens. Can these limits differ?

Example 11.5.5 (A non-Integrable Function). *Galileo: Now you are asking for a bizarre example. However, the following function has the property that all the upper sums equal* $+1$*, while all the lower sums equal* -1*. Thus, it cannot be integrable.*

Definition 11.5.15 (A non-Integrable Function). *Define the function* $f(x) : [0, 1] \to \Re$ *by the rule*

$$f(x) = \begin{cases} -1 & if\ x \quad is\ a\ rational\ number \\ 1 & if\ x \quad is\ not\ a\ rational\ number \end{cases}.$$

Virginia: Yes, I can see that no matter what the choice of the partition, P, it will always be true that $m_k = -1$ *and* $M_k = 1$*.*

Simplicio: How so?

Virginia: Since there will always be a rational number x_k^* *between* x_k *and* x_{k+1}*,* $m_k = -1$*. Thus,* $\underline{S}(P) = -1$ *for any partition P. Since there will always be an irrational number* x_k^* *between* x_k *and* x_{k+1}*,* $M_k = 1$*. Thus,* $\overline{S}(P) = 1$*.*

Galileo: While this example makes the point that we should be careful, we won't use it much. However, it does set the stage for a criterion that guarantees the existence of the integral. The criterion is similar to the Cauchy criterion we had for sequences. In fact, the proof involves the same construction we went through for Cauchy sequences where all but a finite number of terms of a sequence are trapped in a nested sequence of intervals $[a_n, b_n]$, where $b_n - a_n < \frac{1}{n}$.

Galileo: OK, now it is time provide conditions, which guarantee the integral exists.

Simplicio: This discussion will be for the math majors.

Galileo: True, but it will reinforce your understanding of the definition of the integral.

Theorem 11.5.16 (Cauchy Integrability Criterion). *If* $f(x) : [a, b] \to \Re$ *is a bounded function, which has the property that for every* $\epsilon > 0$*, there is a partition P such that* $\overline{S}_f(P) - \underline{S}_f(P) < \epsilon$*, then* $f(x)$ *is integrable.*

Proof. The proof is constructive, where a sequence of partitions $\{P_n\}_{n=1}^{\infty}$ are found with the property that P_{n+1} refines P_n and $\overline{S}_f(P_n) - \underline{S}_f(P_n) < \frac{1}{n}$.

Case n = 1. Let $\epsilon = 1$.

Choose a partition P_1 with the property that $\overline{S}_f(P_1) - \underline{S}_f(P_1) < 1$.

Case n = 2. Let $\epsilon = \frac{1}{2}$.

Choose a partition P_2 with the property that $\overline{S}_f(P_2) - \underline{S}_f(P_2) < \frac{1}{2}$. Since refinement only forces the upper and lower sums to be closer, assume that P_2 refines P_1. (If it doesn't, then add the points of P_1 to P_2.)

Case n = 3. Let $\epsilon = \frac{1}{3}$.

Choose a partition P_3 with the property that $\overline{S}_f(P_3) - \underline{S}_f(P_3) < \frac{1}{3}$. Since refinement only forces the upper and lower sums to be closer, assume that P_3 refines P_2. (If it doesn't, then add the points of P_2 to P_3.)

Continue in this manner for arbitrary integers n to obtain a sequence of partitions with the property that

$$\underline{S}_f(P_1) \leq \underline{S}_f(P_2) \leq \cdots \leq \underline{S}_f(P_n) \leq \overline{S}_f(P_n) \leq \cdots \leq \overline{S}_f(P_2) \leq \overline{S}_f(P_1)$$

and $\overline{S}_f(P_n) - \underline{S}_f(P_n) < \frac{1}{n}$.

Since the sequence $\{\underline{S}_f(P_n)\}_{n=1}^{\infty}$ is bounded increasing, it converges to some number, call it $\int_a^b f(x)\,dx$.

Since the sequence $\{\overline{S}_f(P_n)\}_{n=1}^{\infty}$ is bounded decreasing, it also converges to some number.

Since $\overline{S}_f(P_n) - \underline{S}_f(P_n) < \frac{1}{n}$, the sequence $\{\overline{S}_f(P_n)\}_{n=1}^{\infty}$ also converges to $\int_a^b f(x)\,dx$.

Thus, the function $f(x)$ is integrable.

Simplicio: But, wait a minute. Don't you have to go through the same Challenge, Choose, and Check routine we did before?

Galileo: Of course, you are correct. Since you asked, here it is.

Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Choose an integer n with the property that $n > \frac{1}{\epsilon}$.

Now choose the partition $P = P_n$, where P_n denotes the partition we just constructed.

Step 3. The Check:

Let P' be any refinement of P with x_k^* any choice of points in the interval $[x_k, x_{k+1}]$. Since $\underline{S}_f(P_n) \leq S_f(P') \leq \overline{S}_f(P_n)$ and $\underline{S}_f(P_n) \leq \int_a^b f(x)\ dx \leq \overline{S}_f(P_n)$,

$|S_f(P') - \int_a^b f(x)\ dx| \leq \overline{S}_f(P_n) - \underline{S}_f(P_n) < \frac{1}{n} < \epsilon$.

Simplicio: You told me more than I wanted to know.

Virginia: But the argument really was the same as those given before. Namely, you simply trap the two numbers $S_f(P')$ and $\int_a^b f(x)\ dx$ in the interval $[\underline{S}_f(P_n), \overline{S}_f(P_n)]$. Since the length of this interval is less than ϵ, the two points can't be separated by more than ϵ. Thus, we are done. Think visually.

□

Virginia: What about the converse?

Galileo: The converse is easy because the integral is given to you for free. No infinite process is required.

Proposition 11.5.17. *If $f(x) : [a, b] \to \Re$ is a bounded integrable function, then it has the property that for every $\epsilon > 0$, there is a partition P such that $\overline{S}_f(P) - \underline{S}_f(P) < \epsilon$.*

Proof. Let $\epsilon > 0$ be given.

Since $f(x)$ is integrable with integral $\int_a^b f(x)\ dx$, there is a partition P with the property that if P' is any refinement of P, then $|S_f(P') - \int_a^b f(x)\ dx| < \frac{\epsilon}{2}$. Since the choice of the point x_k^* is arbitrary in the approximating sum $S_f(P) = \sum_{k=0}^{n-1} f(x_k^*)(x_{k+1} - x_k)$, we see that $|\overline{S}_f(P) - \int_a^b f(x)\ dx| < \frac{\epsilon}{2}$ and $|\underline{S}_f(P) - \int_a^b f(x)\ dx| < \frac{\epsilon}{2}$.

Thus,

$$\overline{S}_f(P) - \underline{S}_f(P) = \overline{S}_f(P) - \int_a^b f(x)\ dx + \int_a^b f(x)\ dx - \underline{S}_f(P) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

Galileo: The next proposition provides us with conditions when we know there will never be a problem integrating.

Theorem 11.5.18 (Continuous Functions are Integrable). *If $f(x) : [a, b] \to \Re$ is a function which is continuous at each $x \in [a, b]$, then $f(x)$ is integrable.*

Proof. To prove this proposition all we have to do is check the Cauchy Integrability Criterion. As with the definition of the integral, we have the Challenge, Choice, and Check.

Step 1. The Challenge:
Let $\epsilon > 0$ be given.

Step 2. The Choice:
By Theorem 11.4.2 we can find a $\delta > 0$ with the property that whenever $|x - x'| < \delta$, then $|f(x) - f(x')| < \frac{\epsilon}{b-a}$. Now choose $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ to be any partition with the property that $x_{k+1} - x_k < \delta$, for all $k = 0, 1, \ldots, n-1$.

Step 3. The Check:

By the Extremum Theorem 10.3.1 we know that there are points $x_k^*, x_k^{**} \in [x_k, x_{x+1}]$ with the property that $f(x_k^*) = m_k$ and $f(x_k^{**}) = M_k$.

Thus,

$$\begin{aligned}
\overline{S}_f(P) - \underline{S}_f(P) &= \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k) - \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k) \\
&= \sum_{k=0}^{n-1} (M_k - m_k)(x_{k+1} - x_k) \\
&= \sum_{k=0}^{n-1} (f(x_k^{**}) - f(x_k^{*}))(x_{k+1} - x_k) \\
&\leq \sum_{k=0}^{n-1} \frac{\epsilon}{(b-a)}(x_{k+1} - x_k) \\
&= \frac{\epsilon}{(b-a)} \sum_{k=0}^{n-1} (x_{k+1} - x_k) \\
&= \frac{\epsilon}{(b-a)}(b-a) = \epsilon.
\end{aligned}$$

□

Galileo: There it is.

Virginia: In fact, the argument is virtually the same as for the two examples we discussed earlier. The main difference is that we replaced those tricky summation formulas by Theorem 11.4.2, which actually makes the argument easier.

Galileo: And MUCH more general.

Simplicio: But there is one difference. With the examples we knew the answers before we started. Now we don't.

Galileo: True. However, for the special case when a function is differentiable, we can use Theorem 11.4.1 to help choose your partition. In particular, this theorem provides a tool for measuring the difference $\overline{S}_f(P) - \underline{S}_f(P)$.

Simplicio: How about an example?

Galileo:

Example 11.5.6. *If $f(x) = x^2$ on the interval $[-3, 3]$ and $\epsilon = \frac{1}{10^6}$, then find a partition $P = \{-3 = x_0 < x_1 < \cdots < x_n = 3\}$ with the property that $\overline{S}_f(P) - \underline{S}_f(P) < \epsilon = \frac{1}{10^6}$.*

Simplicio: Let me give this problem a try.

First, compute the first derivative $f'(x) = 2x$.

Second, compute the maximum value of $|f'(x)| = |2x|$ *on the interval* $[-3, 3]$. *For this function the maximum is* $M = 2 * 3 = 6$.

Third, choose $\delta > 0$ *sufficiently small that whenever* $|x - x'| < \delta$, *then* $|f(x) - f(x')| \leq M|x - x'| < M\delta$.

Fourth, the difference

$$\begin{aligned}
\overline{S}_f(P) - \underline{S}_f(P) &= \sum_{k=0}^{n-1} M_k(x_{k+1} - x_k) - \sum_{k=0}^{n-1} m_k(x_{k+1} - x_k) \\
&= \sum_{k=0}^{n-1} (M_k - m_k)(x_{k+1} - x_k) \\
&= \sum_{k=0}^{n-1} M\delta(x_{k+1} - x_k) \\
&\leq M\delta \sum_{k=0}^{n-1} (x_{k+1} - x_k) \\
&= 6\ \delta(3 - (-3)) = 36\ \delta.
\end{aligned}$$

Fifth, if we choose $\delta < \frac{\epsilon}{36} = \frac{1}{36}\epsilon = \frac{1}{36}\frac{1}{10^6}$, *then we guarantee that* $\overline{S}_f(P) - \underline{S}_f(P) < \frac{1}{10^6}$ *for any partition* $P = \{-3 = x_0 < x_1 < \cdots < x_n = 3\}$ *with the property that* $x_{k+1} - x_k < \delta$ *for all* $k = 0, 1, \ldots, n-1$.

Galileo: You should appreciate this control.

Simplicio: It might surprise you, but I do appreciate the ability to measure the error.

Galileo: In the spirit of Professor Polya, let us take a second look at this last example. Note that the key is being able to choose a partition P with the property that $\delta < \frac{\epsilon}{M(b-a)}$. The Mean Value Theorem 11.3.3 tells us the constant M is needed in the denominator.

Simplicio: It still bugs me that only justification for our discussion of Uniform Continuity is one inequality in the middle of Theorem 11.5.18. Mathematicians are neurotic.

Galileo: It is hard to argue with your thought, but they have a need to get it right. At some point your future employer may apply the same test to your performance. If you like neurotic details, you will love this next proposition, which states that if a function is integrable on a closed bounded interval, then it is integrable on any closed bounded subinterval.

Proposition 11.5.19. *If $f(x) : [a,b] \to \Re$ is integrable and $a \leq c \leq d \leq b$, then $\int_c^d f(x)\ dx$ exists.*

Proof. The proof of this proposition depends on Theorem 11.5.16. In order to use this theorem properly, we need to notate the function $f(x)$ restricted to the subinterval $[c,d]$ by $f_R(x) : [c,d] \to \Re$. (i.e.$f_R(x) = f(x)$ for all $x \in [c,d]$.) Now all that is required for the proof is to show that for every $\epsilon > 0$ we can find a partition $P_{f_R} = \{c = x_0 < x_1 < \cdots < \cdots < x_n = d\}$ with the property that $\overline{S(P(f_R))} - \underline{S(P(f_R))} < \epsilon$.

However, since we are assuming that $f(x) : [a,b] \to \Re$ is integrable, we can find a partition P of $[a,b]$ with the property that $\overline{S(P)} - \underline{S(P)} < \epsilon$. Since refinement always makes the upper and lower sums closer together, we might as well assume that the two points c and d are included in P. Now simply create a partition $P(f_R)$ of $[c,d]$ as the members of P with the points less than c and the points larger than d deleted.

Thus, $\overline{S(P(f_R))} - \underline{S(P(f_R))} \leq \overline{S(P))} - \underline{S(P)} < \epsilon$ so we are done. □

OK, it is now time to mention a version of the distributive law for integration.

Proposition 11.5.20. *If $f(x) : [a,b] \to \Re$ is integrable and $c \in [a,b]$, then $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$.*

Proof. Step 1. The Challenge:

Let $\epsilon > 0$ be given.

Step 2. The Choice:

Since we know by the previous proposition the function $f(x)$ is integrable on the interval $[a,c]$, we can find a partition $P_L = \{a = x_0 < x_1 < \cdots < x_n = c\}$ with the property that if P_L' is any refinement of P_L, then $|S(P_L') - \int_a^c f(x)\ dx| < \frac{\epsilon}{2}$.

Similarly, we can find a partition $P_R = \{c = y_0 < y_1 < \cdots < y_m = b\}$ with the property that if P'_R is any refinement of P_R, then $|S(P'_R) - \int_c^b f(x)\ dx| < \frac{\epsilon}{2}$.

Choose $P = P_L \cup P_R = \{a = x_0 < x_1 < \cdots < x_n = c = y_0 < y_1 < \cdots < y_m = b\}$.

Step 3. The Check:

If P' is any refinement of P, then note that the members of P' can be written as $P' = P'_L \cup P'_R$, where P'_L contains all the members of P' to the left of c and P'_R contains all the members of P' to the right of c. create a partition of the left subinterval $[a, c]$ defined by $P'_L = \{a = x_0 < x_1 < \cdots < x_q = c\}$ and a partition of the right subinterval $P'_R = \{c = x_q < x_{q+1} < \cdots < x_n = b\}$. Thus,

$$|S(P') - (\int_a^c f(x)\ dx + \int_c^b f(x)\ dx)| = |S(P'_L) - \int_a^c f(x)\ dx + S(P'_R) - \int_c^b f(x)\ dx| \tag{11.5.1}$$

$$\leq |S(P'_L) - \int_a^c f(x)\ dx| + |S(P'_R) - \int_c^b f(x)\ dx| \tag{11.5.2}$$

$$< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \tag{11.5.3}$$

□

Simplicio:

Exercise Set 11.5.

1. Using the DEFINITION of the integral, show that $\int_1^2 x\ dx = \frac{3}{2}$.
2. Using the DEFINITION of the integral, show that $\int_1^2 x^2\ dx = \frac{7}{3}$.
3. Using the DEFINITION of the integral, show that $\int_0^1 x^3\ dx = \frac{1}{4}$.
4. If $f(x) = x^3 + 3x$ is defined on the interval $[-2, 2]$ and $\epsilon > 0$, then find a partition P with the property that $|S(P) - \int_{-2}^2 f(x)\ dx| < \epsilon$.
5. If $f(x) = x^4 + x$ is defined on the interval $[-3, 3]$ and $\epsilon > 0$, then find a partition P with the property that $|S(P) - \int_{-3}^3 f(x)\ dx| < \epsilon$.

6. If $f(x) = 5|x| + 3|x-1|$ is defined on the interval $[-2, 2]$ and $\epsilon > 0$, then find a partition P with the property that $|S(P) - \int_{-2}^{2} f(x)\, dx| < \epsilon$.

11.6 The Intermediate Value Theorem for Integrals

Galileo: We now turn to the Intermediate Value Theorem for Integrals. Some people call it the Mean Value Theorem for Integrals. Actually, its a bit of both.

Simplicio: Isn't one Intermediate Value Theorem enough?

Galileo: Well no. These theorems provide the key steps in the proofs of the Fundamental Theorem of Calculus and Taylor's Theorem. While you are already familiar with the Fundamental Theorem of Calculus 11.7.3 and 11.7.4, the remainder form of Taylor's Theorem will probably require some work on your part. In my experience, students are only visit Taylor Lite these days.

Virginia: Even for me, it seems like we are a bit over the top on the theory. Why do we need Taylor's Theorem?

Simplicio: Looks like I am beginning to get some support from the rear.

Galileo: The short answer is that this theorem will provide the key step in explaining why the method of Newton/Raphson converges more quickly than the bisection method. When we discuss this topic, we will make numerous computations of roots of functions. For example, we will find that the method of Newton/Raphson will only require six iterations to achieve 14 decimal places of accuracy when approximating $\sqrt{2}$. On the other hand, the bisection method will require more than thirty. Even with today's speedy computer's this difference could become important in a big computational project where these computations must be made millions of times.

The long answer is that Taylor's Theorem will provide a systematic way to numerically compute first, second, and higher order derivatives. These numerical derivatives are used to numerically solve two point boundary value problems in differential equations and partial differential equations. They are also used every where in image and

signal processing. Taylor is a big deal.

Simplicio: While I don't care anything about differential equations, I like the signal processing connection.

Galileo: Just as the derivative detects the amount of change that is taking place with a function, an edge detector is designed to identify those pixels in an image, where rapid change is occurring. Edge detectors are often constructed from numerical first and second derivatives. We now state and prove the Intermediate Value Theorem for Integrals. Note that this theorem is a formal statement of the fact that the area under the curve is the area of a rectangle with base of length $b-a$ and height somewhere between the highest and lowest possible values of the function. For a visual of the geometry see Figure 11.3. Note also that the key idea of the proof is that the mean of the function, $\frac{1}{b-a}\int_a^b f(x)\ dx$, is intermediate between the lowest $(i.e. f(z_1))$ and highest values $(i.e f(z_0))$. Thus, we named it the Intermediate Value Theorem for Integrals.

Theorem 11.6.1 (Intermediate Value Theorem for Integrals). *If $f(x) : [a,b] \to \Re$ is continuous at each point $x \in [a,b]$, then there is a point $z \in [a,b]$ with the property that $\int_a^b f(x)\ dx = f(z)(b-a)$.*

Proof. Since $f(x)$ is continuous at each $x \in [a,b]$, we know it is integrable. Thus, the symbol $\int_a^b f(x)\ dx$ makes sense.

By the Extremum Theorem 10.3.1 there are points $z_0, z_1 \in [a,b]$ with the property that $f(z_1) \leq f(x) \leq f(z_0)$ for all $x \in [a,b]$. Since the numbers $f(z_0)$ and $f(z_1)$ are constants (wrt x), we know by Integral Bounds Corollary 11.5.14 that

$$f(z_1)(b-a) = f(z_1)\int_a^b 1\ dx \leq \int_a^b f(x)\ dx \leq f(z_0)\int_a^b 1\ dx = f(z_0)(b-a).$$

Thus,

$$f(z_1) \leq \frac{1}{b-a}\int_a^b f(x)\ dx \leq f(z_0)$$

so the value $\frac{1}{b-a} \leq \int_a^b f(x)\ dx$ is intermediate between $f(z_1)$ and $f(z_0)$. By the Intermediate Value Theorem 10.2, there is a point $z \in [a,b]$ with the property that $f(z) = \frac{1}{b-a}\int_a^b f(x)\ dx$.

Thus, $\int_a^b f(x)\,dx = f(z)(b-a)$. □

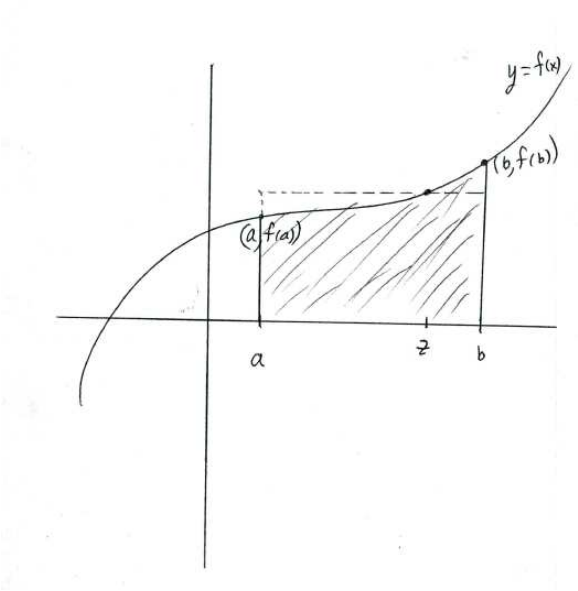


Figure 11.3: The Intermediate Value Theorem for Integrals

Galileo: The next theorem is a generalization of the Intermediate Value Theorem for Integrals.

Simplicio: What!!!? Another one?

Galileo: OK, I know you have had it with all this theory, but this theorem is exactly what we need to prove the error formula for Taylor's Theorem. This error formula is essential to our understanding of the convergence rates of sequences generated by Newton/Raphson. Error formulas guide us when, where, and things go wrong. Remember, the name of the game is control.

Theorem 11.6.2 (Intermediate Value Theorem for Integrals 2). *If $f(t), w(t) : [a,b] \to \Re$ are continuous at each point $t \in [a,b]$ and $w(t) \geq 0$ for all $t \in [a,b]$, then there is a point $z \in [a,b]$ with the property that $\int_a^b f(t)w(t)\,dt = f(z)\int_a^b w(t)\,dt$.*

Proof. Since $f(t)$ is continuous at each $t \in [a,b]$, we know by the Extremum Theorem that there are points $z_0, z_1 \in [a,b]$ with the property that $f(z_1) \leq f(t) \leq f(z_0)$ for

all $t \in [a,b]$. Since $w(t) \geq 0$ for all $t \in [a,b]$, $f(z_1)w(t) \leq f(t)w(t) \leq f(z_0)w(t)$ for all $t \in [a,b]$. Since the numbers $f(z_0)$ and $f(z_1)$ are constants (wrt t), we know

$$\begin{aligned} f(z_1)\int_a^b w(t)\ dt &= \int_a^b f(z_1)w(t)\ dt \\ &\leq \int_a^b f(t)w(t)\ dt \\ &\leq \int_a^b f(z_0)w(t)\ dt \\ &= f(z_0)\int_a^b w(t)\ dt. \end{aligned}$$

Thus,

$$f(z_1) \leq \frac{\int_a^b f(t)w(t)\ dt}{\int_a^b w(t)\ dt} \leq f(z_0)$$

so the value $\frac{\int_a^b f(t)w)t)\ dt}{\int_a^b w(t)\ dt}$ is intermediate between $f(z_1)$ and $f(z_0)$. By the Intermediate Value Theorem 10.2, there is a point $z \in [a,b]$ with the property that

$$f(z) = \frac{\int_a^b f(t)w(t)\ dt}{\int_a^b w(t)\ dt}$$

Thus, $\int_a^b f(t)w(t)\ dt = f(z)\int_a^b w(t)\ dt$. □

Virginia: If you think about it, not only is this last theorem a generalization of the First Intermediate Value Theorem for Integrals, but the proof is the same.

Galileo: Correct.

Virginia: But how are we going to use it to prove Taylor's Theorem?

Galileo: While the function $w(t)$ is completely general, the case most interesting to us is when $w(t) = (x-t)^n$, where $t \in [x_0, x]$.

Simplicio: But if $x_0 > x$, then the interval $[x_0, x]$ has no points in it.

Galileo: Technically, you are correct. However, we only care about values of t between x and x_0.

Virginia: OK, but if the integer n is odd and $x < t < x_0$, then the quantity $x - t$ is negative so that $w(t)$ will be a negative number. The theorem does not apply.

Galileo: Technically, you are again correct. However, if you take a second look at the theorem, you will realize that the theorem is still true if we assume $w(t) \leq 0$ for all t.

Exercise Set 11.6.

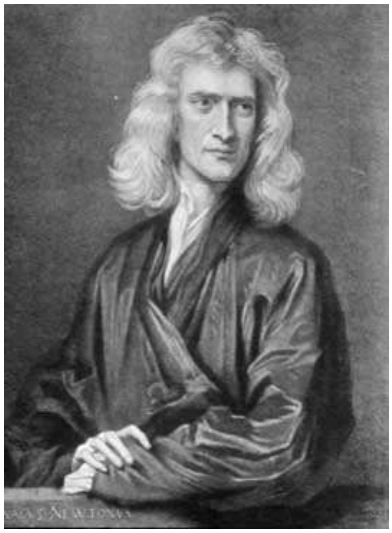
1. If $f(x) = x^2$ for $x \in [0,2]$, find a point $z \in [0,2]$ with the property that $f(z) = \frac{1}{2}\int_0^2 x^2\ dx = \frac{4}{3}$. Draw a graph of the function $y = f(x)$. Indicate the placement of the point $(z, f(z))$ on the graph.

2. If $f(x) = x^3$ for $x \in [0,2]$, find a point $z \in [0,2]$ with the property that $f(z) = \frac{1}{2}\int_0^2 x^3\ dx$. Draw a graph of the function $y = f(x)$. Indicate the placement of the point $(z, f(z))$ on the graph.

3. If $f(x) = x^2$ for $x \in [0,2]$ and $w(x) = (x-2)$, then find a point $z \in [0,2]$ with the property that $\int_0^2 f(t)w(t)\ dt = f(z)\int_0^2 w(t)\ dt$.

4. If $f(x) = x^3$ for $x \in [0,2]$ and $w(x) = (x-2)^2$, then find a point $z \in [0,2]$ with the property that $\int_0^2 f(t)w(t)\ dt = f(z)\int_0^2 w(t)\ dt$.

11.7 The Fundamental Theorem of Calculus



If I have been able to see further, it was only because I stood on the shoulders of giants.-Isaac Newton

Galileo: Let us now introduce our colleague Sir Isaac Newton (1642-1727). Professor Newton made more contributions to our understanding of the world around us than

almost any other scientist. Not only was he an inventor of Calculus, but he also applied it to real physical problems. His Second Law of Motion $F = ma$ is fundamental to the understanding of the motion of a cannonball dropped from the Leaning Tower of Pisa, the orbits of the planets around the sun, the motion of a pendulum, and the motion of a particle through a fluid. His contributions to optics were also remarkable and included building the first reflecting telescope and his recognition that that white light can be refracted into the many beautiful colors we have in the visible spectrum. His Principia (1687) and Opticks (1704) are two of the greatest scientific works ever written.

Newton: You forgot to mention that I served as the Lucasian Professor of Mathematics at the University of Cambridge during the years 1669-1701 and I was president of the Royal Society during the years 1703-1727.

Galileo: Thank you for reminding me of these details. Good sir, could you give us a few insights into the Fundamental Theorem of Calculus?

Newton: The Fundamental Theorem of Calculus provides the bridge that connects the two main themes in calculus: derivatives and integrals.

Simplicio: I must admit that the slope of a tangent line and an integral do not seem to have anything in common.

Newton: But they do. Let us begin our discussion by visualizing the area of a region and the length of its boundary. How about if we begin with a circle?

Simplicio: From Geometry, I know the area of a circle is given by the formula $A = \pi r^2$; the circumference is given by $C = 2\pi r$. So?

Newton: But did you ever notice that $\frac{dA}{dr} = 2\pi r = C$?

Simplicio: Seems like an accident of nature to me.

Newton: Not so. This simple observation points out the general fact that the rate of change of the area of a region is the length of the changing part of the boundary.

Simplicio: Sounds like double talk to me.

Newton: How about a rectangle with height $h = 1$ and base $b = x$. If we think of the area as a function of the length of the base, then the area $A = x$ and $\frac{dA}{dx} = 1$, which

equals the height of the moving edge.

Simplicio: A second accident of nature?

Newton: Actually, these two examples are completely general. For if we have a function $f(t); [a,b] \to \Re$, which is continuous at each $t \in [a,b]$, then the function $F(x) = \int_a^x f(t)\ dt$, computes the area under the curve at each point $x \in [a,b]$. The first part of the Fundamental Theorem of Calculus states that $F'(x) = f(x)$.

Virginia: Which generalizes the example you just presented! Namely, the rate of change of the area under the curve $y = f(t)$ equals the length of the right hand side of the region, namely $f(x)$.

Galileo: Very good.

Newton: But that observation is obvious. The first proposition is exactly what we need to prove the second part of the Fundamental Theorem of Calculus. It basically states that if you have no velocity, then you aren't going anywhere. Maybe some of our students should achieve a little velocity.

Proposition 11.7.1. *If $f(x) : [a,b] \to \Re$ is differentiable at each point $x \in [a,b]$ and $f'(x) = 0$ for all $x \in [a,b]$, then $f(x) = f(a)$ for all $x \in [a,b]$.*

Proof. If $x \in [a,b]$, then by the Mean Value Theorem 11.3.1 we know there is a $z \in [a,b]$ with the property that $f'(z) = \frac{f(x)-f(a)}{x-a}$. Since we are assuming $f'(x) = 0$ for all $x \in [a,b]$, $f'(z) = 0$, which implies the fraction $\frac{f(x)-f(a)}{x-a} = 0$. However, if a fraction equals zero, then the numerator also equals zero. Thus, $f(x) - f(a) = 0$, which implies $f(x) = f(a)$. □

Definition 11.7.2. *If $f(x), F(x) : [a,b] \to \Re$ and $F'(x) = f(x)$ for all $x \in [a,b]$, then the function $F(x)$ is called an antiderivative of $f(x)$.*

Example 11.7.1. *If $F(x) = x^3$ and $f(x) = 3x^2$, then $F(x)$ is an antiderivative of $f(x)$.*

Example 11.7.2. *If $F(x) = x^3 + 1$ and $f(x) = 3x^2$, then $F(x)$ is an antiderivative of $f(x)$.*

Virginia: From these last two examples, we see that a function may have many antiderivatives.

Galileo: Correct.

Newton: The Fundamental Theorem of Calculus shows that there is a close relationship between area and antiderivatives. For convenience, the theorem is split into two parts. The first part relates the derivative of the area under a curve and the height of the changing boundary. The second part is what every Calculus student remembers about computing areas.

Theorem 11.7.3 (Fundamental Theorem of Calculus).

1. *If* $f(t) : [a, b] \to \Re$ *is continuous at each* $t \in [a, b]$ *and* $F(x) = \int_a^x f(t)\ dt$, *then* $F'(x) = f(x)$.

2. *If* $f(t) : [a, b] \to \Re$ *is continuous at each* $t \in [a, b]$ *and* $G(t)$ *is any antiderivative of* $f(t)$, *then* $\int_a^b f(t)\ dt = G(b) - G(a)$.

Proof. Part 1.

If $F(x) = \int_a^x f(t)\ dt$, then there is a $z = z(h)$ (i.e. z depends on h) between x and $x + h$ so that

$$\begin{aligned}
F'(x) &= \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} \\
&= \lim_{h\to 0} \frac{\int_a^{x+h} f(t)\ dt - \int_a^x f(t)\ dt}{h} \\
&= \lim_{h\to 0} \frac{\int_x^{x+h} f(t)\ dt}{h} \\
&= \lim_{h\to 0} \frac{f(z(h)) \int_x^{x+h}\ dt}{h} \\
&= \lim_{h\to 0} \frac{f(z(h))(x+h-x)}{h} \\
&= \lim_{h\to 0} f(z(h)\ )\frac{h}{h} \\
&= \lim_{z\to x} f(z) = f(x).
\end{aligned}$$

Note that we used the Intermediate Value Theorem for Integrals 11.6.1 to justify the equality $\frac{\int_x^{x+h} f(t)\ dt}{h} = \frac{f(z(h)) \int_x^{x+h}\ dt}{h}$.

Simplicio: Why did you write the point as $z = z(h)$?

Newton: Since the point z varies as the point h varies, the point z is actually a function of h. The last equal sign is valid because the function $f(x)$ is continuous at the point x and the values of $z(h)$ converge to x as h converges to 0.

Part 2.

Let $H(x) = G(x) - F(x)$. Since $H'(x) = G'(x) - F'(x) = f(x) - f(x) = 0$ for all x, we know by the previous proposition that $H(x) = H(a)$ for all $x \in [a, b]$. Thus, $G(x) = F(x) + H(a)$ for all $x \in [a, b]$. If $G(t)$ is any antiderivative of $f(t)$, then $G(b) - G(a) = F(b) + H(a) - (F(a) + H(a)) = F(b) - F(a) = F(b) - 0 = \int_a^b f(t)\, dt$. □

Newton: We now give a simplified statement of the Fundamental Theorem of Calculus, which is in the form we will need.

Theorem 11.7.4 (Fundamental Theorem of Calculus 2). *If $x, x_0 \in X$, where X is an interval in $\Re$ and $f(t) : X \to \Re$ is a function with the property that $f'(t)$ is continuous at each $t \in X$, then $\int_{x_0}^{x} f'(t)\, dt = f(x) - f(x_0)$.*

Simplicio: I like simplified.

Virginia: What about Archimedes' formula for the volume of a sphere?

Simplicio: What about it?

Virginia: If $V = \frac{4}{3}\pi r^3$, then $\frac{dV}{dt} = 4\pi r^2$, which just happens to be the surface area of a sphere. Is that an accident?

Newton: And now it becomes obvious where all those theorems in higher dimensional Calculus come from.

Simplicio: Enough of all this theory. How about an example?

Galileo: OK, let's begin with an easy one.

Example 11.7.3. *Compute $\int_0^1 x^4\, dx$.*

Virginia: Since $F(x) = \frac{x^5}{5}$ is an antiderivative of $f(x) = x^4$, we know by the Fundamental Theorem of Calculus 11.7.3 that

$$\int_0^1 x^4\, dx = F(1) - F(0) = \frac{1^5}{5} - \frac{0^5}{5} = \frac{1}{5}.$$

Simplicio: No fancy summations. No partitions. Now I'm in my comport zone. How about another such beast?

Galileo: Don't think those old guys were any less delighted.

Example 11.7.4. *Compute* $\int_0^1 x^n \, dx$.

Virginia: Since $F(x) = \frac{x^{n+1}}{n+1}$ *is an antiderivative of* $f(x) = x^n$, *we know by the Fundamental Theorem of Calculus 11.7.3 that*

$$\int_0^1 x^n \, dx = F(1) - F(0) = \frac{1^{n+1}}{n+1} - \frac{0^{n+1}}{n+1} = \frac{1}{n+1}.$$

Example 11.7.5. *If* $F(x) = \int_0^x t^2 \, dt$, *then compute* $F'(x)$.

Simplicio: I can do this one too. Here goes. Since the function $G(t) = \frac{t^3}{3}$ *is an antiderivative of* $f(t) = t^2$, *we know* $F(x) = \int_0^x t^2 \, dt = G(x) - G(0) = \frac{x^3}{3} - \frac{0^3}{3} = \frac{x^3}{3}$.

Thus, $F'(x) = x^2$.

Virginia: But you forgot to pay attention when we discussed the first part of the Fundamental Theorem of Calculus. You worked much too hard. All you have to do is substitute the upper limit of the integral, namely x, *into the function* $f(t) = t^2$ *to get* $F'(x) = f(x) = x^2$. *You are finished with zero effort.*

Galileo: Theorems are good.

Example 11.7.6. *If* $F(x) = \int_x^0 t^2 \, dt$, *then compute* $F'(x)$.

Virginia: Since $F(x) = \int_x^0 t^2 \, dt = -\int_0^x t^2 \, dt, F'(x) = -x^2$.

Simplicio: I understand that example.

Example 11.7.7. *If* $F(x) = \int_x^a f(t) \, dt$, *then compute* $F'(x)$.

Virginia: Since $F(x) = \int_x^a f(t) \, dt = -\int_a^x f(t) \, dt, F'(x) = -f(x)$.

Example 11.7.8. *If* $F(x) = \int_0^{x^2} t \, dt$, *then compute* $F'(x)$.

Simplicio: An antiderivative of $f(t) = t$ *is the function* $G(t) = \frac{t^2}{2}$, *which of course has derivative* $G'(t) = t$. *Thus, by Theorem 11.7.3* $F(x) = \int_0^{x^2} t \, dt = G(x^2) - G(0)$. *By the Chain Rule for derivatives,* $F'(x) = \frac{dG(x^2)}{dx} - \frac{dG(0)}{dx} = G'(x^2)2x = x^2 2x$.

Virginia: If you notice that the function $F(x)$ can be written as the composition $F(x) = G(H(x))$, where $G(y) = \int_0^y t\ dt$ and $H(x) = x^2$, then $F'(x) = G'(H(x))H'(x)$ and you are done.

Simplicio: Your method was a lot easier.

Virginia: Easy is good. The general method is summarized in the following proposition.

Proposition 11.7.5. *If $f(t) : [a, b] \to \Re$ is continuous at each $t \in [a, b]$, and $F(x) = \int_{g(x)}^{h(x)} f(t)\ dx$, then $F'(x) = f(h(x))h'(x) - f(g(x))g'(x)$.*

Galileo: How about one last example?

Example 11.7.9. *Compute $\int_{x_0}^{x} (x - t)\ dt$.*

Virginia: Since the antiderivative of the function $f(t) = x - t$ is $-\frac{(x-t)^2}{2}$,

$$\int_{x_0}^{x} (x - t)\ dt = -\frac{(x - t)^2}{2}|_{t=x_0}^{x} = 0 - (-\frac{(x - x_0)^2}{2}) = \frac{(x - x_0)^2}{2}.$$

Simplicio: Why did you present this last example?

Galileo: That computation is exactly what we will need for the last step in the proof of Taylor's Theorem.

Simplicio: How about a less abstract example?

Example 11.7.10. *Galileo: OK, how about if we compute $\int_0^\infty e^{-x}\ dx$?*

Simplicio: That's an easy one. By the Fundamental Theorem of Calculus, we know that

$$\int_0^\infty e^{-x}\ dx = -e^{-x}|_{x=0}^{\infty} = 0 - (-1) = 1.$$

Galileo: Very good. we will see that integral again.

Simplicio: How about another easy example?

Example 11.7.11. *Galileo: OK, how about if we compute $\int_{-\pi}^{\pi} \cos^2(x)\ dx$?*

Simplicio: I am not sure about that problem.

Virginia: If you remember your half angle formulas from trigonometry, then you recall that $\cos^2(x) = \frac{1+\cos(2x)}{2}$. *Thus,*

$$\int_{-\pi}^{\pi} \cos^2(x)\ dx = \int_{-\pi}^{\pi} \frac{1+\cos(2x)}{2}\ dx = \int_{-\pi}^{\pi} \frac{1}{2}\ dx + \int_{-\pi}^{\pi} \frac{1}{2}\cos(2x)\ dx = \pi + 0 = \pi.$$

Simplicio: Why did you choose this last example?

Galileo: We just showed that the length of the function $f(x) = \cos(x)$ *on the interval* $[-\pi, \pi]$ *is* $\sqrt{\pi}$.

Simplicio: Interesting. So there actually is a reason for computing this example.

Galileo: This piece of information will provide a key fact when we discuss Fourier Series.

Exercise Set 11.7.

1. Compute $\int_{-\pi}^{\pi} \sin^2(x)\ dx$.

2. If $F(x) = \int_0^x t^9\ dt$, then compute $F'(x)$.

3. If $F(x) = \int_x^0 t^9\ dt$, then compute $F'(x)$.

4. If $F(x) = \int_0^{x^2} t^9\ dt$, then compute $F'(x)$.

5. If $F(x) = \int_0^x \sin(t^2+1)\ dt$, then compute $F'(x)$.

6. If $F(x) = \int_x^{x^3} \sin(t^2+1)\ dt$, then compute $F'(x)$.

7. Compute $\int_{x_0}^{x} (x-t)^2\ dt$.

11.8 Integration By Parts

Brook Taylor (1685 - 1731)

Galileo: Let's now invite Professor Brook Taylor (1685-1731) to remind us about integration by parts. Professor Taylor has many achievements to his credit. Virginia, what can you tell us about Professor Taylor?

Virginia: Professor Taylor was born into a family of culture and means. His father provided him with a fine education in mathematics both at home and later at Cambridge. While his first wife was from a good family, she had little money and his father disapproved of the match. Unfortunately, she died in childbirth. While his father approved of his second marriage, she also died in childbirth.

Simplicio: He suffered a sad life.

Virginia: Life is uncertain.

Galileo: But he achieved great mathematics! In addition to inventing the technique of integration by parts, Professor Taylor also developed methods for approximating functions by polynomials. These methods are now known as Taylor series. As you will see, these methods can be used to numerically approximate derivatives. To this day these methods are used in a multitude of applications from the design of an airfoil to predicting the path of a hurricane. These techniques are now known as finite difference methods. We welcome you Professor Taylor.

Taylor: Let us begin our discussion of integration by parts by remarking that integration generally has fewer tools than differentiation.

Simplicio: How so?

Taylor: With differentiation we have the product, quotient, and chain rules. Unfortunately, integration has no such rules.

Simplicio: Which means there is less to learn. I like that.

Taylor: Maybe so, but then you are left with functions which can be differentiated, but not integrated. For example, try integrating the functions $f(x) = log(x)e^x$, $f(x) = \frac{1}{1+x^6}$, or $f(x) = e^{-x^2}$. While computing the derivatives of these functions is straightforward, they are impossible to integrate using the Fundamental Theorem of Calculus.

Virginia: Is that because you can't compute their antiderivatives?

Taylor: You got it. On the other hand, the technique of integration by parts is an attempt to rescue a product rule for integrals.

Simplicio: What does that mean?

Taylor: Sometimes it works, sometimes it doesn't.

Simplicio: An example please.

Taylor: We will show that the technique works great for the integral $\int_0^\pi x\cos(x)\ dx$ and is helpless for the integral $\int_1^2 log(x)e^x\ dx$.

Galileo: Let's move on to the theorem and its proof.

Taylor: Since we would like to be more formal, we state this method as a theorem with definite integrals. The idea underneath the proof is to simply differentiate the product $u(x)v(x)$ and then manipulate a bit.

Theorem 11.8.1 (Integration by Parts). *If $u(x)$ and $v(x)$ are differentiable functions on an interval $[a,b]$, where $u'(x)$ and $v'(x)$ are continuous at each $x \in [a,b]$, then $\int_a^b u(x)v'(x)dx = u(x)v(x)|_{x=a}^b - \int_a^b v(x)u'(x)dx$.*

Proof. By the Product Rule for Derivatives 11.1.2, we know that

$$\frac{du(x)v(x)}{dx} = u(x)\frac{dv(x)}{dx} + v(x)\frac{du(x)}{dx}.$$

Thus,

$$u(x)\frac{dv(x)}{dx} = \frac{du(x)v(x)}{dx} - v(x)\frac{du(x)}{dx}.$$

Integrating both sides of the equation on the interval $[a, b]$, we find that

$$\int_a^b u(x)\frac{dv(x)}{dx}\,dx = \int_a^b \frac{du(x)v(x)}{dx}\,dx - \int_a^b v(x)\frac{du(x)}{dx}\,dx.$$

Since the function $u(x)v(x)$ is an antiderivative of $\frac{du(x)v(x)}{dx}$, the result follows. □

Simplicio: How about an example?

Taylor: For actual computations, we will simplify the theorem to $\int u\,dv = uv - \int v\,du$, where we understand the functions $u = u(x)$ and $v = v(x)$ depend on x.

Example 11.8.1. *Compute the integral* $\int_0^1 x(x-1)^3\,dx$.

Simplicio: I can do that problem. All you have to do is expand the expression $x(x-1)^3 = x(x^3 - 3x^2 + 3x^1 - 1) = x^4 - 3x^3 + 3x^2 - x$ *and integrate each one of the four terms.*

Taylor: Instead, if we let $u = x$ *and* $dv = (x-1)^3$, *then* $du = dx$ *and* $v = \frac{(x-1)^4}{4}$ *we see that*

$$\int x(x-1)^3\,dx = x\frac{(x-1)^4}{4} - \int \frac{(x-1)^4}{4}\,dx = x\frac{(x-1)^4}{4} - \frac{(x-1)^5}{20}.$$

Thus,

$$\int_0^1 x(x-1)^3\,dx = x\frac{(x-1)^4}{4}|_{x=0}^1 - \frac{(x-1)^5}{20}|_{x=0}^1 = -(-1)\frac{(-1)^5}{20} = -\frac{1}{20}.$$

The worst aspect of the technique is keeping track of the minus signs.

Simplicio: How about another example?

Example 11.8.2. *Compute the integral* $\int_0^\pi x\cos(x)\,dx$.

If we set $u = x$ *and* $dv = \cos(x)$, *then* $du = dx$ *and* $v = \sin(x)$.

Thus,

$$\int_0^\pi x\cos(x)\,dx = x\sin(x)|_{x=0}^\pi - \int_0^\pi \sin(x)\,dx = -(-\cos(x))|_{x=0}^\pi = -2.$$

Example 11.8.3. *Compute the integral* $\int_0^1 x\ e^x\ dx$.

If we set $u = x$ *and* $dv = e^x$, *then* $du = dx$ *and* $v = e^x$.

Thus,

$$\int_0^1 x\ e^x\ dx = xe^x|_{x=0}^1 - \int_0^1 e^x\ dx = e - (e-1) = 1.$$

Example 11.8.4. *Compute the integral* $\int_1^2 log(x)\ e^x\ dx$.

If we set $u = log(x)$ *and* $dv = e^x$, *then* $du = \frac{1}{x}dx$ *and* $v = e^x$.

Thus,

$$\int log(x)\ e^x\ dx = log(x)e^x - \int e^x \frac{1}{x}\ dx.$$

So, what do you do with the integral $\int e^x \frac{1}{x}\ dx$?

Simplicio: I have no clue.

Taylor: Exactly my point. The method provides no useful information.

Virginia: What if you set $u = e^x$ *and* $dv = log(x)$?

Taylor: You end up with an even bigger mess.

Galileo: How about a set of guidelines for using your technique?

Taylor: To reduce the complexity of the integral $\int u\ dv$ for the following examples, make the following choices.

1. If n is a positive integer and $\int x^n \cos(x)\ dx$, then choose $u = x^n$ and $dv = \cos(x)$.
   (This choice will have to be repeated n times.)

2. If n is a positive integer and $\int x^n \sin(x)\ dx$, then choose $u = x^n$ and $dv = \sin(x)$.
   (This choice will have to be repeated n times.)

3. If n is a positive integer and $\int x^n e^x\ dx$, then choose $u = x^n$ and $dv = e^x$.
   (This choice will have to be repeated n times.)

4. If n is a positive integer and $\int x^n log(x)\ dx$, then choose $u = log(x)$ and $dv = x^n$.

5. If $\int e^x \sin(x)\ dx$, then choose $u = e^x$ and $dv = \sin(x)$.
   (This choice will have to be repeated twice.)

6. If $\int e^x \cos(x)\ dx$, then choose $u = e^x$ and $dv = \cos(x)$.
   (This choice will have to be repeated twice.)

Exercise Set 11.8.

1. Compute the integral $\int_0^\pi x \sin(x)\ dx$.
2. Compute the integral $\int_0^\pi x^2 \sin(x)\ dx$.
3. Compute the integral $\int_0^\pi log(x)x\ dx$.

11.9 Taylor's Theorem: Degree One Polynomials

Brook Taylor (1685 - 1731)

Galileo: We now turn to the final topic in our review: Taylor's Theorem.
Simplicio: Does this mean the pain of all this theory will soon lift?
Galileo: Actually, no. Let us now invite Professor Taylor for a second visit. Good sir, could explain your methods for approximating functions by polynomials?
Taylor: The idea behind these approximations is that calculus would be a lot easier if we considered only polynomial functions. As you have noticed, polynomials are attractive because the computation of derivatives and integrals is easy. Unfortunately,

numerous useful functions such as $\cos(x), \sin(x), e^x, \frac{1}{1-x}$, and $ln(x)$ don't quite fit into this setting. The beauty of my theorem is that it provides a strategy for approximating these functions by polynomials.

Simplicio: I like this idea. Calculus would certainly be easier if every function was a polynomial.

Taylor: That is the concept.

Simplicio: Where do we start?

Taylor: The idea is to write a function $f(x) = p_n(x) + E_n(x)$, where $p_n(x)$ is a polynomial of degree n and $E_n(x)$ represents the error. In the next theorem, we approximate a function $f(x)$ by the straight line $y = p_1(x) = f(x_0) + f'(x_0)(x - x_0)$. The error is represented as the integral $E_1(x) = \int_{x_0}^{x} f''(t)(x - t)\ dt$.

Theorem 11.9.1 (Taylor Theorem 1). *If $x, x_0 \in X$, where X is an interval in $\Re$ and $f(t) : X \to \Re$ is a function with the property that $f''(t)$ is continuous at each $t \in X$, then*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x} f''(t)(x - t)\ dt.$$

Proof. The idea of the proof is to apply integration by parts to the last term. In particular, if we let $u(t) = x - t$ and $dv = f''(t)dt$, then $du = -dt$ and $v = f'(t)$. Thus, by parts and the Fundamental Theorem of Calculus, we have the following sequence of equalities.

$$\begin{aligned}\int_{x_0}^{x} f''(t)(x - t)dt &= (x - t)f'(t)|_{t=x_0}^{x} - \int_{x_0}^{x} f'(t)(-dt) \\ &= -(x - x_0)f'(x_0) + f(x) - f(x_0).\end{aligned}$$

Thus, $f(x) = f(x_0) + (x - x_0)f'(x_0) + \int_{x_0}^{x} f''(t)(x - t)\ dt$. □

Simplicio: While the proof of this theorem is easier than I expected, I don't like the formula for the error term.

Galileo: Surprising you should mention this concern. I think you have someone who agrees with you. Let me introduce Professor Joseph Louis Lagrange (1736-1813), who

was a survivor of the French Mathematician. He did much to explain and exploit Professor Taylor's ideas. Welcome Professor Lagrange, but please don't mumble.

Lagrange: I agree that the form of the error term is a nuisance. If you recall the second version of Intermediate Value Theorem for Integrals 11.6.2, then we can present a form for the error that is easier to remember.

Simplicio: You mean we are actually going to use that theorem?

Galileo: We discussed it for a reason.

Lagrange: My version of Taylor's Theorem now becomes:

Theorem 11.9.2 (Lagrange Form of Taylor's Theorem). *If $x, x_0 \in X$, where X is an interval in $\Re$ and $f(t) : X \to \Re$ is a function with the property that $f''(t)$ is continuous at each $t \in X$, then there is a point $z \in X$ so that*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(z)}{2}(x - x_0)^2.$$

Proof. To prove this theorem will apply the Intermediate Value Theorem for Integrals 11.6.2 to the integral $\int_{x_0}^{x} f''(t)(x - t)\, dt$. To be certain we can apply this theorem we have to check the function $w(t) = x - t$ does not change from positive to negative for values of t between x_0 and x. Once we have made this check, the hypotheses hold.

We have two cases to consider.

Case 1. If $x > x_0$, then we are considering $t \in [x_0, x]$.

For this case, the function $w(t) = x - t \geq 0$ for all $t \in [x_0, x]$.

Case 2. If $x \leq x_0$ then we are considering $t \in [x, x_0]$.

For this case, the function $w(t) = x - t \leq 0$ for all $t \in [x, x_0]$.

Now, we can apply the Intermediate Value Theorem for Integrals 11.6.2 to the integral $\int_{x_0}^{x} f''(t)(x - t)\, dt$ and to find a point $z \in [x, x_0]$ so that

$$\int_{x_0}^{x} f''(t)(x - t)\, dt = f''(z) \int_{x_0}^{x} (x - t)\, dt = f''(z) \frac{(x - t)^2}{-2}|_{t=x_0}^{x} = f''(z) \frac{(x - x_0)^2}{2}.$$

□

Lagrange: Notice that we have written the function $f(x)$ in the form $f(x) = p_1(x) +$

$E_1(x)$, where $p_1(x) = f(x_0) + f'(x_0)(x - x_0)$ and $E_1(x) = \frac{f''(z)}{2}(x - x_0)^2$. Thus, the error term now has the form of a second degree polynomial.

Galileo: There it is. Both the statement and proof are elegant and easy to understand.

Simplicio: I agree that this form of the remainder is easier to remember. How about an example?

Galileo:

Example 11.9.1. *Use Taylor's Theorem to compute* $p_1(x) = f(x_0) + f'(x_0)(x - x_0)$ *for the function* $f(x) = \cos(x)$, *where* $x_0 = 0$.

Simplicio: Even I can do this problem. All we have to do is compute $f'(x) = -\sin(x)$ *and notice that* $f(0) = 1$ *and* $f'(0) = 0$.

Thus, $p_1(x) = 1$. *I wish all problems were this easy.*

Galileo: What about a bound on the error?

Virginia: Since $f''(x) = -\cos(x), |f''(x)| \leq 1$ *for all* $x \in \Re$.

Thus, $|E_1(x)| \leq \frac{1}{2}(x - x_0)^2 = \frac{1}{2}x^2$ *for all* $x \in \Re$.

Galileo: You should now understand Taylor.

Simplicio: Wait a minute. You promised that we would approximate a function by a polynomial of degree n. The only polynomial I see is the straight line $p_1(x) = f(x_0) + (x - x_0)f'(x_0)$. Even I can see that a line $y = 1$ is not going to provide a close approximation to the function $f(x) = \cos(x)$.

Galileo: While you are correct, we only need this special case for our discussion of the Newton/Raphson method for computing roots. No worries. We are going to invite Professor Taylor to return when discuss approximation theory. We will definitely see the general case then.

Simplicio: You are making an assumption.

Galileo: Well folks. We have now concluded our discussion of the background material required for tomorrow's gathering.

Virginia: Wait. What is tomorrow's topic?

Galileo: We will show you how to compute roots.

Virginia: We have covered an enormous amount of material today. Could you summarize the essentials of what we need for tomorrow?

Galileo: You must have acquired the following skill set.

1. the ability to comprehend a mathematical argument,
2. the ability to define and apply limit facts,
3. be able to state and apply the Mean Value Theorem 11.3.3, and
4. be able to state and apply Taylor's Theorem 11.9.2.

Tomorrow we will begin to see how all this theory impacts finding the root of a function.

Simplicio: After discussing all these different topics, we are only required to have acquired four skills?

Virginia: Math is easy.

Exercise Set 11.9.

1. Use Taylor's Theorem to compute $p_1(x) = f(x_0)+f'(x_0)(x-x_0)$ for the functions $f(x) = \sin(x), \ln(1-x)$ and e^x at the point $x_0 = 0$.
2. Use Taylor's Theorem to compute $p_1(x)$ for the function $f(x) = \ln(x)$ at the point $x_0 = 1$.
3. If $f(x) = \sin(x)$, for $x \in [-\pi, \pi]$ and $x_0 = 0$, then use Taylor's Theorem to estimate a bound on $E_1(x) = \frac{f''(z)}{2}(x-x_0)^2$. Repeat the exercise for the function $f(x) = e^x$.
4. If $f(x) = \ln(1-x)$ for $x \in [-0.5, 0.5]$, and $x_0 = 0$, then use Taylor's Theorem to estimate a bound on $E_1(x) = \frac{f''(z)}{2}(x-x_0)^2$.

Simplicio: But wait a minute, you never answered my question about approximation by polynomials of degree greater than one.

Taylor: We will address that question at another gathering.

Part V

Day 5. Theory for Root Finding

Chapter 12

Successful Root Finding

Galileo: Our next goal is to establish conditions when our root finding methods "work." In particular, we will show that the method always converges when computing the bisection method, square roots, cube roots, and n^{th} roots. Of course, the square root methods and the cube root methods are special cases of the n^{th} root method, but they are worth doing because the geometry and arguments are so clear. Actually, the three arguments are all based on the idea that a bounded decreasing sequence converges.

Virginia: So that's where the idea for those theorems on convergence sequences came from.

Galileo: Light bulb time.

12.1 The Bisection Method

Galileo: Showing that the bisection method always works is easy. All we have to do is find a bounded increasing sequence or a bounded decreasing sequence.

Virginia: In fact, we have both. For if $[a_n, b_n]$ denotes the interval that has been found at the n^{th} stage of the bisection algorithm, then the sequence of points $\{a_n\}_{n=0}^{\infty}$ is bounded and increasing, while the sequence $\{b_n\}_{n=0}^{\infty}$ is bounded and decreasing. In other words, you have two sequences from which to choose.

Simplicio: But what if they converge to two different points?

Virginia: Remember that the error formula $E_n \leq b_n - a_n = \frac{b-a}{2^n}$, which converges to zero. Thus, $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$.

Simplicio: Good. So the method always works.

Virginia: Well, you do have to remember that the function $f(x)$ is continuous and that $f(a) > 0$ and $f(b) < 0$, or vice versa. Other than that, you are in your comfort zone.

Example 12.1.1. *If $f(x) = x + \sin(x) - 13$, for $x \in [0, 15]$, then we have to check two conditions to make sure that the bisection method will find a root in the interval $[0, 15]$.*

First, we have to check that $f(x)$ is continuous. However, since $f(x)$ is the sum of three continuous functions, it is continuous.

Second, we must check that $f(0)$ and $f(15)$ have opposite signs. However, since $f(0) = -13 < 0$ and $f(15) = 15 + \sin(15) - 13 > 0$, this condition is satisfied and we are done.

Example 12.1.2. *If $f(x) = 3x^2 + 2$, for $x \in [-1, 1]$, the even though $f(x)$ is continuous, the signs of $f(-1)$ and $f(1)$ are the same. Thus, the bisection method does not guarantee a root will be found.*

Simplicio: What about the function $f(x) = 3x^2 - 2$, for $x \in [-1, 1]$?

Galileo: Good point. Despite the fact that the function is continuous, the values of the function at the two endpoints do not have different signs. In fact, we have $f(-1) = f(1) = -1$. Thus, the only problem with applying the bisection method is a poor choice of interval. If we had chosen the interval $[0, 1]$, we would have been fine.

12.2 The Archimedes/Heron Algorithm

Galileo: We now show that the square root method of Archimedes/Heron always produces a bounded decreasing sequence. Recall that when we computed $\sqrt{2}$, our data

showed this property. The proof that this property always holds will be completed in three steps.

1. The geometric mean is less than or equal to the arithmetic mean.

2. The points generated by the algorithm are bounded from below by $\sqrt{K}$.

3. The sequence is always decreasing.

The next three propositions formalize these three statements.

Proposition 12.2.1 (Geometric/Arithmetic Mean). *If $x_1, x_2 \geq 0$ then $\sqrt{x_1 x_2} \leq \frac{x_1+x_2}{2}$.*

Proof. Since $(x_1 - x_2)^2 \geq 0$, the result follows by simply expanding the product and manipulating the factors. □

Proposition 12.2.2 (Boundedness). *If $K > 0, x_0 = 1, x_{k+1} = \frac{x_k + \frac{K}{x_k}}{2}$, and $k \geq 1$, then $x_k \geq \sqrt{K}$.*

Proof. By the previous proposition, $x_{k+1} = \frac{x_k+\frac{K}{x_k}}{2} \geq \sqrt{x_k * \frac{K}{x_k}} = \sqrt{K}$. □

Proposition 12.2.3 (Decreasing). *If $K > 0, x_0 = 1, x_{k+1} = \frac{x_k+\frac{K}{x_k}}{2}, k \geq 0$, and $x_k \geq \sqrt{K}$, then $x_{k+1} \leq x_k$.*

Proof. Since $x_k \geq \sqrt{K}$, $x_k^2 - K \geq 0$. Since $x_{k+1} = \frac{x_k+\frac{K}{x_k}}{2} = x_k - \frac{x_k^2-K}{2x_k}$ and $x_k^2 - K \geq 0$, the result follows. □

Galileo: The next theorem proves that the algorithm of Archimedes/Heron always works.

Theorem 12.2.4 (Square Root Convergence for Archimedes/Heron). *If $K > 0, x_0 = 1, x_{k+1} = \frac{x_k+\frac{K}{x_k}}{2}$, then the sequence $\{x_k\}_{k=1}^{\infty}$ is bounded and decreasing and thus converges. Moreover, if $L = \lim_{k\to\infty} x_k$, then $L = \sqrt[2]{K}$.*

Proof. Since the sequence $\{x_k\}_{k=0}^{\infty}$ bounded and decreasing, it converges to some number L. Thus,

$$L = \lim_{k\to\infty}\{x_{k+1}\} = \frac{\lim_{k\to\infty}\{x_k\} + \frac{K}{\lim_{k\to\infty}\{x_k\}}}{2} = \frac{L + \frac{K}{L}}{2},$$

which implies that $L = \frac{L+\frac{K}{L}}{2}$. Thus, $2L = L + \frac{K}{L}$ and $L^2 = K$. □

Virginia: Now I see why we proved that the limit of the sum is the sum of the limits. This argument is easy.

Simplicio: While I do not have the disposition or time to endure many proofs, I agree that this one isn't too bad.

Exercise Set 12.2.

1. Show the secant method produces a bounded decreasing sequence for the function $f(x) = x^2 - K$, when the algorithm is initialized by the points x_0 and x_1, where $\sqrt{K} < x_1 < x_0$.

12.3 Cube Roots

Joseph-Louis Lagrange (1736-1813)

I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes. It is in the

> works of applications that one must study them; one judges their ability there and one apprises the manner of making use of them.-Joseph-Louis Lagrange

Galileo: We now turn to the problem of showing that the method for computing cube roots always works. While it is virtually the same as the proof of the square root method, it unfortunately has a new technical difficulty.

Virginia: What seems to be the problem?

Galileo: For cube roots the proof that the geometric mean is less than the arithmetic mean becomes a bit more complicated. Let us introduce Joseph-Louis Lagrange (1736-1813). Though self taught, he was able to make significant contributions to the Calculus of Variations, Group Theory, the three body problem, differential equations (the Euler-Lagrange equations), and the theory of constrained maxima and minima. Virginia, could you tell us more about his life?

Virginia: While he is always thought of as French, Professor Lagrange was born in Turin in what is now a part of Italy. In 1755 he began a series of collaborations with Leonhard Euler on problems related to the cycloid. He also worked on the three body problem, the motion of the moon, and the perturbations of the orbits of comets by the planets. He made contributions to algebra and number theory including the first proof of Wilson's theorem: If p is a prime number, then p divides $(p-1)!+1$. In abstract algebra, he proved that the order of a subgroup divides the order of a group.

Galileo: In 1793, he almost lost his life during the French Revolution. If the chemist Lavoisier had not spoken on his behalf, he would have been executed. Unfortunately, Lavoisier was not so lucky since a revolutionary tribunal condemned him to death the next year.

Virginia: Need I reiterate, science seems to be a most dangerous business.

Simplicio: I think I am going to like this guy. He works on real-world problems.

Galileo: I agree. You will also get to meet him again when we discuss the error formulas for Taylor's Theorem and polynomial approximation. Joseph-Louis could you provide us with a bit of insight into your method of constrained maxima and

minima? In particular, we would like to show that the geometric mean never exceeds the arithmetic mean.

Lagrange: While this fact can be shown algebraically, my method of (Lagrange!) multipliers is more elegant. The technique can also be generalized to any number of points.

Proposition 12.3.1 (Geometric/Arithmetic Mean). *If* $x_1, x_2, x_3 \geq 0$, *then*

$$\sqrt[3]{x_1 x_2 x_3} \leq \frac{x_1 + x_2 + x_3}{3}.$$

Proof. An elegant way to prove this result is to recast the problem as a constrained optimization problem, where the function $F(x, y, z) = xyz$ is maximized subject to the constraint $x + y + z = M$. By the method of Lagrange multipliers, we know that the solution to this problem will be found at a critical point of the function $G(x, y, z, \lambda) = F(x, y, z) - \lambda(x + y + z - M)$. In particular, we must solve the system of 4 equations and 4 unknowns:

$$\begin{array}{rcrl} \frac{\partial G}{\partial x} & = & yz - \lambda & = 0 \\ \frac{\partial G}{\partial y} & = & xz - \lambda & = 0 \\ \frac{\partial G}{\partial z} & = & xy - \lambda & = 0 \\ \frac{\partial G}{\partial \lambda} & = & -(x + y + z - M) & = 0. \end{array}$$

From the first 3 equations, we see that the only non-zero solution of this system is when $yz = xz = xy$ or $x = y = z$. From the 4^{th} equation we see that $x + y + z = x + x + x = 3x = M$. Since the maximum value of $F(x, y, z) = xyz$ occurs at $x = y = z = M/3$ and never exceeds $\frac{M}{3} * \frac{M}{3} * \frac{M}{3} = (x+y+z)^3/27$, $xyz \leq (x+y+z)^3/27$. The result follows by taking the cube root of both sides of this expression. □

Simplicio: Unfortunately, I don't remember my Calculus well enough to appreciate that proof. I think I will simply accept this proposition and ask that we move on. At least the statement is easy enough to understand. How did he come up with that complicated proof anyway?

Galileo: He was a smart fellow. In any case, you will be pleased to note that the rest of the argument is virtually the same as the one provided for square roots.

Proposition 12.3.2 (Boundedness). *If* $K > 0, x_0 = 1, x_{k+1} = x_k - \frac{x_k^3 - K}{3x_k^2}$, *then* $x_{k+1} \geq \sqrt[3]{K}$.

Proof. By the previous proposition, $x_{k+1} = \frac{x_k + x_k + \frac{K}{x_k^2}}{3} \geq \sqrt[3]{x_k * x_k * \frac{K}{x_k^2}} = \sqrt[3]{K}$. □

Proposition 12.3.3 (Decreasing). *If* $K > 0, x_0 = 1, x_{k+1} = x_k - \frac{x_k^3 - K}{3x_k^2}, k \geq 0$, *and* $x_k \geq \sqrt[3]{K}$, *then* $x_{k+1} \leq x_k$.

Proof. Since $x_k \geq \sqrt[3]{K}$, $x_k^3 - K \geq 0$. Since $x_{k+1} = x_k - \frac{x_k^3 - K}{3x_k^2}$ and $x_k^3 - K \geq 0$, the result follows. □

We can now use these two propositions to prove the following convergence theorem for the cube root method.

Theorem 12.3.4 (Cube Root Convergence). *If* $K > 0, x_0 = 1, x_{k+1} = x_k - \frac{x_k^3 - K}{3{x_k}^2}$, *then the sequence* $\{x_k\}_{k=1}^{\infty}$ *is bounded and decreasing and thus converges. Moreover,* $\lim_{k \to \infty} x_k = \sqrt[3]{K}$.

Proof. Since the sequence $\{x_k\}_{k=0}^{\infty}$ is bounded and decreasing, it converges to some number L. Thus, we immediately observe that $L = L - \frac{L^3 - K}{3L^2}$ and $L^3 = K$. □

Simplicio: Well, after we passed that initial technical detail, the ideas are not so difficult. In fact, the proof is virtually the same as the one you presented for the square root method.

Galileo: You seem to be getting more comfortable with these proofs. Maybe you should consider becoming a mathematician. You might like the profession.

Simplicio: I fear my economic aspirations are higher than yours.

Galileo: Good family, loyal friends, a glass of red wine, what more is there?

Exercise Set 12.3.

1. Show the secant method produces a bounded decreasing sequence for the function $f(x) = x^3 - K$, when the algorithm is initialized by the points x_0 and x_1, where $\sqrt[3]{K} < x_1 < x_0$.

12.4 n^{th} Roots

Galileo: Just as we were able to determine a method for finding cube roots from the square root method, we can also determine a method for finding n^{th} roots. We have the following recursive algorithm for n^{th} roots of K, where $K > 0$:

$$\begin{aligned} x_0 &= 1, \\ x_{k+1} &= x_k - \frac{x_k^n - K}{n x_k^{n-1}}. \end{aligned}$$

This algorithm leads us to the convergence theorem for the n^{th} root method.

Theorem 12.4.1. *If $K \geq 0, x_0 = 1$, and $x_{k+1} = x_k - \frac{x_k^n - K}{n x_k^{n-1}}$ then the sequence $\{x_k\}_{k=1}^{\infty}$ is bounded and decreasing and thus always converges to $\sqrt[n]{K}$.*

Again, to prove the convergence theorem we use the following three propositions. The first proposition states that the geometric mean is always less than or equal to the arithmetic mean.

Proposition 12.4.2 (Geometric/Arithmetic Mean). *If $x_1, x_2, x_3, \ldots, x_n \geq 0$, then $\sqrt[n]{x_1 x_2 x_3 \ldots x_n} \leq \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$.*

Proof. The proof is the same Lagrange approach to the cube root case. Just more variables. □

Proposition 12.4.3 (Boundedness). *If $K > 0, x_0 = 1, x_{k+1} = x_k - \frac{x_k^n - K}{n x_k^{n-1}}$, then $x_{k+1} \geq \sqrt[n]{K}$.*

Proof. By the definition of the sequence and the previous (i. e. Geometric/Arithmetic Mean) proposition,

$$\begin{aligned} x_{k+1} &= x_k - \frac{x_k^n - K}{n x_k^{n-1}} \\ &= \frac{(n-1)x_k + \frac{K}{x_k^{n-1}}}{n} \\ &\geq \sqrt[n]{x_k^{n-1} * \frac{K}{x_k^{n-1}}} \\ &= \sqrt[n]{K}. \end{aligned}$$

□

Proposition 12.4.4 (Decreasing). *If* $K > 0, x_0 = 1, x_{k+1} = x_k - \frac{x_k^n - K}{n x_k^{n-1}}, k \geq 0$ *and* $x_k \geq \sqrt[n]{K}$, *then* $x_{k+1} \leq x_k$.

Proof. Since $x_k \geq \sqrt[n]{K}$ for all $k \geq 1$, we see that $x_k^n - K \geq 0$. Since $x_{k+1} = x_k - \frac{x_k^n - K}{n x_k^{n-1}}$ and both the numerator and denominator of the expression $\frac{x_k^n - K}{n x_k^{n-1}}$ are both non-negative, $x_{k+1} = x_k -$ non-negative number. Thus, $x_{k+1} \leq x_k$. □

Theorem 12.4.5 (n^{th} Root Convergence). *If* $K > 0, x_0 = 1, x_{k+1} = x_k - \frac{x_k^n - K}{n {x_k}^{n-1}}$, *then the* $\lim_{k\to\infty} x_k$ *exists and* $\lim_{k\to\infty} x_k = \sqrt[n]{K}$.

Proof. Since the sequence $\{x_k\}_{k=0}^{\infty}$ is bounded and decreasing, it converges to some number L. Thus, by the limit theorems we know that $L = L - \frac{L^n - K}{n L^{n-1}}$. Simplifying this expression we see that $L^n = K$ and the result follows. □

Exercise Set 12.4.

1. Show that the method for computing the fifth root of a number always converges. Use your method to compute the 5^{th} root of 10. How does the rate of convergence compare with the rate of convergence when the square roots of these numbers are computed? Repeat for the numbers 100,000 and 0.000001.

12.5 The Newton/Raphson Algorithm

Galileo: We would now like to build on the success of the method of Archimedes/Heron. To do that, we need to consider the key ingredients that guarantee the method will always work.

Virginia: In the discussions of the success of each of the square root, cube root, and n^{th} root methods, we only had to worry about three issues:

1. The geometric mean does not exceed the arithmetic mean.

2. The sequence is bounded from below by the root we are seeking.

3. The sequence is always decreasing.

Galileo: So how do these properties interact?

Virginia: The only reason we need the geometric and arithmetic means is to show that $x_n \geq r$, where $r = \sqrt{K}$ or $r = \sqrt[3]{K}$ is the root. We showed the sequence is decreasing by showing that $x_{n+1} = x_n - Q_n$, where Q_n equals a positive number that becomes smaller for each iteration.

Galileo: How did we show Q_n is positive?

Simplicio: The quantity $Q_n = Q(x_n) = \frac{f(x_n)}{f'(x_n)}$ is positive because both $f(x)$ and $f'(x)$ are positive for all $x > r$.

Galileo: Is this sufficient?

Virginia: I am not sure it will suffice to only have $f(x) > 0$ and $f'(x) > 0$. Think of the example $f(x) = xe^{x^2}$. If we initialize the method of Newton/Raphson with a point just to the left of the bump at $x = \frac{\sqrt{2}}{2}$, then the first iteration x_1 will be negative and be to the left of the root $r = 0$. For example if $x_0 = \frac{\sqrt{2}}{2} - 0.001$, then I suspect we will have a problem.

Galileo: Let's consider the shapes of the curves $y = f(x) = x^2 - K$ and $y = f(x) = xe^{x^2}$. Recall from Calculus that concavity is one measure of the shape of a curve. If $f''(x) > 0$ for all x in some interval X, then the curve $y = f(x)$ is concave up.

Virginia: And thus holds water!

Galileo: Correct. On the other hand, if $f''(x) < 0$ for all x in some interval X, then the curve $y = f(x)$ is concave down.

Simplicio: And thus does not hold water!

Galileo: Note that the first curve is concave up on the interval $[0, \infty)$, while the second is concave down on the interval $[0, \sqrt{3}2)$. Note further that when we use the method of Newton/Raphson to find roots of these functions, the approximations differ.

Virginia: In what way?

Galileo: As we have established, the approximations for the positive root of $f(x) = x^2 - K$ form a decreasing sequence which is bounded from below by the root $r = \sqrt{K}$. However, for the function $f(x) = xe^{x^2}$ with a choice of $x_0 = 0.4$, the sequence of

iterates oscillates between positive and negative estimates. The goal of this discussion is to build on the success of Archimedes/Heron. To this end we first state and prove a small proposition, which states that if the method Newton/Raphson produces a sequence which converges to a number L, then L will be a root of the function.

Proposition 12.5.1 (Newton/Raphson Convergence). *Let $f(x) : [a,b] \to \Re$ be a differentiable function with the property that $|f'(x)| \leq M$ for all $x \in [a,b]$. If a sequence of points $\{x_n\}_{n=1}^{\infty}$ in $[a,b]$ is defined recursively by $x_0 \in [a,b]$, $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, and $\lim_{n\to\infty} x_n = r$, then $f(r) = 0$. (i.e. The point $x = r$ is a root of $f(x)$.)*

Proof. We will prove this theorem by showing that if $\epsilon > 0$, then we can find an integer N with the property that $|f(x_n)| < \epsilon$ for all $n \geq N$.

Step 1. (The Challenge)

Let $\epsilon > 0$ be given.

Step 2. (The Choice of N)

Choose N so that if $n \geq N$, then $| x_n - r| < \frac{\epsilon}{2M}$.

Step 3. (The Check)

Since $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, we begin by subtracting x_n from both sides of the equation and multiplying by $f'(x_n)$ so that $f(x_n) = -f'(x_n)(x_{n+1} - x_n)$. Thus, $|f(x_n)| = |f'(x_n)|\ |x_{n+1} - x_n|$.

However, if $|f'(x)| \leq M$ for all $x \in [a,b]$, then $|f(x_n)| \leq M\ |x_{n+1} - x_n| \leq M(|x_{n+1} - r + r - x_n|) \leq M(|x_{n+1} - r| + |r - x_n|) \leq M(\frac{\epsilon}{2M} + \frac{\epsilon}{2M}) \leq 2\frac{\epsilon}{2} = \epsilon.$ □

Simplicio: Actually, I think I can visualize this proposition in the following way. If this proposition were to be false and $f(r) > 0$, then as the the points x_n get close to r the slope of the tangent lines get steeper and steeper. Thus, the slope of the tangent line at $x = r$ should be infinite.

Virginia: While a good idea, I think you have in mind the special case when $x_n \geq r$ for all n and $f(x) > 0$ and $f'(x) > 0$ for all $x > r$. In this setting, we know that $f(x_n) \geq f(L) > 0$ which I agree would force $f'(r) = +\infty$.

Galileo: The next theorem is a generalization of the proof of the convergence of Archimedes/Heron.

Theorem 12.5.2 (Newton/Raphson Convergence 2). *Let $f(x) : [r, +\infty) \to \Re$ be a function with the following properties:*

1. *$f(x)$ has a root at $x = r$,*
2. *$f(x), f'(x)$, and $f''(x)$ exists for all $x \in (r, +\infty)$,*
3. *x_0 is any point $\in (r, +\infty)$, and*
4. *$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.*

If $f(x) > 0, f'(x) > 0$, and $f''(x) > 0$ for each $x \in (r, +\infty)$, then

1. *$x_{n+1} \leq x_n$ (decreasing),*
2. *$r \leq x_{n+1}$ (bounded below by r), and*
3. *$\lim_{n\to\infty} x_n = r$. (convergence)*

Proof. Step 1.

If $x_n \in (, +\infty)$, then $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{pos}{pos} = x_n - pos \leq x_n$. Thus, $x_{n+1} \leq x_n$ and the sequence is decreasing.

Step 2. If we suppose that $x_n > r$, then we must show that $x_{n+1} \geq r$.

If $x_n > r$, then the vertical distance between the curve $y = f(x)$ and the tangent line $y = f(x_n) + f'(x_n)(x - x_n)$ at the point $x = x_n$ is $d_n = f(x_n) + f'(x_n)(r - x_n) - f(r) = f(x_n) - f(r) + f'(x_n)(r - x_n)$. But, by the Mean Value Theorem, there is a point $z \in [r, x_n]$ with the property that $f(x_n) - f(r) = f'(z)(x_n - r)$.

Thus, $d_n = f'(z)(x_n - r) + f'(x_n)(r - x_n) = f'(z)(x_n - r) - f'(x_n)(x_n - r) = (f'(z) - f'(x_n))(x_n - r) = -f''(z_2)(x_n - z)(x_n - r) < 0$. Thus, the tangent line is a negative number at the point $x = r$ and the approximation x_{n+1} must be between r and x_n.

Step 3.

Since the sequence $\{x_n\}_{n=1}^{\infty}$ is bounded from below and decreasing, it converges to some number L. By the previous proposition, we know that $f(L) = 0$. Since we are assuming $f(x) > 0$, for all $x > r$, then we have a contradiction if $f(L) > 0$. Thus, it must be true that $L = r$. □

Simplicio: Despite your motivation, that proof was a bit over my head. How about an example?
Galileo: Sure, how about the polynomial $p(x) = x^3 + x - 1$?

Example 12.5.1. *If $p(x) = x^3 + x - 1$, then note that*

1. $p(0) = -1$,
2. $p(1) = 1$,
3. $p'(x) = 3x^2 + 1 > 0$ *for all* x, *and*
4. $p''(x) = 6x > 0$ *for all* $x > 0$.

Thus, the polynomial has a root $x = r$ between 0 and 1. Since both the first and second derivatives are positive for $x > 0$, we know above theorem applies. Thus, if we initialize Newton/Raphson with any point $x_0 \geq 1$, the method will always converge.

Virginia: I was just thinking about the proof of the Proposition you just presented. If you apply the proof to the function $f(x) = x^2 + 1$, then we know the sequence derived from Newton/Raphson cannot possibly converge. As we showed by computing millions of terms, the sequence bounces all over the place. The inequality $|f(x_n)| \leq M\ |x_{n+1} - x_n|$ is useful here because with our function we know that $f(x) \geq 1$ for all $x \in \Re$. Thus, if we restrict our attention to a particular interval, say $[-1, 1]$, then $f'(x) = 2x$ so that $|f'(x)| \leq 2 = M$ for all $x \in [-1, 1]$. Thus, $1 \leq f(x_n) \leq 2\,|x_{n+1} - x_n|$, which implies that no two consecutive terms of the sequence can be within $\frac{1}{2}$ of one another.
Simplicio: Hmmm.

Exercise Set 12.5.

1. If $p(x) = x^5 + x^3 - 1$, then show that the method of Newton/Raphson can always be used to compute the positive real root.

2. If p and q are positive numbers and $p(x) = x^3 + px - q$, then show that the method of Newton/Raphson can always be used to compute the positive real root.

Chapter 13

Convergence Rates For Sequences

Galileo: While we have mentioned linear and quadratic convergence, we now turn to the problem of making these ideas precise.

Simplicio: You mean you want to know *why* the method of Archimedes/Heron takes 5 or 6 iterations to compute $\sqrt{2}$, while the bisection method takes more than 30?

Galileo: Correct.

Virginia: I think it is interesting that it might be possible to make these ideas precise. It seems like you would only be able to compute a few simple examples and then hope they are representative when ou are confronted by a new problem.

Galileo: I think you will be surprised how easy it is to understand the difference.

Simplicio: Easy is good.

Virginia: What do we have to know?

Galileo: The Mean Value Theorem will be the key for linear convergence, Taylor's Theorem will be the key for quadratic convergence,

13.1 Linear Convergence

Galileo: While the next discussion may appear a bit annoying at first, we now need to define the Newton/Raphson method in terms of functions instead of sequences. The reason for this increase in difficulty is to provide a context so we can present a

careful discussion of the convergence rate.

Example 13.1.1. *Galileo: Let us begin with the simple example* $T(x) = \frac{1}{2}x$. *Note with this example, we have a root at* $x = 0$. *Better yet, we can find that root by letting* $x_0 = 1$ *and making the following computations:*

1. $x_1 = T(x_0) = \frac{1}{2}$,
2. $x_2 = T(x_1) = \frac{1}{2}x_1 = \frac{1}{2^2}, and$
3. $x_3 = T(x_2) = \frac{1}{2}x_2 = \frac{1}{2^3}$.

What do you notice about this sequence?
Simplicio: Well, it is obviously converging to zero.
Galileo: Sure, but how fast?
Simplicio: The error seems to be cut in half at each iteration.
Galileo: Your observation is on target.

Example 13.1.2. *Galileo: Now, we present a slight variation on the previous example by defining* $T(x) = \frac{2}{3}x$?
Simplicio: Well if we let $x_0 = 1$ *and iterate, we see that*

1. $x_1 = T(x_0) = \frac{2}{3}$,
2. $x_2 = T(x_1) = \frac{2}{3}x_1 = (\frac{2}{3})^2, and$
3. $x_3 = T(x_2) = \frac{2}{3}x_2 = (\frac{2}{3})^3$.

Thus, the sequence $\{x_k\}_{k=0}^{\infty}$ *converges to zero. However, this time the error is reduced by only* 33% *at each iteration.*

Galileo: Now I think you can see that these examples lead us to the following definition.

Definition 13.1.1 (Linear Convergence). *If a sequence* $\{x_k\}_{k=0}^{\infty}$ *converges to a number* L*, then the rate of convergence is called linear or* $(1^{st} - order)$ *if there are constants* $K > 0$ *and* $0 \leq M < 1$ *and an integer* N *with the property that if* $k \geq N$*, then* $|x_k - L| \leq KM^k$.

Galileo: In the examples given above, note that the limit $L = 0, K = 1$. In the first example, $M = \frac{1}{2}$, while in the second $M = \frac{2}{3}$. Note also for these examples that $lim_{n\to\infty} M^n = 0$. The next proposition will show that if $0 \leq M < 1$, then this will always be true. Actually, this proposition will be used on a number of different occasions during our future discussions. In particular, we will need this fact when we discuss the convergence of the Geometric Series.

Proposition 13.1.2. *If* $|M| < 1$*, then* $lim_{n\to\infty} M^n = 0$.

Proof. If $M = 0$, then the proof is easy so let us assume that $M \neq 0$.

Step 1. (The Challenge) Let $\epsilon > 0$ be given.

Step 2. (The Choice of N.) Choose $N > \frac{log(\epsilon)}{log(|M|)}$.

Step 3. (The Check that N is sufficiently large.) If $n \geq N$, then $n > \frac{log(\epsilon)}{log(|M|)}$.

Since $|M| < 1, log(|M|) < 0$. Note that the inequality changes signs when we multiply both sides by $log(|M|)$.

Thus, we know $nlog(|M|) < log(\epsilon)$. By the properties of logarithms, $log(|M|^n) < log(\epsilon)$ and we are done. □

Simplicio: I hate to be annoying, but which log function did you use?
Galileo: I guess I was a bit sloppy on that point, but it really doesn't matter. Remember that all log functions are the same up to some constant multiple.

The purpose of the next proposition is to establish sufficient conditions for when we know a sequence converges linearly. Sometimes mathematicians actually use this criterion as the definition for linear convergence. Since our first goal will be to show that the bisection method produces a sequence which converges linearly to a root and since it is not obvious that this criterion is satisfied for the bisection method, we will use the weaker definition given above.

Proposition 13.1.3 (Test for Linear Convergence for a Sequence). *If $|M| < 1$ and $\{x_k\}_{k=0}^{\infty}$ is a sequence with the property that $|x_{k+1} - L| \leq M|x_k - L|$ for all $k \geq 0$, then the sequence $\{x_n\}_{n=0}^{\infty}$ converges linearly to L. In particular, $|x_n - L| \leq |x_0 - L|M^n$ for all $n \geq 0$.*

Proof. Since $|x_{k+1} - L| \leq M|x_k - L|$ for all k, we know

1. If $k = 0$, then $|x_1 - L| \leq M|x_0 - L|$.

2. If $k = 1$, then $|x_2 - L| \leq M|x_1 - L| \leq M^2|x_0 - L|$.

3. If $k = 2$, then $|x_3 - L| \leq M|x_2 - L| \leq M^3|x_0 - L|$.

4. If $k = n - 1$, then $|x_n - L| \leq M|x_{n-1} - L| \leq M^n|x_0 - L|$.

In the definition of linear convergence, note that $K = |x_0 - L|$.

Since $|M| < 1$, we know that $lim_{n\to\infty} M^n = 0$. Thus, $lim_{n\to\infty} x_n = L$. □

Virginia: Can you give us an example of a sequence, which converges but does not converge linearly?

Example 13.1.3. *Galileo: While the sequence $x_k = \frac{1}{k}$ converges to zero at a reasonable rate, it does not converge linearly.*

To show this we actually have to give a short proof by contradiction.

Proof. By way of contradiction, assume there are constants K and M so that $0 \leq M < 1$ and $|\frac{1}{k} - 0| \leq KM^k$ for all $k = 1, 2, \ldots$.

However, if this is true, then by computing the logarithms of both sides, we see that

$$log(\frac{1}{k}) \leq log(K) + k\ log(M)$$

or

$$-log(M) \leq \frac{log(k)}{k} + \frac{log(K)}{k}.$$

Since $0 \leq M < 1, -log(M) > 0$.

Since $\lim_{k\to\infty} \frac{log(k)}{k} = 0$ and $\lim_{k\to\infty} \frac{log(K)}{k} = 0$, we conclude that

$$0 < -log(M) \leq 0,$$

a contradiction. □

Virginia: So I guess this sequence is in the "slow" group.

Galileo: You got it.

Exercise Set 13.1.

1. Determine whether or not the sequence $x_n = \frac{1}{n!}$ converges linearly to zero.
2. Determine whether or not the sequence $x_n = \frac{1}{n^n}$ converges linearly to zero.
3. Prove: The sequence $x_k = \frac{1}{k^2}$ does NOT converge linearly to zero.
4. Prove: If $lim_{n\to\infty}|\frac{x_{n+1}-L}{x_n-L}| = M < 1$, then the sequence $\{x_n\}_{n=0}^{\infty}$ converges linearly to L.

13.2 Linear Convergence for the Bisection Method

Galileo: Now let us now show that the bisection method converges linearly. All we have to do is show that our error formula satisfies the definition for linear convergence.

Proposition 13.2.1 (Linear Convergence for the Bisection Method). *Let $f(x) : [a, b] \to \Re$ be a function, which is continuous at each $x \in [a, b]$ and either $f(a) > 0$ and $f(b) < 0$ or $f(a) < 0$ and $f(b) > 0$. If $[a_n, b_n]$ denotes a sequence of intervals defined by the Bisection Method, r is a root of $f(x)$ with the property that $r \in [a_n, b_n]$, for all n, and $E_n = r - a_n$ denotes the error between a_n and r, then $|E_n| = |a_n - r| \leq (b - a)\frac{1}{2^n}$.*

Proof. Let $[a_0, b_0] = [a, b]$. Since $r \in [a_n, b_n]$ for all n, we know

$|E_1| = |a_1 - r| \leq (b_1 - a_1) \leq (b_0 - a_0)\frac{1}{2}$.

$|E_2| = |a_2 - r| \leq (b_2 - a_2) \leq (b_1 - a_1)\frac{1}{2} \leq (b_0 - a_0)\frac{1}{2^2}$.

$|E_n| = |a_n - r| \leq (b_n - a_n) \leq (b_{n-1} - a_{n-1})\frac{1}{2} \leq (b_0 - a_0)\frac{1}{2^n}$. □

Galileo: Actually, we could have made a slightly smarter choice for the approximation to the root if we had chosen the midpoint $m_n = \frac{a_n+b_n}{2}$. With this choice we see that then $|E_n| \leq (b-a)(\frac{1}{2})^{n+1}$.

Simplicio: Is that all there is to it?

Galileo: Some topics are easy.

13.3 Linear Convergence For Newton/Raphson

Galileo: I chose these examples because they provide insight into why the square root and cube root algorithms converge.

Proposition 13.3.1 (Linear Convergence for Archimedes/Heron). *If $K > 0, x_0 \geq \sqrt{K}$, and $x_{n+1} = \frac{x_n+\frac{K}{x_n}}{2}$, then the sequence $\{x_n\}_{n=0}^{\infty}$ converges linearly to $\sqrt{K}$. Moreover, $|x_n - \sqrt{K}| \leq (\frac{1}{2})^n |x_0 - \sqrt{K}|$.*

Proof. If $f(x) = x^2 - K$, where $K > 0$, the square root algorithm is given by the function $T(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2-K}{2x} = \frac{1}{2}x + \frac{K}{2x}$. Since $T(x) \geq \sqrt{K}$ for all $x \geq \sqrt{K}$, the domain and range of this function can both be taken to be the interval $[\sqrt{K}, +\infty)$. Since $T'(x) = \frac{1}{2} - \frac{K}{2x^2} \in [0, \frac{1}{2}]$ for all $x \geq \sqrt{K}$, we can apply the Mean Value Theorem to the function $T(x)$ at the values $a = x_k$ and $b = x_{k+1}$ to get $x_{k+1} - \sqrt{K} = T(x_k) - T(\sqrt{K}) = T'(z)(x_k - \sqrt{K})$.

Thus, if we initialize our algorithm with a choice of $x_0 \geq \sqrt{K}$, then for all integers $k \geq 1$ we see that $|x_{k+1} - \sqrt{K}| = |T'(z)(x_k - \sqrt{K})| \leq \frac{1}{2}|x_k - \sqrt{K}|$. Thus, by the Test for Linear Convergence we see that the sequence $\{x_n\}_{n=0}^{\infty}$ converges linearly to $\sqrt{K}$ and $|x_n - \sqrt{K}| \leq (\frac{1}{2})^n |x_0 - \sqrt{K}|$. □

Thus, the difference between the $(n)^{th}$ estimate and $\sqrt{K}$ is less than 50% of the difference between the previous estimate and $\sqrt{K}$ for all n.

Simplicio: I noticed that you suddenly changed the assumption in the Archimedes/Heron algorithm from $x_0 = 1$ to $x_0 \geq \sqrt{K}$. What is going on here?

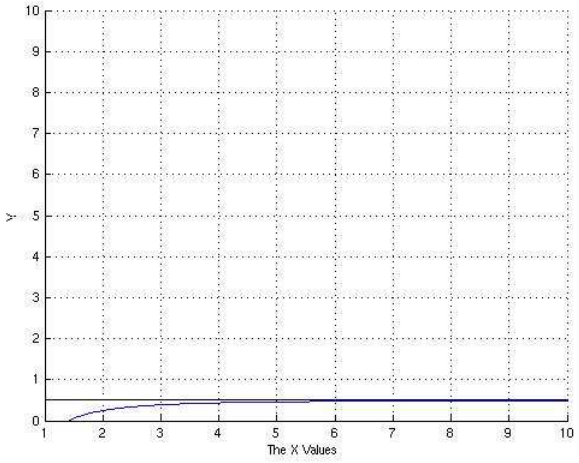


Figure 13.1: The Graph of $y = T'(x) = \frac{1}{2} - \frac{K}{2x^2}$ when $T(x) = \frac{1}{2}x + \frac{K}{2x}$

Galileo: I tried to slip that past you, but you caught me. The reason is that $|T'(1)|$ may exceed 1. Even though it will always be true that $x_1 = T(x_0) \geq \sqrt{K}$, the statement of the proposition is cleaner if we assume $x_0 \geq \sqrt{K}$. Maybe we should have always initialized the algorithm with $x_0 = \frac{K}{2}$. If $K \geq 4$, we will always know that $x_0 \geq \sqrt{K}$.

Simplicio: What about the cube root algorithm?

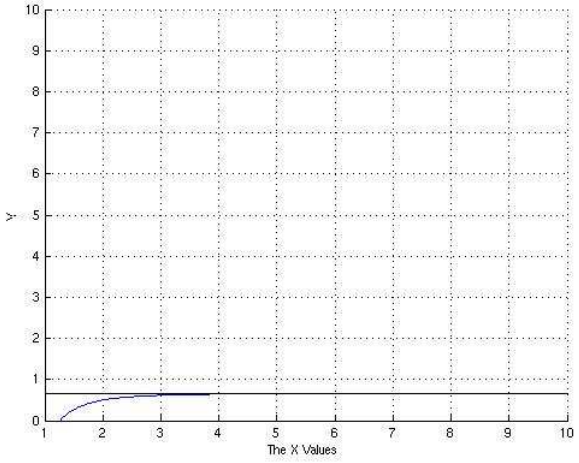


Figure 13.2: The Graph of $y = T'(x) = \frac{2}{3} - \frac{K}{x^3}$ when $T(x) = \frac{2}{3}x + \frac{K}{3x^2}$

Galileo: Same game. Begin by letting $f(x) = x^3 - K$. If we initialize our algorithm with a choice of $x_0 \geq \sqrt[3]{K}$, then for all integers $n \geq 1$ we see that

$$T(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^3 - K}{3x^2} = \frac{2}{3}x + \frac{K}{3x^2}.$$

Thus, $T'(x) = \frac{2}{3} - \frac{2K}{3x^3}$. By looking at the graph of the function we see that $T'(x) \in [0, \frac{2}{3}]$ for all $x \geq \sqrt[3]{K}$. By the Mean Value Theorem we can again apply the Linear Convergence Criterion to make the estimate $|x_n - \sqrt[3]{K}| = |T(x_{n-1}) - T(\sqrt[3]{K})| \leq \frac{2}{3}|x_{n-1} - \sqrt[3]{K}| \leq (\frac{2}{3})^n |x_0 - \sqrt[3]{K}|$ so that the sequence converges linearly to $\sqrt[3]{K}$.

Example 13.3.1. *In our example where $f(x) = (x - 1000)^2$ and $x_0 = 1$, recall that the sequence of Newton/Raphson iterates converged to the root $r = 1000$. If we once again let $T(x) = x - \frac{f(x)}{f'(x)} = \frac{x}{2} + 500$, then note that $T(1000) = 1000$ and $T'(x) = \frac{1}{2}$. Thus, $|T'(x)| = \frac{1}{2} < 1$ for all $x \in \Re$. By the Mean Value Theorem, we can see that if x_n denotes the n^{th} iterate generated by the method of Newton/Raphson, then $|x_n - r| = |x_n - 1000| = |T(x_{n-1}) - T(1000)| = |\frac{1}{2}||x_{n-1} - 1000| = \frac{1}{2}|x_{n-1} - 1000|$ for all n. Thus,*

1. $|x_1 - 1000| = \frac{1}{2}|x_0 - 1000|$,
2. $|x_2 - 1000| = \frac{1}{2}|x_1 - 1000| = (\frac{1}{2})^2|x_0 - 1000|$,
3. $|x_3 - 1000| = \frac{1}{2}|x_2 - 1000| = (\frac{1}{2})^3|x_0 - 1000|$,
4. $|x_4 - 1000| = \frac{1}{2}|x_3 - 1000| = (\frac{1}{2})^4|x_0 - 1000|$,
5. $\vdots$
6. $|x_n - 1000| = \frac{1}{2}|x_{n-1} - 1000| = (\frac{1}{2})^n|x_0 - 1000|$,

Thus, our error is reduced by 50% for each iteration. Note also that the closer the initial guess is to the final answer, the better the approximation. This example should help make the Theorem on Linear Convergence for Newton/Raphson more concrete.

Let X be an interval in $\Re$. If the function $f(x) \in C^2(X)$, then define a new transformation by the rule $T(x) = x - \frac{f(x)}{f'(x)}$. If $f'(x) \neq 0$ for all $x \in X$, then $T(x)$ will be well-defined for all $x \in X$. We assume $f(x) \in C^2(X)$ because we will want to compute $T'(x)$ and $f''(x)$ appears as a factor in the formula for $T'(x)$. Also, if r is a root of $f(x)$, then $T(r) = r$. Conversely, if $T(r) = r$, then $f(r) = 0$. Note also that the sequence of points generated by the method of Newton/Raphson can be written as $x_{k+1} = T(x_k)$. For example, if $f(x) = x^2 - K$, then $T(x) = x - \frac{x^2-K}{2x}$.
Galileo: The first step is to compute the derivatives of $T(x)$. This information is stored in the following proposition.

Proposition 13.3.2. *Let X be an interval in $\Re$. Let $T(x) : X \to \Re$ be defined by the formula $T(x) = x - \frac{f(x)}{f'(x)}$, where $f(x) \in C^2(X)$ and $f'(x) \neq 0$ for all $x \in X$, then $T'(x) = \frac{f(x) \cdot f''(x)}{[f'(x)]^2}$ for all $x \in X$.*

Proof. Use the quotient rule from Calculus to compute the derivative of $T(x)$. □

Galileo: Note in the previous proposition that the minus sign in the formula $T(x) = x - \frac{f(x)}{f'(x)}$ is the key to the simplification.
Simplicio: The minus sign?
Galileo: Note that if $f(x) = x^2 - K$ and $T(x) = x - \frac{f(x)}{f'(x)} = \frac{1}{2}x + \frac{K}{2x}$, then the domain and range of $T(x)$ are the intervals $[\sqrt{K}, \infty)$. Thus, $T(x) : [\sqrt{K}, \infty) \to [\sqrt{K}, \infty)$. The first derivative is $T'(x) = \frac{1}{2} - \frac{K}{2x^2}$, which has the property that $0 \leq T'(x) < \frac{1}{2}$ for all $x \in [\sqrt{K}, \infty)$. We showed earlier that if $x_0 \in [\sqrt{K}, \infty)$ and $x_{k+1} = T(x_k)$, then the sequence $\{x_k\}_{k=0}^{\infty}$ converges to $\sqrt{K}$.

The next proposition provides general conditions which guarantee that the Newton/Raphson sequence will converge to a root. While it may appear a bit forbidding at first, it is not so difficult to remember if you keep the previous examples in mind when you read it. Better yet, the proof is no more difficult than the these examples already discussed.

Theorem 13.3.3 (Linear Convergence for Newton/Raphson). *Let X be an interval in $\Re$. Let $f(x) : X \to X$ be a function with the property that the functions*

$f(x), f'(x)$, and $f''(x)$ are continuous at each $x \in X$. If

1. *$x = r$ is a root of $f(x)$,*
2. *$f'(x) \neq 0$ for all $x \in X$,*
3. *$T(x) = x - \frac{f(x)}{f'(x)}$,*
4. *$T(x) \in X$ for all $x \in X$, and*
5. *$|T'(x)| \leq M < 1$ for all $x \in X$,*

then for any choice of $x_0 \in X$ the sequence defined by $x_{n+1} = T(x_n)$ converges linearly to the root r. Moreover, $|x_n - r| \leq M^n |x_0 - r|$ for all n.

Proof. Let $x_0 \in X$.

For any integer n we know by the Mean Value Theorem that there is a point z between x_0 and r such that $T(x_n) - T(r) = T'(z)(x_n - r)$.

Since $T(x_n) = x_{n+1}$ and $T(r) = r$, $x_{n+1} = r + T'(z)(x_n - r)$. Since $|T'(z)| \leq M < 1$, $|x_{n+1} - r| \leq M|x_n - r|$ so that x_{n+1} is not only between r and x_n, but closer to r than the previous estimate.

Since $x_{n+1} - r = T'(z)(x_n - r)$, $|x_1 - r| \leq M|x_0 - r|$ so that $|x_2 - r| \leq M|x_1 - r| \leq M^2|x_0 - r|$, $|x_3 - r| \leq M|x_2 - r| \leq M^3|x_0 - r|$, etc. Thus, the general pattern emerges that for all n $|x_n - r| \leq M^n|x_0 - r|$. Since $M < 1$, the sequence $\{M^n\}_{n=0}^{\infty}$ converges to zero. Consequently the sequence $\{x_n\}_{n=0}^{\infty}$ converges to r. □

Simplicio: OK, the examples helped in following the proof, but what if $M = 0.99$?
Galileo: If $M = 0.99$, then number of computations required to achieve a reasonable degree of accuracy could be quite large. For example, if we would like to find the number of iterations required to guarantee accuracy of 0.1, then we have to find an integer n so that $(0.99)^n < 0.1$. Solving for n we find that $n > \frac{-log(10)}{log(0.99)} = 229.1053$. If we would like accuracy of less than 0.01, then we would have to choose $n > \frac{-log(100)}{log(0.99)} = 458.2106$.
Simplicio: What if $M > 1.0$?

Galileo: First, the proposition doesn't allow for this case so from a technical point of view your question is irrelevant. However, if the function $T(x)$ has the property that $|T'(x)| > 1.0$ for numerous points $x \in X$, then the iterates may even diverge.

Example 13.3.2. *Galileo: If* $f(x) = x^{\frac{1}{3}}$, *then* $T(x) = -2x$. *Thus, if* $x_0 = 1$ *and* $x_{n+1} = T(x_n)$, *then we obtain the following sequence of iterates.*

x_0	1.000000000000000
x_1	-2.00000000000000
x_2	4.00000000000000
x_3	-8.00000000000000
x_4	16.00000000000000
x_5	-32.00000000000000
x_6	64.00000000000000

Table 13.1: Six Computations of $x_{n+1} = T(x_n) = -2x$

Simplicio: Even I can see that this sequence is oscillating to $\pm\infty$.

Virginia: Now that we have discussed all this theory, how about a simple question we can all understand. In particular, we know that the Newton/Raphson method works for all cubic polynomials of the form $f(x) = x^3 - K$. Right?

Galileo: Correct.

Virginia: But what if we ask: Does Newton/Raphson work for any cubic polynomial? In fact, let us make the question even easier by restricting our attention to polynomials of the form $f(x) = x^3 + px + q$, where $p > 0$. Since we know that the Cardano formulas can be used to write down an answer, it would be reassuring to know that Newton/Raphson will also produce an answer. We also know by the examples we have discussed that Newton/Raphson may fail. The reason the question interests me is because if $f(x) = x^3 + px + q$, then

$$T(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^3 + px + q}{3x^2 + p}.$$

Thus,

$$T'(x) = \frac{f(x)f''(x)}{f'(x)^2} = \frac{(x^3 + px + q)(6x)}{(3x^2 + p)^2}.$$

For large x we know that $|T'(x)| \leq \frac{6}{9} + \epsilon =\leq \frac{6}{9} + \frac{1}{9} = \frac{7}{9} < 1$ so this problem seems to fit the above Proposition if x is "out near infinity." Of course, if x is near zero, $T'(x)$ could be quite large so the condition that $|T'(x)| \leq \frac{7}{9} < 1$ will not always be satisfied. Galileo: I don't know the answer immediately.

Exercise Set 13.3.

1. If $f(x) = x^5 - K$, then find $T(x)$.

2. If $f(x) = x^5 - K$ and $x_0 = 1$, then show that the Newton/Raphson algorithm converges linearly to the root $\sqrt[5]{K}$.

3. If $f(x) = x^7 - K$ and $x_0 = 1$, then show that the Newton/Raphson algorithm converges linearly to the root $\sqrt[7]{K}$.

4. If $f(x) = (x - 10,000)^2$ and $x_0 = 1$, then show that the Newton/Raphson algorithm converges linearly to the root $r = 10,000$. How much is the error reduced for each iteration?

5. If $f(x) = (x - 10,000)^3$ and $x_0 = 1$, then show that the Newton/Raphson algorithm converges linearly to the root $r = 10,000$. How much is the error reduced for each iteration?

6. If $f(x) = xe^{-x^2}$, then find an interval $(-a, a)$ so that the function $T(x) = x - \frac{f(x)}{f'(x)}$ has the property that $|T'(x)| \leq 1.0$ for all $x \in (-a, a)$. Show also that the Newton/Raphson algorithm converges linearly to the root $x = 0$ in that interval.

7. Compute $T'(x)$ for the functions $f(x) = x^5 - K, f(x) = x^7 - K$, and $f(x) = x^n - K$. What do you notice about $T'(x)$ when $x \geq r$, where $r = \sqrt[n]{K}$ is a root?

13.4 Quadratic Convergence For Newton/Raphson

Galileo: We now address two key issues associated with the Newton/Raphson method. Since our computational experiments indicate that it converges rapidly, our first goal is to understand exactly what the phrase "rapid convergence" means. Since the method fails (with a poor choice of initial point) for functions as easy to define as $f(x) = xe^{-x^2}$, the second issue is to determine an interval of convergence for the method.

Again, we call on our friend Taylor to explain the issues involved with this analysis. Taylor: We begin by defining two key functions, which generate sequences exhibiting the difference between linear and quadratic convergence.

$$\begin{aligned} T_1(x) &= \tfrac{1}{2}x \\ T_2(x) &= \tfrac{1}{2}x^2 \end{aligned}$$

Example 13.4.1. *Sequences generated by $T_1(x)$ converge linearly to zero.*

Using the function $T_1(x)$ and a real number x_0, define the following sequence:

$$\begin{aligned} x_1 &= T_1(x_0) &&= \tfrac{1}{2}x_0 \\ x_2 &= T_1(x_1) &&= \tfrac{1}{2}x_1 &&= \tfrac{1}{4}x_0 \\ x_3 &= T_1(x_2) &&= \tfrac{1}{2}x_2 &&= \tfrac{1}{8}x_0 \\ x_4 &= T_1(x_3) &&= \tfrac{1}{2}x_3 &&= \tfrac{1}{16}x_0 \\ &\vdots \\ x_{k+1} &= T_1(x_k) &&= \tfrac{1}{2}x_k &&= \tfrac{1}{2^{k+1}}x_0 \end{aligned}$$

Thus, for any choice of x_0 the limit $\lim_{k\to\infty} x_k = 0$. If we define $T_1(x) = Mx$, where $M \in (-1, 1)$, then the resulting sequence also converges to zero. The closer M is to zero, the faster the sequence converges. If the value of M is close to 1.0 (e.g. $M = 0.99$), then the sequence converges slowly.

Example 13.4.2. *Sequences generated by $T_2(x)$ converge quadratically to zero.*

Using the function $T_2(x)$ and a real number x_0 define the following sequence:

$$\begin{aligned}
x_1 &= T_2(x_0) &&= \tfrac{1}{2}x_0^2 \\
x_2 &= T_2(x_1) &&= \tfrac{1}{2}x_1^2 &&= \tfrac{1}{8}x_0^4 \\
x_3 &= T_2(x_2) &&= \tfrac{1}{2}x_2^2 &&= \tfrac{1}{2^7}x_0^8 \\
x_4 &= T_2(x_3) &&= \tfrac{1}{2}x_3^2 &&= \tfrac{1}{2^{15}}x_0^{16} \\
&\vdots \\
x_{k+1} &= T_2(x_k) &&= \tfrac{1}{2}x_k^2 &&= \tfrac{1}{2^{2^{k+1}-1}}x_0^{2^k}
\end{aligned}$$

Note that if $x_0 \in (-2, 2)$, *then* $\lim_{k\to\infty} x_k = 0$. *If* $x_0 = \pm 1$, *then* $\lim_{k\to\infty} x_k = 2$. *If* $|x_0| > 2$, *then the sequence* $\{|x_k|\}_{k=0}^{\infty}$ *becomes arbitrarily large and thus does not converge.*

Simplicio: OK, let's see some numbers.

Galileo: Note that if $x_0 = 0.1$, *then the sequence will be within single precision accuracy (i.e. within* 10^{-10}*) after only 3 iterations and within double precision accuracy (i.e. within* 10^{-14}*) after only 4 iterations.*

x	$x_k = T_1(x_{k-1})$	$x_k = T_2(x_{k-1})$
x_0	1.00000000000000	1.00000000000000
x_1	0.50000000000000	0.50000000000000
x_2	0.25000000000000	0.12500000000000
x_3	0.12500000000000	0.00781250000000
x_4	0.06250000000000	0.00003051757812
x_5	0.03125000000000	0.00000000046566
x_6	0.01562500000000	0.00000000000000

Table 13.2: Six Computations of $x_{n+1} = T_1(x_n) = \frac{1}{2}x$ and $x_{n+1} = T_2(x_n) = \frac{1}{2}x^2$

Galileo: How about those numbers?

Simplicio: They sure look familiar. In fact, they are almost the same as the sequence we computed for $\sqrt{2}$.

Galileo: You got it.

Galileo: Let us summarize these two examples by making the following observations for more general choices of the initial value x_0.

1. If $x_0 \in \Re$ and $T_1(x) = \frac{1}{2}x$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated recursively by $x_{n+1} = T_1(x_n)$ always converges to zero.

2. If $|x_0| < 2$ and $T_2(x) = \frac{1}{2}x^2$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated recursively by $x_{n+1} = T_2(x_n)$ always converges to zero.

3. If $|x_0| > 2$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated by the function $T_2(x)$ always diverges.

4. If $x_0 = 2$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated by the function $T_2(x)$ converges to one.

5. If $x_0 = -2$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated by the function $T_2(x)$ oscillates between 1 and -1 (and thus diverges).

6. If $|x_0| < 2$, then the sequence of points generated by $T_2(x)$ converges to zero faster than the one generated by $T_1(x)$.

The rate of convergence associated with $T_2(x)$ is called *quadratic* (or 2^{nd}-order) convergence.

Taylor: We formalize the above concepts in the following definitions.

Definition 13.4.1 (Quadratic Convergence). *If a sequence $\{x_n\}_{n=0}^{\infty}$ converges to a number L, then the rate of convergence is called quadratic (or $2^{nd} - order$) if there is a constant M and an integer N such that if $n \geq N$, then $|x_{n+1} - L| \leq M|x_n - L|^2$.*

Example 13.4.3. *Galileo: Let's begin by showing the method of Archimedes/Heron generates a quadratically converging sequence. Note the similarity between this discussion and the sequence generated by $T_2(x)$.*

If $K > 1$, $f(x) = x^2 - K$ and $x_0 > \sqrt{K}$, then the the method of Archimedes/Heron generates a sequence, which converges quadratically to $\sqrt{K}$.

Let $T(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2-K}{2x}$. *As we have noted many times before, the point* $r = \sqrt{K}$ *is a root of* $f(x)$.

If $x > \sqrt{K}$ *and* $x_0 = r$, *then by Taylor's Theorem we know there exists a point* $z \in [\sqrt{K}, \infty)$ *with the property that*

$$T(x) = T(r) + T'(r)(x-r) + \frac{T'(z)}{2}(x-r)^2.$$

Since $r = \sqrt{K}$,

$$T(x) = T(\sqrt{K}) + T'(\sqrt{K})(x-\sqrt{K}) + \frac{T'(z)}{2}(x-\sqrt{K})^2.$$

Since $T(x) = x - \frac{x^2-K}{2x} = \frac{1}{2}x + \frac{K}{2x}$, *note that*

1. $T'(x) = \frac{1}{2} - \frac{K}{2x^2}$ *and*

2. $T''(x) = \frac{K}{x^3}$.

Thus,

1. *if* $x \in [\sqrt{K}, \infty)$, *then* $T(x) \in [\sqrt{K}, \infty)$,

2. $T(\sqrt{K}) = \sqrt{K}$,

3. $T'(\sqrt{K}) = \frac{1}{2} - \frac{K}{2(\sqrt{K})^2} = \frac{1}{2} - \frac{1}{2} = 0$, *and*

4. *if* $K > 1$, *then* $|T''(x)| \leq |T''(\sqrt{K})| = \frac{K}{K^{\frac{3}{2}}} = \frac{1}{\sqrt{K}} \leq 1$.

Also, since $T''(x) = \frac{K}{x^3}$, *we see that for any* $x \in [\sqrt{K}, +\infty)$,

$$|T''(x)| \leq |T''(\sqrt{K})| = \frac{K}{K^{\frac{3}{2}}} = \frac{1}{\sqrt{K}} \leq 1.$$

Thus, the constant $M = 1$ *will have the property that* $|T''(x)| \leq M = 1$, *for any* $x \in [\sqrt{K}, \infty)$.

Thus, by Taylor's Theorem there is a point $z \in [\sqrt{K}, +\infty)$ *with the property that*

$$\begin{aligned} T(x) &= T(\sqrt{K}\ ) + T'(\sqrt{K})(x - \sqrt{K}\ ) + \frac{T''(z)}{2}(x - \sqrt{K}\ )^2 \\ &= \sqrt{K} + \frac{T''(z)}{2}(x - \sqrt{K}\ )^2. \end{aligned}$$

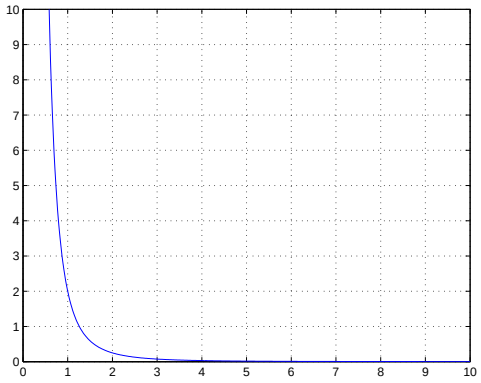


Figure 13.3: The Graph of $y = T''(x) = \frac{2}{x^3}$.

If n is any integer, $x = x_n$, and $x_{n+1} = T(x_n)$, then there is a point $z = z_n \in [\sqrt{K}, +\infty)$ so that

$$\begin{aligned} |x_{n+1} - \sqrt{K}| &= |T(x_n) - T(\sqrt{K})\ | \\ &= \frac{|T''(z_n)|}{2}\ (x_n - \sqrt{K}\ )^2 \leq\ \frac{1}{2}(x_n - \sqrt{K}\ )^2. \end{aligned}$$

To illustrate the power of what we have achieved, let's consider the special case when $K = 3^2 = 9$. Of course, this choice implies that the root $r = \sqrt{9} = 3$. If the initial guess is $x_0 = 4$, then

$$\begin{aligned} |x_1 - 3| \leq& \frac{1}{2}|x_0 - 3|^2 = && \frac{1}{2}(4-3)^2 = \frac{1}{2} \\ |x_2 - 3| \leq& \frac{1}{2}|x_1 - 3|^2 \leq && \frac{1}{2}(x_1 - 3)^2 = \frac{1}{2}\ (\frac{1}{2})^2 = (\frac{1}{2})^3 \\ |x_3 - 3| \leq& \frac{1}{2}|x_2 - 3|^2 \leq && \frac{1}{2}(x_2 - 3)^2 = \frac{1}{2}\ ((\frac{1}{2})^3)^2 = (\frac{1}{2})^7 \\ |x_4 - 3| \leq& \frac{1}{2}|x_3 - 3|^2 \leq && \frac{1}{2}(x_3 - 3)^2 = \frac{1}{2}\ ((\frac{1}{2})^7)^2 = (\frac{1}{2})^{15} \end{aligned}$$

In general,

$$|x_n - 3| \leq (\frac{1}{2})^{2^n - 1}.$$

Simplicio: But wait a minute, what if I choose the initial guess to be $x_0 = 5$? With

this choice, we see that

$$\begin{aligned}
|x_1 - 3| \leq & \frac{1}{2}|x_0 - 3|^2 = & \frac{1}{2}(5-3)^2 = \frac{1}{2}\ 2^2 = 2,\\
|x_2 - 3| \leq & \frac{1}{2}|x_1 - 3|^2 \leq & \frac{1}{2}(x_1-3)^2 = \frac{1}{2}\ 2^2 = 2,\\
|x_3 - 3| \leq & \frac{1}{2}|x_2 - 3|^2 \leq & \frac{1}{2}(x_2-3)^2 = \frac{1}{2}\ 2^2 = 2,\\
|x_4 - 3| \leq & \frac{1}{2}|x_3 - 3|^2 \leq & \frac{1}{2}(x_3-3)^2 = \frac{1}{2}\ 2^2 = 2.
\end{aligned}$$

Galileo: Thus, if our initial guess that is far from the root, then these inequalities do not provide any useful information.

Virginia: But the same is true of our function $T_2(x) = \frac{1}{2}x^2$. If we choose $x_0 = 2$, then the sequence $x_{n+1} = T_2(x_n)$ diverges. Mr. Simplicio, you have simply pointed out that poor initial choices lead to evil outcomes.

Galileo: The next theorem shows that this example generalizes to any function $f(x)$.

Simplicio: This theorem looks complicated.

Galileo: Even though it has 6 separate hypotheses, they all say something you would want to have happen with the function and its first and second derivatives.

Theorem 13.4.2 (Quadratic Convergence for Newton/Raphson). *Let X be a closed interval in $\Re$ and let $f(x) : X \to X$ be a function. If*

1. *$f(x), f'(x), f''(x)$, and $f'''(x)$ are all continuous at each $x \in X$,*
2. *$x = r \in X$ is a root of $f(x)$,*
3. *$f'(x) \neq 0$ for all $x \in X$,*
4. *$T(x) = x - \frac{f(x)}{f'(x)} \in X$ for all $x \in X$*
5. *$|T'(x)| \leq M_1 < 1$ for all $x \in X$, and*
6. *$|T''(x)| \leq M_2$ for all $x \in X$,*

then for any choice of $x_0 \in X$ *the sequence defined by* $x_{n+1} = T(x_n)$ *converges quadratically to the root* r. *In fact, for all* n *we know that* $|x_{n+1} - r| \leq \frac{M_2}{2}|x_n - r|^2$.

Proof. If $f(x) : [a, b] \to \Re$ is a function with the property that $f(x)$, $f'(x)$, and $f''(x)$ are all continuous at each $x \in [a, b]$ and $f'(x) \neq 0$ for all $x \in X$, then $T(x) = x - \frac{f(x)}{f'(x)}$ is differentiable and

$$T'(x) = 1 - \frac{f'(x)f'(x) - f''(x)f(x)}{(f'(x)\ )^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Since $|T'(x)| \leq M_1 < 1$ for all $x \in X$, we know by the Mean Value Theorem that the sequence defined by $x_{n+1} = T(x_n)$ converges linearly to the root $x = r$.

By Taylor's Theorem we know that there is a point $z \in X$ such that

$$T(x) = T(r) + T'(r)(x - r) + \frac{T''(z)}{2}(x - r)^2.$$

Since $f(r) = 0, T(r) = r - \frac{f(r)}{f'(r)} = r - 0 = r$. Since $T'(x) = \frac{f(x)f''(x)}{(f'(x)\ )^2}$, $T'(r) = 0$. (Thus, if r is a root of $f(x)$, then r is a fixed point of $T(x)$ and also a root of $T'(x)$.)

Thus,

$$T(x) = T(r) + T'(r)(x - r) + \frac{T'(z)}{2}(x - r)^2 = r + \frac{T'(z)}{2}(x - r)^2.$$

Hence, for any $x \in X$, there is a point $z \in X$ so that

$$T(x) - r = \frac{T'(z)}{2}(x - r)^2.$$

If $|T''(x)| \leq M$ for all $x \in X$, then

$$|T(x) - r| \leq \frac{M}{2}(x - r)^2 \text{ for all } x \in X.$$

If n is any integer, $x = x_n$, and $x_{n+1} = T(x_n)$, then just as in the special case with Archimedes/Heron we see that

$$|x_{n+1} - r| = |T(x_n) - r| \leq \frac{M_2}{2}(x_n - r)^2.$$

Since the sequence $\{x_n\}_{n=0}^{\infty}$ converges to r, the convergence is quadratic. □

Galileo: Actually, we can now compute an error formula for Newton/Raphson the same way we did for the sequence $\{x_n\}_{n=0}^{\infty}$ generated by the function $T_2(x)$.

Corollary 13.4.3 (Quadratic Error Formula for Newton/Raphson). *If the hypotheses of the Quadratic Convergence Theorem for Newton/Raphson are all satisfied and n is any integer $n \geq 0$, then*

$$|x_n - r| \leq \frac{2}{M_2}[\frac{M_2}{2}(x_0 - r)]^{2^n}.$$

Proof. Since

$$|x_{n+1} - r| = |T(x_n) - r| \leq \frac{M_2}{2}(x_n - r)^2 \text{ for all } n,$$

$$\begin{aligned}
|x_1 - r| &\leq \frac{M_2}{2}(x_0 - r)^2 \\
|x_2 - r| &\leq \frac{M_2}{2}(x_1 - r)^2 \leq \frac{M_2}{2}(\frac{M_2}{2}(x_0 - r)^2)^2 = \frac{2}{M_2}[\frac{M_2}{2}(x_0 - r)]^4 \\
|x_3 - r| &\leq \frac{M_2}{2}(x_2 - r)^2 \leq \frac{M_2}{2}(\frac{2}{M_2}[\frac{M_2}{2}(x_0 - r)]^4)^2 = \frac{2}{M_2}[\frac{M_2}{2}(x_0 - r)]^8
\end{aligned}$$

□

Example 13.4.4. *Galileo: Note that we have already discussed this error formula for the function $f(x) = x^2 - 9$ with initial guesses of $x_0 = 4$ and $x_0 = 5$. In general, if $K \geq 1, f(x) = x^2 - K$, and $x_0 > \sqrt{K}$ is arbitrary, then we still notice that the constant $M_2 = 1$ will dominate the 2^{nd} derivative of $T(x)$. Thus, we see that the root $r = \sqrt{K}$ and*

$$|x_n - \sqrt{K}| \leq \frac{2}{M_2}[\frac{M_2}{2}(x_0 - \sqrt{K})]^{2^n} \leq 2[\frac{1}{2}(x_0 - \sqrt{K})]^{2^n}.$$

Simplicio: So if we are smart enough to choose x_0 close enough to $\sqrt{K}$ so that $|\frac{1}{2}(x_0 - \sqrt{K}) < 1$, then the error estimate will tell us that the sequence will converge rapidly to the root.

Galileo: Correct.

Example 13.4.5. *Galileo: If $K \geq 1$, $f(x) = x^3 - K$, and $x_0 > \sqrt[3]{K}$ is arbitrary, then $T(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^3-K}{3x^2} = \frac{2}{3}x + \frac{K}{3x^2}$. Thus, $T'(x) = \frac{2}{3} - \frac{2K}{3x^3}$ and $T''(x) = \frac{2K}{x^4}$. Thus, $|T''(x)| \leq \frac{2K}{\sqrt[3]{K^4}} = \frac{2}{\sqrt[3]{K}} \leq 2$. Thus, we see that the constant $M_2 = 2$ will dominate the second derivative $|T''(x)|$ and*

$$|x_n - r| = |x_n - \sqrt[3]{K}| \leq \frac{2}{M_2}[\frac{M_2}{2}(x_0 - \sqrt[3]{K})]^{2^n} \leq [(x_0 - \sqrt[3]{K})]^{2^n}.$$

Simplicio: Again, If we are smart enough to choose x_0 close enough to $\sqrt[3]{K}$ so that $|(x_0 - \sqrt[3]{K}) < 1$, then the error estimate will tell us that the sequence will converge rapidly to the root.

Galileo: Correct again.

Simplicio: OK, I understand this error formula now. However, I would like to ask one simple question about that Quadratic Convergence Theorem.

Galileo: Yes.

Simplicio: Why do we have all those hypotheses? Can't we just say that the convergence is always quadratic?

Galileo: Actually, I am sorry to report that the answer to your question is: "No!"

Example 13.4.6. *For example, the polynomial $p(x) = (x-5)^2$ has a double root at $x = 5$. If we apply Newton/Raphson to find this root, we discover that*

$$T(x) = x - \frac{p(x)}{p'(x)} = x - \frac{(x-5)^2}{2(x-5)} = x - \frac{1}{2}(x-5) = \frac{1}{2}x + \frac{5}{2}.$$

While it is easy to show that the convergence rate is linear, the convergence rate fails to be quadratic. The root cause of the problem (pardon the pun) is that the first derivative $p'(x) = 2(x-5)$ happens to also have a root at $x = 5$. Thus, $p'(5) = 0$ and hypothesis 3 in the Quadratic Convergence Theorem is violated.

Simplicio: So?

Galileo: While the error is reduced by 50% at each iteration, the convergence never speeds up the way it does for Archimedes/Heron. Make a few computations and you will see that I am correct.

Virginia: Murphy strikes once again!

Galileo: We now define the term *simple root* to make this distinction. For Newton/Raphson, the bottom line is that we are on firm ground as long as we have simple roots.

Definition 13.4.4. *If $f(x)$ is a differentiable function defined on the interval (a, b) with root $x = r \in (a, b)$, then r is called a simple root if $f'(r) \neq 0$.*

Taylor: Note that if $K > 0$, then the roots of $p(x) = x^n - K$ are simple.

Simplicio: Since $p'(x) = nx^{n-1}$, I can see that $p'(\sqrt[n]{K}) = n\sqrt[n]{K}^{n-1} \neq 0$.

Taylor: In general, a polynomial $p_n(x)$ will have a simple root if and only if it is not repeated. For example, if $p_n(x)$ has a repeated root $x = r$ implies the function $p(x)$ has a factor of $(x-r)^2$. If the root is repeated three times, then $p(x)$ has a factor of $(x-r)^3$. The Fundamental Theorem of Algebra states that any polynomial can be completely factored. Gauss provided five different proofs of this intuitively obvious theorem several hundred years ago. The proofs involve a knowledge of complex variables–a beautiful subject you should know.

Theorem 13.4.5 (Fundamental Theorem of Algebra). *If $a_{n-1}, a_{n-2}, \ldots, a_1, a_0$ are complex numers and $p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0$, then there are complex numbers $r_1, r_2, \ldots, r_n$ with the property that $p(x) = (x - r_1)(x - r_2)\ldots(x - r_n)$.*

Taylor: The next proposition characterizes polynomials, which have a simple root at $x = r$. In particular, a polynomial $p(x)$ has a simple root if and only if it is divisible by the factor $(x - r)$ and not by $(x - r)^2$.

Proposition 13.4.6. *If $a_{n-1}, a_{n-2}, \ldots, a_1, a_0$ are complex numers and $p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0$, then $p(x)$ has a simple root at $x = r$ if and only if $p(x) = (x - r)g(x)$, where $g'(r) \neq 0$.*

Proof. By the Fundamental Theorem of Algebra we know that $p(x) = (x - r_1)(x - r_2)\ldots(x - r_n)$ so that $p(x) = (x - r)g(x)$. By the product rule from Calculus, we know $p'(x) = (x - r)g'(x) + g(x)$. Thus, $p'(r) = g(r)$ so that $p'(r) \neq 0$ if and only if $g(r) \neq 0$. □

Taylor: I hope you agree that we now completely understand the role of simple roots and quadratic convergence when we use the method of Newton/Raphson to compute roots of functions.

Virginia: Yes, I do. However, I have one question. Namely, when used Newton/Raphson to compute a root of $f(x) = x^2 - 0.000001$, the convergence rate was noticeably slower than when we computed a root of $f(x) = x^2 - 2$. This function has simple roots. What is going on here?

Taylor: Excellent question. I think you will understand the answer when if you simply compute the constant M_2. Give it a try.

Exercise Set 13.4.

1. Determine whether or not the sequence $x_n = \frac{1}{n!}$ converges quadratically to zero.

2. Determine whether or not the sequence $x_n = \frac{1}{n^n}$ converges quadratically to zero.

3. Show: If $x_0 \in \Re$ and $T_1(x) = \frac{1}{2}x$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated recursively by $x_{n+1} = T_1(x_n)$ always converges linearly to zero.

4. Show: If $|x_0| < 1$ and $T_2(x) = \frac{1}{2}x^2$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated recursively by $x_{n+1} = T_1(x_n)$ always converges quadratically to zero.

5. Show: If $x_0 \in \Re$ and $T_1(x) = \frac{1}{2}x$, then the sequence of points $\{x_n\}_{n=0}^{\infty}$ generated recursively by $x_{n+1} = T_1(x_n)$ fails to converge quadratically to zero. (Hint: This problem requires a short proof by contradiction.)

6. Determine the rate of convergence for the sequence $x_k = \frac{1}{7^k}$. More specifically, first show the sequence converges linearly to zero, then decide whether or not it converges quadratically to zero. Repeat this exercise for the sequence $x_k = \frac{1}{3^{2^k}}$.

7. Prove: If $T(x) : \Re \to \Re$ is differentiable for each $x \in \Re$, $x_0 \in \Re$, $M \in [0, 1)$, the sequence $x_{n+1} = T(x_n)$ converges to L, and $|T'(x)| \leq M$ for all $x \in \Re$, then the sequence $\{x_n\}_{n=0}^{\infty}$ converges linearly to L.

8. Show: If $K > 1$ and $x_0 > \sqrt[5]{K}$, then the method of Newton/Raphson produces a sequence which converges quadratically to the root $r = \sqrt[5]{K}$ of the function $f(x) = x^5 - K$. (Compute the constants M_1 and M_2.) Note that if $K = 32$, then the root $r = 2$. If $x_0 = 3$, then compute the constant $\frac{M_2}{2}|x_0 - 2|$. How close does the initial guess x_0 have to be chosen to the root $r = 2$ to guarantee that $\frac{M_2}{2}|x_0 - 2| < 1$?

9. If $f(x) = x^3 + 3x + 1$, then show that the method of Newton/Raphson converges quadratically to a root in the interval $[-1, 0]$. (Suggestion: Use a graphing program to show that $|T'(x)| \leq 0.9$ for all $x \in [-10, 10]$.)

10. If $f(x) = (x - 1000)^2$ and $x_0 = 1$, then show that the method of Newton/Raphson does NOT converge quadratically to the root $r = 1000$. Why doesn't the Quadratic Convergence Theorem apply? Which hypothesis is not satisfied?

11. If $f(x) = (x - 1000)^3$ and $x_0 = 1$, then show that the method of Newton/Raphson does NOT converge quadratically to the root $r = 1000$. Why doesn't the Quadratic Convergence Theorem apply? Which hypothesis is not satisfied?

12. If $f(x) = x^2$ or x^3 and $T(x) = x - \frac{f(x)}{f'(x)}$, then show the sequence defined by $x_0 = 1$, $x_{k+1} = T(x_k)$ converges to 0 at a linear, but not quadratic rate. Do these examples contradict the quadratic convergence of the Newton/Raphson method?

13. If $f(x) = x^2 - 0.00001$, then use the method of Newton/Raphson to compute the constant M_2. What do you conclude about the Quadratic Error Formula for Newton/Raphson?

Chapter 14

The Contraction Mapping Theorem

Stefan Banach (1892-1945)

> Mathematics is the most beautiful and most powerful creation of the human spirit. Mathematics is as old as Man.-Stefan Banach

Galileo: We now turn to Stefan Banach's (1892-1945) Contraction Mapping Theorem.
Simplicio: Who was this Banach guy?
Galileo: He was a hard drinking, heavy smoker, who liked to socialize with his friends late into the night at the Scottish Café in Lvov, Ukraine. You probably would have enjoyed his company.

Simplicio: I think I should.

Galileo: His theorem constitutes an amazing generalization of Archimedes/Heron and Newton/Raphson. Not only can this method be used to compute roots of non-linear equations, but it also has applications to areas you would never expect.

Simplicio: Like what?

Galileo: The method can be used to solve a system of linear equations.

Simplicio: We have the technique of row operations. Isn't that good enough?

Galileo: While row operations work fine for small systems, these alternative methods work much better for large sparse systems.

Simplicio: What does "sparse" mean?

Galileo: A matrix is sparse if most of its entries equal zero. Recall that the idea behind row operations is to transform the given matrix into an upper triangular (or even diagonal) form. Thus, the goal is to generate a new matrix with most entries equal to zero. Two problems may arise if the original matrix has most entries equal to zero. The first problem is that we may be wasting our time if we make an entry zero when it is already zero. If we are not careful, we might actually transform zero entries into non-zero entries.

Simplicio: OK, how about another application?

Galileo: The Contraction Mapping Theorem can be used to show the existence and uniqueness of solutions of differential equations.

Simplicio: I don't want to hear math talk about existence and uniqueness.

Galileo: What if the problem you are trying to solve has no solution? You might want to know if a solution exists. If you know a solution exists, you might want to know if there is more that one solution. Uniqueness is useful because once you find a solution, you can go home.

Simplicio: But I don't like differential equations.

Galileo: Unfortunately, many of the most important real-world applications require a differential equation as part of their model. If change occurs, a good bet is that there is a differential equation lurking nearby. How about fractals?

Simplicio: What is a fractal?

Galileo: Fractals are sets with the property that any part of the set is similar to the whole set. More specifically, the entire set can be translated, rotated, and shrunk to fit on top of any subset. In other words, the set is self similar. Fractal techniques can be used to produce beautiful pictures. The wallpaper in my bath is of fractal origin.

Virginia: I have seen the snowflake and the fern and agree they are captivating.

Galileo: Fractal methods can also be used to compress images.

Simplicio: Now that is an application even I can appreciate.

Galileo: As it turns out, the Contraction Mapping Theorem can often be used to solve a problem written in the form $T(x) = x$, where $|T'(x)| < M < 1$, for all x. The solution of such an equation will be a fixed point of $T(x)$.

Simplicio: What is a fixed point?

Galileo: A point $x = F$ is a fixed point for a function $T(x)$ if $T(F) = F$.

Virginia: Just as $F = \sqrt{K}$ is a fixed point of the function $T(x) = x - \frac{x^2-K}{2x}$!

Galileo: Correct.

Virginia: I now understand why you began our discussion with the method of Archimedes/Heron. The ideas of yesterday are the ideas of today.

Galileo: Correct again.

Simplicio: So how do we solve for this fixed point?

Virginia: How about if we begin by making an initial guess $x = x_0$ and then iterate by setting $x_{n+1} = T(x_n)$. That strategy worked before. My hunch would be that the sequence $\{x_n\}_{n=0}^{\infty}$ converges to the point F.

Galileo: You should be teaching this seminar.

Simplicio: What about the convergence rate? I like quadratic.

Galileo: While the convergence rate for Newton/Raphson usually turns out to be quadratic, the convergence rate for the Contraction Mapping Theorem usually turns out to be linear. The contraction factor M controls the rate of convergence. If $T'(F) = 0$, then the argument we used for Newton/Raphson can be used to show the convergence rate is quadratic.

14.1 Contraction Mapping Examples

Galileo: We now turn to a more detailed discussion of the Contraction Mapping Theorem.

Simplicio: How about if we begin with a simple example?

Galileo: Let us begin with the problem that you are to solve the equation $x = \frac{1}{2}x + 3$.

Simplicio: But this problem is too easy. Obviously, the answer is $x = 6$.

Galileo: The answer is easy because you have an excellent understanding of algebra. Remember that more than 1000 years passed between the geometry of the ancient Greeks and the appearance of the commutative, associative, and distributive laws from algebra.

Example 14.1.1. *Solve the equation* $x = \frac{1}{2}x + 3$.

If we let $T(x) = \frac{1}{2}x + 3$, *and* $x_0 = 0$, *then we can iterate in the same way we did for the method of Newton/Raphson. Note that the last computation, namely 5.9766, is beginning to approach the correct answer.*

$$
\begin{aligned}
x_1 = T(x_0) &= 3 \\
x_2 = T(x_1) &= \frac{1}{2}3 + 3 = 4.5 \\
x_3 = T(x_2) &= \frac{1}{2}4.5 + 3 = 5.25 \\
x_4 = T(x_3) &= \frac{1}{2}5.25 + 3 = 5.625 \\
x_5 = T(x_4) &= \frac{1}{2}5.625 + 3 = 5.8125 \\
x_6 = T(x_5) &= \frac{1}{2}5.8125 + 3 = 5.9062 \\
x_7 = T(x_6) &= \frac{1}{2}5.9062 + 3 = 5.9531 \\
x_8 = T(x_7) &= \frac{1}{2}5.9531 + 3 = 5.9766
\end{aligned}
$$

Simplicio: This method is too much work. After a million iterations, we still won't have the exact answer. I prefer using the laws of algebra for this problem.

Galileo: We now repeat this technique to solve a simple non-linear equation.

Example 14.1.2. *Solve the equation* $x = \frac{1}{2}\sin(x) + 13$.

If we let $T(x) = \frac{1}{2}\sin(x) + 13$, *and* $x_0 = 0$, *then we can iterate in the same way we did for the method of Newton/Raphson. Note that the sequence* $\{x_k\}_{k=0}^{\infty}$ *seems to be converging to a number approximately equal to 13.35.*

$$
\begin{aligned}
x_1 &= T(x_0) = 13 \\
x_2 &= T(x_1) = \frac{1}{2}\sin(13) + 13 = 13.21 \\
x_3 &= T(x_2) = \frac{1}{2}\sin(13.21) + 13 = 13.30 \\
x_4 &= T(x_3) = \frac{1}{2}\sin(13.30) + 13 = 13.33 \\
x_5 &= T(x_4) = \frac{1}{2}\sin(13.33) + 13 = 13.35
\end{aligned}
$$

Galileo: Note that no algebraic manipulation of the expression $x = \frac{1}{2}\sin(x) + 13$ can be used to solve this equation for x.

Simplicio: Now I see the point of this example.

Galileo: One final remark is in order. Namely, the method is constructive.

Simplicio: What do you mean by constructive?

Galileo: The method doesn't simply say a solution exists. Instead, the technique provides a procedure to approximate the desired answer. As you might expect, engineers vastly prefer methods where you simply make a guess, compute, and the answer magically appears. The Contraction Mapping Theorem fits that mold exactly.

In fact, the technique can be implemented in the following four lines of computer code:

Let $x = x_0$ be the initial guess.

for n = 0, 1, ..., N

    x = T(x);

end

Exercise Set 14.1.

1. Use the above iterative technique to approximate a solution of the equation $x = \frac{1}{2}\cos(x) + 3$. Begin the process with $x_0 = 0$.

2. Use the above iterative technique to approximate a solution of the equation $x = e^{-x}$. Begin the process with $x_0 = 0$.

3. Use the above iterative technique to approximate a solution of the equation $x = e^{x}$. Begin the process with $x_0 = 0$.

14.2 The Contraction Mapping Theorem in $\Re$

Simplicio: That discussion contained many more technical details than I can tolerate. Let's move on to something more understandable.
Galileo: It isn't as bad as you think, but OK. let's get back to the Contraction Mapping Theorem.
Cauchy: We now check a few technical propositions, which will be used to prove the contraction mapping theorem. The first proposition is the familiar formula for summing a finite geometric series.

The next proposition provides a bound on the difference between two successive terms in a sequence.

Proposition 14.2.1. *If* $|x_{k+1} - x_k| \leq M|x_k - x_{k-1}|$ *for all* $k \geq 1$, *then* $|x_{k+1} - x_k| \leq M^k|x_1 - x_0|$.

Proof. If $k = 1$, then $|x_2 - x_1| \leq M^1|x_1 - x_0|$.

If $k = 2$, then $|x_3 - x_2| \leq M^2|x_1 - x_0|$.

If $k = 3$, then $|x_4 - x_3| \leq M^3|x_1 - x_0|$.

If $k = 4$, then $|x_5 - x_4| \leq M^4|x_1 - x_0|$.

Inductively, $|x_{k+1} - x_k| \leq M^k|x_1 - x_0|$. □

The next proposition provides a bound on the difference between any two terms in a sequence. This proposition is fundamental to proving the contraction mapping

theorem. It is also the key to unlocking the rate of convergence, which is important in real applications.

Proposition 14.2.2 (The Contraction Mapping Error Estimate). *If* $0 \leq M < 1$ *and* $|x_{k+1} - x_k| \leq M|x_k - x_{k-1}|$ *for all* $k \geq 1$, *then whenever* $n \geq N$, $|x_n - x_N| \leq \frac{M^N}{1-M}|x_1 - x_0|$.

Proof. By the triangle inequality and successive applications of the previous proposition, we know that

$$|x_n - x_N| = |x_n - x_{n-1} + x_{n-1} - x_{n-2} + x_{n-2} - \ldots + x_{N+1} - x_N|$$

$$\leq |x_n - x_{n-1}| + |x_{n-1} - x_{n-2}| + |x_{n-2} - x_{n-3}| + \ldots + |x_{N+1} - x_N|$$

$$\leq M^{n-1}|x_1 - x_0| + M^{n-2}|x_1 - x_0| + \ldots + M^N|x_1 - x_0|$$

$$= (M^{n-1} + M^{n-2} + \ldots + M^N)|x_1 - x_0|$$

$$= M^N(M^{n-N-1} + M^{n-N-2} + \ldots + M + 1)|x_1 - x_0|.$$

Since $0 \leq M < 1$,

$$|x_n - x_N| \leq M^N \frac{1 - M^{n-N}}{1 - M}|x_1 - x_0| \leq \frac{M^N}{1 - M}|x_1 - x_0|.$$

□

Proposition 14.2.3. *If* $0 \leq M < 1$ *and* $|x_{k+1} - x_k| \leq M|x_k - x_{k-1}|$ *for all* $k > 0$, *then there exists a unique real number* L *such that* $lim_{k\to\infty} x_k = L$.

Proof. Step 1. Let $\epsilon > 0$ be given.

Step 2. Choose N large enough that $\frac{M^N}{1-M}|x_1 - x_0| < \epsilon$, for all $i \geq j \geq N$.

Step 3. By the previous proposition, we know $|x_n - x_N| \leq \frac{M^N}{1-M}|x_1 - x_0| < \epsilon$.

Thus, the sequence $\{x_k\}_{k=1}^{\infty}$ is Cauchy. Since every Cauchy sequence converges, there is a unique real number L such that $lim_{k\to\infty} x_k = L$. □

Definition 14.2.4. *If* X *is a closed interval in* $\Re$ *and* $T : X \to X$, *then* $T(x)$ *is called a* **contraction** *if there is a number* $0 \leq M < 1$ *such that* $|T(x) - T(y)| \leq M|x - y|$ *for all* $x, y \in X$.

The constant M is called the **contraction factor** of $T(x)$.

Simplicio:: So what is this contraction factor?

Galileo: The intuitive idea of a contraction is exactly what the word implies. Namely, if given any two points $x, y \in X$, then the function $T(x)$ always moves the two points so that they are closer together. Since the absolute value function always produces a measure of distance we know that $dist(x, y) = |x - y|$ and $dist(T(x), T(y)) = |T(x) - T(y)|$. This, if $M < 1$, then $dist(T(x), T(y)) \leq Mdist(x, y)$. Thus, the points x and y are moved closer together. If $M = \frac{1}{2}$, then they will be 50% closer than they were before.

Simplicio:: What if the contraction factor equals 2?

Galileo: If $|T'(x) \geq 2$ for many values of x, then we have an expansion rather than a contraction. While these functions are sometimes studied, we will not consider them.

Simplicio:: How do we tell whether or not a function is a contraction?

Galileo: The purpose of the next proposition is to present a criterion for when a function can be identified as a contraction. The answer is to simply compute the first derivative and check to see if it is always less (in absolute value) than 1. Note that this proposition already appeared in the discussion on the method of Newton/Raphson.

Proposition 14.2.5. *If X is a closed interval in $\Re$ and $T(x)$ is a differentiable function $T : X \to X$ with the property that $|T'(x)| \leq M < 1$ for all $x \in X$, then $T(x)$ is a contraction with contraction factor M.*

Proof. If $x, y \in X$, then by the Mean Value Theorem we know that there is a point $z \in X$ such that $T'(z) = \frac{T(x)-T(y)}{x-y}$. Since $|T'(z)| \leq M < 1$, $|\frac{T(x)-T(y)}{x-y}| \leq M$. Thus, $|T(x) - T(y)| \leq M|x - y|$. □

Galileo: Before we turn to the next idea, we need to prove that contractions are actually continuous functions. This detail will be needed in the proof of the Contraction Mapping Theorem, where we need to know that limits commute with continuous functions. In particular, we need to know that if $lim_{n\to\infty} x_n = P$, then $lim_{n\to\infty} T(x_n) = T(lim_{n\to\infty} x_n) = T(P)$. Another way to phrase this fact is to state

that if a function is continuous at a point P, then limits can be evaluated at P by simply substituting the point P in the function.

Simplicio: In other words, we didn't need limits in the first place.

Galileo: You could say that.

Proposition 14.2.6 (Contractions are Continuous). *If X is an interval and $T(x) : X \to \Re$ is a contraction with contraction factor $0 \leq M < 1$, then $T(x)$ is continuous at every $x \in X$.*

Proof. Let $\overline{x} \in X$.

Step 1. Let $\epsilon > 0$ be given.

Step 2. Choose $\delta = \epsilon$.

Step 3. Since $T(x)$ is a contraction, we know that if $|x - \overline{x}| < \delta$, then $|T(x) - T(\overline{x}) \leq M|x - \overline{x}| < |x - \overline{x}| < \delta = \epsilon$. □

Galileo: We now turn to the second idea embedded in the Contraction Mapping Theorem.

Definition 14.2.7. *If $T : X \to X$ is a function and $T(F) = F$ for some $F \in X$, then the point $F \in X$ is said to be a* **fixed point** *for $T(x)$.*

Galileo: Consider the following examples.

Example 14.2.1. *If $T_1(x) = \frac{1}{2}x$, then $F = 0$ is a fixed point of $T_1(x)$. Note that $T_1(x)$ has exactly one fixed point,*

Example 14.2.2. *If $T(x) = x + 5$, then $T(x)$ has no fixed points.*

Example 14.2.3. *If $T_2(x) = x^2$, then $F = 0$ and $F = 1$ are fixed points for $T_2(x)$. Note that $T_2(x)$ has two fixed points.*

Example 14.2.4. *If $T_3(x) = x^3$, then $F = 0, F = 0$, and $F = 1$ are all fixed points for $T_3(x)$. Note that $T_3(x)$ has three fixed points.*

Example 14.2.5. *If* $T(x) = x - \frac{x^2-K}{2x}$, *then* $T(\sqrt{K}) = \sqrt{K}$. *Thus,* $T(x)$ *has* $\sqrt{K}$ *for a fixed point. In Figure 14.1, this fixed point is displayed as the intersection of the curves* $y = x$ *and* $y = T(x) = x - \frac{x^2-K}{2x}$.

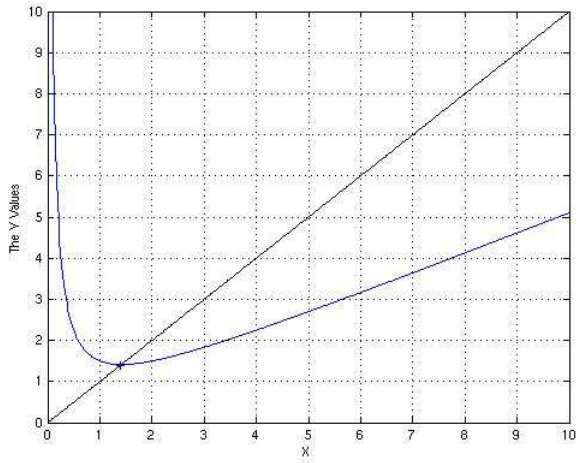


Figure 14.1: The Fixed Point for the Function $T(x) = x - \frac{x^2-K}{2x}$.

Example 14.2.6. *If* $T(x) = x - \frac{x^3-K}{3x^2}$, *then* $T(\sqrt[3]{K}) = \sqrt[3]{K}$. *Thus,* $T(x)$ *has* $\sqrt[3]{K}$ *for a fixed point. In Figure 14.2, this fixed point is displayed as the intersection of the curves* $y = x$ *and* $y = T(x) = x - \frac{x^3-K}{3x^2}$.

Example 14.2.7. *If we want to solve the equation If* $x = T(x) = \frac{1}{2}\sin(x) + 13$, *then the solution is the fixed point* F *of* $T(x)$. *In Figure 14.3, this fixed point is displayed as the intersection of the curves* $y = x$ *and* $y = T(x) = \frac{1}{2}\sin(x) + 13$.

Example 14.2.8. *If* $T(x) = x - \frac{f(x)}{f'(x)}$, *where* $f(r) = 0$, *then* $T(r) = r$. *Thus,* $T(x)$ *has* $x = r$ *as a fixed point.*

Galileo: We now prove the Contraction Mapping Theorem. Note that the proof mirrors exactly what we have discussed with the root finding method of Newton/Raphson. Namely, begin with an initial guess x_0 and create a sequence of numbers by iteratively

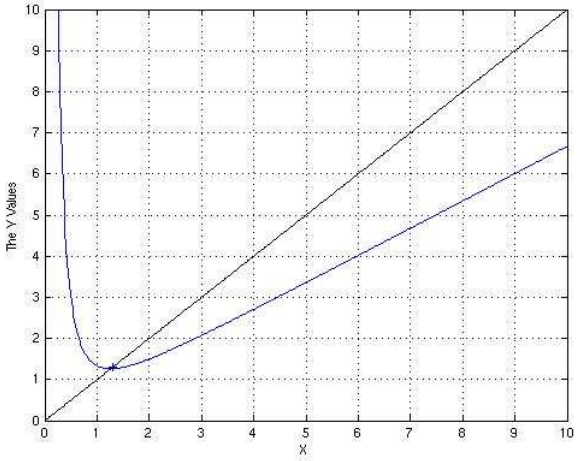


Figure 14.2: The Fixed Point for the Function $T(x) = x - \frac{x^3 - K}{3x^2}$.

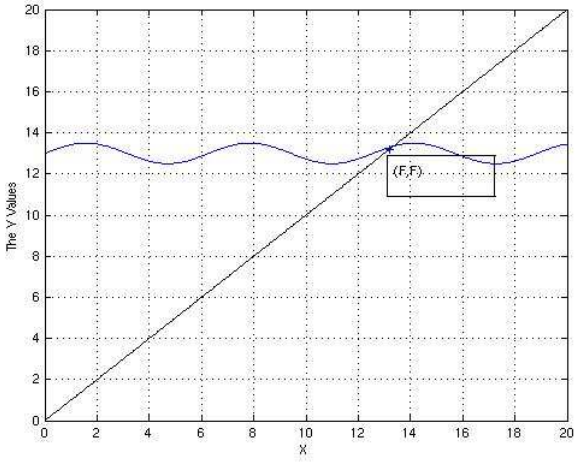


Figure 14.3: The Fixed Point for the Function $T(x) = \frac{1}{2}\sin(x) + 13$.

computing $T(x_n)$ and defining $x_{n+1} = T(x_n)$. Note that we actually produce a unique fixed point.

Simplicio: But why should I care if I only have one fixed point?

Galileo: If you only have one fixed point, then you only have to compute once.

Theorem 14.2.8 (The Contraction Mapping Theorem). *If X is a closed interval in $\Re$ and $T(x) : X \to X$ is a contraction, then $T(x)$ has a unique fixed point $F \in X$. Moreover, if the contraction factor for $T(\mathbf{x})$ is M, x_0 is any initial point in X, and $x_k = T(x_{k-1})$, then the error at the n^{th} iteration is given by the formula $|x_n - F| \leq \frac{M^n}{1-M}|x_1 - x_0|$.*

Proof. Let x_0 be any point $\in X$. Let $x_{k+1} = T(x_k)$ for all $k \geq 0$. Since $T(x)$ is a contraction, $|x_{k+1} - x_k| \leq M|x_k - x_{k-1}|$ for all $k \geq 1$. Thus, the sequence $\{x_k\}_{k=1}^{\infty}$ converges to some point F. Since the interval X is closed, the point $F \in X$. Since $x_{k+1} = T(x_k)$ and $T(x)$ is a continuous function, $F = lim_{k\to\infty}\{x_k\} = lim_{k\to\infty}\{x_{k+1}\} = lim_{k\to\infty}\{T(x_k)\} = T(lim_{k\to\infty}\{x_k\}) = T(F)$. Thus, F is a fixed point for $T(x)$.

The fact that the fixed point is unique follows from the fact that the function $T(x)$ is a contraction. In particular, if F_1 and F_2 are two distinct fixed points of $T(x)$, then $|F_1 - F_2| = |T(F_1) - T(F_2)| \leq M|F_1 - F_2| < |F_1 - F_2|$, which is a contradiction. Thus, $T(x)$ has exactly one fixed point. The error estimate follows from the contraction mapping error estimate. □

Simplicio: So what is the important information that I need to remember from this discussion?

Galileo: Remember this:

1. The mapping $T(x)$ MUST be a contraction. (You can usually check this fact by showing $|T'(x)| \leq M < 1$ for all x.)

2. The choice of initial point x_0 is arbitrary.

3. A sequence is created by computing $x_k = T(x_{k-1})$.

4. The sequence $\{x_n\}_{n=0}^{\infty}$ always converges to some number F, which is a fixed point of $T(x)$.

5. The convergence rate of the sequence $\{x_n\}_{n=0}^{\infty}$ is linear and controlled by the inequality: $|x_n - F| \leq \frac{M^n}{1-M}|x_1 - x_0|$. (Thus the error can be precontrolled.)

Example 14.2.9. *We will now show how the Contraction Mapping Theorem can be used to solve the equation* $x = \frac{1}{2}\sin(x) + 13$ *with the given prescribed accuracy of* 0.00001. *We begin by defining* $T(x) = \frac{1}{2}\sin(x)+13$. *To show that* $T(x)$ *is a contraction, all we have to do is to notice that* $T'(x) = \frac{1}{2}\cos(x)$ *so that* $|T'(x)| \leq \frac{1}{2}$ *for all* $x \in \Re$. *Thus,* $T(x)$ *is a contraction with contraction constant* $M = \frac{1}{2}$. *If* $x_0 = 0$, *then* $x_1 = T(x_0) = T(0) = 13$ *so that* $|x_0 - x_1| = 13$. *Thus, to find an integer* n *with the property that*

x_n *is within* 0.00001 *of the solution* $F = \frac{1}{2}\sin(F) + 13$ *all we need to do is to find an integer* n *with the property that* $|x_n - F| < \frac{M^n}{1-M} * |x_0 - x_1| = \frac{(\frac{1}{2})^n}{1-\frac{1}{2}} * 13 = (\frac{1}{2})^{n+1} * 13 < 0.00001$.

Taking natural logarithms of both sides of this last inequality we see that we should choose n *large enough that* $n + 1 > \frac{\ln(0.00001/13)}{-\ln(2)} = \frac{-14.0779}{-0.6931} = 20.3115$. *Thus, we must choose* $n > 20.3115 - 1 = 19.3115$.

Simplicio: So, the bottom line is that the formula tells us we get faster convergence if we make a smart choice of x_0 and we are blessed with a small value for M.
Galileo: Correct.

Exercise Set 14.2.

1. Use the Contraction Mapping Theorem to solve the equation $x = \frac{1}{3}\cos(2x) - 5$ with error less than 0.000001. If $x_0 = 0$, then how many iterative steps are required to guarantee that the required accuracy.

2. Use the Contraction Mapping Theorem to solve the equation $x = e^{-\frac{1}{2}x}$ with error less than 0.00001. If $x_0 = 0$, then how many iterative steps are required to guarantee that the required accuracy.

14.3 The Contraction Mapping Theorem in $\Re^n$

Galileo: We begin this section with an example, which demonstrates an iterative method for solving a system of linear equations. Compare this method with the row operations you learned in linear algebra. Remember that this example is for demonstration purposes only. In a real application, the matrix might be as large as 1000×1000 or even larger.

Example 14.3.1. *Solve the following system.*

$$2x + y = 3$$
$$x + 2y = 3$$

Note that the answer is: $x = 1, y = 1$.

To solve the problem using the technique of the contraction mapping theorem, we begin by manipulating the equation until it is in the form $\mathbf{x} = \mathbf{T}(\mathbf{x})$, where $\mathbf{x}$ is a 2-dimensional vector. This task can be completed by solving the first equation for x and the second for y. When we do this manipulation, we obtain 2 equations: $x = \frac{3-y}{2}$ and $y = \frac{3-x}{2}$. These 2 equations can be written in vector/matrix form as:

$$\mathbf{T}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{3}{2} - \frac{y}{2} \\ \frac{3}{2} - \frac{x}{2} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \end{pmatrix}.$$

If we initialize the process be letting $\mathbf{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, then $\mathbf{x}_1 = \mathbf{T}(\mathbf{x}_0) = \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \end{pmatrix}$ $\mathbf{x}_2 = \mathbf{T}(\mathbf{x}_1) = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{4} \end{pmatrix}$ and $\mathbf{x}_3 = \mathbf{T}(\mathbf{x}_2) = \begin{pmatrix} \frac{9}{8} \\ \frac{9}{8} \end{pmatrix}$.

If we let $\mathbf{x}_{n+1} = \mathbf{T}(\mathbf{x}_n)$, then the sequence of vectors $\{\mathbf{x}_n\}_{n=0}^{\infty}$ seems to be converging to the vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Simplicio: Magic!! This technique looks good to me.

Galileo: I am glad you like this method. Now let's take a look at another example.

Example 14.3.2. *Solve the following system.*

$$x + 2y = 3$$
$$2x + y = 3$$

Note again that the answer is: $x = 1, y = 1$.

Simplicio: But we just solved this problem.

Galileo: Solving for the variables x and y, we can again find the function $\mathbf{T}(x)$.

$$\mathbf{T}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3-2y \\ 3-2x \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ -2 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ 3 \end{pmatrix}.$$

We can again initialize the iterative process with the vector $\mathbf{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

When we compute $\mathbf{x_1} = \mathbf{T}(\mathbf{x_0}), \mathbf{x_2} = \mathbf{T}(\mathbf{x_1}), \mathbf{x_3} = \mathbf{T}(\mathbf{x_2})$,, etc, notice what happens to the sequence of vectors.

Simplicio: I see that $\mathbf{x_1} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$,

$\mathbf{x_2} = \begin{pmatrix} -3 \\ -3 \end{pmatrix}$, $\mathbf{x_3} = \begin{pmatrix} 9 \\ 9 \end{pmatrix}$, and $\mathbf{x_4} = \begin{pmatrix} -15 \\ -15 \end{pmatrix}$.

The sequence of vectors seem to be oscillating their way out to infinity.

Galileo: Excellent observation.

Example 14.3.3. *Now consider a system of three equations and three unknowns. In particular, solve the following system.*

$$4x + y = 5$$
$$x + 4y + z = 6$$
$$y + 4z = 5$$

Note that the answer is: $x = 1, y = 1, z = 1$.

Again, these equations can be written in vector/matrix form as:

$$\mathbf{T}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} \frac{5}{4} \\ \frac{6}{4} \\ \frac{5}{4} \end{pmatrix}.$$

If we initialize the method with $\mathbf{x}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ and define $\mathbf{x}_{n+1} = \mathbf{T}(\mathbf{x}_n)$, then the sequence of vectors again seems to converge to the correct answer.

Galileo: The beauty of the contraction mapping theorem is that it is valid in a multitude of different settings. In particular, it works in $\Re^n$ as well as abstract settings suitable for differential equations and fractals.

Even better, the proof just provided for the 1-dimensional case can be immediately translated to a proof in any dimension. To accommodate the new setting in $\Re^n$, the only changes that need to be implemented are:

1. The closed interval X must be replaced by a closed subset of $\Re^n$. (Thus, we need to define what it means for a set to be closed.)

2. The absolute value sign must be changed to a norm appropriate for the setting. (Thus, we need to define what a norm is.)

Note that while norms can be defined in many different ways and can be quite abstract, the underlying idea is always the same: measure the distance between two points. Thus, if P_1 and P_2 are two points in $\Re^n$, then the distance between them is the norm of $P_1 - P_2$. This distance is usually written in an expression of the form $dist(P_1, P_2) = \|P_1 - P_2\|$.

While the definition of a contraction can now be defined in terms of norms, it will be helpful if we can establish a criterion, which can be used to show a given function is a contraction. Since the condition $|T'(x)| \leq M < 1$ implies the function $T(x)$ is a contraction for functions of one variable, the analogue for $\Re^n$ is the norm of the

derivative $dT(x)$, where $dT(x)$ denotes the $n \times n$ matrix of derivatives. (Recall that the matrix of derivatives is nothing but the matrix of partial derivatives.

To keep the discussion simple, let's not waste mental energy defining what it means for a subset of $\Re^n$ to be *closed.* Instead, let us consider only the set $\Re^n$ and then remark that it is, in fact, closed. While numerous different norms can be defined on $\Re^n$, let us consider the one defined as the maximum of the absolute values of the n coordinates. The next definition formalizes this in a more mathematical way.

Definition 14.3.1. *If* $\mathbf{x} \in \Re^{\mathbf{n}}$, *then* $\|\mathbf{x}\|_\infty = \mathbf{max}\{|\mathbf{x_k}| : \mathbf{x_k}$ *is the k-th coordinate of* $\mathbf{x}\}$.

Simplicio: I don't like this notation, could you give me a simple example?
Galileo: The $\infty-$norm of the vector $(1, -2, 3, -4)$ is 4.
Simplicio: Why are we interested in knowing about norms?
Galileo: Because we can use them to compute the distance between two vectors (or points) in $\Re^n$. In particular, if $\mathbf{x}, \mathbf{y} \in \Re^n$, then the distance between $\mathbf{x}$ and $\mathbf{y}$ is $\|\mathbf{x}-\mathbf{y}\|_\infty$. Once we have the distance between two vectors defined, then we can define what it means for a sequence to converge. In particular, with the $\infty-$norm it is easy to show that a sequence of vectors converges to a particular vector if and only if it converges in each coordinate. Thus, all the hard work we did in the $1-$dimensional case is immediately transferable to the setting in $\Re^n$.

We now define the term contraction for a function $T(\mathbf{x}) : \Re^n \to \Re^n$. This definition is given in terms of the $\infty-$norm.

Definition 14.3.2. *If* $T(\mathbf{x}) : \Re^n \to \Re^n$, *then* $T(\mathbf{x})$ *is called a contraction if there is a real number* $M \in [0, 1)$ *with the property that* $\|T(\mathbf{x}) - T(\mathbf{x}')\|_\infty \leq M\|\mathbf{x} - \mathbf{x}'\|_\infty$ *for all* $\mathbf{x}, \mathbf{x}' \in \Re^n$.

Simplicio: But how do I recognize a contraction when I see one?
Galileo: You simply show the norm of the function (or more formally "the operator") is less than one.
Simplicio: But what is the norm of an operator?

Galileo: You ask the right questions. We begin with the definition of the norm of a matrix.

Definition 14.3.3. *If* $\mathbf{A} \in \Re^{m\times n}$, *then the* $\infty-$*norm of* $\mathbf{A}$ *is defined by*

$$\|\mathbf{A}\|_\infty = max\{\|\mathbf{a}_1\|_1, \|\mathbf{a}_2\|_1, \ldots, \|\mathbf{a}_n\|_1\},$$

where $\mathbf{a}_k$ *denotes the* k^{th} *row of* $\mathbf{A}$ *and* $\|\mathbf{a}_k\|_1 = |a_{k1}| + |a_{k2}| + \ldots + |a_{km}|$.

Proposition 14.3.4. *If* $\mathbf{A} \in \Re^{m\times n}$, $\|\mathbf{A}\|_\infty = M$, *and* $T(\mathbf{x}) = \mathbf{A}x + \mathbf{b}$, *then for all* $\mathbf{x}, \mathbf{x}' \in \Re^n$ $\|T(\mathbf{x}) - T(\mathbf{x}')\|_\infty \leq M\|\mathbf{x} - \mathbf{x}'\|_\infty$.

Proof. This proof is left as an exercise. □

Simplicio: And we can see from this proposition that the matrix given in the previous exercise has $\infty-$norm equal to $\frac{1}{2}$ and is thus a contraction.
Galileo: Very good. Now you are ready for a bit of formalism from Professor Cauchy. First we give the definition of what it means for a sequence to converge. Second, we give the definition of a *Cauchy* sequence. As in the $1-$dimensional setting, these two ideas are equivalent.

Definition 14.3.5. *A sequence of vectors,* $\{\mathbf{x}_k\}_{k=0}^\infty$ *in* $\Re^n$ *is said to converge to a vector* $\mathbf{x}_L \in \Re^n$ *if for every* $\epsilon > 0$ *there is an integer* N, *such that if* $k \geq N$, *then* $||\mathbf{x}_k - \mathbf{x}_L||_\infty < \epsilon$.

Definition 14.3.6. *A sequence of vectors* $\{\mathbf{x}_k\}_{k=1}^\infty$ *in* $\Re^n$ *is said to be* **Cauchy** *if for every* $\epsilon > 0$ *there is an integer* N, *such that if* $n \geq N$, *then* $||\mathbf{x}_n - \mathbf{x}_N||_\infty < \epsilon$.

Theorem 14.3.7. *If a sequence of vectors* $\{\mathbf{x}_k\}_{k=1}^\infty$ *in* $\Re^n$ *converges to a vector* $\mathbf{x}_L \in \Re^n$, *then it is Cauchy. Conversely, if a sequence of vectors* $\{\mathbf{x}_k\}_{k=1}^\infty$ *in* $\Re^n$ *is Cauchy, then it converges to some vector* $\mathbf{x}_L \in \Re^n$.

Proof. While the proof of the first statement in the proposition is straightforward. In particular, it is left as an exercise. The proof of the second statement is left for another day. □

Galileo: Thus, if a sequence of vectors in $\Re^n$ is Cauchy, then it is Cauchy on each coordinate. Since the sequence of vectors converges on each coordinate, it converges.

Theorem 14.3.8 (The Contraction Mapping Theorem in $\Re^n$). *If $T : \Re^n \to \Re^n$ is a contraction, then $T(\mathbf{x})$ has a unique fixed point $\mathbf{x}_L$ in $\Re^n$. Moreover, if the contraction factor for $T(\mathbf{x})$ is M, $\mathbf{x}_0$ is any initial vector, and $\mathbf{x}_k = T(\mathbf{x}_{k-1})$, then the error at the n^{th} iteration is given by the formula $\|\mathbf{x}_n - \mathbf{x}_L\|_\infty \leq \frac{M^n}{1-M}\|\mathbf{x}_0 - \mathbf{x}_1\|_\infty$.*

Proof. Let $\mathbf{x_0} \in \Re^\mathbf{n}$ and $\mathbf{x_{k+1}} = \mathbf{T}(\mathbf{x_k})$ for all $k \geq 0$. Since the same argument used in the 1-dimensional version can be used to show that the sequence $\{\mathbf{x_k}\}_{\mathbf{k=0}}^{\infty}$ is Cauchy in $\Re^n$, the sequence is Cauchy in each coordinate. Since the sequence converges on each coordinate, it converges. The proof of the error estimate is virtually the same as the proof given in the $1-$dimensional case. The only difference is that each absolute value sign must be replaced by the symbol for the infinity norm. □

Simplicio: Hey, I think I am beginning to get the hang of this theorem for $\Re^n$, but I already know how to solve systems of linear equations using the method of row operations or Gaussian elimination. Why would I want to bother with this new method?

Exercise Set 14.3.

1. Use the Contraction Mapping Theorem to solve the system of equations

$$4x + y = 5$$
$$x + 4y = 5.$$

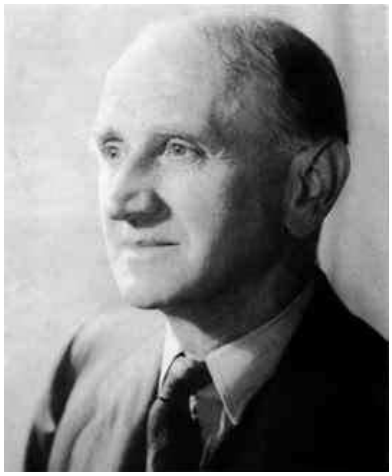
Initialize the method with the vector

$$\mathbf{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

How many iterations are required to guarantee an accuracy of less than 0.00001 on each coordinate?

Chapter 15

Aitken's Method



Alexander Craig Aitken (1895-1967)

> Ever the road beneath
> Changes: now night begins to fall,
> And I see the last long road of all,
> The road to dusty death.-Alexander Craig Aitken

Galileo: The purpose of the technique presented in this section is to speed up the rate of convergence of a given sequence.

Simplicio: While the idea seems reasonable, how can that be possible?

Galileo: Alexander Craig Aitken (1895 - 1967) came up with the idea that if a sequence converges linearly, then we can give it boost towards the ultimate answer. If we could

improve the convergence rate from linear to quadratic, we would be quite satisfied with the technique.

Simplicio: Who was this Aitken fellow?

Galileo: Professor Aitken was born in Dunedin, New Zealand and attended the University of Otago. He had an incredible memory being able to recite π to 2000 places. He could also instantly multiply and divide large numbers. An excellent memory is not always a blessing. He had trouble forgetting all the bad things that happened in his life.

Simplicio: I can see the dark side in his poetry. I am not sure I want to compete with him in any way.

Galileo: His idea is the following. If we assume the sequence $\{x_n\}_{n=1}^{\infty}$ converges to L (i.e. $lim_{n\to\infty} x_n = L$) and for large n enjoys the property

$$\frac{x_{n+1} - L}{x_n - L} \approx M < 1,$$

then we know the convergence will be linear. Thus, this condition is a bit stronger than linear convergence. In any case, if $lim_{n\to\infty} \frac{x_{n+1}-L}{x_n-L} = M < 1$, then both the quotient $\frac{x_{n+1}-L}{x_n-L}$ and the quotient $\frac{x_{n+2}-L}{x_{n+1}-L}$. will be approximately equal to M.

If we make this assumption about the two quotients, then we see that

$$\frac{x_{n+1} - L}{x_n - L} \approx \frac{x_{n+2} - L}{x_{n+1} - L},$$

which implies that

$$(x_{n+1} - L)^2 \approx (x_{n+2} - L)(x_n - L)$$

or

$$x_{n+1}^2 - 2\ x_{n+1}L + L^2 \approx x_{n+2}x_n - (x_n + x_{n+2})L + L^2$$

or

$$x_{n+1}^2 - 2\ x_{n+1} \cdot L \approx x_{n+2} \cdot x_n - (x_n + x_{n+2})L.$$

Therefore,

$$L(-x_{n+2} + 2x_{n+1} - x_n) \approx x_{n+1}^2 - x_{n+2}x_n$$

and

$$L \approx \frac{x_{n+1}^2 - x_{n+2} \cdot x_n}{-x_{n+2} + 2x_{n+1} - x_n} = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

Therefore, we can (hopefully) accelerate convergence to L if we define a new sequence by the rule:

Definition 15.0.9 (Aitken's Method). *If $\{x_n\}_{n=0}^{\infty}$ is a sequence of numbers, then the Aitken's Method for accelerating the convergence is given by $\hat{x}_n = x_n - \frac{(x_{n+1}-x_n)^2}{x_{n+2}-2x_{n+1}+x_n}$.*

Definition 15.0.10. *If $\{x_n\}_{n=1}^{\infty}$ is a sequence, then the forward difference formula is given by $\Delta x_n = x_{n+1} - x_n$. Higher powers are defined inductively by $\Delta^k x_n = \Delta(\Delta^{k-1} x_n)$.*

Virginia: Is there any connection between this formula and the first derivative? They look similar.

Galileo: In fact it is. If you think of the first derivative as a limit of the quotients $\frac{f(x+\Delta)-f(x)}{\Delta}$, then the "derivative" of a sequence should be the "limit" of $\frac{x_{n+1}-x_n}{n+1-n} = \frac{x_{n+1}-x_n}{1} = x_{n+1} - x_n$. Of course, we can't compute limits because we have a discrete set of points. Instead, we simply think of the two points x_{n+1} and x_n as "close" to one another.

Example 15.0.4. *The only reason we need higher powers of the forward difference formula for Aitken's Method is to compute the second forward difference $\Delta^2 x_n = \Delta(\Delta x_n) = \Delta(x_{n+1} - x_n) = x_{n+2} - 2x_{n+1} + x_n$*

Virginia: This formula should represent the 2^{nd} derivative. Correct?

Galileo: You are correct.

Proposition 15.0.11 (Aitken's Method). *If $\{x_n\}_{n=0}^{\infty}$ is a sequence of numbers, then the Aitken's Method for accelerating the convergence is given by $\hat{x}_n = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n}$.*

Simplicio: This formula looks suspiciously familiar.

Galileo: It should. Note the similarity between this formula and the formula $T(x) = x - \frac{f(x)}{f'(x)}$ given by Newton/Raphson. This association should help you remember Aitken's formula.

Example 15.0.5. *Let us begin by applying Aitken's method to the linearly convergent sequence* $x_n = \frac{1}{2^n}$. *With this special case, we see that*

$$\begin{aligned}
\hat{x}_n &= x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} \\
&= \frac{1}{2^n} - \frac{(\frac{1}{2^{n+1}} - \frac{1}{2^n})^2}{\frac{1}{2^{n+2}} - 2\frac{1}{2^{n+1}} + \frac{1}{2^n}} \\
&= \frac{1}{2^n} - \frac{2^{n+2}(\frac{1}{2^{n+1}} - \frac{1}{2^n})^2}{1 - 4 + 4} \\
&= \frac{1}{2^n} - \frac{2^{n+2}}{2^{2n+2}} \\
&= \frac{1}{2^n} - \frac{1}{2^n} \\
&= 0.
\end{aligned}$$

Thus, Aitken's Method converts a linearly convergent sequence to one that converts instantly!!

Simplicio: Hey, this method works great. Does it give any relief for the bisection method?

Galileo: To answer your question properly, we must first decide how we are going to implement the method. In the previous example, we were given a formula for the n^{th} term of the sequence. Unfortunately, nature is not so kind. The algorithm of Johan Steffensen (1873-1961) computes two terms of the sequence and then makes an Aitken's computation. Try integrating this idea into a bisection algorithm and see how it does when you compute $\sqrt{2}$.

Exercise Set 15.1.

1. Apply Aitken's method to the sequence $x_n = \frac{1}{3^n}$. How many steps does it take to converge to zero?

2. Apply Aitken's method to the sequence $x_n = 3^n$. What number does the sequence converge to? How many steps does it take to converge?

3. Apply Aitken's method to the sequence $x_n = \frac{1}{n}$. Do you find any benefit by applying Aitken's method?

4. Apply Aitken's method to the sequence $x_n = \frac{1}{2^{2^n}}$. How many steps does it take to converge?

5. Devise a hybrid Bisection/Aitken's Method to find the positive root of the function $f(x) = x^2 - K$, where $K > 1$ and the initial interval is $[1, K]$. Apply your algorithm to the function when $K = 10^{10}$. Does your algorithm provide a significant improvement in the rate of convergence? While there are a multitude of different ways to create a hybrid algorithm, you might begin by alternating the two methods.

6. Devise a hybrid Newton/Raphson/Aitken's Method to find the positive root of the function $f(x) = x^2 - K$, where $K > 1$ and the initial guess is $x_0 = K$. Apply your algorithm to the function when $K = 10^{10}$. Does your algorithm provide a significant improvement in the rate of convergence? While there are a multitude of different ways to create a hybrid algorithm, you might begin by alternating the two methods.

7. Devise a hybrid Contraction Mapping Theorem/Aitken's Method to solve the equation $x = \frac{1}{2}\sin(x) + 13$. find the root of the function $f(x) = x^2 - K$, where $K > 0$. Apply your algorithm to the function when $K = 10^{10}$. Does your algorithm provide a significant improvement in the rate of convergence? While there are a multitude of different ways to create a hybrid algorithm, you might begin by alternating the two methods.

Part VI

Day 6. Linear Algebra Review

Giuseppe Peano (1858-1932)

> Ambiguity of language is philosophy's main source of problems. That is why it is of the utmost importance to examine attentively the very words we use. -Giuseppe Peano

Galileo: Linear Algebra is probably the most important prerequisite for applications.

Simplicio: I took Linear Algebra from Professor Poubelle. All we did was solve systems of equations using row operations. It was easy.

Galileo: Unfortunately, Linear Algebra is probably the most important mathematics course you will ever take.

Virginia: More important than Calculus?

Galileo: Man has been making observations and measurements since the beginning of written history. This data leads to conjectures. Conjectures lead to mathematical models. Whenever you model a problem, your first instinct is to make it linear. Linear models are easy to understand and compute.

Simplicio: How about an example?

Galileo: If I paid twice as much for a house as you did, then mine ought to be twice as big. In other words, if you double the price, then you should double the size. These ideas go back several thousand years to the ancient Greeks with their discussions of similar triangles and proportions.

Simplicio: But what if your house is on a Florida beach and costs twice as much as

my student ghetto trash littered dump, then my house might still be the same size as yours. Mine might even be larger.

Galileo: That's correct. Life is often nonlinear.

Virginia: How about a more scientific example?

Galileo: My colleague, Aristotle (384-322B.C.E.), asserted a linear relationship between the distance a dropped object travels and the time of flight. In other words, if the time of flight is doubled, then the distance should also be doubled. Unfortunately, my data showed that his speculation was not correct.

Virginia: So what do we need to know from Linear Algebra?

Galileo: Despite Euclid's concern for detail, the story of Geometry wasn't completed until Hermann Grassmann (1809-1877), George Cantor (1845-1918), Bernhard Riemann (1826-1866), Giuseppe Peano (1858-1932), David Hilbert (1862-1943), Kurt Gödel (1906-1978), Bertrand Russell (1872-1970), and others finally reduced all the mathematical and logical issues to the axioms of set theory. (While not quite accurate, I refer to these fellows as the "grumpy, 19^{th} century, German mathematicians.") Thus, this effort to "get it right" took several thousand years to unfold.

Simplicio: Weren't we talking about Linear Algebra? Why have we digressed once again to Geometry?

Galileo: Every geometric idea corresponds with an algebraic expressions in Linear Algebra. Do you remember Euclid's 14 axioms?

Virginia: I remember a couple of them:

1. A point is that which has no part.

2. A line is breadthless length.

Galileo: Very good. Now, do you remember Peano's 10 axioms for a vector space?

Simplicio: Not a chance.

Galileo: While you might prefer that all of Linear Algebra was limited to a discussion of $\Re^n$, remember that the definition is given more abstractly. Namely, a vector space V is a set V together with two operations addition, denoted by $+$, and scalar

multiplication, denoted by $\cdot$. These operations satisfy a number of rules including the associative, commutative, and distributive laws. We also have an additive identity (namely 0) and additive inverses.

Simplicio: Why all this unnecessary abstraction?

Galileo: Because when we study approximation theory, we need geometric ideas expressed in algebraic language. Notice that the idea of a vector in a vector space is Euclid's idea of a point. All the scalar multiples of that point produce a line. All the positive multiples of a non-zero vector produce a ray so we are ready to talk about the angle between two rays emanating from the same point.

Virginia: What about triangles and parallelograms?

Galileo: We can build a triangle by taking linear combinations of two sides.

Simplicio: The same idea works for parallelograms.

Virginia: Except we have to be careful the sides of the figure are linearly independent. Otherwise, we will end up with a ray. In fact, we need the idea of linear independence to generate $n-$dimensional figures. If I remember correctly, a vector space has dimension n if it has a basis with n elements. We spent a lot of time in our Linear Algebra class showing that any two bases have the same number of elements.

Simplicio: Why did you do that? Isn't that obvious?

Virginia: I found those proofs to be difficult.

Galileo: Peano's axioms are exactly what you need to slug through those proofs. Let me now turn your attention back to Euclid's idea of a point. Note that his definition contains the implicit assumption the reader already knows a point should lie in the plane. The definition is rather negative because it does not tell you what it is, but rather what it is not. Rene Descartes (1596-1650) recognized that a point can be thought of iin a more positive way as a pair of real numbers (x, y). Cantor and Peano realized that Euclid's definitions of a point was totally inadequate for modern applications. In particular, they realized that the functions could be thought of as points.

Simplicio: Your kidding? Functions are't points. They are defined for points in their

domain.

Galileo: It is an interesting leap forward, isn't it? In any case, they decided is worthwhile to abstract the idea of a point to such functions as $1, x, x^2, \ldots, x^n$.

Virginia: And note that these functions (or should I say points) are linearly independent. Thus, the vector space they span is $n+1$-dimensional.

Galileo: Similarly, Jean Baptiste Joseph Fourier (1768-1830) recognized that for any positive integer n, the functions $1, \cos(x), \cos(2x), \ldots, \cos(nx), \sin(x), \sin(2x), \ldots, \sin(nx)$ represent $2n+1$ linearly independent functions (or points!) which span a $2n+1$-dimensional vector space. During our tutorial we will also discuss orthogonal polynomials, splines and wavelets. These new sets of functions all form vector spaces in a natural way. While the definition of a vector space is a bit abstract when you first encounter it, the beauty is its generality. In other words, you don't have to keep reiterating the same definitions and theorems over and over again. Think of it as a well-written subroutine for some computer program you are writing. The software should be concise so it is simple to comprehend, but it should also be general so it can be used in as many different settings as possible.

Simplicio: What you said is interesting. I will have to think about it.

Virginia: I guess Euclid's concept of a line is similarly limited.

Simplicio: I would have to agree.

Galileo: We should now move on to the geometric ideas of distance, angles, and projections. These ideas were distilled and abstracted into a single concept: the inner product.

Definition 15.0.12. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ *are vectors in* $\Re^n$, *then the inner product of* $\mathbf{u}$ *and* $\mathbf{v}$ *is defined as* $< \mathbf{u}, \mathbf{v} >= \mathbf{u}^t\mathbf{v} = \sum_{k=1}^{n} u_k v_k$.

Simplicio: Why did you write the superscript ^{t} on the vectors?

Galileo: In the culture of Linear Algebra, we prefer to think of points as column vectors. Unfortunately, in the culture of publishing, it is more convenient to write row vectors to save space on the page. The superscripts ^{t} denotes the transpose, which flips a row vector to a column vector and vice versa. Thus, the inner product

$< \mathbf{u}, \mathbf{v} >$ is simply equal to the matrix product of the row vector $\mathbf{u}^t$ and the column vector $\mathbf{v}$.

Simplicio: So you have simply defined the dot product of two vectors.

Galileo: Exactly. Now let's turn to the problem of defining length and distance.

Definition 15.0.13. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *is a vector in* $\Re^n$, *then the length (or* $2-$*norm) of* $\mathbf{u}$ *is* $\|\mathbf{u}\|_2 = \sqrt{< \mathbf{u}, \mathbf{u} >}$.

Simplicio: What is that little subscript 2 doing there?

Galileo: Actually, I must apologize, but that subscript comes from the Pythagorean Theorem. As it turns out, there are a multitude of different norms. In fact, for any real number $p \geq 1$, there is a $p-$norm. However, as someone interested in applications, the only values of p that you might ever use are $p = 1, 2, \infty$.

The $1-$norm is sometimes called the taxi-cab metric and is defined as follows:

Definition 15.0.14. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *is a vector in* $\Re^n$, *then the* $1-$*norm of* $\mathbf{u}$ *is* $\|\mathbf{u}\|_1 = \sum_{k=1}^{n} |u_k|$.

The $\infty-$norm is sometimes called the *sup* norm and is defined but the following rule. This metric

Definition 15.0.15. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *is a vector in* $\Re^n$, *then the* $\infty-$*norm of* $\mathbf{u}$ *is* $\|\mathbf{u}\|_\infty = max\{|u_1|, |u_2|, \ldots, |u_n|\}$.

Note that the taxi-cab and sup metrics do not involve computing a square root. Thus, they are faster and easier to compute than the $2-$norm.

We now use the $2-$norm to define the distance between two vectors.

Definition 15.0.16. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ *are vectors in* $\Re^n$, *then the distance between* $\mathbf{u}$ *and* $\mathbf{v}$ *is* $\|\mathbf{u} - \mathbf{v}\|_2$.

Simplicio: Since I saw these definitions in Calculus, I am comfortable with these ideas.

Galileo: Good. Now we are ready to define the cosine of the angle between two vectors and the projection of one vector onto another.

Definition 15.0.17. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ *are vectors in* $\Re^n$, *then the cosine of the angle* θ *between* $\mathbf{u}$ *and* $\mathbf{v}$ *is* $\cos(\theta) = \frac{<\mathbf{u},\mathbf{v}>}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}$

While people in applications expect a method to "always work," they may not be so fortunate. As the patient reader will see, most techniques have to be used with caution. The purpose of many of the theorems is to provide conditions and guidelines when the techniques will provide useful estimates. Remarkably, one key to a multitude of stable methods is the concept of orthogonality, which is nothing more than another word for right angle. Thus, numerical techniques look to Geometry as a source of ideas for methods that always work. We will see this theme throughout these notes.

Galileo: You might be surprised to learn you have an ally in the mathematician Pafnuty Chebyshev (1821-1894), who once remarked: "To isolate mathematics from the practical demands of the sciences is to invite the sterility of a cow shut away from the bulls."

Simplicio: I bet Professor Chebyshev and I would get along just fine.

Chapter 16

Stable Techniques: The Role of Orthogonality

Galileo: I am a believer in "Applications driven mathematics." However, before we move on, I must add that the concept of orthogonality is essential to the success of a multitude of numerical methods. To say two vectors are orthogonal is just a fancy way of saying they are perpendicular. A triangle is called a right triangle if its two shorter edges are perpendicular. As Virginia just noted, my colleague Pythagoras has a lot to say about right triangles. More recently, Professor Chebyshev showed that his polynomials also have special orthogonality properties.

Simplicio: But why should I care?

Galileo: Some techniques are stable, while others are unstable.

Simplicio: Stable? Unstable? I don't get it.

Galileo: Will no one rid me of this meddlesome fellow?

Virginia: OK, OK. I think it is time to relax here.

Galileo: Think of a mathematical technique as a black box that produces answers for given types of inputs. An example of such a black box is a calculator. Anyone working in applications should worry about whether or not a technique produces "reasonable" outputs when given "reasonable" inputs. Many techniques lack this important property–at least some of the time.

Simplicio: An example please.

Galileo: Suppose $\mathbf{A}$ is a 2×2 matrix and we are suppose to solve the system of linear equations $\mathbf{Ax} = \mathbf{b}$. In this example, the inputs are the the matrix $\mathbf{A}$ and the 2-dimensional vector $\mathbf{b}$. The output is the 2-dimensional vector $\mathbf{x}$.

Simplicio: No problem. I remember the formula that solves such a system.

Virginia: If I remember correctly, the formula requires that you divide by the determinant of $\mathbf{A}$. Thus, if $det(\mathbf{A}) = 0$, there may be a problem.

Galileo: Yes, Virginia. You have pointed out an important hypothesis to that theorem. Namely, we must assume $det(\mathbf{A}) \neq 0$.

Simplicio: I still don't see the problem.

Galileo: Consider the following two systems of linear equations in the plane:

System 1:

$$\begin{aligned} 1.001x + y &= 2.001 \\ x + y &= 2 \end{aligned}$$

Note that the equations of these two lines are close to being parallel. Solving the system we find $x = 1$ and $y = 1$.

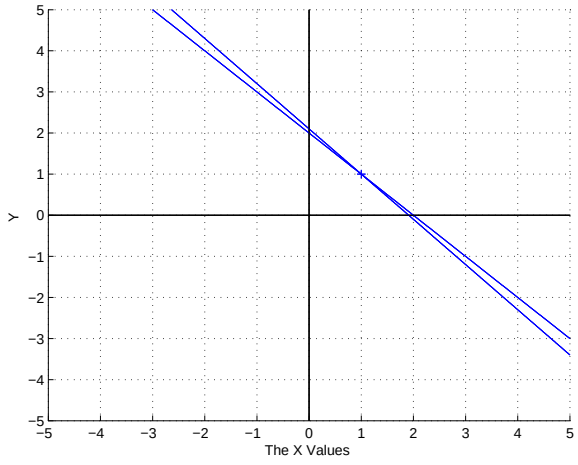


Figure 16.1: The Almost Parallel Lines for System 1

Now consider a slight modification of this system of equations.

System 2:

$$\begin{aligned} 1.001x + y &= 2 \\ x + y &= 2 \end{aligned}$$

Solving this new system we find $x = 0$ and $y = 2$. Thus, we have modified only one entry in the vector $\mathbf{b}$ by the minuscule amount 0.001.

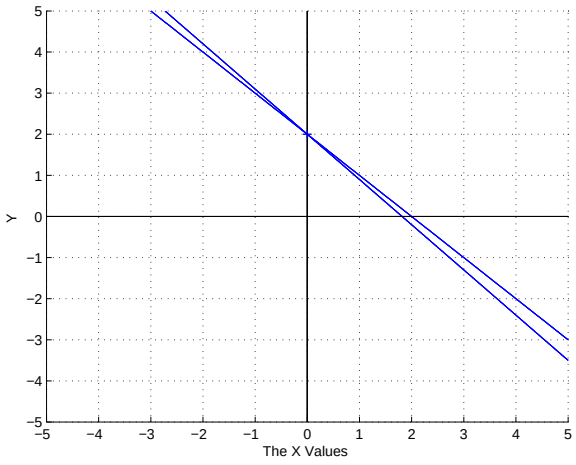


Figure 16.2: The Almost Parallel Lines for System 2

This change has led to a difference of 1 in both entries of the answer. If we define the coefficient matrix by

$$\mathbf{A} = \begin{pmatrix} 1.001 & 1 \\ 1 & 1 \end{pmatrix},$$

then note that $det(\mathbf{A}) = 0.001 \neq 0$. Thus, the matrix equation can be solved by row operations. However, considering the size of the change in the inputs, the size of the change in the outputs is large. This is evil.

A second phrasing of stability is: If given two different sets of inputs which are close together, then the outputs should also be close together. Our little example fails to possess this important property. If you are an engineer, you come to avoid unstable methods because they produce weird untrustworthy answers. If a technique

lacks this property, then engineers won't use it. This issue appears repeatedly in a multitude of numerical techniques.

Simplicio: But we are talking about row operations here! People still use this method every day. I have solved dozens of problems using row operations and have never observed this problem. How come nobody ever warned me about this problem before? What is the problem?

Galileo: Its all in the hypotheses. Note that the two columns of $\mathbf{A}$ are almost parallel, which implies that the matrix $\mathbf{A}$ is mildly ill-conditioned. I am willing to bet that the matrix equations you solved in your previous courses all had integer entries. Your professors were being easy on you so you could compute the answer without having to keep track of a lot of decimal places. In real-life applications, don't expect integer entries. Let me finish the discussion of this example by remarking that this problem disappears if the columns of $\mathbf{A}$ are orthogonal (or almost orthogonal). We will revisit this issue numerous times in our future discussions. We will find that matrices associated with polynomial approximations of data are evil, while matrices associated with Fourier series, spline, and wavelet approximations are good. As you will see, the mantra for numerical techniques is: "The name of the game is control."

Simplicio: OK, let's get back to a discussion of the prerequisites for this tutorial.

Galileo: Of course, you also need to have a solid background in Calculus. I use the phrase "solid background" to mean that you either remember the material or are willing to make an effort to review it on your own. At a minimum, you should be able to compute derivatives of functions using the sum, product, quotient, and chain rules. You should also be able to compute easy integrals using substitution. We will review integration by parts, the Fundamental Theorem of Calculus, and Taylor's Theorem. Recall that the big concept in Differential Calculus is that the tangent line at a given point on a curve is the line that best approximates the curve at that point. The slope of this line is computed as the derivative of the function. By the way, Brook Taylor (1685-1731) was a British mathematician, whose ideas are used everywhere in Numerical Analysis. We will see a lot of him.

Simplicio: I think I can handle the Calculus prerequisite.

16.1 Linear Algebra = Geometry + Algebra

Galileo: As for Linear Algebra, I can only say it is probably the most important course you will ever take in mathematics. The first instinct of an engineer is to transform a given problem into a linear one–at least over a short time span. General interest in this magnificent subject is easy to understand: With only two techniques you have the ability to solve virtually any linear problem. The first technique is the method of Gaussian elimination, otherwise known as row operations. The second is the diagonalization of a matrix using eigenvalues and eigenvectors.

Simplicio: Until a few minutes ago, I had no problem with row operations. However, I must admit that I have always been a bit insecure when it comes to eigenvalues. Professor Poubelle covered the topic at the end of the semester and we ran out of time and energy.

Galileo: And so it is. You learned one of the two big ideas.

Virginia: I will agree with Mr. Simplicio. I had Professor Picky Picky Picky for Linear Algebra, While he was a good teacher, we rarely computed anything. We also had the problem that we got bogged down in lots of definitions, theorems, and proofs. The good Professor said the purpose of the course was to teach us about abstract mathematics. I worked hard and enjoyed the material, but was never quite able to master the topic of diagonalizing a matrix. Some how, I always got the transition matrix backwards. That stuff at the end of the semester was very confusing.

Galileo: And there it is: the psychotic bifurcation of a beautiful subject. My view is that deep down Linear Algebra is the fusion of geometry and algebra. If you will, it is the "algebratization" of Euclid's Geometry. Maybe it should have been called Linear Algebraic Geometry. Of course, it is too late now. The beauty of Linear Algebra is that algebraic expressions and formulas are provided for each geometric concept. Points, rays, and lines can be represented by vectors; angles and distances

can be computed using the inner product; areas and volumes can be computed with the determinant function; and congruences can be represented by the combination of orthogonal and translational matrices. This strong connection between the two subjects is no accident. For the 100 years of the 19^{th} Century, mathematicians worked incessantly to get Geometry "right."

In fact, the subject matter you studied in your Calculus, Linear Algebra, and Vector Analysis courses is a direct result of this effort to algebratize geometry. The first reason was to make geometry rigorous; the second was to facilitate the incorporation of geometric ideas into the modeling of real-life applications. In the process of proving the Fundamental Theorem of Algebra, Gauss recognized that complex numbers could be represented as vectors in the plane. Sir William Rowan Hamilton (1805-1865) generalized the idea of the complex numbers to the quaternions, which provide an algebraic structure for $\Re^4$. This structure satisfies the associative and distributive laws, but the multiplication fails to be commutative. He also invented the word "vector."

Simplicio: Can't the complex numbers be thought of as a subset of the quaternions?

Galileo: Correct. Three other mathematicians, who contributed to this search for the est blend of geometry and algebra were Hermann Grassmann (1809-1877), Arthur Cayley (1821-1895), and Josiah Willard Gibbs (1839-1903). While Grassmann contributed to many aspects of the subject, his efforts were focused on making the subject as abstract and general as possible. In particular, he formalized the terms inner and outer product in terms of their properties instead of their formulas. These ideas will become important when we investigate Fourier series. Cayley worked with Hamilton on matrix algebra. In fact, he invented the term. Do you know what the word *matrix* means in Latin?

Simplicio: No clue.

Galileo: Womb.

Simplicio: Oh.

Galileo: Gibbs was the first high quality American mathematician. Trained in Eu-

rope, he studied thermodynamics and heat transfer. Entropy and enthalpy were his ideas.

Simplicio: Why would I care about thermodynamics? Thank heavens I never had to study that difficult subject.

Galileo: Thermodynamics is a subject that grew out of the invention of the steam engine. You drive a car, don't you?

Simplicio: Sure.

Galileo: Gibbs was also one of the founding fathers of Vector Analysis. He even invented the notation for the dot product and the cross product. His Vector Analysis emerged as the winner over Hamilton's quaternions for most applications. Giuseppe Peano (1858-1932) was a clear-thinking Italian, who has numerous credits to his name including the formal definition of induction on the integers, the construction of continuous functions which raise dimension, and the formal definition of an abstract vector space. Peano is responsible for that abstract definition you should have learned in your first course in linear algebra. The reason for the abstraction was to get away from the idea of a fixed coordinate system in Euclidean $n-$dimensional space $\Re^n$.

My colleagues Professors Giuseppe Peano (1858-1930) and David Hilbert (1862-1943) were instrumental in setting up the axioms for a vector space so they fit a multitude of different applications. It seems that different research groups were all doing the same thing. They just didn't realize it. Peano and Hilbert completed this work at the end of the 19^{th} Century. At the same time they put Euclid on a solid foundation.

Simplicio: OK, that is enough history.

Galileo: Let us now turn to the idea of a vector space, which plays a fundamental role in the topics we will discuss.

Simplicio: What is a vector space? Professor Poubelle never discussed that topic.

Virginia: A vector space is a set together with two operations: addition and scalar multiplication. The axioms include associative, commutative, and distributive laws as well as additive identity and inverses. The plane $\Re^2$, three space $\Re^3$, and $\Re^n$ are

examples of vector spaces, where the set of scalars is the set of real numbers. The elegant feature of a vector space is that a function $f(x)$ continuous at each $x \in [a, b]$ can be thought of as a vector. If we denote the collection of all continuous functions on $[a, b]$ by $C^0[a, b]$, then it is straightforward (but boring, boring, boring) to show that $C^0[a, b]$ is a vector space.

Simplicio: When I think of vectors, I think of little arrows with a poison tip. Which way does $f(x)$ point?

Galileo: Just as a vector in the plane has two coordinates and a vector in three space has three coordinates, a function $f(x)$ has a coordinate for each $x \in [a, b]$. Thus, the space $C^0[a, b]$ is looking like an infinite dimensional vector space. (Galileo sips from his goblet.)

Simplicio: This infinite dimensional stuff is just mathematical games to keep you guys off the streets. We live in three space. I say three space is as high as we need to go.

Galileo: What about time?

Simplicio: OK, I will concede four dimensions.

Galileo: What about phase space in Physics? Those guys like to have a particle move around in six dimensional space: three coordinates for position and three for velocity.

Simplicio: OK, six.

Galileo: String theory puts us at 11. Actually, a signal with n terms can be thought of as a vector in $\Re^n$. Similarly, the set of all digital images in bitmap format with 256 rows and 256 columns lie in a $256 \times 256 = 2^{2^8}$ dimensional space so you might as well concede the point. You get one dimension for each pixel location.

Simplicio: Those last two examples make this discussion more interesting. I like images.

Galileo: What about a careful definition of this notion of dimension?

Simplicio: It is what you said. The definition of dimension is simply the number of coordinates. I don't see the problem.

Galileo: Virginia, do you have any thoughts on this matter?

Virginia: Dimension is a tricky concept to make mathematically rigorous. If I re-

member correctly from Professor Picky's class, we first defined the notions of linear combination and linear dependence. Once we had defined linear dependence, the definition of independence is easy. Namely, a set of vectors is independent if it is not dependent. A basis is defined as any subset of a given vector space with the property that it is both linearly independent and maximal with respect to the property of being independent. The definition of dimension for a vector space can now be defined as the number of vectors in a basis. Professor Picky also defined the notion of a spanning set and then proved that any minimal spanning set is a basis.

Simplicio: That definition doesn't sound too bad. What's the problem.

Virginia: The problem is that there can be a zillion different bases for a given vector space. Professor Picky spent several days going through some kind of exchange argument, which showed that any two bases have the same cardinality.

Simplicio: Cardinality? What is that?

Virginia: The cardinality of a set is the number of points in the set.

Simplicio: Well why didn't you say so? Professor Poubelle was right. All this math stuff is rubbish. You make the easiest ideas difficult for absolutely no reason. You math people are all neurotically obsessed by details. Boring, boring, boring. So I guess I should be polite and ask why should we care whether or not two bases have the same cardinality?

Virginia: Consider your favorite vector space $\Re^2$. It is easy to check that the standard basis $B = \mathbf{e}_1 = (1,0), \mathbf{e}_2 = (0,1)$ is a basis. From the discussion I just gave, we now know that any other basis will also have two members. Similar comments apply to $\Re^3$.

Simplicio: Big deal. The standard basis is good enough for me. Why should I transform something easy into something complicated?

Galileo: Actually, the opposite is true. Do you happen to remember my colleague Apollonius (262-190 B.C.E.)?

Virginia: He was the one with the conic sections. Right?

Galileo: And what did his theorem say?

Virginia: If you cut (or intersect) a plane with a cone, then you get either a parabola, a hyperbola, or an ellipse. Actually, you can also get some less interesting cases such as a point, a line, two lines, and even the empty set.

Galileo: Very good. Now what did my friend Rene Descartes (1596-1650) show?

Virginia: I am not sure I remember.

Galileo: Descartes showed that if you consider the subset of the plane defined by $S = \{(x, y) : Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0\}$, then the set must be a conic section. The argument proceeds in two steps. The first is to translate the x and y axes so that in the new coordinate system the constants $D = E0$. This step is easy and leaves us with the expression $Ax^2 + Bxy + Cy^2 + F = 0$. The second step is to rotate the coordinates so that $B = 0$. This rotation is carried out by the matrix

$$\mathbf{S} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix},$$

where $c = cos(\theta)$ and $s = sin(\theta)$ for an appropriately chosen angle θ. With these two transformations, we end up with the same seven cases you just mentioned. If we define the discriminant by the formula $\Delta = B^2 - 4AC$, then the three interesting cases become:

1. If $\Delta = 0$, then the set S is a parabola defined by $y = ax^2$.
2. If $\Delta > 0$, then the set S is a hyperbola defined by $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$.
3. If $\Delta < 0$, then the set S is an ellipse defined by $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

This process has reduced a complicated expression to one in standard form, where the three interesting cases can be identified by simply computing Δ.

Virginia: In other words, the discriminant discriminates!

Simplicio: OK, all well and good, but what employer is going to pay me a worthy salary for knowing this little theorem about conic sections?

Virginia: You can always teach.

Simplicio: Fat chance of that *ever* happening.

Galileo: While I will admit that this theorem may seem a bit old fashioned, it provides the concept for a multitude of applications. In particular, an employer will be interested in whether or not you are knowledgeable about Fourier Series. Fourier didn't invent Fourier series and he never got the math right, but he did manage to draw attention to a technique that works over a broad range of applications. He knew it worked.

Simplicio: Who is this Fourier and how did he get started?

Galileo: Jean Baptiste Joseph Fourier (1768-1830) accompanied Napoleon Bonaparte to Egypt as his chief scientist in 1798. While Fourier enjoyed the sunny weather, it seems that the English did not particularly appreciate the French having control of this important region. As a result, Lord Horatio Nelson attacked and defeated the French in the Battle of the Nile in 1798. With his subsequent return to France, Fourier chose to live in Grenoble, where the winters are long, cold, and miserable. While he turned up the heat in his apartment and put on extra coats, he was unable to keep out the winter chill. He suffered mightily. In his misery, he began his investigations into the heat equation.

Simplicio: Why are you telling me this sad story about poor old Mr. Fourier? I care nothing about his heat equation.

Galileo: You might reconsider that statement. While Fourier investigated the heat equation, his series continue to be used in a multitude of applications even 200 years after their invention. Two reasons for this longevity come to mind. First, Fourier series are simply linear combinations of sines and cosines so they should probably be considered when modeling any phenomenon associated with the motion of a wave. If you think about it, waves are involved in a multitude of application areas including optics, electromagnetic waves, communications, acoustics, and speech recognition. For an electrical engineer or computer scientist, Fourier also provides the basis for the acquisition, transmission, compression, and filtering of signals and images. When one speaks of "applications driven mathematics," Fourier series should be one of the first topics to come to mind. In any case, Fourier is a core subject for students in

both pure and applied mathematics. I apologize for the extended soliloquy.

Simplicio: Now you have my full attention.

Galileo: Before we move on to Fourier, let's take one more look at Descartes' rotation matrix $\mathbf{S}$. The new coordinate system can be described by the basis formed by the two column vectors of $\mathbf{S}$. More specifically, the two new basis vectors are:

$$\begin{pmatrix} c \\ s \end{pmatrix} \textit{and} \begin{pmatrix} -s \\ c \end{pmatrix}.$$

These vectors have two fundamentally important properties. First, they are orthogonal. Second, they both have length one. In the language of modern Linear Algebra, we have diagonalized the matrix

$$\mathbf{M} = \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix},$$

which allows us to write the quadratic expression $Ax^2 + Bxy + Cy^2$ as the matrix product:

$$Ax^2 + Bxy + Cy^2 = (x, y) \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

If λ_1 and λ_2 are the eigenvalues of $\mathbf{M}$, then

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c, \end{pmatrix} \begin{pmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

In other words, if $\lambda_1 > 0$ and $\lambda_2 > 0$, then the conic section is an ellipse; if $\lambda_1 > 0$ and $\lambda_2 < 0$, then the section is a hyperbola; and if either $\lambda_1 = 0$ or $\lambda_2 = 0$, then the section is a parabola. Thus, the eigenvalues can also be used to distinguish the three cases.

Simplicio: OK, OK, that ugly word *orthogonal* is now looking better.

Galileo: The question now becomes: How do you compute lengths, distances, and angles in a vector space?

Virginia: There is nothing in the definition of a vector space that says anything about either property.

Galileo: You are, in fact, correct.

Simplicio: So what do we do?

Galileo: Leave it to another grumpy German geek from the 19^{th} Century, one Hermann Grassmann (1809-1877), to invent the idea of an inner product. The virtue of this idea is that it solves all three problems at the same time. In particular, this device can be used to make abstract definitions for length (also called norm), distance, and angle. His ideas were so ahead of his time, nobody could understand what he was talking about.

Simplicio: Why would anyone want all this abstraction? Why not keep it understandable?

Galileo: Think about your computer software classes. When you write a subroutine or procedure to make some computation, you should make it as generally useful as possible. If you are sloppy and write a new subroutine for each new situation, your software will expand out of control. The same strategy has existed in mathematics since the ancient Greeks, where it has taken the best minds to recognize that a hodgepodge of different special cases sometimes fall under the same umbrella. As an inexperienced beginner, the problem becomes a lack of familiarity with all the relevant special cases that gave rise to the abstract definition. The problem with modern mathematical pedagogy is that we begin with the final product. This approach tends to be elegant, but sterile.

Simplicio: Don't think I haven't noticed.

16.2 Linear Algebra: The Role of Inner Products

Figure 16.3: Hermann Grassmann (1809-1877)

Galileo: To begin our discussion of inner product spaces, let us begin with the special case of the inner product defined on $\Re^n$.

Definition 16.2.1. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$, *are column vectors in* $\Re^n$, *then the inner product is* $< \mathbf{u}, \mathbf{v} >= \mathbf{u}^t\mathbf{v} = \sum_{k=1}^{n} u_k v_k$.

Simplicio: Remind me about that little t next to the vector $\mathbf{u}$.

Galileo: That exponent t indicates the transpose of the column vector to a row vector. While publishers would like all vectors to be written horizontally, we would like to think of them as column vectors.

Simplicio: What useful purpose does this serve?

Galileo: We would like to consider a matrix as a particulary useful type of function, whose domain consists of all the column vectors in $\Re^n$ and whose range consists of all the column vectors in $\Re^m$. These column vectors will be considered to be points. This function function can be computed by the rules of matrix multiplication.

Simplicio: Thus, the product $\mathbf{u}^t\mathbf{v}$ simply indicates the usual dot product. For example, if $\mathbf{u} = (u_1, u_2)^t$ and $\mathbf{v} = (v_1, v_2)^t$, then $< \mathbf{u}, \mathbf{v} >= \mathbf{u}^t\mathbf{v} = u_1v_1 + u_2v_2$.

Galileo: Yes, you are correct. Now, using this inner product, we can define the length of the vector.

Definition 16.2.2. $\|\mathbf{u}\|_2 = \sqrt{< \mathbf{u}, \mathbf{u} >}$.

Simplicio: Wait a minute. What is this notation $\|\mathbf{u}\|_2$? More to the point. What is that little subscript 2 doing there?

Virginia: Once again, I bet it is Pythagoras lurking around.

Galileo: We can also define the distance between two vectors.

Definition 16.2.3. *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$, *are column vectors in* $\Re^n$, *then the distance between* $\mathbf{u}$ *and* $\mathbf{v}$ *is* $dist(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2$.

Simplicio: In other words, the distance between two vectors is the length of their difference.

Galileo: Correct. In addition to length, the notion of inner product allows us to compute the cosine of the angle between two vectors $\mathbf{u}$ and $\mathbf{v}$.

Definition 16.2.4. *The cosine of the angle* θ *between two vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *is defined by the formula*

$$\cos(\theta) = \frac{< \mathbf{u}, \mathbf{v} >}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}.$$

Thus, we can now compute the angle θ by the arccosine function. We can also check to see if two vectors are 90 degrees (or orthogonal) by simply computing the inner product $< \mathbf{u}, \mathbf{v} >$. If this quantity equals zero, they are orthogonal. For example, if $\mathbf{u} = (c, -s)^t$ and $\mathbf{v} = (s, c)^t$, then $< \mathbf{u}, \mathbf{v} >= cs - cs = 0$. Thus, the vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal.

Simplicio: What is that little subscript 2 doing on the length formula $\|\mathbf{u}\|_2$?

Galileo: We put a subscript there to remind you to compute the square root of the sum of the squares of the coordinates of $\mathbf{u}$. As it turns out, we will sometimes find it convenient to compute $\|\mathbf{u}\|_p$. This symbol represents the p^{th} root of the sum of the p^{th} powers of the coordinates of $\mathbf{u}$. You computer types tend to like it when $p = 1$ or $p = \infty$.

Simplicio: In God's green earth, how can $p = \infty$? If $p = \infty$, then we are summing the infinite power of a bunch of numbers.

Galileo: The case $p = \infty$ denotes the maximum of the absolute values of all the coordinates. Don't worry. We will return to that point. OK, what can we observe about our rotation matrix **S**?

Simplicio: The two columns are orthogonal.

Virginia: And hence we won't have the problem with stability we had with our matrix **A**!

Galileo: You got it. Looking ahead, you might also like to know that the 2×2 Fourier matrix is defined by the equation:

$$\mathbf{F}_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

so this matrix has orthogonal columns each with unit length. In fact, if we interchange the two columns of F_2, the matrix represents a rotation of $\Theta = -45$. Note that $det(F_2) = -1$, which implies that there is a "flip" across some line in the plane. The beauty of the general Fourier matrix F_n is that it will have orthogonal columns of unit length.

Simplicio: This stuff is OK.

Galileo: Unfortunately, I have bad news for you. The situation deteriorates a little from here.

Simplicio: How so?

Galileo: When we begin approximating a function $f(x)$ on an interval $[a, b]$, we will have many different bases to chose from. For example, we can approximate the function by linear combinations of functions from the basis $B_P = \{1, x, x^2, \ldots, x^n\}$. This type of approximation is by polynomials. We have a number of different techniques including Taylor and polynomial interpolation. We can also approximate $f(x)$ by linear combinations of functions in the basis

$B_F = \{1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots, \cos(nx), \sin(nx)\}$. As it turns out for most applications, the second basis is preferred to the first. First, as we mentioned a few minutes ago, there are a multitude of applications involving some kind of wave phenomena. Since the functions $\cos(nx)$ and $\sin(nx)$ certainly look like waves, they

should provide a good model. Second, as we shall see in a moment, these functions have marvelously stable mathematical properties.

Simplicio: How so?

Galileo: I hate to tell you but the answer once again is, you guessed it, orthogonality.

Simplicio: But wait a minute. How the heck can two functions be orthogonal? That makes no sense.

Galileo: Now we are back to grumpy Grassmann, who recognized that we can compute the inner product of two continuous functions $f(x)$ and $g(x)$ defined on an interval $[a, b]$ by simply compute the integral. In other words, simply define the inner product by the formula:

$$< f(x), g(x) >= \int_a^b f(x)g(x)\ dx.$$

If you think of the integral as simply a fancy summation symbol and the values of x as coordinates, this formula is just an extension of the dot product. Thus, the functions $f(x)$ and $g(x)$ are orthogonal if $\int_a^b f(x)g(x)\ dx = 0$. In particular, if we consider the functions $\cos(mx)$ and $\sin(nx)$ to be defined on the interval $[-\pi, \pi]$, then it will turn out that $\int_{-\pi}^{\pi} \cos(mx)\sin(nx)\ dx = 0$. Thus, these two trigonometric functions are orthogonal. Are you back in your comfort zone yet?

Simplicio: I am getting there.

Virginia: If I hear you correctly, we can now compute the length of a function $f(x)$ by the formula

$$\|f(x)\|_2 = \sqrt{< f(x), f(x) >} = \sqrt{\int_a^b f(x)^2\ dx}.$$

We can also compute the cosine of the angle between two functions $f(x)$ and $g(x)$ by the formula

$$\cos(\theta) = \frac{< f(x), g(x) >}{\|f(x)\|_2 \|g(x)\|_2}.$$

Galileo: Correct.

Simplicio: But what does it mean to talk about the length of a function? What sense does it make to talk about the angle between two functions? In particular, what is the angle between the functions x^m and x^n?

Galileo: Unfortunately, the news here is not good. While we can easily compute the required integrals on any interval, say $[-1, 1]$, the cosine of the angle between them can be arbitrarily close to one implying the functions are close to being parallel. As we have already observed, this situation can lead to evil.

Virginia: I can visualize the problem here.

Galileo: Before we leave the topic of inner product, let's mention one more special case. In particular, if the functions $f(x)$, $g(x)$, and $\omega(x)$ are continuous on the interval $[a, b]$ and $\omega(x) > 0$ for all $x \in [a, b]$, then we can define an inner product by the rule

$$< f(x), g(x) > = < f(x), g(x) >_{\omega(x)} = \int_a^b f(x)\ g(x)\ \omega(x)\ dx.$$

The function $\omega(x)$ can be thought of as a weighting function.

Simplicio: Once again, I see this definition as just one more playground for the math geeks. It looks to me like abstraction for the sake of abstraction.

Galileo: Unfortunely, I think Professors Adrien-Marie Legendre (1752-1833), Charles Hermite (1822-1901), Pafnuty Chebyshev (1821-1894), and Edmund Nicholas Laguerre (1834-1886) might beg to differ. They each contributed to the study of *orthogonal polynomials*. As the name orthogonal polynomials suggests, these fellows studied polynomials, which by the appropriate choice of weighting function also happen to be orthogonal on some interval $[a, b]$. In each case, their method provides an elegant new basis for $C^0[a, b]$. Professor Legendre studied the case when the weighting function $\omega(x) = 1$ for all $x \in [-1, 1]$. Professor Hermite studied the case when $\omega(x) = e^{-x^2}$ for all $x \in (-\infty, \infty)$. Professor Chebyshev studied the case when $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ for all $x \in [-1, 1]$. Professor Laguerre studied the case when $\omega(x) = e^{-x}$ for all $x \in [0, \infty)$. For Professor Legendre the first few basis vectors are

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= x, \\ L_2(x) &= x^2 - \frac{1}{3}, \\ L_3(x) &= 5x^3 - 3x,\ etc. \end{aligned}$$

For Professor Hermite the first few basis vectors are

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= 2x, \\ H_2(x) &= 4x^2 - 2, \\ H_3(x) &= 8x^3 - 3x, \ etc. \end{aligned}$$

For Professor Chebyshev the first few basis vectors are

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \ etc. \end{aligned}$$

Finally, for Professor Laguerre the first few basis vectors are

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= -x + 1, \\ L_2(x) &= \frac{1}{2}(x^2 - 4x + 2), \\ L_2(x) &= \frac{1}{6}(-x^3 + 9x^2 - 18x + 6), \ etc. \end{aligned}$$

These polynomials can be computed using their definition, integration by parts, and the Gram-Schmidt orthogonalization process you learned in a beginning Linear Algebra course.

Since the weighted integral $\int_a^b f(x)g(x)\omega(x)\ dx$ is an inner product, anytime a fact is demonstrated about an inner product space, it will also be true for these orthogonal polynomials. The Pythagorean Theorem is the most notable example. These orthogonal polynomials not only have notable mathematical properties, but also have applications to differential equations and Physics. Legendre polynomials are closely associated with Laplace's equation, heat transfer, and the topic of spherical harmonics in physics. The literature written on these topics is vast.

Simplicio: I never studied these properties in my physics class. These applications sound difficult.

Galileo: We are now in a position to understand the virtues of orthogonal projections.

Simplicio: I hate to think.

Galileo: Well then, visualize for a moment that you are holding two cannonballs in your hands. If you let go, they drop to the floor. If they were close together at the beginning, they will land side-by-side when they hit the floor. Note that the angle between the flight-path of each ball and the floor is 90 degrees. In other words, any vector lying in the floor is orthogonal to the flight-path of each ball.

Simplicio: I see.

Galileo: On the other hand, suppose I fling the cannonballs sideway towards the edge of the room.

Simplicio: So?

Galileo: Even if they are close together when they are in your hands, they may still strike the floor at points far apart. If the room is large, they may land very far apart. The virtue of Fourier series approximation is that it amounts to an orthogonal projection from an infinite dimensional space into a finite dimensional space. Small errors in measurement at the beginning remain small. This is good.

Virginia: It is better to drop than fling?

Galileo: You have it. Moreover, the technique of linear least squares is also based on this same concept. Statisticians give daily thanks to the Greek Goddess Orthogonal.

Simplicio: This is more than I can stand.

Galileo: Since it took people decades to understand Grassmann, it is not too surprising you might have to think about these ideas for a minute or two. However, let me comment that an inner product can be defined on an abstract vector space by simply stating four simple properties. The whole process is amazingly elegant and simple. (Galileo sips.)

Virginia: I like these ideas. Simple is good.

Simplicio: I think I am going to have a bad hangover.

Galileo: No drinking for you. Alcohol kills brain cells you cannot afford to lose. In fact, we now summarize the connections between Geometry and Linear Algebra in Table 16.2.

Geometry	$\rightarrow$	Linear Algebra
point	$\rightarrow$	vector
line	$\rightarrow$	vector
ray	$\rightarrow$	vector
distance	$\rightarrow$	inner product
angle	$\rightarrow$	inner product
right angle	$\rightarrow$	orthogonality (inner product)
area	$\rightarrow$	determinant
volume	$\rightarrow$	determinant
congruence	$\rightarrow$	linear transformation
similarity	$\rightarrow$	linear transformation

Table 16.1: The Connections Between Geometry and Linear Algebra

.

Virginia: So the key ideas of Geometry are encapsulated in the four concepts: vector, inner product, determinant, and linear transformation.

Galileo: You got it. Better yet, it is rigorous and set up for making computations. The subject is perfect for our computer guys.

16.3 A Linear Algebra Version of Pythagoras

David Hilbert (1862-1943)

David Hilbert: "He who seeks for methods without having a definite problem in mind seeks in the most part in vain."

Galileo: Let us introduce David Hilbert, an expert in Linear Algebra. He wrote a classic work on the foundations of geometry, where his mission was to formulate the logical structure of geometry into the most mathematically correct framework possible. He also enjoyed applications as well as dancing on Saturday nights. We will ask him to present a more modern version of the Pythagorean Theorem.

Simplicio: Well, at least we don't have to deal with that logic guy. He was a downer.

Virginia: You try my patience.

Hilbert: In the interest of keeping the discussion accessible and concrete, we begin with an example.

Example 16.3.1. *Let*

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Note that the square of the length of the vector $\mathbf{u}$ *equals 2. Note that the square of the length of the vector* $\mathbf{v}$ *also equals 2.*

Note that the square of the length of the vector

$$\mathbf{u} + \mathbf{v} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

equals 4. Since $2 + 2 = 4$*, we have proved a special case of the Algebraic Version of the Pythagorean Theorem.*

Simplicio: Even I can handle that computation.

Hilbert: We now generalize this example by proving the theorem for column vectors in $\Re^n$ which have the form

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix},$$

where each $u_j \in \Re$.

Theorem 16.3.1 (Algebraic Version of Pythagoras). *If* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ *are two orthogonal vectors in* $\Re^n$*, then* $\|\mathbf{u} + \mathbf{v}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$.

Proof. By the properties listed in the previous proposition combined with the assumption that $< \mathbf{u}, \mathbf{v} >=< \mathbf{v}, \mathbf{u} > = 0$, we see that

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|_2^2 &= < \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} > \\ &= < \mathbf{u}, \mathbf{u} > + < \mathbf{u}, \mathbf{v} > + < \mathbf{v}, \mathbf{u} > + < \mathbf{v}, \mathbf{v} > \\ &= \|\mathbf{u}\|_2^2 + 0 + 0 + \|\mathbf{v}\|_2^2 \\ &= \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2. \end{aligned}$$

□

Simplicio: While the proof is short, I still don't like this unnecessary abstraction.

Hilbert: Do you understand how the theorem applies to the examples?

Simplicio: No problem, for the vectors $\mathbf{u} = (1, 1)^t$ and $\mathbf{v} = (1, -1)^t$, we simply observe that $\mathbf{u} + \mathbf{v} = (2, 0)^t$ and $4 = 2 + 2$.

For the vectors $\mathbf{u} = (3, 4)^t$ and $\mathbf{v} = (-4, 3)^t$, we simply observe that $\mathbf{u} + \mathbf{v} = (-1, 7)^t$ and $\|\mathbf{u} + \mathbf{v}\|_2^2 = 1 + 49 = 5^2 + 5^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$.

Hilbert: Good. To help you visualize the theorem in two dimensions, we have included a diagram in Figure 16.4.

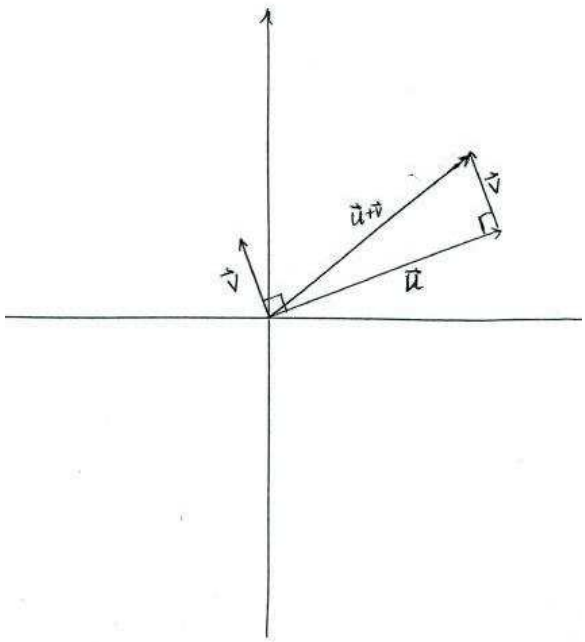


Figure 16.4: The Linear Algebra Version of the Pythagorean Theorem

Hilbert: Unfortunately, the abstraction gets worse. However, before we move in that direction, I would like to point out that the proof only used the properties of the inner product we showed in the proposition. You have now seen Hermann Grassmann at work. Namely, first identify and isolate the key properties associated with an idea and then prove as much as you can about the properties. A benefit of this process is that complicated summation notation is replaced by a pair of brackets. Once you

get used to this method of doing business, the ideas underlying the technique become more transparent. Later we will reprove the theorem for Fourier series, which live in a more general inner product space. In an effort to prepare you for Fourier Series, how about if we restate the Pythagorean Theorem the vector $\mathbf{w}$ is a linear combination of two orthogonal vectors $\mathbf{u}$ and $\mathbf{v}$, which have the same length L. For Fourier series the constant $L = \sqrt{\pi}$. Note also, that if the constant $L = 1$, then the statement is the same as the theorem we just presented.

Example 16.3.2. *Let*

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

As we noted before, the length of the vectors $\mathbf{u}$ *and* $\mathbf{v}$ *both equal* $\sqrt{2}$.

If $\mathbf{w} = a\mathbf{u} + b\mathbf{v}$, *then* $\|\mathbf{w}\|_2^2 = 2(a^2 + b^2)$. *Note that this observation is a special case of the next Theorem.*

Theorem 16.3.2 (Algebraic Version of Pythagoras 2). *Let* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ *and* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ *be two orthogonal vectors in* $\Re^n$ *with the property that* $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = L$. *If* $a, b \in \Re$ *and* $\mathbf{w} = a\mathbf{u} + b\mathbf{v}$, *then* $\|\mathbf{w}\|_2^2 = L^2(a^2 + b^2)$.

Proof. Since $\mathbf{w} = a\mathbf{u} + b\mathbf{v}$,

$$\begin{aligned}
\|\mathbf{w}\|_2^2 = \; & = < \mathbf{w}, \mathbf{w} > \\
& = < a\mathbf{u} + b\mathbf{v}, a\mathbf{u} + b\mathbf{v} > \\
& = \; a^2 < \mathbf{u}, \mathbf{u} > + ab < \mathbf{u}, \mathbf{v} > + ba < \mathbf{v}, \mathbf{u} > + b^2 < \mathbf{v}, \mathbf{v} > \\
& = \; a^2 < \mathbf{u}, \mathbf{u} > + 0 + 0 + b^2 < \mathbf{v}, \mathbf{v} > \\
& = a^2 L^2 + b^2 L^2 \\
& = L^2(a^2 + b^2).
\end{aligned}$$

□

Simplicio: OK, why should I care about this last theorem? Why don't we just leave these ideas in Geometry where they belong?

Hilbert: We are looking ahead to Fourier series, which we will be discussing at a later date. The 2×2 matrix

$$\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

is the matrix representing the discrete Fourier transform. Note that the columns of this matrix are the vectors $\mathbf{u}$ and $mathbfV$ we just discussed in the previous example. The advantage of Peano's abstract definition of vector space is that functions can also be thought of as vectors. In particular, trigonometric functions such as $1, \cos(x)$, and $\sin(x)$ can now be considered vectors in a very large (i.e infinite dimensional) space. While you may think of the inner product as the dot product of two vectors in $\Re^n$, we can also define the inner product of two functions as the integral of their product over some interval $[a, b]$. When we discretize these functions, we end up with the 3×3 Fourier matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 1 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix}.$$

Note that the columns of this matrix are pairwise orthogonal.

Simplicio: So?

Hilbert: This observation is important because the Pythagorean Theorem can now be applied to these three column vectors to inforce stability.

Simplicio: I see that the columns are orthogonal. Infinite dimensional spaces?

Example 16.3.3. *Hilbert: The idea of an infinite dimensional space is not so strange when you realize that each point $x \in [a, b]$ can be thought of as a coordinate for a function $f(x)$. Since trigonometric functions are usually defined on the interval $[-\pi, \pi]$ (or $[0, 2\pi]$), the inner product becomes*

$$< f(x), g(x) > = \int_{-\pi}^{\pi} f(x)g(x)\ dx.$$

The make the connection between Fourier series and Pythagoras, let $h(x) = a\cos(x) + b\sin(x)$. *Since it is an easy exercise from calculus to show that*

1. $< \cos(x), \cos(x) >=< \sin(x), \sin(x) >= \pi$ *and*

2. $< \cos(x), \sin(x) >=< \sin(x), \cos(x) >= 0,$

we observe that

$$\begin{aligned}\int_{-\pi}^{\pi} (h(x))^2 \, dx &= < a\cos(x) + b\sin(x), a\cos(x) + b\sin(x) > \\ &= a^2 < \cos(x), \cos(x) > + 2ab < \cos(x), \sin(x) > \\ &+ b^2 < \sin(x), \sin(x) > \\ &= a^2\pi + 2 * 0 + b^2\pi = \pi(a^2 + b^2).\end{aligned}$$

Grassmann would be proud to see his abstract definition of the inner product becoming a central focus in this important application.

Virginia: I see the potential here for some interesting mathematical ideas.

Exercise Set 16.1.

1. If $\mathbf{u}, \mathbf{v}$, and $\mathbf{w}$ represent the columns of the matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 1 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix},$$

   then show that $\|\mathbf{u} + \mathbf{v} + \mathbf{w}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 + \|\mathbf{w}\|_2^2$.

2. Prove the Pythagorean Theorem for three vectors: If $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$, $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$, and $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$ are three vectors in $\Re^n$ with the property that $\mathbf{u} \perp \mathbf{v}$, $\mathbf{u} \perp \mathbf{w}$, and $\mathbf{v} \perp \mathbf{w}$, then $\|\mathbf{u} + \mathbf{v} + \mathbf{w}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 + \|\mathbf{w}\|_2^2$.

3. Prove the following theorem: Let $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t, \mathbf{v} = (v_1, v_2, \ldots, v_n)^t$, and $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$ be pairwise orthogonal vectors in $\Re^n$ with the property that $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = L$. If $a, b, c \in \Re$ and $\mathbf{z} = a\mathbf{u} + b\mathbf{v} + c\mathbf{w}$, then $\|\mathbf{z}\|_2^2 = L^2(a^2 + b^2 + c^2)$.

4. Prove the parallelogram law: If $\mathbf{u} = (u_1, u_2, \ldots, u_n)^t$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$, then $\|\mathbf{u} + \mathbf{v}\|_2^2 + \|\mathbf{u} - \mathbf{v}\|_2^2 = 2\|\mathbf{u}\|_2^2 + 2\|\mathbf{v}\|_2^2$.

5. Prove the Law of Cosines.

Part VII

Day 5. Approximation Theory

Chapter 17

Taylor Polynomials

Brook Taylor

Galileo: We now ask Professor Taylor to rejoin us so he can explain the general version of the theorem that made him famous. While strictly speaking it is not a method of interpolation, it does provide an entry point into the topic of polynomial interpolation. Professor Taylor, tell us your theorem.

Taylor: Actually, it is a straight forward generalization of what we presented when we showed the method of Newton/Raphson converges quadratically. Again, the concept is to write a given function $f(x) = p_n(x) + E_n(x)$, where $p_n(x)$ is a polynomial of degree n and $E_n(x)$ represents the error. In other words, a smooth function can be

written as the sum of a polynomial and an error. With luck, the error term will be small.

Theorem 17.0.3 (Taylor). *If $a < x, x_0 < b$ and $f(x) \in C^{n+1}[a,b]$, then*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n \, dt.$$

Proof. The proof employs the same technique as the one given for the case $n = 1$. The idea is to always attack the error term using the technique of integration by parts, where we integrate $f^{(n+1)}(t)$ and differentiate $(x-t)^n$. Since we have already demonstrated the proof for $n = 1$, we will prove the next case when $n = 2$.

For $n = 2$ we let $u = (x-t)^2$ and $dv = f'''(t)\, dt$ so that $du = -2(x-t)dt$ and $v = f''(t)$. When we apply integration by parts, we get the following reduction.

$$\begin{aligned}\int_{x_0}^{x} f'''(t)(x-t)^2 dt &= (x-t)^2 f''(t)|_{t=x_0}^{x} - \int_{x_0}^{x} f''(t)(-2)(x-t)dt \\ &= -(x-x_0)^2 f''(x_0) + 2\int_{x_0}^{x} f''(t)(x-t)dt \\ &= -(x-x_0)^2 f''(x_0) + 2[f'(x_0)(x-x_0) + f(x) - f(x_0)].\end{aligned}$$

Thus, $f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \frac{1}{2}\int_{x_0}^{x} f'''(x)(x-t)^2 \, dt$.

The general case is proved by applying the technique of integration by parts n times to the integral $\int_{x_0}^{x} f^{n+1}(t)(x-t)^n dt$, where $u = (x-t)^n$ and $dv = f^{n+1}(t)dt$. If $n = 47$, then the technique will have to be applied 47 times.

□

Taylor: Note that any polynomial of the form $p_2(x) = a_0 + a_1 x + a_2 x^2$ is a Taylor series. Similarly, any n^{th} degree polynomial $p_n(x) = \sum_{k=0}^{n} a_k x^k$ represents a Taylor series. Here is an example to work out.

Example 17.0.4. *Compute the first $n+1$ non-zero terms of the Taylor series for the function $f(x) = \cos(x)$ at the point $x = x_0 = 0$. Since $\cos(x) = \cos(-x)$, for all $x \in \Re$, the function $\cos(x)$ is an even function. This fact should tip you off that the Taylor series expansion will only have even powers of x represented.*

Solution: When we compute the derivatives of $f(x)$*, we find that* $f(0) = 1, f'(0) = 0, f''(0) = -1, f'''(0) = 0, f^{(4)}(0) = 1, etc.$

Thus, the series expansion at $x = 0$ *is*

$$\cos(x) \approx 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 + \cdots + (-1)^k \frac{1}{(2n)!}x^{2n} = \sum_{k=0}^{n} (-1)^k \frac{1}{(2k)!}x^{2k}.$$

Simplicio: Where did the formula for the polynomial
$p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$ come from? How did you ever think of that?
Taylor: You can figure out the formula for yourself. Just work out the next example.

Example 17.0.5.

Problem: If $p_2(x) = 2 + 3x + 5x^2$ *and* $x_0 = 7$*, then find constants* A, B, C *so that* $p_2(x) = A + B(x - 7) + C(x - 7)^2$.

Solution: If $p_2(x) = A + B(x - 7) + C(x - 7)^2$*, then* $p_2(7) = A$.

Since $p_2'(x) = B + 2C(x - 7)$, $p_2'(7) = B$.

Since $p_2''(x) = 2C$, $p_2'(7) = \frac{C}{2}$.

Thus, $p_2(x) = p_2(7) + p_2'(7)(x - 7) + \frac{p_2''(7)}{2}(x - 7)^2$.

Taylor: This last exercise provides formulas for a polynomial when expanded about an arbitrary point x_0. This formula is exactly my theorem for the special case that the function is a polynomial.
Simplicio: But what about the error term?
Taylor: Since you have begun the process with a polynomial, the errors are all zero. In other words, there is no error term.
Simplicio: And of course, I have to complain about the proof. While I understand integration by parts, I must say I am curious about how you came up with that idea.
Galileo: Now you are asking the more difficult question: How does the creative process take place in your brain? While the answer will probably never be known, hard work and careful thought are definitely prerequisites.
Lagrange: I would like to intercede a second time to insist that I have a more elegant form of this theorem, where the integral in the error term is replaced by a derivative similar to the ones in the polynomial part.

Theorem 17.0.4 (Taylor). *If $a < b$ and $f(x) \in C^{n+1}[a,b]$, then for every pair of points $x, x_0 \in (a,b)$ there is a point $z \in (a,b)$ such that*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{(n+1)}.$$

Proof. As before, we will only prove this form of the theorem for the integer $n = 2$. In this case, we have the equation $f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \frac{1}{2}\int_{x_0}^{x} f'''(t)(x-t)^2\,dt$. Again, by the Intermediate Value Theorem for Integrals, we see that there is a point z between x_0 and x such that

$$\begin{aligned}\frac{1}{2}\int_{x_0}^{x} f'''(x)(x-t)^2\,dt &= \frac{1}{2}f'''(z)\int_{x_0}^{x}(x-t)^2\,dt \\ &= \frac{1}{2}f'''(z)\frac{(x-x_0)^3}{3} \\ &= \frac{f'''(z)}{6}(x-x_0)^3\end{aligned}$$

□

Simplicio: While all this mathematics is quite lovely, could you give me one useful application.

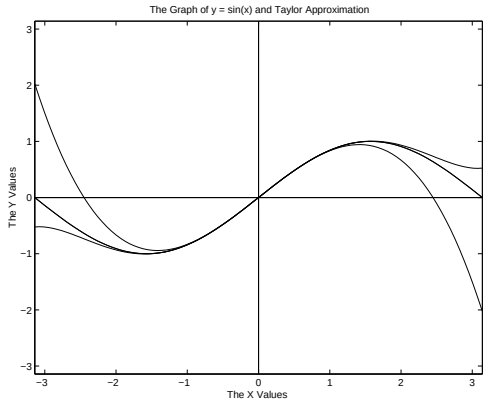


Figure 17.1: Successive Taylor Approximations of $f(x) = \sin(x)$

Galileo: This request is not a problem. Consider the question of designing a calculator. On that calculator you would like to have a button, which computes the value of $\sin(x)$ for a given value of x.

Simplicio: That feature would be convenient.

Galileo: So how do you think you might design such a device?

Simplicio: Since there is no formula for $\sin(x)$, I have no earthly idea.

Galileo: While Taylor's theorem does not provide a formula for the exact value of $\sin(x)$, it does manage to provide a formula for an approximation to an accuracy as close as you wish. In particular, the strategy can be described in the following steps:

1. Decide the accuracy you require. For single precision, this requirement is $\frac{1}{10^7}$. For double precision, this requirement is $\frac{1}{10^{14}}$.

2. Decide the size of the interval (a, b) you would like to compute. Since $\sin(x)$ is 2π periodic, this interval might be $[-\pi, \pi]$.

3. Find an integer n so that the error term $E_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x - x_0)^{n+1}$ is less than the required accuracy for all $x \in (a, b)$.

Simplicio: That strategy sounds reasonable.

Galileo: Well then, here are some problems to practice on.

Exercise Set 17.1.

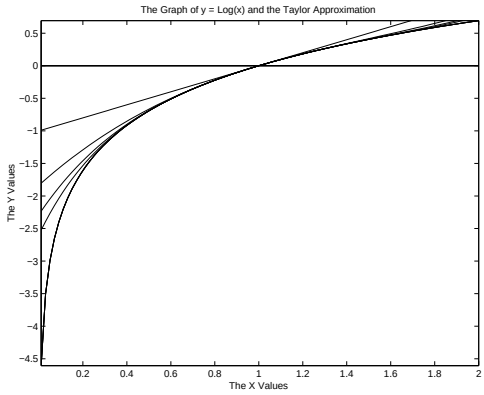


Figure 17.2: Successive Taylor Approximations of $f(x) = \ln(x)$

1. Compute the first 5 non-zero terms of the Taylor series for the functions $\sin(x)$, $\frac{1}{1+x^2}$, $\tan(x$ and e^x at the point $x = x_0 = 0$. Note that while the function $\frac{1}{1+x^2}$ requires only even powers of x, the functions $\sin(x)$ and $\tan(x)$ have only odd powers of x represented. Do you remember the definition of what it means for a function to be *even* or *odd*?

2. Compute the first 5 non-zero terms of the Taylor series for the function $ln(x)$ at the point $x = x_0 = 1$.

3. Given the function $f(x) = \sin(x)$ defined on the interval $[-\pi, \pi], x_0 = 0$, and a tolerance $tol = \frac{1}{10,000}$ determine an integer n with the property that the polynomial $p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$ has the property that $|f(x)-p_n(x)| \leq tol$ for all $x \in [-\pi, \pi]$. Could this technique be effectively programmed into a calculator to estimate the function $\sin(x)$ to single precision? What if we reduce the size of the interval to $[-\pi/2, \pi/2]$? What about double precision?

4. Given the function $f(x) = \cos(x)$ defined on the interval $[-\pi/2, \pi/2]$ and a tolerance $tol = \frac{1}{10,000,000}$ determine how many terms of the Taylor series will be required to guarantee that the error between the function and the Taylor series is less than the tolerance.

5. Given the function $f(x) = e^x$ defined on the interval $[-1, 1]$ and a tolerance $tol = \frac{1}{10,000,000}$ determine how many terms of the Taylor series will be required to guarantee that the error between the function and the Taylor series is less than the tolerance.

6. Given the function $f(x) = ln(1 + x)$ defined on the interval $[\frac{-1}{2}, \frac{1}{2}], x_0 = 0$, and a tolerance $tol = \frac{1}{10,000}$ determine an integer n with the property that the polynomial $p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$ has the property that $|f(x)-p_n(x)| \leq tol$ for all $x \in [\frac{-1}{2}, \frac{1}{2}]$. Could this technique be effectively programmed into a calculator to estimate the $ln(x)$ to single precision? What if we reduce the size of the interval to $[\frac{-1}{4}, \frac{1}{4}]$? What about double precision?

7. What about the Taylor Series error formula for the functions $\frac{1}{1+x^2}$ and $\tan(x)$? (Assume that $x_0 = 0$.)

8. Show that the Mean Value Theorem is a special case of Taylor's Theorem when $n = 0$.

9. Use Taylor's Theorem to compute the square root of 3. (Hint: Let $f(x) = \sqrt{x}, x_0 = 4$, and $x = 3$.) How many terms of the Taylor series will be needed to to guarantee an accuracy of less than 0.00001? What happens when $x_0 = 1$, and $x = 2$ are used to compute $\sqrt{2}$?

10. If $p_2(x) = 2+3x+5x^2$ and $x_0 = 7$, then show that $p_2(x) = p_2(7)+p_2'(7)(x-7)+\frac{p_2''(7)}{2}(x-7)^2$. (Hint: Let $p_2(x) = A+B(x-7)+C(x-7)^2$, compute derivatives, substitute $x = 7$, and solve for A, B, and C.)

11. If $p_2(x) = a_0 + a_1x + a_2x^2$ and x_0 is any real number, then show that $p_2(x) = p_2(x_0) + p_2'(x_0)(x - x_0) + \frac{p_2''(x_0)}{2}(x - x_0)^2$.

12. If $p_n(x) = \sum_{k=0}^{n} a_k x^k$ and x_0 is any real number, then show that $p_n(x) = \sum_{k=0}^{n} \frac{p_n^{(k)}}{k!}(x - x_0)^k$.

13. Given the function $f(x) = ln(x)$ defined on the interval $[\frac{3}{4}, \frac{5}{4}], x_0 = 1$, and a tolerance $tol = \frac{1}{10{,}000}$ determine an integer n with the property that the polynomial $p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$ has the property that $|f(x)-p_n(x)| \leq tol$ for all $x \in [\frac{3}{4}, \frac{5}{4}]$.

Chapter 18

Polynomial Interpolation

Galileo: Let us begin with an investigation of three simple and easy to understand sequences.

The first sequence begins $1, 3, 5, 7, 9$.

The second sequence begins $1, 2, 4, 8, 16$.

The third sequence begins $1, 1, 2, 3, 5, 8$.

The question 8 year old kids are often asked in their elementary mathematics classes is: What is the next term in each of these sequences?

Simplicio: Even I can answer these questions. Since the first sequence is clearly arithmetic, the next term will be 2 more than the last and thus 11. Since the second sequence is clearly geometric and each term is twice the previous, the next term will equal 32. The last sequence is clearly Fibonacci where the rule is that each term is the sum of the previous two terms, the answer is 13.

Galileo: Not so fast. I contend that the next term for each sequence should equal 47.

Simplicio: Impossible!

Galileo: Actually, you provided evidence that indicates your prejudice when you identified the sequences as arithmetic, geometric, and Fibonacci. I did not provide that information. Thus, you read a structure into the problem that was not present.

Simplicio: But that is what I did when I was a kid and I always got the right answers then. Are you telling me that the rules of mathematics have changed?

Galileo: No. I am telling you that the structure of the answer was implied, but not explicitly provided, which means that a a unique answer is not forced. In particular, your teachers were sufficiently sloppy that any answer is correct. This problem could have been avoided if they had been more specific when they asked the question. Of course, you have worked enough of these problems that you instinctively know which answer the teacher wants so you have no problem.

Simplicio: To think that I have been misled all these years. This turn of events is quite disturbing.

Galileo: No worries. We will now show how to produce a formula to interpolate any set of data using the technique of polynomial interpolation. The idea is the following. For any given finite set of data points (x_k, y_k), $k = 0, 1, 2, \ldots, n$, where the x_k's are distinct, we can find a degree n polynomial $p_n(x)$ such that $p_n(x_k) = y_k$ for all $k = 0, 1, 2, \ldots, n$. In particular, we can find a 5^{th} degree polynomial $p_5(x)$, which interpolates the data $(0, 1), (1, 3), (2, 5), (3, 7), (4, 9)$, and $(5, 47)$.

Simplicio: I will be interested to see that polynomial $p_5(x)$.

Galileo: In this presentation we will provide three different techniques used to perform the interpolation, a statement of the Lagrange error formula, and the classical example of Runge, which indicates that polynomial interpolation can be dangerous. Note that while three different techniques are presented, they all produce the same answer.

Simplicio: If the method is dangerous, then why would we play with it?

Galileo: Because the methods are easy to understand and they give insight into least squares, Fourier, and spline methods, which are used on a daily basis in today's world of high technology.

18.1 The Method of Lagrange

Joseph Louis-Lagrange: "As long as algebra and geometry have been separated, their progress has been slow and their uses limited; but when these two sciences have been united, they have lent each mutual forces, and have marched together towards perfection."

Galileo: The first interpolation technique to be presented is the method of Lagrange polynomial interpolation. While Joseph Louis-Lagrange (1736-1813) made numerous contributions to algebra, analysis, and differential equations, his observations concerning polynomial interpolation also bear his name. Napoleon named Lagrange to the Legion of Honour and Count of the Empire in 1808. He summarized his life's work with the quote "I do not know."

We begin the discussion of our first technique for polynomial interpolation with a definition.

Note that the following statements are easy to check.

Proposition 18.1.1. *If* (x_k, y_k), $k = 0, 1, 2, \ldots, n$ *are given points with the* x_k*'s distinct and* $w_k(x) = (x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)$, *then the functions* $L_k(x) = \frac{w_k(x)}{w_k(x_k)}$ *satisfy the following relations:*

1. $\deg(L_k(x)) = n$,

2. $L_k(x_k) = 1$, *and*

3. $L_k(x_j) = 0$ *if* $j \neq k$.

Definition 18.1.2. *If the data points* (x_k, y_k), $k = 0, 1, 2, \ldots, n$ *have the property that the* x_k*'s are distinct, then Lagrange interpolating polynomials are defined by the formula* $p_n(x) = \sum_{k=0}^{n} y_k \cdot L_k(x)$.

Proposition 18.1.3 (The Method of Lagrange). *If points* (x_k, y_k), $k = 0, 1, 2, \ldots, n$ *are given, where the* x_k*'s are distinct, then the polynomial* $p_n(x) = \sum_{k=0}^{n} y_k \cdot L_k(x)$ *has the property that* $p_n(x_k) = y_k$ *for all* $k = 0, 1, 2, \ldots, n$.

Proof. This proposition is immediate since $L_k(x_k) = 1$ and $L_k(x_j) = 0$, if $j \neq k$. □

Exercise Set 18.1.

1. Use the method of Lagrange to find a quadratic polynomial $p_2(x)$ such that $p_2(1) = 3, p_2(2) = 5$, and $p_2(3) = 7$.

2. Use the method of Lagrange to find a cubic polynomial $p_3(x)$ such that $p_3(1) = 3, p_3(2) = 5, p_3(3) = 7$, and $p_3(4) = 11$.

3. Find a 5^{th} degree polynomial $p_5(x)$, which interpolates the data $(0, 1), (1, 3), (2, 5), (3, 7), ($ and $(5, 47)$.

18.2 The Technique of Newton Divided Differences

Galileo: Sir Isaac Newton (1642-1727) was an English mathematician, who made historic contributions to mathematics, optics, and celestial mechanics. The *Principia* is recognized as the greatest scientific book ever written. In this monumental work he analyzed the motion of the bodies in resisting and non-resisting media under the action of centripetal forces. The results were applied to orbiting bodies, projectiles,

pendulums, and free-fall near earth. He further demonstrated that the planets were attracted toward the sun by a force varying as the inverse square of the distance. He also explained the eccentric orbits of comets, the tides, and the precession of the earth's axis. While the invention of the calculus may have been his greatest contribution to mathematics, his method of divided differences provides a computationally efficient technique for implementing polynomial interpolation.

Let us begin the discussion by solving the following simple problem. Given three points $(x_0, y_0), (x_1, y_1), (x_2, y_2)$, find constants c_0, c_1, c_2 with the property that the polynomial $p_2(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)$ has the property that $p_2(x_0) = y_0, p_2(x_1) = y_1$, and $p_2(x_2) = y_2$. A quick check shows that $c_0 = y_0$ and $c_1 = \frac{y_1 - y_0}{x_1 - x_0}$. A not so quick check shows that

$$c_2 = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0}$$

We can begin to understand these formulas if we assume the data is generated by a function $y = f(x)$. In particular, if $y_k = f(x_k)$ for all $k = 0, 1, 2$.

Definition 18.2.1. *Let $f(x)$ be a function defined on the interval $[x_0, x_n]$.*

$$\begin{array}{lll} x_0 & f[x_0] = f(x_0) & \\ & & f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} \\ x_1 & f[x_1] = f(x_1) & \qquad\qquad f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}. \\ & & f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1} \\ x_2 & f[x_2] = f(x_2) & \end{array}$$

Note that such a construction is called a "cascade."

Proposition 18.2.2. *If $y = f(x)$ is a function and $y_0 = f(x_0), y_1 = f(x_1)$, and $y_2 = f(x_2)$, where the points x_0, x_1, and x_2 are distinct, then the polynomial $p_2(x)$ defined by $p_2(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$ has the property that $p_2(x_0) = y_0, p_2(x_1) = y_1$, and $p_2(x_2) = y_2$.*

Proof. It is easy to check that $p_2(x_0) = y_0$ and $p_2(x_1) = y_1$. A bit of algebra can be used to show that $p_2(x_2) = y_2$. □

Thus, the previous proposition can be used to show that formulas exist for the constants c_0, c_1, and c_2; namely, the top entry of each column in the cascade.

We now indicate how this technique can be applied to any set of data points by making the following definition.

Definition 18.2.3. *The k^{th} divided difference relative to $x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}$ is given by*

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, \ldots x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

While it is easy to check that formula $p_2(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$ interpolates the data, it is more tedious to check the general case.

Proposition 18.2.4 (Newton Divided Differences). *If $x_0, x_1, x_2, \ldots, x_n$ are distinct points and $y_k = f(x_k)$ for all $k = 0, 1, 2, \ldots, x_n$, then the polynomial*

$$p_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, \ldots, x_k](x - x_0) \ldots (x - x_{k-1})$$

has the property that $p_n(x_i) = f(x_i)$ for all $i = 0, 1, 2, \ldots, n$.

Simplicio: So, why should I waste my time learning this second method?

Galileo: Let us inquire what our friendly expert, Isaac Newton, has to say on this matter.

Newton: Well, if you had bothered to work out the previous two exercises, you would have noticed that computing the cubic polynomial $p_3(x)$ is only slightly more work than computing the quadratic polynomial $p_2(x)$. If you want to really appreciate my method, then use the method of Lagrange to compute these two polynomials.

Simplicio: But I did use the method of Lagrange to compute $p_2(x)$ and $p_3(x)$. It didn't seem bad at all.

Newton: Did you simplify your answer for $p_3(x)$ so that it is in the form $p_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$?

Simplicio: No.

Newton: After you do this simple exercise, then you can complain about my method. Not until.

Exercise Set 18.2.

1. Use the method of Newton divided differences to find a quadratic polynomial $p_2(x)$ such that $p_2(1) = 3, p_2(2) = 5$, and $p_2(3) = 7$.

2. Use the method of Newton divided differences to find a cubic polynomial $p_3(x)$ such that $p_3(1) = 3, p_3(2) = 5, p_3(3) = 7$, and $p_3(4) = 11$.

3. Show that the Newton divided difference formula works for cubic polynomials.

18.3 The Technique of Vandermonde

Galileo: Alexandre Theophile Vandermonde (1735-1796) was a French mathematician, whose first love was music. He only turned to mathematics when he was 35 years old. His mathematical interests were in the theory of equations and the theory of determinants.

The technique of Vandermonde evolves in a natural way from the problem: Given a set of data points $(x_0, y_0), (x_1, y_1)$, and (x_2, y_2), where the x_k's are distinct, find a quadratic polynomial of the form
$p_2(x) = a_0 + a_1 x + a_2 x^2$ such that $p_2(x_0) = y_0, p_2(x_1) = y_1$, and $p_2(x_2) = y_2$. Thus, a system of three equations and three unknowns must be solved. In matrix format this system becomes:

$$\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

More generally, given a set of data points (x_k, y_k), $k = 0, 1, 2, \ldots, n$, where the x_k's are distinct, find an $n-$degree polynomial of the form $p_n(x) = \sum_{k=0}^{n} a_k x^k$ such that $p_n(x_k) = y_k$, for all $k = 0, 1, 2, \ldots, n$. The answer to this question is the solution to the following system of equations:

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Thus, the constants $a_0, a_1, \dots, a_n$ can be found if the system can be solved. The following proposition shows the system can be solved as long as the x_k's are distinct.

Proposition 18.3.1. *If*

$$V_n = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix},$$

then $det(V_n) = \prod_{0 \le i < k \le n} (x_k - x_i)$.

Proof. If we let

$$V_n(x) = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \dots & x_{n-1}^n \\ 1 & x & x^2 & \dots & x^n \end{pmatrix},$$

then notice that $det(V_n(x))$ is a polynomial of degree n with roots $x_0, x_1, x_2, \dots, x_{n-1}$. Thus, $det(V_n(x)) = C_n(x - x_0)(x - x_1)\dots(x - x_{n-1})$, where C_n is some constant. While a straightforward induction argument can be used to show that $C_n = det(V_{n-1}(x_{n-1}))$, the proof is best understood by simply computing the special cases when $n = 1$ and $n = 2$.

□

A matrix of the form given in the previous proposition is called a Vandermonde matrix.

The previous proposition shows that if the x_k's are pair-wise distinct, then the determinant is different from zero and the system cannot only be solved, but the solution is unique. The matrix V is called a *Vandermonde* matrix.

Proposition 18.3.2. *If (x_k, y_k), $k = 0, 1, 2, \ldots, n$ are $n+1$ distinct points, then the polynomial defined by $p_n(x) = \sum_{k=0}^{n} a_k x^k$, where the values of $a_0, a_1, \ldots, a_n$ are computed as the solution of the equation*

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

has the property that $p_n(x_k) = y_k$ for all $k = 0, 1, 2, \ldots, n$.

Proof. Note that this proposition is a simple restatement in matrix form that $p_n(x_k) = y_k$ for all $k = 0, 1, 2, \ldots, n$. □

Galileo: The next proposition is a uniqueness theorem.

Simplicio: Why would I possibly care about uniqueness?

Galileo: Well, we have shown you three different techniques to compute the interpolating polynomial. You might wonder if you might get three different answers. In fact, the next proposition shows that all techniques will result in the same answer.

Proposition 18.3.3. *Let (x_k, y_k), $k = 0, 1, 2, \ldots, n$ be a set of $n+1$ points. If the x_k's are distinct and $p_n(x) = \sum_{k=0}^{n} a_k x^k$ and $q_n(x) = \sum_{k=0}^{n} b_k x^k$ are polynomials such that $p_n(x_k) = q_n(x_k)$ for all $k = 0, 1, \ldots, n$, then $a_k = b_k$ for all $k = 0, 1, \ldots, n$.*

Proof. If $p_n(x_k) = q_n(x_k)$ for all $k = 0, 1, \ldots, n$, then we have a system of linear equations of the form $\mathbf{Va} = \mathbf{Vb}$, where $\mathbf{V}$ is a Vandermonde matrix, $\mathbf{a} = (a_0, a_1, \ldots, a_n)^t$,

and $\mathbf{b} = (b_0, b_1, \ldots, b_n)^t$. Since the x_k's are distinct, the determinant of the Vandermonde matrix is different from zero so that the matrix V has an inverse. Thus, $a_k = b_k$ for all $k = 0, 1, \ldots, n$. □

Simplicio: But wait a minute! How am I going to solve a system of 3 equations and 3 unknowns or 4 equations and 4 unknowns? These computations will be required to get the final answer?

Galileo: That is why Babbage invented the computer.

Simplicio: Who was Babbage?

Galileo: Charles Babbage (1791-1871) was the designer of the difference engine, which implemented Newton's method of divided differences. Together with a bit of help from his lady friend, Augusta Ada King, countess of Lovelace (1815-1852), he also designed (but never built) the forerunner of the modern electronic computer. If you want to see a reconstruction of his difference engine, visit the Science Museum in London. It weighs a mere 3 tons.

Simplicio: Not a calculator you could strap to your belt.

Exercise Set 18.3.

1. Use the method of Vandermonde to find a quadratic polynomial $p_2(x)$ such that $p_2(1) = 3, p_2(2) = 5$, and $p_2(3) = 7$.

2. Use the method of Vandermonde to find a cubic polynomial $p_3(x)$ such that $p_3(1) = 3, p_3(2) = 5, p_3(3) = 7$, and $p_3(4) = 11$.

3. Use the method of Vandermonde to find a cubic polynomial $p_3(x)$ such that $p_3(-1) = 2, p_3(0) = 5, p_3(2) = 7$, and $p_3(-2) = 3$.

18.4 Error Estimation for Polynomial Interpolation

Galileo: We now turn to the problem of computing the error between a function and its polynomial interpolation. While we have three different techniques for polynomial interpolation (Lagrange, Newton, and Vandermonde), we saw at the end of the last section that they all produce the same answers. The focus of the next discussion is to provide a formula for the error. Since the three techniques all produce the same polynomial approximation $p_n(x)$, we only need one error formula.

Simplicio: While one error formula is good news, I can tell that more theory is on the way. I would appreciate it if we could keep the discussion simple.

Galileo: Professor Lagrange could you help us?

Lagrange: If a function $f(x)$ is differentiable at every point in an interval $[x_0, x]$, then we know by the Mean Value Theorem that there is a point $z \in [x_0, x]$ so that $f(x) = f(x_0) + f'(z)(x - x_0)$. Just as Rolle's Theorem can be used to prove the Mean Value Theorem, the generalized Rolle's Theorem can be used to prove the error formula for polynomial interpolation.

Simplicio: But I don't remember Rolle's Theorem.

Lagrange: The way to visualize Rolle's Theorem is to imagine throwing a ball in the air and catching it when it comes down. What can you say about the velocity of the ball at its highest point?

Simplicio: Since the ball is changing direction from upward to downward motion, obviously the velocity is zero.

Lagrange: Your observation is correct. Now take that observation one step further by throwing a ball into the air and instead of catching it on the way down, let it hit the ground and bounce back up into your hand. If this experiment is conducted carefully, there will be three different moments in time, where the height of the ball is the same (i.e. the height of your hand above the ground). What can you conclude?

Simplicio: The ball will now have two different moments in time, where the velocity

is zero. I don't get it.

Lagrange: Well, if the velocity is zero at two different points in time, then what can you say about the acceleration?

Simplicio: It seems like the acceleration must be zero at some moment in time between when the velocities are zero.

Lagrange: You are correct. Now you are ready to understand a general theorem, which we now state. We indicate a proof for the cases when $n = 1$ and $n = 2$.

Theorem 18.4.1 (The Generalized Rolle's Theorem). *If $f(x) \in C^n[a, b]$, $a \leq x_0 < x_1 < \cdots < x_n \leq b$, and $f(x_k) = 0$ for all $k = 0, 1, 2, \ldots, n$, then there exists a point $z \in (a, b)$ such that $f^{(n)}(z) = 0$.*

Proof. If $n = 1$, then we have two distinct points x_0 and x_1 so that $f(x_0) = 0$ and $f(x_1) = 0$. By the Rolle's Theorem you endured in your first calculus course, there is a point z between x_0 and x_1 so that $f'(z) = 0$. In particular, when $n = 1$, the Generalized Rolle's Theorem is exactly Rolle's Theorem.

If $n = 2$, then we have three distinct points x_0, x_1, and x_2 so that $f(x_0) = f(x_1) = f(x_2) = 0$. Thus, a point z_1 can be found in the interval (x_0, x_1) such that $f'(z_1) = 0$ and a point z_2 can be found in the interval (x_1, x_2) such that $f'(z_2) = 0$. Applying the familiar form of Rolle's Theorem a third time, we can find a point $z \in (z_1, z_2)$ such that $f''(z) = 0$.

The general form of this theorem is proved by employing the familiar form of Rolle's Theorem multiple times. For example, if $n = 3$, then Rolle's Theorem will have to be cited $3 + 2 + 1 = 6$ times. □

Lagrange: We now use this general theorem to prove the error formula for polynomial interpolation. Note that the Mean Value Theorem is a special case of this theorem. Note also, the error term is identical with the error term for Taylor's Theorem if we allow all the points $x_0, x_1, x_2, \ldots, x_n$ to equal one another. Thus, in a very real sense, this theorem generalizes Taylor's Theorem. However, a different proof is required.

Theorem 18.4.2 (The Lagrange Error Formula for Interpolating Polynomials). *If $f(x) \in C^{n+1}[a,b]$, $a \leq x_0 < x_1 < x_2 < \cdots < x_n \leq b$, $p_n(x)$ is the unique polynomial of degree n such that $p_n(x_k) = f(x_k)$ for $k = 0, 1, 2, \ldots, n$, then for each $x \in [a,b]$, there exists a $z \in [a,b]$ such that*

$$f(x) = p_n(x) + \frac{f^{(n+1)}(z)}{(n+1)!}(x - x_0)(x - x_1)\ldots(x - x_n).$$

Proof. Let $x \in [a,b]$. Since the theorem is obviously true if $x = x_k$ for some k, we assume $x \neq x_k$ for all $k = 0, 1, \ldots, n$.

Let

$$G(t) = f(t) - p_n(t) - (f(x) - p_n(x)) \cdot \frac{w_n(t)}{w_n(x)},$$

where $w_n(t) = \prod_{k=0}^{n}(t - x_k)$.

1. $G(x) = 0$,
2. $G(x_k) = 0$ for $k = 0, 1, 2, \ldots, n$, and
3. $G^{(n+1)}(t) = f^{(n+1)}(t) - 0 - (f(x) - p_n(x)) \cdot \frac{(n+1)!}{w_n(x)}$.

Thus, we have shown that we have $n + 2$ distinct points $x, x_0, x_1, \ldots, x_n$ with the property that $G(x) = 0$.

By the Generalized Rolle's Theorem, there exists a $z \in [a,b]$ such that $G^{(n+1)}(z) = 0$ so that

$$0 = f^{(n+1)}(z) - (f(x) - p_n(x))\frac{(n+1)!}{w_n(x)}.$$

□

Lagrange: Note the similarity between the error formula for interpolating polynomials and Taylor's Theorem. You might find the next proposition useful in working the following exercises.

Proposition 18.4.3. *If $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ are equally spaced points, $h = \frac{b-a}{n}$, and $\omega_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$, then $|\omega_n(x)| \leq n!h^{n+1}$ for all $x \in [a,b]$.*

Proof. The graph of the 10-degree polynomial $\omega_{10}(x) = (x-1)(x-2)\dots(x-10)$ is displayed in 18.1. Note that the maximum (in absolute value) occurs between 1 and 2 and between 9 and 10. This fact is true in general. Thus, if $h = \frac{b-a}{n}$ and $x \in [x_0, x_1]$, then $|\omega_n(x)| \leq hh(2h)(3h)\dots(nh) = n!h^{n+1}$. □

Exercise Set 18.4.

1. Let $f(x) = \sin(\pi x)$ for $x \in [-1, 1]$. Let $p_n(x)$ be the Lagrange Interpolating polynomial for $f(x)$ using the evenly spaced points $-1 = x_0 < x_1 < x_2 < \cdots < x_n = 1$. Find an integer n with the property that $|p_n(x) - \sin(\pi x)| \leq 10^{-3}$ for all $x \in [-1, 1]$.

2. If $f(x) = e^x, x \in [-1, 1]$ and $tol = 10^{-7}$, then how many equally spaced points $-1 = x_0 < x_1 < x_2 < \cdots < x_n = 1$ must be computed to guarantee that the interpolating polynomial $p_n(x)$ will differ by less than tol from $f(x) = e^x$ for all $x \in [-1, 1]$.

3. If $f(x) = \ln(1-x), x \in [-\frac{1}{2}, \frac{1}{2}]$ and $tol = 10^{-7}$, then how many equally spaced points $-\frac{1}{2} = x_0 < x_1 < x_2 < \cdots < x_n = \frac{1}{2}$ must be computed to guarantee that the interpolating polynomial $p_n(x)$ will differ by less than tol from $f(x) = \ln(1-x)$ for all $x \in [-\frac{1}{2}, \frac{1}{2}]$.

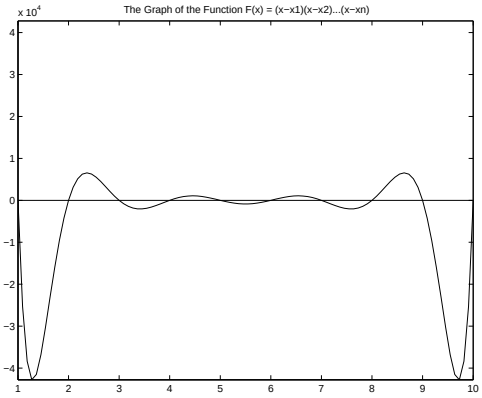


Figure 18.1: The Graph of the Function $\omega_{10}(x) = (x-1)(x-2)\dots(x-10)$

18.5 Polynomial Interpolation: The Runge Example

Simplicio: So we have completed our introduction to statistics. I am secure in my knowledge of these methods.

Galileo: Not so fast. We now introduce the German Mathematician, Carle Runge (1856-1927), who showed quite clearly why polynomials are a disaster. Professor Runge could you explain your classic example illustrating this problem?

Runge: While polynomial interpolation is easy to understand and straightforward to implement, it is dangerously unstable for uniformly spaced data sets containing as few as 20 points. If we define the function $f(x) = \frac{1}{1+x^2}$ on the interval $[-\pi, \pi]$, and take the points $-\pi = x_0 < x_1 < \cdots < x_n = \pi$ to be uniformly spaced points, then you would think that the approximation by the interpolating polynomials would get better and better as the degree of the polynomial n becomes larger.

Simplicio: That conclusion seems only reasonable since the error terms for the functions $f(x) = \sin(x)$ and $f(x) = e^x$ decrease rapidly as n is increased.

Runge: The bad news is that many situations exist where this desirable property fails to hold. For example, let us consider the function $f(x) = \frac{1}{1+x^2}$ be defined on the interval $[-\pi, \pi]$. The graph of this function is displayed in Figure 18.2.

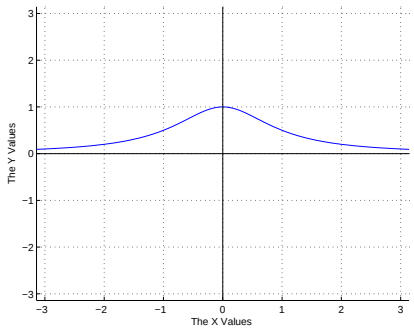


Figure 18.2: The Graph of the Function $f(x) = \frac{1}{1+x^2}$ for $x \in [-\pi, \pi]$

Runge: Now let $x_k = -\pi + k\frac{2\pi}{n}$ be equally spaced points between $-\pi$ and π and define points (x_k, y_k), where $y_k = f(x_k) = \frac{1}{1+x_k^2}$ for $k = 0, 1, \ldots, n$. The next step is to approximate $f(x)$ by polynomials interpolating the points (x_k, y_k). To illustrate what happens, in Figures 18.3,

we graph the $6, 20$, and 22 degree polynomial interpolants along with the function $f(x)$. Even though the approximations are accurate in the middle of the interval, the approximations at the endpoints become worse and worse. In fact, the difference between the polynomials $p_n(x)$ and the function $f(x)$ converges to infinity.

Simplicio: These graphs are disturbing. Even I can understand that if the data is

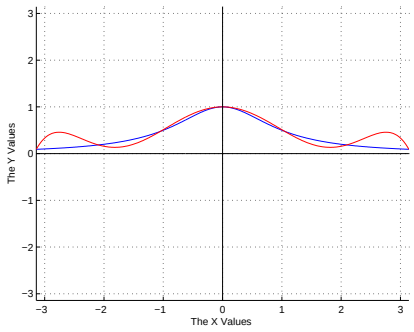


Figure 18.3: The 6^{th} Degree Polynomial Approximation of $f(x) = \frac{1}{1+x^2}$

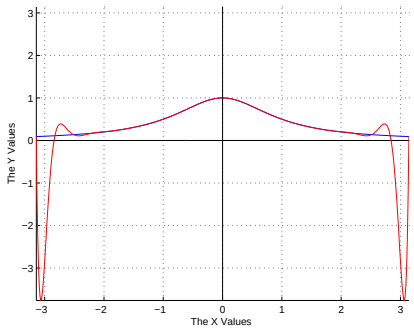


Figure 18.4: The 20^{th} Degree Polynomial Approximation of $f(x) = \frac{1}{1+x^2}$

as smooth as those supplied by the function $f(x) = \frac{1}{1+x^2}$, then the approximations should improve. Why are the results so terrible?

Galileo: For insight into the cause, think about the Pythagorean Theorem and the example we discussed many days ago, where the lines were almost parallel.

Simplicio: I am not sure what example you mean.

Galileo: Recall the system of two equations and two unknowns and its slight modification:

System 1:

$$\begin{aligned} 1.001x + y &= 2.001 \\ x + y &= 2 \end{aligned}$$

Note that these equations are close to being parallel. Solving the system we find $x = 1, y = 1$.

System 2:

$$\begin{aligned} 1.001x + y &= 2 \\ x + y &= 2 \end{aligned}$$

The solution to this system of equations is $x = 0, y = 2$.

Simplicio: Now I remember.

Galileo: When a mathematical method is unstable, it is often the case that a set of

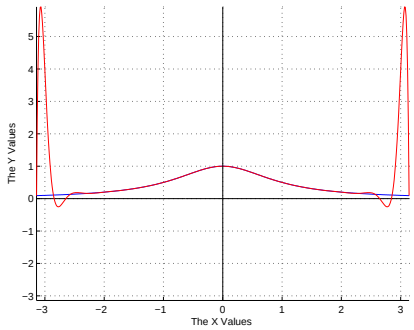


Figure 18.5: The 22^{th} Degree Polynomial Approximation of $f(x) = \frac{1}{1+25x^2}$

basis vectors are close to parallel. Note the computations in the following example and you might attain a better understanding of the problem.

Example 18.5.1. *If*

$$V = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix},$$

then let's compute the angles between the column vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.

If θ_{12} *represents the angle between the first two column vectors, then*

$$\cos(\theta_{12}) = \frac{< \mathbf{v}_1, \mathbf{v}_2 >}{\|\mathbf{v}_1\| \, \|\mathbf{v}_2\|} = \frac{1+2+3}{\sqrt{3}\sqrt{1+4+9}} = \frac{6}{\sqrt{3}\sqrt{14}} = 0.93.$$

Thus, the angle $\theta_{12} = 22.2$ *degrees.*

If θ_{13} *represents the angle between the first and third column vectors, then*

$$\cos(\theta_{13}) = \frac{< \mathbf{v}_1, \mathbf{v}_3 >}{\|\mathbf{v}_1\| \, \|\mathbf{v}_3\|} = \frac{1+4+9}{\sqrt{3}\sqrt{1+16+81}} = \frac{14}{\sqrt{3}\sqrt{98}} = 0.82.$$

Thus, the angle $\theta_{13} = 35.3$ *degrees.*

If θ_{23} *represents the angle between the second and third column vectors, then*

$$\cos(\theta_{23}) = \frac{< \mathbf{v}_2, \mathbf{v}_3 >}{\|\mathbf{v}_2\| \, \|\mathbf{v}_3\|} = \frac{1+8+27}{\sqrt{1+4+9}\sqrt{1+16+81}} = \frac{36}{\sqrt{14}\sqrt{98}} = 0.97.$$

Thus, the angle $\theta_{23} = 13.6$ *degrees.*

Simplicio: While the angle between the first and second columns are over 22 degrees, the angle between the 2^{nd} and 3^{rd} is about 13 degrees so they are almost parallel. I now can see that this computation means that the solution of a polynomial interpolation problem has the potential to have very poor results. It looks like the last two columns are the most parallel. Is that true in general?

Galileo: Make some more computations and see for yourself.

Simplicio: Hmmm.

Galileo: I think you are beginning to understand why our friend Fourier searched for a better method.

Exercise Set 18.5.

1. Form the 5×5 Vandermonde matrix generated by the vector $\mathbf{v} = [1, 2, 3, 4, 5]$. Compute the angles between the first and second, first and fourth, and fourth and fifth columns. Which pair of vectors is closest to being parallel?

2. Compute the angle between the vectors $\mathbf{v}_1 = [1.001, 1]$ and $\mathbf{v}_2 = [1, 1]$.

18.6 Linear Least Squares Approximation

Galileo: We begin our discussion of linear least squares with the problem of finding a straight line through three given data points $(x_0, y_0), (x_1, y_1), (x_2, y_2)$. If we try to "solve" this problem, we are in the position of trying to solve a linear system of 3 equations and 2 unknowns. In particular, if the equation of the line is in the form $y = a_0 + a_1 x$, then we have to find constants a_0 and a_1 such that $y_k = y(x_k) = a_0 + a_1 x_k$, for $k = 0, 1, 2$. Thus, we have to solve the matrix equation given by:

$$\begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

Simplicio: But how can you solve a system of 3 equations and 2 unknowns?

Galileo: Yes, a problem does exist with having more equations (or constraints) than unknowns. While a solution may exist, it is unlikely. In fact, the probability is zero.

Despite this problem, note that the system can be written $\mathbf{Aa} = \mathbf{y}$, where $\mathbf{A} = \mathbf{V^t}$ and $\mathbf{V}$ is a Vandermonde matrix.

Since this task is usually impossible, we are in the position of identifying the line of the form $y = a_0 + a_1 x$ with the property that the sum of the squares of the distances from each point to the corresponding point on the line is minimized. In particular, if

$d_k = a_0 + a_1 x_k - y_k$, then we need to minimize the residual:

$$\begin{aligned} R = &\ R(a_0, a_1) \\ = &\ d_0^2 + d_1^2 + d_2^2 \\ = &\ (a_0 + a_1 x_0 - y_0)^2 + (a_0 + a_1 x_1 - y_1)^2 + (a_0 + a_1 x_2 - y_2)^2. \end{aligned}$$

Since the function R is minimized at a critical point, we compute the derivatives:

$$\begin{aligned} \frac{\partial R}{\partial a_0} &= 2(a_0 + a_1 x_0 - y_0) + 2(a_0 + a_1 x_1 - y_1) + 2(a_0 + a_1 x_2 - y_2) &= 0 \\ \frac{\partial R}{\partial a_1} &= 2(a_0 + a_1 x_0 - y_0)x_0 + 2(a_0 + a_1 x_1 - y_1)x_1 + 2(a_0 + a_1 x_2 - y_2)x_2 &= 0, \end{aligned}$$

which leads to the 2×2 matrix equation:

$$\begin{pmatrix} 3 & x_0 + x_1 + x_2 \\ x_0 + x_1 + x_2 & x_0^2 + x_1^2 + x_2^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} y_0 + y_1 + y_2 \\ x_0 y_0 + x_1 y_1 + x_2 y_2 \end{pmatrix}.$$

Galileo: Note that this matrix can be easily solved to find the line that "best fits" the data. Note that if the matrix equation $\mathbf{V^t a} = \mathbf{y}$ is multiplied on both sides by the matrix

$$\mathbf{V} = \begin{pmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \end{pmatrix},$$

then the resulting 2×2 system is exactly the same as the 2×2 system discovered by computing partial derivatives and searching for a critical point.

Simplicio: But what if we are presented an arbitrary numer of points? How does the discussion change?

Galileo: Simply add up more terms. In other words, if we would like to fit a straight line through the points $(x_0, y_0), (x_1, y_1), \ldots (x_{n-1}, y_{n-1}), (x_n, y_n)$, where $x_0 < x_1 < \cdots < x_{n-1} < x_n$, then the matrix equation becomes:

$$\begin{pmatrix} n+1 & \sum_{k=0}^n x_k \\ \sum_{k=0}^n x_k & \sum_{k=0}^n x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^n y_k \\ \sum_{k=0}^n x_k y_k \end{pmatrix}.$$

Simplicio: This formula looks like nothing but a fancy way of averaging numbers to me.

Galileo: Why is that?

Simplicio: Well if the parameter a_1 happens to be zero, then a_0 is simply the average of the $y-$values.

Galileo: That is correct. We now repeat this discussion to conduct a search for a quadratic polynomial $p_2(x) = a_0 + a_1 x + a_2 x^2$, which "best fits" a given data set of 4 points

$(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3)$. For an exact fit of the data we would have to be able to solve the system of 4 equations and 3 unknowns given by:

$$\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

Again, since this system cannot be solved, we minimize the residual:

$$\begin{aligned} R = & \ R(a_0, a_1, a_2) \\ = & \ d_0^2 + d_1^2 + d_2^2 + d_3^2 \\ = & \ \sum_{k=0}^{3} (a_0 + a_1 x_k + a_2 x_k^2 - y_k)^2. \end{aligned}$$

The critical point where the minimum value of R occurs can be found by computing the partial derivatives of R with respect to the variables a_0, a_1, a_2.

$$\begin{aligned} \frac{\partial R}{\partial a_0} &= 2 \sum_{k=0}^{3} (a_0 + a_1 x_k + a_2 x_k^2 - y_k) &&= 0, \\ \frac{\partial R}{\partial a_1} &= 2 \sum_{k=0}^{3} (a_0 + a_1 x_k + a_2 x_k^2 - y_k) x_k &&= 0, \\ \frac{\partial R}{\partial a_2} &= 2 \sum_{k=0}^{3} (a_0 + a_1 x_k + a_2 x_k^2 - y_k) x_k^2 &&= 0. \end{aligned}$$

The resulting system of 3 equations and 3 unknowns is:

$$\begin{pmatrix} 4 & \sum_{k=0}^{3} x_k & \sum_{k=0}^{3} x_k^2 \\ \sum_{k=0}^{3} x_k & \sum_{k=0}^{3} x_k^2 & \sum_{k=0}^{3} x_k^3 \\ \sum_{k=0}^{3} x_k^2 & \sum_{k=0}^{3} x_k^3 & \sum_{k=0}^{3} x_k^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^{3} y_k \\ \sum_{k=0}^{3} x_k y_k \\ \sum_{k=0}^{3} x_k^2 y_k \end{pmatrix}.$$

If $x_0 < x_1 < x_2 < \cdots < x_m$ are $m+1$ distinct points and $y_0, y_1, \ldots, y_m$ are any $m+1$ values, then the polynomial $p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$ with the property that $p(x_i) = y_i$ can be found when $m = n$ by solving the equation $\mathbf{V}^t\mathbf{x} = \mathbf{b}$, where

$$\mathbf{V} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_0^n & x_1^n & x_2^n & \ldots & x_m^n \end{pmatrix},$$

$$\mathbf{x} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}, \text{ and } \mathbf{b} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

If $m > n$, then $m + 1 > n + 1$, which implies the "solution" of this equation will be in the least squares sense. In particular, the equation

$$(\mathbf{V}^t)^t\mathbf{V}^t\mathbf{x} = (\mathbf{V}^t)^t\mathbf{b} \text{ or}$$

$$\mathbf{V}\mathbf{V}^t\mathbf{x} = \mathbf{V}\mathbf{b}$$

must be solved. The matrix $\mathbf{V}\mathbf{V}^t$ is $(n+1) \times (n+1)-$dimensional. Since $\det \mathbf{V} = \prod_{i<j}(x_j - x_i)$, the rank of the matrix $(\mathbf{V}) = n + 1$ whenever the points x_k are distinct. Since $\mathbf{V}\mathbf{V}^t$ has rank $n + 1$, it can be shown that the matrix $\mathbf{V}\mathbf{V}^t$ is invertible, which implies that the least squares problem always has a unique solution.

If $\mathbf{A} \in R^{m\times n}$ and $\mathbf{x}, \mathbf{b} \in R^n$, then the general linear least squares problem is to "solve" the matrix equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ even if $m > n$. In this formulation we would like to find the vector $\mathbf{x}$ which minimizes the function $\mathbf{r}(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$. In the case

when the rank $(\mathbf{A}) = n$, the solution to this problem can be obtained by computing the gradient of $\mathbf{r}(\mathbf{x})$ with respect to the variables x_k. It is an easy exercise to show that the unique critical point of this function is the solution of the equation:

$$\mathbf{A}^t\mathbf{A}\mathbf{x} = \mathbf{A}^t\mathbf{b}.$$

If $\mathbf{A}$ has rank $= n$, then the matrix $\mathbf{A}^t\mathbf{A}$ is symmetric and positive definite. In particular, it is invertible. If the Cholesky factorization is used to solve the linear system of equations $\mathbf{A}^t\mathbf{A}\mathbf{x} = \mathbf{A}^t\mathbf{b}$, then this technique for solving this least squares problem is referred to as the method of "normal equations".

Exercise Set 18.6.

1. Find the equation of the line $y = p_1(x) = a_0 + a_1x$, which provides the least squares best fit for the data $(1, 2), (2, 3), (3, 5)$.

2. Find the parabola $y = p_2(x) = a_0 + a_1x + a_2x^2$, which provides the least squares best fit for the data $(1, 2), (2, 3), (3, 5), (4, -1)$.

18.7 Linear Classifiers

Galileo: One of the consequences of the recent proliferation of technology and computers is the incredible amount of data that is generated daily. In fact, the data is generated so rapidly that it is impossible to analyze and interpret without numerical techniques. One of the most important areas of study in statistics is the development of automated techniques that classify into two or more groups. For example, the people in the military would like to be able to reliably differentiate between a school bus and a tank, while a physician would like assistance in automated diagnosis. In many approaches, you would like to train your technique with data, where you already know the answers. The training process often involves the estimation of parameters for some function or distribution, which can be used to classify a new data set. If

your method provides reasonable answers over a wide range of data, then it can be considered a success. If not, then it will be ignored.

Simplicio: How do we start?

Galileo: As usual, let's start small. For example, if we would like to classify a set of points (x_k, y_k), for $k = 0, 1, 2, \ldots, n$ into two categories, then a linear classifier can be formulated as a least squares fit to the data $(x_k, y_k, 1)$ for one group of the set and $(x_k, y_k, -1)$ for the other group. In other words, find the parameters α_0, α_1 and α_2, which "solve" the system

$$\begin{aligned}
\alpha_0 + \alpha_1 x_0 + \alpha_2 y_0 &= z_0 \\
\alpha_0 + \alpha_1 x_1 + \alpha_2 y_1 &= z_1 \\
\alpha_0 + \alpha_1 x_2 + \alpha_2 y_2 &= z_2 \\
\vdots \qquad\qquad\quad &\vdots \\
\alpha_0 + \alpha_1 x_n + \alpha_2 y_n &= z_n,
\end{aligned}$$

where $z_k = -1$ or 1. Does this setting look familiar?

Simplicio: It certainly does. But, what if you have four or five categories? It seems to me you have several different ways to make the computations.

Galileo: True, but let us just keep it simple and consider only two categories. Let us now compute an example, where one set of points in the plane is defined by $S_1 = \{(0,0), (1,0), (0,1), (1,1)\}$ and a second is defined by $S_2 = \{(-1,-1)\}$. We then find a linear function $z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ with the property that the line $0 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ separates S_1 and S_2. To this end simply create the linear system:

$$\begin{aligned}
\alpha_0 + \alpha_1 0 + \alpha_2 0 &= 1 \\
\alpha_0 + \alpha_1 1 + \alpha_2 0 &= 1 \\
\alpha_0 + \alpha_1 0 + \alpha_2 1 &= 1 \\
\alpha_0 + \alpha_1 1 + \alpha_2 1 &= 1 \\
\alpha_0 + \alpha_1(-1) + \alpha_2(-1) &= -1,
\end{aligned}$$

The matrix equation becomes:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}.$$

The transpose of the coefficient matrix is:

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

Thus, multiplying both sides of the matrix equation by the transpose, we get

$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}.$$

The linear function becomes $z = 0.3913 + 0.5217x + 0.5217y$. The line separating the two sets is: $0.0 = 0.3913 + 0.5217x + 0.5217y$ or $y = -x - 0.75$.

Simplicio: While I notice that this line is perpendicular to the line through the midpoints of the two sets, I would have thought it would be the perpendicular bisector. What is going on?

Galileo: Since the set S_1 has a higher variance (or standard deviation) than the set S_2, the line is shifted closer to the set S_2, which makes sense from a geometrical point of view.

Simplicio: Even I can understand that idea. For if one set has a small variance and the other has a low variance, place the line closer to the set with low variance.

Galileo: Exactly.

Exercise Set 18.7.

1. Given two data sets $S = \{(-1, 1), (-1, -1), (0, 0)\}$ and $T = \{(1, 1), (1, -1)\}$, find a line L of the form $\alpha_0 + \alpha_1 x + \alpha_2 y = 0$ with the property that L separates the set S from T.

2. Given two data sets $S = \{(1, 1), (0, 0)\}$ and $T = \{(1, 0), (0, 1)\}$, find a line L of the form $\alpha_0 + \alpha_1 x + \alpha_2 y = 0$ with the property that L separates the set S from T.

Chapter 19

Fourier Interpolation

> Jean Baptiste Joseph Fourier: "The differential equations of the propagation of heat express the most general conditions, and reduce the physical questions to problems of pure analysis, and this is the proper object of theory." Analytical Theory of Heat

Galileo: We now turn to the problem of interpolation by trigonometric series of the form $T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)]$. While this type of series is typically referred to as a Fourier series, after the French mathematician Jean Baptiste Joseph Fourier (1768-1830), others had computed these series many years before. In particular, Euler had used one such series to show such identities as $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. While these identities are interesting and curious to mathematicians, it was Fourier who

showed their usefulness in modeling heat flow through a medium.

Let us ask him how he formed his insights.

Fourier: I joined Napoleon's army when he invaded Egypt in 1798. While we enjoyed warm weather and great success for a while, Lord Nelson destroyed the French fleet in the Battle of the Nile on August 1, 1798. Since this event brought an end to the sun and fun, I returned to Grenoble, where I was forced to endure cold, dreary winters with freezing temperatures. In an effort to deal with this state of affairs, I began an investigation of the heat equation. In 1807, I completed the memoir "On the Propagation of Heat in Solid Bodies," where I presented these ideas in manuscript form. This work was presented to the Paris Institute on 21 December 1807 and reviewed by a committee consisting of Lagrange, Laplace, Monge, and Lacroix. The members of this committee were unhappy with the work because of unresolved questions concerning the expansions of functions as trigonometric series. My colleague, Biot, was also unhappy because he felt he should have been referenced for the work he did on this topic in 1804. While I found this review unfair, I could do nothing about it. In 1811, I submitted an extension of this work to a second competition and actually won the prize.

Simplicio: That sounds great!

Fourier: Well, there was only one other entry. Worse yet, the report of the committee (whose members were Lagrange, Laplace, Malus, Hauy, and Legendre) was not completely favorable since it objected to the lack of rigor in the treatment of the mathematics. My paper, "Theorie analytique de la chaleur," was finally published in 1822. Even then Biot continued to claim priority.

Simplicio: Well, don't take it so hard. Your ideas are appreciated.

Galileo: But it did take 100 years to get all the mathematical issues sorted out with his series.

Simplicio: Which issues?

Galileo: Since Linear Algebra had not yet been invented, such ideas as linear independence, basis, inner product, and orthogonality had not yet been formulated. Since

the definition of limit had not yet been invented, the understanding of convergence was also murky. Eventually, the mathematicians parsed convergence into a number of different types including: uniform convergence, pointwise convergence, and convergence in the mean. Trigonometric series live best in Hilbert Space, where Pythagoras rules.

Simplicio: I am confused about this convergence concept.

Galileo: If you remember the error formulas for Taylor series and polynomial interpolation, they can be used to check for uniform and pointwise convergence.

Simplicio: How so?

Galileo: If you recall the error formula for Taylor is

$$E_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x - x_0)^{n+1}$$

while the error formula for polynomial interpolation is

$$E_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x - x_0)(x - x_1)\dots(x - x_n).$$

In the problems you were assigned, you were given and $\epsilon > 0$ and then were expected to find an integer n with the property that $|E_n(x)| < \epsilon$ for all $x \in [a, b]$. When you solve this kind of problem, you are showing that the sequence of polynomials are converging uniformly to the given function $f(x)$.

Simplicio: So what is pointwise convergence?

Virginia: Let me guess. Pointwise convergence is when you begin the problem by restricting your attention to a particular point x.

Galileo: Correct.

Virginia: So with the Runge example $f(x) = \frac{1}{1+x^2}, x \in [-\pi, \pi]$, we have pointwise convergence for any particular choice of x, but we do not have uniform convergence because the approximations fly off to infinity near the boundaries of the interval $[-\pi, \pi]$. In other words, For a given $\epsilon > 0$ (such as $\epsilon = 0.00001$), we cannot find an integer n which works for all x in the interval.

Galileo: For pointwise convergence, each point x requires its own individualized integer n.

Simplicio: What is convergence in the mean?

Galileo: While we haven't yet observed this type of convergence, it is a distance metric based on the integral of the difference of two functions. In particular, $d(f(x), g(x)) = \sqrt{\int_a^b (f(x) - g(x))^2 \, dx}$. While this metric is tuned to work well with Fourier series, the bad news is that it is weaker than uniform convergence. Surprisingly, this distance formula is closely connected with Pythagoras.

Simplicio: Life would be easier if we only had one type of convergence.

Galileo: Sorry, Mother Nature won't allow it. In fact, She insists we consider them all.

19.1 Fourier Interpolation: Introductory Examples

Simplicio: How about if you give me the short lecture on trigonometric series.

Fourier: Since you probably never appreciated partial differential equations, let us begin by solving the following simple interpolation question:

Given three points y_0, y_1, and y_2 and three angles x_0, x_1, and x_2, find a trigonometric polynomial of the form $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ with the property that $T_1(x_0) = y_0, T_1(x_1) = y_1$, and $T_1(x_2) = y_2$.

Simplicio: The answer to that problem is easy. All you have to do is solve the system of three equations and three unknowns.

$$\begin{pmatrix} 1 & \cos(x_0) & \sin(x_0) \\ 1 & \cos(x_1) & \sin(x_1) \\ 1 & \cos(x_2) & \sin(x_2) \end{pmatrix} \begin{pmatrix} \frac{a_0}{2} \\ a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

But when do you ever have to consider data in the form (x_k, y_k), where x_k is an angle?

Fourier: If you are modeling the temperature of a metal rod of length L, then the point x_k can be used to represent the position along the rod. If the points x_0, x_1, x_2

are equally spaced, then $x_0 = 0, x_1 = \frac{L}{2}$, and $x_2 = L$. In general, if we have $n+1$ equally spaced points $0 = x_0 < x_1 < \cdots < x_n = L$, then we can define the points by $x_k = \frac{kL}{n}$ for $n = 0, 1, \ldots, n$.

Simplicio: But these points don't represent angles!

Fourier: No problem, we will simply replace each x_k by the angle $\frac{2k\pi x_k}{L}$. Note that these angles vary between 0 and 2π. We will get to that aspect of the heat equation, but let's keep it simple for the moment and restrict our attention to the question about interpolation.

Simplicio: How about the equally spaced angles $x_0 = 0, x_1 = \pi$, and $x_2 = 2\pi$?

Fourier: Well, the idea is right, but $\cos(0) = \cos(2\pi)$ and $\sin(0) = \sin(2\pi)$ so the first and third row of the coefficient matrix are the same.

Virginia: Thus, the determinant of the matrix is zero, which implies the solution may not exist. We may not be able to solve for the constants a_0, a_1, and b_1.

Fourier: A better choice is $x_0 = 0, x_1 = 2\pi/3$, and $x_2 = \frac{4\pi}{3}$, which leads to the matrix:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 1 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix}.$$

Do you notice anything special about this matrix?

Simplicio: I can't say that I do.

Fourier: If you take a careful look at the three columns of this matrix, you will notice that they are pairwise perpendicular.

Simplicio: You mean check if the dot product of any two of these columns are zero?

Fourier: Precisely.

Simplicio: But, why should I care about this detail? Why would anyone care?

Fourier: If you multiply the matrix A by its transpose $\mathbf{A}$, you get the following product:

$$\mathbf{A}^t\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 1 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & \frac{3}{2} \end{pmatrix}.$$

Simplicio: I think I am beginning to understand. It seems like the product $\mathbf{A}^t\mathbf{A}$ is a diagonal matrix. Is that always true?

Fourier: If you make a smart choice of angles, the columns of the matrix will be perpendicular. However, before we consider that question, let us solve for the three constants a_0, a_1, and b_1.

Simplicio: No problem, the answers are the solutions to the matrix equation $\mathbf{A}^t\mathbf{A}\mathbf{x} = \mathbf{A}^t\mathbf{y}$, where

$$\mathbf{x} = \begin{pmatrix} a_0 \\ a_1 \\ b_1 \end{pmatrix}$$

and

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

Dividing through by the constants on the diagonal, we uncover formulas for a_0, a_1, and b_1 :

$$\begin{aligned} a_0 &= \tfrac{2}{3}(y_0 + y_1 + y_2), \\ a_1 &= \tfrac{2}{3}(y_0\cos(x_0) + y_1\cos(x_1) + y_2\cos(x_2)), \\ b_1 &= \tfrac{2}{3}(y_0\sin(x_0) + y_1\sin(x_1) + y_2\sin(x_2)). \end{aligned}$$

Virginia: Thus, we can summarize this discussion by saying that while the matrix equation can be solved for a multitude of choices of x_0, x_1, x_2, a "smart" choice is $x_0 = 0, x_1 = \frac{2\pi}{3}, x_2 = 2\frac{2\pi}{3}$.

Galileo: Correct! With a clever choice of x_0, x_1, x_2, we can easily solve the matrix equation.

Simplicio: And we don't even need row operations!

Galileo: Correct again.

Simplicio: But wait a minute. Once we have the formulas for the coefficients a_0, a_1, b_1, then can't we simply throw away the matrices?

Galileo: You are thinking like an engineer. To implement the method all you need are the formulas. However, let us consider the question: Why are these Fourier still

used today?

Virginia: I bet it is because of Pythagoras.

Galileo: Correct again. The Fourier matrices avoid all the serious stability issues exhibited by the Runge example. While polynomial interpolation has serious stability issues, the Fourier methods are always stable. In fact, the columns of the Fourier matrix have a Linear Algebra version of the Pythagorean Theorem that simply doesn't exist for a general Vandermonde matrix.

Fourier: How about if we step through the process again with the number of points increased from three to five? If we choose the angles to be equally spaced, we again see that a workable choice is:

$$\begin{aligned} x_0 &= 0, \\ x_1 &= 1 * 2\pi/5, \\ x_2 &= 2 * 2\pi/5, \\ x_3 &= 3 * 2\pi/5, \\ x_4 &= 4 * 2\pi/5. \end{aligned}$$

To keep the determinant of the coefficient matrix from being equal to zero, note that the angle x_4 has been chosen to be different from 2π. Now, if we have been given five points y_0, y_1, y_2, y_3, and y_4, we are then expected to find five constants a_0, a_1, a_2, b_1, and b_2, with the property that the trigonometric polynomial

$$T_2(x) = \frac{a_0}{2} + a_1 \cos(x) + a_2 \cos(2x) + b_1 \sin(x) + b_2 \sin(2x)$$

has the property that $T_2(x_k) = y_k$ for all $k = 0, 1, 2, 3, 4$.

Simplicio: And the answer to this problem is going to be another one of those matrix equations?

Fourier: Yes, and this time the matrix equation becomes

$$\begin{pmatrix} 1 & \cos(x_0) & \cos(2x_0) & \sin(x_0) & \sin(2x_0) \\ 1 & \cos(x_1) & \cos(2x_1) & \sin(x_1) & \sin(2x_1) \\ 1 & \cos(x_2) & \cos(2x_2) & \sin(x_2) & \sin(2x_2) \\ 1 & \cos(x_3) & \cos(2x_3) & \sin(x_3) & \sin(2x_3) \\ 1 & \cos(x_4) & \cos(2x_4) & \sin(x_4) & \sin(2x_4) \end{pmatrix} \begin{pmatrix} \frac{a_0}{2} \\ a_1 \\ a_2 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}.$$

When we compute the matrix for the given angles, we get:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0.3090 & -0.8090 & 0.9511 & 0.5878 \\ 1 & -0.8090 & 0.3090 & 0.5878 & -0.9511 \\ 1 & -0.8090 & 0.3090 & -0.5878 & 0.9511 \\ 1 & 0.3090 & -0.8090 & -0.9511 & -0.5878 \end{pmatrix}.$$

What do you notice about the columns?

Simplicio: Once again, each entry in the first column equals 1.

Fourier: What else?

Simplicio: Since each entry in the first column equals 1, the dot product of the first column and any other column will equal the sum of the entries in that particular column. Since the sum of the entries in each of these 4 columns equals zero, the first column will be perpendicular to each of the other four. In general, it appears that any two columns are once again perpendicular.

Fourier: You have made an important and fundamental insight because we are again in the position to easily solve the matrix equation $\mathbf{Ax} = \mathbf{y}$. In particular, what happens when we multiply both sides of this equation by the transpose $\mathbf{A}^t$?

Simplicio: Of course we get the equation $\mathbf{A}^t\mathbf{Ax} = \mathbf{A}^t\mathbf{y}$.

Virginia: Since the matrix product $\mathbf{A}^t\mathbf{A}$ is always a diagonal matrix, it will be easy to solve as soon as we compute the diagonal entries.

Fourier: Simplicio, let's compute the product $\mathbf{A}^t\mathbf{A}$.

Simplicio: No problem, the answer is the matrix:

$$\mathbf{A}^t\mathbf{A} = \begin{pmatrix} 5.0000 & -0.0000 & -0.0000 & 0 & 0 \\ -0.0000 & 2.4999 & 0.0001 & 0 & -0.0000 \\ -0.0000 & 0.0001 & 2.4999 & 0 & -0.0000 \\ 0 & 0 & 0 & 2.5002 & 0 \\ 0 & -0.0000 & -0.0000 & 0 & 2.5002 \end{pmatrix}.$$

It looks like a diagonal matrix except for two off-diagonal entries which equal 0.0001.

Fourier: If we had kept a few more digits of precision, those terms would have disappeared when we computed the matrix $\mathbf{A}$..

Virginia: In fact, it looks like the off-diagonal entries should equal 0.0000 and all the diagonal entries other than the first should equal 5/2.

Fourier: Those thoughts are correct. The next concern is to point out that these ideas are completely general. For this benefit we need to expend a bit of effort to organize the necessary mathematical facts that will make the ideas precise.

Virginia: I am beginning to wonder about all that Gaussian elimination stuff we learned in Linear Algebra. This Fourier approach is so much easier. No row operations required.

Exercise Set 19.1.

1. Given data y_0, y_1, y_2, y_3, y_4 compute the values for the coefficients a_0, a_1, a_2, b_1, b_2.

2. Given data $y_0 = 1, y_1 = 2, y_2 = 3, y_3 = 4, y_4 = 5$ compute the values for the coefficients a_0, a_1, a_2, b_1, b_2. Check that the function $T_2(x)$ actually interpolates the data by showing that $T_2(x_2) = y_2 = 3$.

19.2 Fourier Interpolation: Coefficient Formulas

Fourier: The next goal is to show that the interpolation technique we have just discussed for 3 data points and five data points can be generalized to any odd number of points. In particular, we will discuss the general case when we are given $2n+1$ data

points (x_k, y_k) for $k = 0, 1, 2 \ldots, 2n$. In this setting, our coefficient matrix will have $2n+1$ rows and $2n+1$ columns. The first proposition states that any two columns of the coefficient matrix $\mathbf{A}$ will always be perpendicular. As before this will imply that the matrix product $\mathbf{A}^t\mathbf{A}$ will be a diagonal matrix. The second proposition states that the diagonal entries of the matrix $\mathbf{A}^t\mathbf{A}$ are $\frac{2n+1}{2}$.

Virginia: Except, of course, for the first entry which is $2n + 1$. Right?

Fourier: Correct! Finally, the third proposition presents formulas for the coefficients $a_0, a_1, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$. The formulas follow easily from these key properties of $\mathbf{A}$ and $\mathbf{A}^t\mathbf{A}$.

Fourier: Since the first proposition is written in mathematically technical language with five different summations (all equal to zero), we will begin by discussing the implications of each part. In particular, we make the following observations:

1. The summation $\sum_{k=0}^{2n} \cos(mx_k) = 0$ implies the inner product (i.e. dot product) of the 1^{st} column and the m^{th} $\cos(x)$ column equals zero. Thus, the first column will always be perpendicular to any $\cos(x)$ column.

2. The summation $\sum_{k=0}^{2n} \sin(mx_k) = 0$ implies the 1^{st} column is perpendicular to the m^{th} $\sin(x)$ column. Thus, the first column will always be perpendicular to any $\sin(x)$ column.

3. The summation $\sum_{k=0}^{2n} \cos(jx_k) \cdot \sin(mx_k) = 0$ implies the j^{th} $\cos(x)$ column is perpendicular to the m^{th} $\sin(x)$ column.

4. The summation $\sum_{k=0}^{2n} \sin(jx_k) \cdot \sin(mx_k) = 0$ implies the j^{th} $\sin(x)$ column is perpendicular to the m^{th} $\sin(x)$ column.

5. The summation $\sum_{k=0}^{2n} \cos(jx_k) \cdot \cos(mx_k) = 0$ implies the j^{th} $\cos(x)$ column is perpendicular to the m^{th} $\cos(x)$ column.

Simplicio: But if $j = m$ in the last two remarks, then we are computing the dot product of a column with itself. That doesn't sound right to me.

Galileo: Your observation is correct. In the proposition, we will assume that $j \neq m$ to insure that the columns are actually different.

Virginia: And these 5 pieces of information imply that any two columns of the coefficient matrix are perpendicular. Isn't that right?

Galileo: Correct.

Proposition 19.2.1 (Orthogonality for Discrete Fourier). *If $0 < m, j \leq 2n$ are integers and $x_k = \frac{k}{2n+1}2\pi$ for $k = 0, 1, \ldots, 2n$, then the following statements hold:*

1. *If $m \leq 2n$, then $\sum_{k=0}^{2n} \cos(mx_k) = 0$.*
2. *If $m \leq 2n$, then $\sum_{k=0}^{2n} \sin(mx_k) = 0$.*
3. *If $j \leq n$, and $m \leq n$, then $\sum_{k=0}^{2n} \cos(jx_k) \cdot \sin(mx_k) = 0$.*
4. *If $j \leq n, m \leq n$, and $j \neq m$, then $\sum_{k=0}^{2n} \sin(jx_k) \cdot \sin(mx_k) = 0$.*
5. *If $j \leq n, m \leq n$, and $j \neq m$, then $\sum_{k=0}^{2n} \cos(jx_k) \cdot \cos(mx_k) = 0$.*

Proof. The underlying idea behind this proof is that the formula for the geometric series works just as well for complex numbers as it does for real numbers. In particular, if $z \neq 1$, then

$$\sum_{k=0}^{n} z^k = \frac{1 - z^{n+1}}{1 - z}.$$

The proof is exactly the same as before. All you need to know is that all the usual associative, commutative, and distributive rules apply.

However, we will also need Euler's formula, which can be stated as follows.

Lemma 19.2.2. *If $i = \sqrt{-1}$, then $e^{ix} = \cos(x) + i\sin(x)$.*

This formula can be proved by Taylor series, Calculus, or Differential Equations. If we consider the special case when $x = \pi$, then we have Euler's famous identity $e^{i\pi} = -1$. Note that this identity combines three remarkable constants e, π, i into

the familiar number -1. However, we should not get distracted from the business at hand.

If we let $z = e^{\frac{2\pi i}{2n+1}} = \cos(\frac{2\pi}{2n+1}) + i\sin(\frac{2\pi}{2n+1})$, then note that whenever $m \leq n$, then $(z^m)^{2n+1} = (z^{2n+1})^m = (e^{2\pi i})^m = 1^m = 1$.

By the geometric series formula we now observe that since $x_k = k\frac{2\pi}{2n+1} = \frac{2\pi k}{2n+1}$, $mx_k = \frac{2\pi km}{2n+1}$,

$$\begin{aligned} Z &= \sum_{k=0}^{2n} \cos(mx_k) + i\sum_{k=0}^{2n} \sin(mx_k) \\ &= \sum_{k=0}^{2n} e^{imx_k} = \sum_{k=0}^{2n} e^{i\frac{2\pi km}{2n+1}} = \sum_{k=0}^{2n} (e^{i\frac{2\pi}{2n+1}})^{mk} \\ &= \sum_{k=0}^{2n} z^{mk} = \sum_{k=0}^{2n} (z^m)^k = \frac{1-(z^m)^{2n+1}}{1-z^m} = \frac{1-(z^{2n+1})^m}{1-z^m} = \frac{1-1^m}{1-z^m} = 0. \end{aligned}$$

Thus, both the real and imaginary parts of Z equal zero. Since the real part of Z equals zero, $\sum_{k=0}^{2n} \cos(mx_k) = 0$. Since the imaginary part of Z equals zero, $\sum_{k=0}^{2n} \sin(mx_k) = 0$. Thus, the first two statements in the proposition are verified.

Virginia: But why did we assume that $m \leq n$?

Fourier: If m is an integer multiple of $2n+1$, then $z^m = 1$, which mean that we are dividing by zero and thus can't apply the geometric series.

The other three statements follow from the identities:

$$\begin{aligned} \cos(A)\sin(B) &= \frac{1}{2}[\sin(A+B) + \sin(B-A)], \\ \cos(A)\cos(B) &= \frac{1}{2}[\cos(A+B) + \cos(A-B)], \\ \sin(A)\sin(B) &= \frac{1}{2}[\cos(A-B) - \cos(A+B)]. \end{aligned}$$

In particular, if $A = jx_k$ and $B = mx_k$, then

$$\begin{aligned} \cos(jx_k)\sin(mx_k) &= \frac{1}{2}[\sin((j+m)x_k) + \sin((m-j)x_k)], \\ \cos(jx_k)\cos(mx_k) &= \frac{1}{2}[\cos((j+m)x_k) + \cos((j-m)x_k)], \\ \sin(jx_k)\sin(mx_k) &= \frac{1}{2}[\cos((j-m)x_k) - \cos((j+m)x_k)]. \end{aligned}$$

Thus, the third summation can be written as

$$\begin{aligned}\sum_{k=0}^{2n}\cos(jx_k)\cdot\sin(mx_k) &= \frac{1}{2}\sum_{k=0}^{2n}[\sin((j+m)x_k)+\sin((m-j)x_k)]\\ &= \frac{1}{2}\sum_{k=0}^{2n}[\sin((j+m)x_k)]+\frac{1}{2}\sum_{k=0}^{2n}[\sin((m-j)x_k)]\\ &= 0+0=0.\end{aligned}$$

Fourier: Note that this argument requires the fact that you have already proved statement 2 in the proposition and that the sums $j+m \leq 2n$ and $j-m \leq 2n$.

Statements 4 and 5 in the proposition follow from the same argument we just gave for statement 3.

□

Fourier: In the next proposition, we compute the values of the entries along the diagonal of the matrix product A^tA. This computation is equivalent to the computation of the norms (i.e lengths) of the columns of A. The first statement in the Equal Length Formulas can be used to show that the first diagonal entry of the matrix product will equal $2n+1$; the second two items can be used to show that all the other diagonal entries will equal $\frac{2n+1}{2}$.

Proposition 19.2.3 (Equal Length Formulas for Discrete Fourier). *If* $0 < m \leq n$ *are integers and* $x_k = \frac{k}{2n+1}2\pi$ *for* $k = 0, 1, \ldots, 2n$, *then*

1. $\sum_{k=0}^{2n} 1 = 2n+1$,
2. $\sum_{k=0}^{2n}\cos^2(mx_k) = \frac{2n+1}{2}$, *and*
3. $\sum_{k=0}^{2n}\sin^2(mx_k) = \frac{2n+1}{2}$.

Proof. The relations follow from the half angle formulas and the orthogonality proposition. Recall that the half angle formulas are:

1. $\cos^2(\theta) = \frac{1+cos(2\theta)}{2}$

2. $\sin^2(\theta) = \frac{1-cos(2\theta)}{2}$

Note that these formulas have the virtue that they can be used to reduce expressions of the form $\cos^2(\theta)$ and $\sin^2(\theta)$ to linear expressions.

In particular, the first half angle formula combined with statement 1 from the previous proposition can be used to make the following reduction:

$$\begin{aligned}\sum_{k=0}^{2n}\cos^2(mx_k) =& \sum_{k=0}^{2n}\frac{1+\cos(2mx_k)}{2}\\ =& \frac{1}{2}\sum_{k=0}^{2n}\{1+\cos(2mx_k)\}\\ =& \frac{1}{2}\sum_{k=0}^{2n}1+\frac{1}{2}\sum_{k=0}^{2n}\cos(2mx_k)\\ =& \frac{1}{2}\sum_{k=0}^{2n}1+\ 0\\ =& \frac{2n+1}{2}.\end{aligned}$$

Similarly, the second half angle formula can be used to prove the third equation in the proposition:

$$\begin{aligned}\sum_{k=0}^{2n}\sin^2(mx_k) =& \sum_{k=0}^{2n}\frac{1-\cos(2mx_k)}{2}\\ =& \frac{1}{2}\sum_{k=0}^{2n}\{1-\cos(2mx_k)\}\\ =& \frac{1}{2}\sum_{k=0}^{2n}1=\frac{2n+1}{2}.\end{aligned}$$

□

Fourier: We now present the key formulas for the coefficients for the trigonometric series. These formulas allow you to interpolate any given data set $y_0, y_1, y_2, \ldots, y_{2n}$ with a function of the form

$$T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}[a_k\cos kx + b_k\sin kx].$$

Note that the argument is the same strategy, where we solve a matrix equation $A\mathbf{a} = \mathbf{y}$ by multiplying both sides of the equation by A^t resulting in the equation $D\mathbf{a} = A^tA\mathbf{a} = A^t\mathbf{y}$, where D is a diagonal matrix.

Theorem 19.2.4 (Fourier Coefficients: Discrete Case). *If $x_k = \frac{k}{2n+1}2\pi$, for $k = 0, 1, \ldots, 2n$ and $y_0, y_1, y_2, \ldots, y_{2n}$ are $2n+1$ given data values, then constants a_k and b_k can be found so that the trigonometric polynomial*

$$T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}[a_k \cos kx + b_k \sin kx]$$

has the property that $T_n(x_k) = y_k$ for all $k = 0, 1, \ldots, 2n$.

In particular, the formulas are:

$$a_k = \frac{2}{2n+1}\sum_{j=0}^{2n} y_j \cos(kx_j) \text{ for } k = 0, 1, 2, \ldots, n,$$

and

$$b_k = \frac{2}{2n+1}\sum_{j=0}^{2n} y_j \sin(kx_j) \text{ for } k = 1, 2, \ldots, n.$$

Proof. Fourier: Since we require the property $T_n(x_k) = y_k$ for all $k = 0, 1, \ldots, 2n$, we must solve a $(2n+1) \times (2n+1)$ dimensional linear system of the form $A\mathbf{a} = \mathbf{y}$, where A is the coefficient matrix consisting of various sines and cosines, $\mathbf{a} = (a_0, a_1, \ldots, a_n, b_1, b_2, \ldots, b_n)$, and $y = (y_0, y_1, \ldots, y_{2n})$.

For the special case $n = 2$ the requirement that $T_2(x_k) = y_k$ leads to a system of 5 equations and 5 unknowns:

$$\begin{aligned}
\frac{a_0}{2} + a_1\cos(x_0) + a_2\cos(2x_0) + b_1\sin(x_0) + b_2\sin(2x_0) &= y_0\\
\frac{a_0}{2} + a_1\cos(x_1) + a_2\cos(2x_1) + b_1\sin(x_1) + b_2\sin(2x_1) &= y_1\\
\frac{a_0}{2} + a_1\cos(x_2) + a_2\cos(2x_2) + b_1\sin(x_2) + b_2\sin(2x_2) &= y_2\\
\frac{a_0}{2} + a_1\cos(x_3) + a_2\cos(2x_3) + b_1\sin(x_3) + b_2\sin(2x_3) &= y_3\\
\frac{a_0}{2} + a_1\cos(x_4) + a_2\cos(2x_4) + b_1\sin(x_4) + b_2\sin(2x_4) &= y_4.
\end{aligned}$$

Simplicio: So we can once again write a matrix equation $A\mathbf{a} = \mathbf{y}$, where $A, \mathbf{a}$, and $\mathbf{y}$ are the usual suspects.

Virginia: If we multiply both sides of the equation by the transpose A^t, the resulting equation is $A^tA\mathbf{a} = A^t\mathbf{y}$. Since any two columns of A are orthogonal, the matrix A^tA is diagonal.

Simplicio: Even I can see this is true by the Orthogonality Proposition.

Virginia: By the Equal Lengths Formulas, the first entry on the diagonal is $2n + 1$, while the remaining diagonal entries equal $\frac{2n+1}{2}$.

Simplicio: Thus, the coefficient formulas are simply the result of multiplying the vector $A^t\mathbf{y}$ by the inverse of the diagonal matrix $D = A^tA$. In symbols, $\mathbf{a} = (A^tA)^{-1}A^t\mathbf{y}$.

Fourier: You got it.

Simplicio: Actually, I rather liked that proof.

Fourier: Then how about a second proof?

Simplicio: Sorry, but one proof is plenty for me.

Fourier: Well, let's call it an observation then.

Virginia: Let's see it.

Fourier: What we are really doing here is searching for constants (given by the vector $\mathbf{a}$) which allow us to write the vector

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

as a linear combination of the sines and cosines. In particular, if we write the linear system $A\mathbf{a} = \mathbf{y}$ as a linear combination of the columns of the coefficient matrix A, we have

$$\frac{a_0}{2}\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}+a_1\begin{pmatrix}\cos(x_0)\\\cos(x_1)\\\cos(x_2)\\\cos(x_3)\\\cos(x_4)\end{pmatrix}+a_2\begin{pmatrix}\cos(2x_0)\\\cos(2x_1)\\\cos(2x_2)\\\cos(2x_3)\\\cos(2x_4)\end{pmatrix}+b_1\begin{pmatrix}\sin(x_0)\\\sin(x_1)\\\sin(x_2)\\\sin(x_3)\\\sin(x_4)\end{pmatrix}+b_2\begin{pmatrix}\sin(2x_0)\\\sin(2x_1)\\\sin(2x_2)\\\sin(2x_3)\\\sin(2x_4)\end{pmatrix}=\begin{pmatrix}y_0\\y_1\\y_2\\y_3\\y_4\end{pmatrix}.$$

If we let

$$\mathbf{u}_k=\begin{pmatrix}\cos(kx_0)\\\cos(kx_1)\\\cos(kx_2)\\\cos(kx_3)\\\cos(kx_4)\end{pmatrix}\quad\text{and}\quad\mathbf{v}_k=\begin{pmatrix}\sin(kx_0)\\\sin(kx_1)\\\sin(kx_2)\\\sin(kx_3)\\\sin(kx_4)\end{pmatrix},$$

then we can write the vector $\mathbf{y}$ as a linear combination

$$\mathbf{y}=\frac{a_0}{2}\mathbf{u}_0+a_1\mathbf{u}_1+a_2\mathbf{u}_2+b_1\mathbf{v}_1+b_2\mathbf{v}_2.$$

Virginia: I see what you are after. You have changed the basis for the vector space $\Re^5$ from the usual basis vectors $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\mathbf{e}_4,\mathbf{e}_5$, to a new basis $\mathbf{u}_0,\mathbf{u}_1,\mathbf{u}_2,\mathbf{v}_1,\mathbf{v}_2$. The constants $\frac{a_0}{2},a_1,a_2,b_1,b_2$ are simply the result of the usual change of basis formulas.

Fourier: Very good.

Simplicio: So.

Fourier: If we want to compute the coefficient a_m, we simply compute the inner product $<\mathbf{u}_m,\mathbf{y}>$. Since the orthogonality lemma for discrete Fourier shows that the columns are pairwise orthogonal, we know $<\mathbf{u}_m,\mathbf{u}_k>=0$ for all $k\neq m$. Since also know $<\mathbf{u}_m,\mathbf{v}_k>=0$ for all k,

$$<\mathbf{u}_m,\mathbf{y}>=\frac{a_0}{2}\mathbf{u}_0+a_1\mathbf{u}_1+a_2\mathbf{u}_2+b_1\mathbf{v}_1+b_2\mathbf{v}_2=a_m<\mathbf{u}_m,\mathbf{u}_m>=a_m\frac{5}{2}.$$

Virginia: Thus, we see immediately that

$$\begin{aligned}a_m=&\frac{2}{5}<\mathbf{u}_m,\mathbf{y}>\\=&\frac{2}{5}(y_0\cos(mx_0)+y_1\cos(mx_1)+y_2\cos(mx_2)+y_3\cos(mx_3)+y_4\cos(mx_4)).\end{aligned}$$

The same argument can be used to show $b_m = \frac{2}{5}\sum_{j=0}^{4} y_j \sin(mx_j)$. In both cases the Orthogonality and Equal Lengths properties are crucial.

Simplicio: But why would we want to go to this extra trouble?

Fourier: Because we will see this argument again in several other settings. First, we will see this exact same argument in the proof of the Pythagoras/Parseval formula for trigonometric series. Second, we will see it again when we discuss least squares for trigonometric series. Third, we will use this exact same argument for the continuous case, where we write a function $f(x) : [-\pi, \pi] \to \Re$ as an infinite series of the form

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos(kx) + b_k \sin(kx)).$$

In this last case, the inner product relevant to the discussion becomes an integral rather than a dot product.

Simplicio: But why would you want to do that?

Fourier: You know how to hurt a guy. This technique is exactly what I used to solve the heat equation.

□

Fourier: These thoughts can be summarized in the following proposition. Note that we have not even assumed that the matrix $\mathbf{A}$ is square. Note this proposition well. We will revisit it soon.

Proposition 19.2.5. *If $\mathbf{A} \in \Re^{m\times n}$ is a matrix with m rows and n columns, which has the property that every pair of column vectors are perpendicular, then the product $\mathbf{D} = \mathbf{A}^t\mathbf{A}$ is an $n \times n$ diagonal matrix. Moreover, the j^{th} diagonal entry of $\mathbf{D}$ is the square of the length of the j^{th} column of $\mathbf{A}$.*

Proof. The proof is simply the observation that the $(i, j)^{th}$ entry of $\mathbf{D}$ is the dot product of the i^{th} and j^{th} columns of $\mathbf{A}$. □

Simplicio: Even I can understand this one.

Example 19.2.1. *Problem: Given the data* $y_0 = 3, y_1 = 2, y_2 = 2$, *compute the Fourier coefficients* a_0, a_1, b_1.

By the coefficient formulas,

$$\begin{aligned} a_0 &= \frac{2}{3}(y_0 + y_1 + y_2) &&= \frac{2}{3}(3+2+2) &&= \frac{14}{3}. \\ a_1 &= \frac{2}{3}(y_0 - \frac{1}{2}\ y_1 - \frac{1}{2}\ y_2) &&= \frac{2}{3}(3-1-1) &&= \frac{2}{3}. \\ b_1 &= \frac{2}{3}(\frac{\sqrt{3}}{2}\ y_1 - \frac{\sqrt{3}}{2}\ y_2) &&= \frac{2}{3}(\frac{\sqrt{3}}{2}2 - \frac{\sqrt{3}}{2}2) &&= 0. \end{aligned}$$

Since $b_1 = 0$, $T_1(x) = \frac{7}{3} + \frac{2}{3}\cos(x)$. *In particular, the basis function* $\sin(x)$ *is unnecessary.*

Example 19.2.2. *Problem: Given the data* $y_0 = 0, y_1 = 2, y_2 = -2$, *compute the Fourier coefficients* a_0, a_1, b_1.

By the coefficient formulas,

$$\begin{aligned} a_0 &= \frac{2}{3}(y_0 + y_1 + y_2) &&= \frac{2}{3}(0+2-2) &&= 0. \\ a_1 &= \frac{2}{3}(y_0 - \frac{1}{2}\ y_1 - \frac{1}{2}\ y_2) &&= \frac{2}{3}(0-1+1) &&= 0. \\ b_1 &= \frac{2}{3}(\frac{\sqrt{3}}{2}\ y_1 - \frac{\sqrt{3}}{2}\ y_2) &&= \frac{2}{3}(\frac{\sqrt{3}}{2}2 + \frac{\sqrt{3}}{2}2) &&= 4\frac{\sqrt{3}}{3}. \end{aligned}$$

Since $a_0 = 0$ *and* $a_1 = 0$, $T_1(x) = 4\frac{\sqrt{3}}{3}\sin(x)$. *In particular, the basis functions* 1 *and* $\cos(x)$ *are unnecessary.*

Example 19.2.3. *Problem: Given the data* $y_0 = 0, y_1 = 2, y_2 = 3, y_3 = -3, y_4 = -2$, *compute the Fourier coefficients* a_0, a_1, a_2, b_1, b_2.

If

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{2\pi}{5} \\ 2\frac{2\pi}{5} \\ 3\frac{2\pi}{5} \\ 4\frac{2\pi}{5} \end{pmatrix} = \begin{pmatrix} 0 \\ 1.2566 \\ 2.5133 \\ 3.7699 \\ 5.0265 \end{pmatrix},$$

then

$$\cos(x) = \begin{pmatrix} \cos(x_0) \\ \cos(x_1) \\ \cos(x_2) \\ \cos(x_3) \\ \cos(x_4) \end{pmatrix} = \begin{pmatrix} 1.0000 \\ 0.3090 \\ -0.8090 \\ -0.8090 \\ 0.3090 \end{pmatrix}, \cos(2x) = \begin{pmatrix} \cos(2x_0) \\ \cos(2x_1) \\ \cos(2x_2) \\ \cos(2x_3) \\ \cos(2x_4) \end{pmatrix} = \begin{pmatrix} 1.0000 \\ -0.8090 \\ 0.3090 \\ 0.3090 \\ -0.8090 \end{pmatrix}$$

and

$$\sin(x) = \begin{pmatrix} \sin(x_0) \\ \sin(x_1) \\ \sin(x_2) \\ \sin(x_3) \\ \sin(x_4) \end{pmatrix} = \begin{pmatrix} 0.0000 \\ 0.9511 \\ 0.5878 \\ -0.5878 \\ -0.9511 \end{pmatrix}, \sin(2x) = \begin{pmatrix} \sin(2x_0) \\ \sin(2x_1) \\ \sin(2x_2) \\ \sin(2x_3) \\ \sin(2x_4) \end{pmatrix} = \begin{pmatrix} 0.0000 \\ 0.5878 \\ -0.9511 \\ 0.9511 \\ -0.5878 \end{pmatrix}.$$

Thus, by the coefficient formulas, we see that

$$\begin{aligned}
a_0 &= \frac{2}{5}(y_0 + y_1 + y_2 + y_3 + y_4) \\
&= \frac{2}{5}(0 + 2 + 3 - 3 - 2) = 0, \\
a_1 &= \frac{2}{5}(y_0 \cos(x_0) + y_1 \cos(x_1) + y_2 \cos(x_2) + y_3 \cos(x_3) + y_4 \cos(x_4)) \\
&= \frac{2}{5}(0 + 2 * 0.3090 - 3 * 0.8090 + 3 * 0.8090 - 2 * 0.3090) = 0, \\
a_2 &= \frac{2}{5}(y_0 \cos(2x_0) + y_1 \cos(2x_1) + y_2 \cos(2x_2) + y_3 \cos(2x_3) + y_4 \cos(2x_4)) \\
&= \frac{2}{5}(0 - 2 * 0.8090 + 3 * 0.3090 - 3 * 0.3090 + 2 * 0.8090) = 0, \\
b_1 &= \frac{2}{5}(y_0 \sin(x_0) + y_1 \sin(x_1) + y_2 \sin(x_2) + y_3 \sin(x_3) + y_4 \sin(x_4)) \\
&= \frac{2}{5}(0 + 2 * 0.9511 + 3 * 0.5878 + 3 * 0.5878 + 2 * 0.9511) = 2.9324, \\
b_2 &= \frac{2}{5}(y_0 \sin(2x_0) + y_1 \sin(2x_1) + y_2 \sin(2x_2) + y_3 \sin(2x_3) + y_4 \sin(2x_4)) \\
= &= \frac{2}{5}(0 + 2 * 0.5878 - 3 * 0.9511 - 3 * 0.9511 + 2 * 0.5878) = -1.3422.
\end{aligned}$$

Thus, $a_0 = a_1 = a_2 = 0$, *which implies that* $T_1(x) = 2.9324 \sin(x) - 1.3422 \sin(2x)$.

In particular, the basis functions $1, \cos(x)$, *and* $\cos(2x)$ *are unnecessary.*

Simplicio: I hate to be difficult, but I have just one quick question that has begun to eat at me.

Fourier: Sure.

Simplicio: In all the discussions you have given so far, you have always assumed that you are given $2n+1$ points in your vector $\mathbf{y}$. If $\mathbf{y} = (y_0, y_1, \ldots, y_{2n})$, then it has an odd number of coordinates. What do you do if you are given an even number of points?

Fourier: The short answer is that "It depends on the problem." The longer answer is that certain problems only require the $\cos(mx)$ functions and certain other problems only require the $\sin(mx)$ functions. For example, the heat equation only requires the $\sin(mx)$ functions and the JPEG compression techniques only require $\cos(mx)$ functions. With a compression problem, we will "double" the data from $n+1$ points to $2n+1$ points in such a way that the coefficients $b_k = 0$ for all k. Thus, we will always have an odd number of points and we will never need the basis functions $\sin(mx)$.

Simplicio: Interesting. In other words the basis functions of the form $\cos(mx)$ will suffice.

Fourier: Precisely. Here are a few exercises, which you should not find particularly challenging.

Exercise Set 19.2.

1. Show that the quantity $\frac{a_0}{2}$ always represents the average of the given data set $y_0, y_1, \ldots, y_{2n}$.

2. Given the data $y_0 = 1, y_1 = 2, y_2 = 3$, compute the Fourier coefficients a_0, a_1, b_1. Plot the data and the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ on the same graph.

3. Given the data $y_0 = \sin(0), y_1 = \sin(\frac{2\pi}{3}), y_2 = \sin(\frac{4\pi}{3})$ compute the Fourier coefficients a_0, a_1, b_1. Plot the data and the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ on the same graph.

4. Given the data $y_0 = \cos(0), y_1 = \cos(\frac{2\pi}{3}), y_2 = \cos(\frac{4\pi}{3})$ compute the Fourier coefficients a_0, a_1, b_1. Plot the data and the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ on the same graph.

5. Given the data $y_0 = 0, y_1 = 2, y_2 = -2$, compute the Fourier coefficients a_0, a_1, b_1. Which coefficients equal zero?

6. Given the data $y_0 = 3, y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 1$, compute the Fourier coefficients a_0, a_1, a_2, b_1, b_2. Which coefficients equal zero?

7. Given the data $y_0 = 0, y_1 = 1, y_2 = 2, y_3 = -2, y_4 = -1$, compute the Fourier coefficients a_0, a_1, a_2, b_1, b_2. Plot the data and the function $T_2(x) = \frac{a_0}{2} + a_1 \cos(x) + a_2 \cos(2x) + b_1 \sin(x) + b_2 \sin(2x)$ on the same graph. Which coefficients equal zero?

8. Given the data $y_0 = 0, y_1 = A, y_2 = -A$, where A is an arbitrary number, compute the Fourier coefficients a_0, a_1, b_1. Which coefficients equal zero?

9. Given the data $y_0 = A, y_1 = B, y_2 = B$, where A and B are arbitrary numbers. Compute the Fourier coefficients a_0, a_1, b_1. Which coefficients equal zero?

10. Given the data $y_0 = 0, y_1 = A, y_2 = B, y_3 = B, y_4 = A$, where A and B are arbitrary, compute the Fourier coefficients a_0, a_1, a_2, b_1, b_2. Which coefficients equal zero?

19.3 Fourier Least Squares

Fourier: Let's go back to the just completed discussion of interpolation and change the rules to the setting, where we have more data points than coefficients. In other words, let's consider the problem:

Given data y_0, y_1, y_2, y_3, y_4 and equally spaced points $x_0 = 0, x_1 = \frac{2\pi}{5}, x_2 = 2\frac{2\pi}{5}, x_3 = 3\frac{2\pi}{5}, x_4 = 4\frac{2\pi}{5}$, find the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ "best fits the data".

Virginia: Let me guess. We are once again confronted by the same the setting we had for polynomial least squares. Namely, we simply "solve" the set of equations:

$$\begin{aligned}
T_1(x_0) &= \frac{a_0}{2} + a_1\cos(x_0) + b_1\sin(x_0) = && y_0\\
T_1(x_1) &= \frac{a_0}{2} + a_1\cos(x_1) + b_1\sin(x_1) = && y_1\\
T_1(x_2) &= \frac{a_0}{2} + a_1\cos(x_2) + b_1\sin(x_2) = && y_2\\
T_1(x_3) &= \frac{a_0}{2} + a_1\cos(x_3) + b_1\sin(x_3) = && y_3\\
T_1(x_4) &= \frac{a_0}{2} + a_1\cos(x_4) + b_1\sin(x_4) = && y_4.
\end{aligned}$$

This set of equations morphs into equation $A\mathbf{a} = \mathbf{y}$, where

$$A = \begin{pmatrix} 1 & \cos(x_0) & \sin(x_0)\\ 1 & \cos(x_1) & \sin(x_1)\\ 1 & \cos(x_2) & \sin(x_2)\\ 1 & \cos(x_3) & \sin(x_3)\\ 1 & \cos(x_4) & \sin(x_4) \end{pmatrix}, \mathbf{a} = \begin{pmatrix} \frac{a_0}{2}\\ a_1\\ b_1 \end{pmatrix}, \text{ and } \mathbf{y} = \begin{pmatrix} y_0\\ y_1\\ y_2\\ y_3\\ y_4 \end{pmatrix}.$$

Sadly, this equation is once again overdetermined. However, the good news is that we can solve it by multiplying both sides of the equation by the transpose A^t to obtain $A^tA\mathbf{a} = A^t\mathbf{y}$. The beauty of this matrix equation is that

$$D = A^tA = \begin{pmatrix} 5 & 0 & 0\\ 0 & \frac{5}{2} & 0\\ 0 & 0 & \frac{5}{2} \end{pmatrix}.$$

Thus, the matrix equation $A\mathbf{a} = \mathbf{y}$ is easily "solved" for the coefficients a_0, a_1, b_1. Better yet, the formulas for these coefficients are exactly the same as those we presented moments ago.

Simplicio: How did she know that?

Virginia: Math is easy.

Fourier: Actually, when I first began working on these series, I didn't understand this issue all that well either. However, let's be sure to mention that this technique

is equivalent to the problem of minimizing the residual $R = \sum_{k=0}^{4}(\frac{a_0}{2} + a_1\cos(x_k) + b_1\sin(x_k) - y_k)^2$ with respect to the parameters a_0, a_1, b_1.

Simplicio: Even I can see that this quantity R can be found by computing the gradient

$$\nabla R = \begin{pmatrix} \frac{\partial R}{\partial a_0} \\ \frac{\partial R}{\partial a_1} \\ \frac{\partial R}{\partial b_1} \end{pmatrix},$$

setting each coordinate equal to zero, and solving three equations and three unknowns.

Fourier: The matrix equation is $2(A^tA\mathbf{a} - A^t\mathbf{y}) = 0$, which is obviously equivalent to our friend $A^tA\mathbf{a} = A^t\mathbf{y}$.

Fourier: We summarize our discussion with the following theorem.

Theorem 19.3.1 (Fourier Coefficients: Linear Least Squares). *If $x_k = \frac{k}{2n+1}2\pi$, for $k = 0, 1, \ldots, 2n$ and $y_0, y_1, y_2, \ldots, y_{2n}$ are $2n+1$ given data values, then for any integer $N \leq n$, constants a_k and b_k can be found so that the trigonometric polynomial*

$$T_N(x) = \frac{a_0}{2} + \sum_{k=1}^{N}[a_k\cos kx + b_k\sin kx]$$

has the property that the function $T_N(x)$ provides a best least squares fit to the data y_k for all $k = 0, 1, \ldots, 2n$.

Moreover, the coefficients can found by be computing the following formulas:

$$a_k = \frac{2}{2n+1}\sum_{j=0}^{2n} y_j\cos(kx_j) \text{ for } k = 0, 1, 2, \ldots, N,$$

and

$$b_k = \frac{2}{2n+1}\sum_{j=0}^{2n} y_j\sin(kx_j) \text{ for } k = 1, 2, \ldots, N.$$

Simplicio: That WAS easy.

Fourier: Now its time to work some problems. Since the formulas are the same as for interpolation, these problems should provide no challenge.

Exercise Set 19.3.

1. Given the data $y_0 = 3, y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 1$, compute the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ which best fits the data in the sense of least squares. Plot the data and the function.

2. Given the data $y_0 = 0, y_1 = 1, y_2 = 2, y_3 = -2, y_4 = -1$, compute the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ which best fits the data in the sense of least squares. Plot the data and the function.

3. Given the data $y_0 = 1, y_1 = -1, y_2 = 1, y_3 = -1, y_4 = 1, y_5 = -1, y_6 = 1$, compute the function $T_2(x) = \frac{a_0}{2} + a_1 \cos(x) + a_2 \cos(2x) + b_1 \sin(x) + b_2 \sin(2x)$ which best fits the data in the sense of least squares. Plot the data and the function.

19.4 Fourier Interpolation: The Runge Example Revisited

Fourier: To illustrate the benefit of using trigonometric interpolation, let's revisit our friend Carl Runge. Recall that polynomial interpolation is a disaster when we approximate the curve $y = f(x) = \frac{1}{x^2+1}$, for $x \in [-\pi, \pi]$.

Simplicio: That's right. We saw those rabbit ears pop up near the boundary points of the interval $x = -\pi$ and π. The graphs of the approximations fly off to infinity.

Fourier: Let's apply our new interpolation method to this same curve. In particular, let's approximate the curve $f(x) = \frac{1}{x^2+1}$ by the trigonometric polynomials $T_n(x)$ on the interval $[-\pi, \pi]$. The results of these experiments (for the integers $n = 1, 2$, and $n = 20$) are displayed in Figures 19.1, 19.2, and 19.3. Note that in Figure 19.3 it is impossible to distinguish between the original curve and the approximation by $T_{20}(x)$. Unlike polynomial interpolation, the approximations provide improved approximations of the original curve when more points are added.

Simplicio: I am glad to see that we now have a reliable method we can count on to

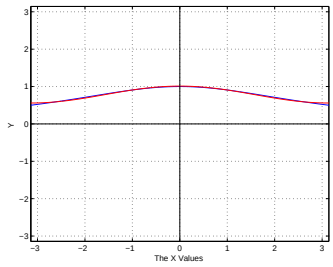


Figure 19.1: Fourier Interpolation of $f(x) = \frac{1}{1+x^2}$ by $T_1(x), x \in [-\pi, \pi]$

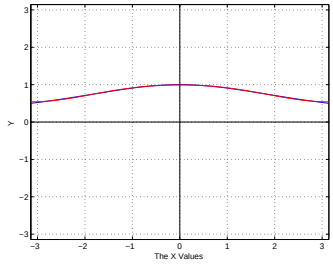


Figure 19.2: Fourier Interpolation of $f(x) = \frac{1}{1+x^2}$ by $T_2(x), x \in [-\pi, \pi]$

always give the results we want.

Fourier: In a moment, we will discuss the trigonometric approximation of the function $y = f(x) = x$, where we again encounter a slight "blip" at the endpoints of the interval. However, this time the problem is not as violent as is the case with polynomial interpolation.

Simplicio: I have an unimportant question, which has been nagging me. Namely, while all the data (x_k, y_k) mention in the theorems we have proved is designed so that $x_k \in [0, 2\pi]$, you took the domain of the function $f(x) = \frac{1}{1+x^2}$ to be the interval $[-\pi, \pi]$. Thus, the points x_k must lie in the interval $[-\pi, \pi]$. I know this is a small difference, but since we are being picky, I thought ....

Fourier: You should have been a mathematician. You spotted a bit of sloppiness on my part. Actually, since the functions $\cos(x)$ and $\sin(x)$ are both 2π periodic, we could go through the same analysis for any interval of length 2π. While the Orthogonality and Equal Lengths propositions will hold, the coefficients will be different.

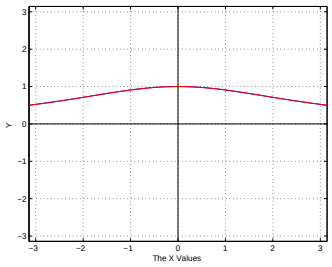


Figure 19.3: Fourier Interpolation of $f(x) = \frac{1}{1+x^2}$ by $T_{20}(x), x \in [-\pi, \pi]$

Virginia: But the coefficients a_k, b_k will be different only because the entries in the coefficient matrix will have been permuted around. Right?

Fourier: For example, if $x_0 = -\pi = -180$ (deg), $x_1 = -\pi + \frac{2\pi}{3} = -60$ (deg), $x_2 = -\pi + 2\frac{2\pi}{3} = 60$ (deg), then the coefficient matrix becomes

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 1 & \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}.$$

Simplicio: Looks like Orthogonality and Equal Lengths are OK to me.

Exercise Set 19.4.

1. Compute the Fourier coefficient matrix for five points on the interval $[-\pi, \pi]$. (i.e. Compute the matrix A when $x_0 = -\pi, x_1 = -\pi + \frac{2\pi}{5}, x_2 = -\pi + 2\frac{2\pi}{5}, x_3 = -\pi + 3\frac{2\pi}{5}, x_4 = -\pi + 4\frac{2\pi}{5}$.

2. Let $y = f(x) = \frac{1}{1+x^2}$, for $x \in [-\pi, \pi]$. Write a program to approximate $f(x)$ by the functions $T_0(x)$ for various evenly evenly spaced points $-\pi = x_0, x_1, \ldots, x_{2n} < \pi$. Plot the functions $y = f(x)$ and $y = T_0(x)$ on the same graph. How good are the approximations? What do you notice?

19.5 Fourier Interpolation: Gibbs' Phenomenon

Fourier: We now provide a short discussion of the famous Gibbs' phenomenon. Josiah Willard Gibbs(1839-1903) was the first outstanding American mathematician. His

contributions were in a wide range of areas including vector analysis, the orbits of comets, the thermodynamics of fluids, electromagnetic radiation, and statistical mechanics. His investigations into thermodynamics involved the mathematics surrounding the words energy, entropy, and enthalpy.

Simplicio: Doesn't entropy involve the ideas of order and disorder?

Virginia: If I remember correctly, entropy always increases. Isn't that the idea behind the Second Law of Thermodynamics?

Fourier: Very good. Gibbs was a dedicated natural scientist, who developed sophisticated mathematical ideas to model real physical phenomena. He continued the investigations into the study of the steam engine begun by Sadi Carnot (1796-1832). Their research led to the modern theory of Thermodynamics. Gibbs brought more mathematics to the table. In any case, we are not going to discuss this topic today. Instead, we are going to mention Gibbs' contribution to Trigonometric series.

Simplicio: And, ...

Fourier: While the phenomenon appears in many different disguises, we will demonstrate it only for the function $f(x) = x$ defined on the interval $[-\pi, \pi]$. If we let $T_n(x) = 2\sum_{k=1}^{n} \frac{(-1)^{k+1}}{k} \cdot \sin(kx)$ on $[-\pi, \pi]$ for $n = 4, 8$, and 20, then note the graphs of the approximations in Figures 19.4, 19.5, and 19.6.

Simplicio: Since we are computing on the interval $[-\pi, \pi]$, won't we once again encounter a modified version of the Fourier matrix the way we just did with the Runge example? In other words, when we compute the entries in the coefficient matrix A, we will use the points $x_0 = -\pi, x_1 = -\pi + \frac{2\pi}{2n+1}, x_2 = -\pi + 2\frac{2\pi}{2n+1}, x_3 = -\pi + 3\frac{2\pi}{2n+1}, \ldots x_{2n} = -\pi + 2n\frac{2\pi}{2n+1}$.

Fourier: Correct. In any case, these examples lead to the well-known *Gibbs* phenomenon, where a slight "blip" appears at the endpoints $-\pi$ and π. Note that this "blip" continues to appear even for a 20 degree polynomial. This blip is about 9% of the difference between $+\pi - (-\pi) = 2\pi$ and thus about 0.56.

Simplicio: And once again we have a setting, where an approximation of good data leads to mediocre results.

Virginia: But at least the blips don't go off to infinity. I would say that is an improvement.

Fourier: That's why we call them Fourier series.

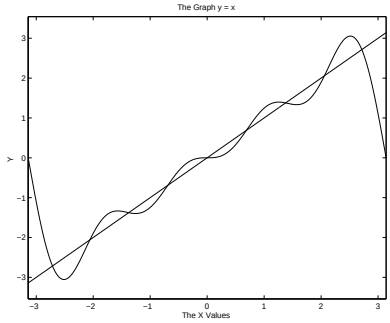


Figure 19.4: The Gibbs Effect When Approximating $f(x) = x$ by $T_4(x)$

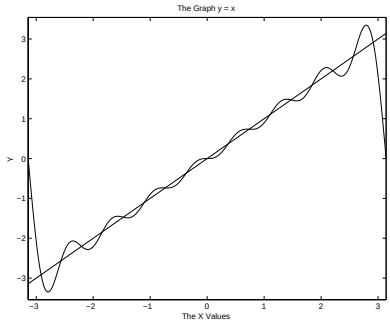


Figure 19.5: The Gibbs Effect When Approximating $f(x) = x$ by $T_8(x)$

Exercise Set 19.5.

1. Compute and plot the trigonometric series approximation of the function defined by

$$f(x) = \begin{cases} -1 & \text{if } x \in [-\pi, 0) \\ 1 & \text{if } x \in [0, \pi] \ . \end{cases}$$

Approximate the blip at $x = 0$ for the integers $n = 4, 6$, and 10. Where do you find the blips? How big are they?

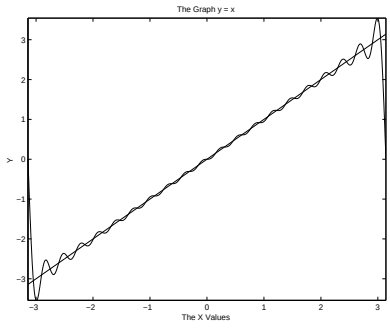


Figure 19.6: The Gibbs Effect When Approximating $f(x) = x$ by $T_{20}(x)$

2. Compute and plot the trigonometric series approximation of the function defined by

$$f(x) = \begin{cases} 0 & \text{if } x \in [0, \pi) \\ 1 & \text{if } x \in [\pi, 2\pi] \end{cases}.$$

Approximate the blip at $x = 0$ for the integers $n = 4, 6$, and 10. Where do you find the blips? How big are they?

19.6 Fourier Interpolation: Pythagoras/Parseval

Galileo: We are now in a position to prove the theorem of Pythagoras one more time. How about an explanation Professor Fourier?

Fourier: The only difference between this theorem and the one Professor Hilbert gave you previously is that the word vector will be replaced by the word function. In particular, we will prove the theorem for functions of the form $T_n(x)$. If you understood Professor Hilbert's Linear Algebra proof, you will understand this one with no problem. The data vector $\mathbf{y} = (y_0, y_1, y_2, \ldots, y_{2n})^t$ can be visualized at the diagonal in an n-dimensional parallelepiped (i.e. rectangle). If $\mathbf{x} = (x_0, x_1, \ldots, x_{2n})^t$, then the sides of the box can be visualized as represented by a set of orthogonal vectors of the form

$\mathbf{u}_0 = (1, 1, \ldots, 1)^t, \mathbf{u}_1 = \cos(\mathbf{x})^t, \mathbf{u}_2 = \cos(2\mathbf{x})^t, \ldots, \mathbf{u}_n = \cos(n\mathbf{x})^t, \mathbf{v}_1 = \sin(\mathbf{x})^t, \mathbf{v}_2 = \sin(2\mathbf{x})^t, \ldots, \mathbf{v}_n = \sin(n\mathbf{x})^t$. We understand that the notation $\mathbf{u}_1 = \cos(\mathbf{x})^t$ simply

means that $\mathbf{u}_1 = \cos(\mathbf{x})^t = (\cos(x_0, \cos(x_1), \ldots, \cos(x_{2n})^t$. The proof works because the vectors $\mathbf{u}_m = \cos(m\mathbf{x})^t$ and $\mathbf{v}_k = \sin(k\mathbf{x})^t$ are mutually perpendicular.

Virginia: Interesting.

Simplicio: Groan.

Theorem 19.6.1 (Pythagoras/Parseval). *If $x_k = \frac{k}{2n+1}2\pi$, for $k = 0, 1, \ldots, 2n$, and $y_0, y_1, y_2, \ldots, y_{2n}$ are $2n+1$ given data values, and*

$$T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}[a_k \cos kx + b_k \sin kx]$$

has the property that $T_n(x_k) = y_k$ for all $k = 0, 1, \ldots, 2n$, then

$$\frac{2}{2n+1}\sum_{k=0}^{2n} y_k^2 = \frac{a_0^2}{2} + \sum_{k=1}^{n}(a_k^2 + b_k^2).$$

Proof. As in the previous theorem, we only give the proof for the case when $n = 2$.

If we set $\mathbf{y} = (y_0, y_1, y_2, y_3, y_4)^t, \mathbf{u}_0 = (1, 1, 1, 1, 1)^t, \mathbf{u}_1 = \cos(\mathbf{x})^t, \mathbf{u}_2 = \cos(2\mathbf{x})^t, \mathbf{v}_1 = \sin(\mathbf{x})^t$, and $\mathbf{v}_2 = \sin(2\mathbf{x})^t$, then $T_2(x) = \frac{a_0}{2}\mathbf{u}_0 + a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + b_1\mathbf{v}_1 + b_2\mathbf{v}_2$. Thus, by the Orthogonality and Equal Lengths Properties, we see that

$$\begin{aligned}
< \mathbf{y}, \mathbf{y} > &= < \frac{a_0}{2}\mathbf{u}_0 + a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + b_1\mathbf{v}_1 + b_2\mathbf{v}_2, \frac{a_0}{2}\mathbf{u}_0 + a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + b_1\mathbf{v}_1 + b_2\mathbf{v}_2 > \\
&= < \frac{a_0}{2}\mathbf{u}_0, \frac{a_0}{2}\mathbf{u}_0 > + < a_1\mathbf{u}_1, a_1\mathbf{u}_1 > + < a_2\mathbf{u}_2, a_2\mathbf{u}_2 > + \\
&\quad < b_1\mathbf{v}_1, b_1\mathbf{v}_1 > + < b_2\mathbf{v}_2, b_2\mathbf{v}_2 > \\
&= (\frac{a_0}{2})^2 < \mathbf{u}_0, \mathbf{u}_0 > + a_1^2 < \mathbf{u}_1, \mathbf{u}_1 > + a_2^2 < \mathbf{u}_2, \mathbf{u}_2 > + \\
&\quad b_1^2 < \mathbf{v}_1, \mathbf{v}_1 > + b_2^2 < \mathbf{v}_2, \mathbf{v}_2 > \\
&= 5(\frac{a_0}{2})^2 + \frac{5}{2}a_1^2 + \frac{5}{2}a_2^2 + \frac{5}{2}b_1^2 + \frac{5}{2}b_2^2.
\end{aligned}$$

Thus,

$$\frac{2}{5}\sum_{k=0}^{4} y_k^2 = \frac{a_0^2}{2} + \sum_{k=1}^{2}(a_k^2 + b_k^2).$$

The reason the argument works is because orthogonality implies that all the inner products $< \mathbf{u}_m, \mathbf{v}_k >, < \mathbf{u}_m, \mathbf{u}_k >$, and $< \mathbf{v}_m, \mathbf{v}_k >$ equal zero except for the special cases when $< \mathbf{u}_0, \mathbf{u}_0 >= 5, < \mathbf{u}_1, \mathbf{u}_1 >= \frac{5}{2}, < \mathbf{u}_2, \mathbf{u}_2 >= \frac{5}{2}, < \mathbf{v}_1, \mathbf{v}_1 >= \frac{5}{2}, < \mathbf{v}_2, \mathbf{v}_2 >= \frac{5}{2}$. Orthogonality does the trick. □

Fourier: As I think you can see, the proof of the general theorem is going to be the same as the $n = 2$ case. The only difference is that we will need to sum more terms.

Simplicio: Even I can see that. In fact, this proof looks familiar.

Fourier: It should. When we proved the coefficient formula, we used almost the same argument.

Simplicio: Ok, but what is all this theory good for? How about an example?

Fourier: It is a bit difficult to give an interesting example for this theorem because if I give you a vector $\mathbf{y} = (y_0, y_1, y_2, \ldots, y_{2n})^t$, note that all you are going to do is compute the sum $\frac{2}{2n+1}\sum_{k=0}^{2n} y_k^2$ and the sum $\frac{a_0^2}{2} + \sum_{k=1}^{2n}(a_k^2 + b_k^2)$ and check if they are equal. Do you want me to bore you?

Simplicio: Not today.

Virginia: But the theorem is a lovely extension of Pythagoras's ideas. I really like this theorem.

Simplicio: I am sure you do.

19.7 A Fourier Application: Signal Compression

Fourier: How about an application?

Simplicio: An application would be appreciated.

Fourier: How about signal and image compression?

Simplicio: I must admit that I find image compression interesting.

Fourier: The first piece of information to mention is that the discrete cosine transform is an integral component of the JPEG and MPEG file formats that are used to display images on the internet. At least that was true until the year 2000.

Simplicio: What happened then?

Fourier: The techniques were upgraded from trigonometric series to wavelets.

Simplicio: What is a wavelet?

Fourier: While there are a multitude of technicalities with wavelets, the basic idea is to build a collection of basis functions that have the same orthogonality properties

as sines and cosines, but which don't oscillate up and down forever. In other words, outside of some finite interval (e.g. $[0, 1]$), they always equal zero. These functions are preferred in a multitude of real applications because data collection and computer memory are necessarily finite.

Galileo: Professor Fourier you digress.

Fourier: While there are a number of ways to compress a signal, the idea we will explore uses the Theorem of Pythagoras/Parseval to measure how many coefficients will be required to produce an accurate reconstruction of the signal.

Simplicio: What does it mean to reconstruct a signal?

Fourier: By the Coefficient Formula for Trigonometric Interpolation, we can always solve the problem: If given $x_k = \frac{k}{2n+1}2\pi$, for $k = 0, 1, \ldots, 2n$ and $y_0, y_1, y_2, \ldots, y_{2n}$ are $2n + 1$ given data values, then constants a_k and b_k can be found so that the trigonometric polynomial

$$T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}[a_k \cos kx + b_k \sin kx]$$

has the property that $T_n(x_k) = y_k$ for all $k = 0, 1, \ldots, 2n$.

Since $T_n(x_k) = y_k$ for all $k = 0, 1, \ldots, 2n$, we have perfect reconstruction. If we throw away some of the coefficients (i.e. set some a_k's or b_k's $= 0$), then we can no longer expect the equalities $T_n(x_k) = y_k$ to always hold. This issue leads to the concept of imperfect reconstruction and provides a fundamental technique for lossy compression.

Simplicio: In other words, the idea is to replace the given data values $y_0, y_1, y_2, \ldots, y_{2n}$ by the coefficients $a_0, a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$. If you retain all the coefficients, you have perfect reconstruction because you can always compute $y_k = T_n(x_k)$. The problem with perfect reconstruction is that you have no savings when you store your data on your hard drive. For lossy compression, you can "reconstruct" the data by computing $\hat{y}_k = \hat{T}_n(x_k)$, where the formula for $\hat{T}_n(x)$ is the same as for $T_n(x)$ except that some of the coefficients have been set equal to zero.

Fourier: Very good. First of all, when we set certain coefficients $a_k = 0$ and $b_k = 0$,

for $k > N$, then we are simply computing the best least squares approximation of the data using the function

$$\hat{T}_n(x) = T_N(x) = \frac{a_0}{2} + \sum_{k=1}^{N} [a_k \cos(kx) + b_k \sin(kx)],$$

where $N < n$.

Second, as we remarked in our discussion of least squares, the coefficient formulas are the same as before. In particular,

$$a_k = \frac{2}{2n+1} \sum_{j=0}^{2n} y_j \cos(kx_j) \text{ for } k = 0, 1, 2, \ldots, N,$$

and

$$b_k = \frac{2}{2n+1} \sum_{j=0}^{2n} y_j \sin(kx_j) \text{ for } k = 1, 2, \ldots, N.$$

Virginia: But wait a minute. I detect a potential problem here. What if $\hat{y}_k = T_N(x_k)$ is not close to the original value y_k?

Fourier: We should now call in our friend Pythagoras. He would be proud to know his ideas are still being discussed after all these years. In his place, let us remark that we know by Pythagoras/Parseval that

$$\frac{2}{2n+1} \sum_{k=0}^{2n} y_k^2 = \frac{a_0^2}{2} + \sum_{k=1}^{n} (a_k^2 + b_k^2).$$

If we form the fraction

$$Q = \frac{\frac{a_0^2}{2} + \sum_{k=1}^{n} (a_k^2 + b_k^2)}{\frac{2}{2n+1} \sum_{k=0}^{2n} y_k^2},$$

then Q most definitely equals 1.

Simplicio: No argument on this point.

Fourier: If $N \leq n$ and we form the fraction

$$Q_N = \frac{\frac{a_0^2}{2} + \sum_{k=1}^{N} (a_k^2 + b_k^2)}{\frac{2}{2n+1} \sum_{k=0}^{2n} y_k^2},$$

then $Q \leq 1$.

Simplicio: We have simply discarded some positive terms in the numerator. No argument on this point as well.

Fourier: Let me now ask you an old question in Physics. What is the formulas for kinetic energy?

Virginia: Of course, $KE = \frac{1}{2}mv^2$.

Fourier: Notice that kinetic energy has the velocity squared. If you think about it, the fraction Q_N can be thought of as providing a measure of the energy in the coefficients used divided by the total energy of the data. If $N = n$, the two measures of energy are in balance and $Q_N = 1$. In this case, we have lossless compression. We now ask the following question: If we want 90% of the information in the data, then how do we choose N?

Virginia: I bet I can guess. How about if we simply choose N to be an integer less than n with the property that $Q_N > 0.90$. We will achieve the greatest compression, if we choose N to be the smallest such integer.

Fourier: You got it.

Simplicio: How about an example?

Fourier:

Example 19.7.1. *Given seven data points* $-3, -2, -1, 0, 1, 2, 3$, *compute the smallest integer* N *with the property that* $Q_N > 0.85$, *where oefficients* N *are needed so that the quotient*

$$Q_N = \frac{\frac{a_0^2}{2} + \sum_{k=1}^{N}(a_k^2 + b_k^2)}{\frac{2}{2n+1}\sum_{k=0}^{2n} y_k^2}.$$

If we make the computations, we find $Q_1 = 0.6640$ *and* $Q_2 = 0.8685$. *Thus, we can choose* $N = 2$.

Simplicio: Seems OK to me.

Fourier: Now lets conduct a little experiment, where we "double the data" before we compute the coefficients and the quotient Q_N. What I mean by doubling the data is to take the 7-dimensional vector $\mathbf{y} = (-3, -2, -1, 0, 1, 2, 3)$ and extend it to the 13-dimensional vector

$\mathbf{y}_2 = (-3, -2, -1, 0, 1, 2, 3, 3, 2, 1, 0, -1, -2)$. When we do this, we find that all the coefficients $b_k = 0$, for all $k = 1, 2, 3, 4, 5, 6$

Simplicio: Why didn't you make $\mathbf{y}_2$ into a 14-dimensional vector with last coordinate equal to -3?

Fourier: If we use that strategy, we don't quite achieve the key symmetry relation that makes the terms cancel out. Since the functions $\sin(kx)$ are "anti-symmetric" about the vertical line $x = \pi$ (i.e. $f(\pi + x) = -f(\pi - x)$ for all $x \in \Re$) and since $\pi - x_j = x_{13-j} - \pi$, we know that

$$\begin{aligned}
\sin(kx_7) &= -\sin(kx_6),\\
\sin(kx_8) &= -\sin(x_5),\\
\sin(kx_9) &= -\sin(x_4),\\
\sin(kx_{10}) &= -\sin(x_3),\\
\sin(kx_{11}) &= -\sin(x_2),\\
\sin(kx_{12}) &= -\sin(x_1).
\end{aligned}$$

Note that we may need the sum formulas for $\sin(x)$ to make this argument complete.

Thus,

$$\begin{aligned}
y_6\sin(kx_7) + y_6\sin(kx_6) &= 0,\\
y_5\sin(kx_8) + y_5\sin(kx_5) &= 0,\\
y_4\sin(kx_9) + y_4\sin(kx_4) &= 0,\\
y_3\sin(kx_{10}) + y_3\sin(kx_3) &= 0,\\
y_2\sin(kx_{11}) + y_2\sin(kx_2) &= 0,\\
y_1\sin(kx_{12}) + y_1\sin(kx_1) &= 0
\end{aligned}$$

and thus

$$\begin{aligned}
b_k = {} & y_0\sin(0) + y_1\sin(kx_1) + y_2\sin(kx_2) + y_3\sin(kx_3) +\\
& y_4\sin(kx_4) + y_5\sin(kx_5) + y_6\sin(kx_6) + y_6\sin(kx_7) +\\
& y_5\sin(kx_8) + y_4\sin(kx_9) + y_3\sin(kx_{10}) + y_2\sin(kx_{11}) +\\
& y_1\sin(kx_{12}) = 0.
\end{aligned}$$

We have just proved a special case of the following proposition.

Proposition 19.7.1. *If* $\mathbf{y} = (y_0, y_1, \ldots, y_n, y_{n+1}, \ldots, y_{2n})$ *is a* $(2n+1)-$*dimensional vector with the property that* $y_{n+1} = y_n, y_{n+2} = y_{n-1}, \ldots, y_{2n} = y_1$, *then* $b_k = 0$, *for any integer* $k = 1, 2, \ldots, n$.

Proof. Since $x_j = j\frac{2\pi}{2n+1}$ and $x_{2n+1-j} = (2n+1-j)\frac{2\pi}{2n+1}$, a bit of arithmetic can be used to show that $\pi - x_j = x_{2n+1-j} - \pi$.

By the sum formula for $\sin(x)$, we know

$$\sin(k(\pi - x_j)) = \sin(k\pi)\cos(kx_j) - \sin(kx_j)\cos(k\pi) = -\sin(kx_j)\cos(k\pi)$$

and

$$\begin{aligned}\sin(k(x_{2n+1-j} - \pi)) &= \sin(kx_{2n+1-j})\cos(k\pi) - \sin(k\pi)\cos(kx_{2n+1-j}) \\ &= \sin(kx_{2n+1-j})\cos(k\pi).\end{aligned}$$

Thus,

$$-\sin(kx_j)\cos(k\pi) = \sin(kx_{2n+1-j})\cos(k\pi).$$

Dividing both sides of this equation by $\cos(k\pi)$, we see that

$$-\sin(kx_j) = \sin(kx_{2n+1-j}).$$

Since $y_{2n+1-j} = y_j$, for all j,$-y_j \sin(kx_j) = y_{2n+1-j}\sin(kx_{2n+1-j})$ for all j. Since the coefficient

$$b_k = \frac{2}{2n+1}\sum_{j=1}^{n}(y_j \sin(kx_j) + y_{2n+1-j}\sin(kx_{2n+1-j}) = 0,$$

we are done.

□

Simplicio: That last proof was a bit technical. How about an example.

Fourier: No problem How about if we repeat the same problem we discussed a few minutes ago?

Simplicio: I'm easy.

Example 19.7.2. *Given the same seven data points (so $n = 6$)* $-3, -2, -1, 0, 1, 2, 3$, *we double the data to the 13 point set* $-3, -2, -1, 0, 1, 2, 3, 3, 2, 1, 0, -1, -2$. *We now would like to interpolate this data with the function $T_6(x)$. By the proposition, we know that $b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = 0$ so we only need to compute the coefficients a_k's. As before, we would like to find the smallest integer $N \leq 6$ with the property that $Q_N > 0.85$, where*

$$Q_N = \frac{\frac{a_0^2}{2} + \sum_{k=1}^{N} a_k^2}{\frac{2}{2n+1}(y_0^2 + 2\sum_{k=1}^{6} y_k^2)}.$$

If we make the computations, we find $Q_1 = 0.9839 > 0.85$. Thus, we can choose $N = 1$.

Simplicio: $N = 1$! Hey, that trick worked much better than when the b_k's were involved. With the previous computations, we saw that $Q_1 = 0.6640$ and $Q_2 = 0.8685$, where the value of Q_1 requires the three coefficients a_0, a_1, b_1 and the value of Q_2 requires the five coefficients a_0, a_1, a_2, b_1, b_2. By doubling the data and saving only the coefficients a_0 and a_1, we get far better reconstruction than before. I am getting a bit more interested. What's going on here?

Virginia: I bet there is a theorem lurking here somewhere.

Fourier: You got it. The problem with the data set is that the first value $y_0 = -3$ and the last data point $y_6 = 3$. In particular, the values y_0 does not equal y_6.

Simplicio: In fact, the data is essentially a straight line between the two points $(0, -3)$ and $(2\pi, 3)$ so Gibbs is sure to haunt you.

Fourier: The Gibbs problem disappears if you approximate a continuous function $f(x) : [-\pi, \pi] \to \Re$, which is blessed with the additional property that $f(-\pi) = f(\pi)$. An even function always has this desired property.

Virginia: So this theorem will show that the approximations of the Runge example $f(x) = 11 + x^2$ by functions $T_n(x)$ will converge with no blips?

Fourier: Correct.

Simplicio: So, even is good, odd is evil.

Fourier: Not quite.

with the property that $f(-\pi) = f(\pi)$. Similarly, the Gibbs problem disappears if you are approximating a continuous function $f(x) : [0, 2\pi] \to \Re$ with the property that $f(0) = f(2\pi)$. While we always manage to ge pointwise convergence for any continuous function, we manage

Fourier: How about working a few problems to check your understanding?

Exercise Set 19.7.

1. Given seven data points $0, 2, 3, 5, 7, 11, 13$, compute the coefficients $a_0, a_1, a_2, a_3, b_1, b_2, b_3$. How many coefficients $N \leq 3$ are needed so that the quotient
$$Q = \frac{\frac{a_0^2}{2} + \sum_{k=1}^{N}(a_k^2 + b_k^2)}{\frac{2}{2n+1}\sum_{k=0}^{6} y_k^2} \geq 0.90?$$

2. Redo the previous problem after the data has been doubled.

3. If the data $y_0, y_1, \ldots, y_{2n}$ has the property that it is anti-symmetric about the middle value (i.e. $y_n = -y_{n+1}, y_{n-1} = -y_{n+2}$, etc.) and $y_0 = 0$, then show that all the coefficients $a_k = 0$. Thus, if a data set has this property, then the coefficients a_k do not have to be computed.

19.8 Complex Numbers: A Brief Review

Galileo: In preparation of our discussion of the complex Fourier transform, we now are forced to consider (and understand!) complex numbers.

Virginia: Mother Nature insists!

Galileo: This transform arises from polynomial interpolation, where the points (z_k, y_k) are chosen with the restriction that the points z_k lie uniformly spaced on the unit circle in the complex plane.

Simplicio: Wait a minute. Our previous discussion of the discrete Fourier transform seems just fine to me. Why would we complicate the discussion by introducing complex numbers? I am only interested in real data. In any case, I am out of my comfort zone here.

Virginia: They have always seemed a bit imaginary to me as well.

Galileo: If our goal is to solve equations, Mother Nature won't allow us to ignore complex numbers. Recall that the equation $x^2 = -1$ does not have a real solution. In general, even though your problem can be stated in terms of complex numbers, the solution may not.

Simplicio: Onward.

Galileo: While the ancient Greeks had the concepts of distance, numbers, addition, subtraction, multiplication, and division, they had only a limited understanding of Algebra. In fact, not only were negative numbers unknown to them, they didn't even have the concept of zero.

Simplicio: I understand zero dollars!

Galileo: No problem arose when the ancients wanted to solve an equation of the form $x+2 = 3$ or a proportion of the form $\frac{x}{1} = \frac{1-x}{x}$. However, this truncated understanding of Algebra led to trouble when they tried to solve equations like $x + 3 = 2$.

Virginia: Where the answer is negative and you are forced to consider negative numbers?

Galileo: Correct. Even much later the Father of Algebra, Muhammad ibn Mūsā al-Khwārizmī (780-850) avoided negative numbers. For example, he would write the expression $ax^2 - bx = 0$ as $ax^2 = bx$.

Simplicio: But negative numbers are easy. You just add, subtract, multiply, and divide the same way you manipulate positive numbers. No problem.

Galileo: Very good. but for the ancients, negative numbers were just as virtual as complex numbers are for you. What does it mean for you to have $-\$100.00$ in your pocket?

Virginia: Your last purchase was charged to your credit card!

Galileo: Exactly my point. Credit cards are virtual. How about a second question: Why does $(-1)(-1) = +1$?

Simplicio: I'm not sure. Actually, I never did like that rule.

Galileo: The underlying force behind that equation is the desire to solve different

types of equations. These equations can be linear, quadratic, cubic, or worse. It doesn't matter.

Simplicio: I don't see the connection between the rule $(-1)(-1) = +1$ and solving equations.

Galileo: While the subject of Algebra has taken several millenia to unfold, we are now clear that the essence of an algebraic structure is a set of points X together with one or two operations such as addition and/or multiplication. The points in the set X are usually represented by the letters a, b, c and $x, y, z, etc.$, while the operations are usually represented by the symbols $+, -, *, /$. It wasn't until Nicole d' Oresme (1323-1382), Johannes Widmann (1460-1524), William Oughtred (1574-1660), and Gottfried Wilhelm Leibniz (1646-1716) came along that people began to realize these mathematical operations deserve their own symbology. By using different symbols for points and operations, the implicit message is that they are indeed different. Did you realize that the symbol for addition " + " is derived from the Latin word "et."

Virginia: Which, of course, means "and."

Galileo: Not only were the ancient Greeks not quite clear about points and operations, but even the Father of Algebra, the Medieval Indian mathematicians, and Leonardo of Pisa (1188-1250) (otherwise known as Fibonacci) were also not quite clear. The modern view is that the starting point should be the set of natural numbers $N = 1, 2, \ldots, n, \ldots$ together with the operations of addition $(+)$ and multiplication $(*)$. These operations should satisfy both the associative and commutative laws for both addition and multiplication. The distributive law is the force that binds addition and multiplication. If you didn't have the distributive law, then you could study these two operations separately. The whole numbers are the slightly larger set $W = 0, 1, 2, \ldots, n, \ldots$ with the same two operations.

Simplicio: So what about negative numbers?

Galileo: The discussion becomes clear when we consider the whole numbers as a subset of the integers $Z = \ldots, -n, \ldots, -3, -2, -1, 0, 1, 2, \ldots, n, \ldots$, where the minus sign $(-)$ functions in two ways. First, this sign indicates a new symbol to be added

to the set N. Second, it acts as a new operation, which is the inverse operation for addition. Not only does the set Z contain both N and W, but the operations of addition and multiplication can be extended so the associative, commutative, and distributive laws continue to hold. The fact that the number zero continues as the additive identity (along with the rules $n + 0 = n, n + (-n) = 0$, and $n - n = 0$) is crucial.

Simplicio: So, why do I want all these laws?

Galileo: Because you can now solve equations by repeated applications of just a few simple laws. In other words, once you know these laws, you can manipulate the equations with no fear of getting an incorrect answer. This process actually make slife easier. Colin Maclaurin (1698-1746) understood this strategy. He always considered a negative quantity to be no less real than a positive one.

Virginia: Didn't you forget the additive and multiplicative identities?

Galileo: Oops! You are correct. You need to know:

1. $n + 0 = n$,
2. $n * 1 = n$,
3. $n * 0 = 0$, and
4. $n + (-n) = 0$.

Simplicio: Ok, so why is $(-1) * (-1) = +1$?

Galileo: By rearranging, we can write $(-1) * (-1) = +1$ as $(-1) * (-1) - 1 = 0$, which is equivalent to

$$(-1) * (-1) + (-1) * (+1) = (-1) * (-1 + 1) = (-1) * 0 = 0.$$

Thus, if we decide to extend the distributive law to the integers Z, then Mother Nature gives us no choice other than to make the rule $(-1) * (-1) = +1$.

Simplicio: So how do these remarks apply to complex numbers?

Galileo: While the ancient Greeks were well aware of the quadratic formula and while Cardano extended (with the help of others!) extended the formula to cubics, it wasn't

until Rafael Bombelli (1526-1572) that clarity emerged. In his text, *Algebra*, he presented our now familiar rules for addition and multiplication of complex numbers.
Virginia: It is remarkable that such simple ideas took so long to unfold.
Simplicio: No wonder I have always hated Algebra and found it so difficult. They took 1500 years to figure it out.
Galileo: We have seen that before. Think about the Contraction Mapping Theorem. It is simple theorem to state and prove, but remarkably general.
Simplicio: I would say abstract.
Virginia: So Bombelli had the idea that the real numbers can be thought of as a subset of a larger set of numbers with the property that the equation $x^2 = 1$ can be solved. Better yet, the operations of addition and multiplication can be extended to this larger set in such a way that the associative, commutative, and distributive laws continue to hold.
Simplicio: I worry.
Galileo: If we assume that that the real numbers are well understood (and that is not at all obvious), we can write a complex number in two different ways. The first is as a sum $z = a + bi$, where $i = \sqrt{-1}$. From this vantage point, we can add two numbers by the following rule:

Definition 19.8.1 (Complex Addition). *If* $a, b, c, d \in \Re, i = \sqrt{-1}$, $z_1 = a + bi$, *and* $z_2 = c + di$, *then* $z_1 + z_2 = (a + b) + (c + d)i$.

We can also multiply two complex numbers by the rule:

Definition 19.8.2 (Complex Multiplication). *If* $a, b, c, d \in \Re, i = \sqrt{-1}$, $z_1 = a + bi$, *and* $z_2 = c + di$, *then* $z_1 * z_2 = (ac - bd) + (ad + bc)i$.

The advantage of complex numbers is that Euler's formula $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ allows you to consolidate two trigonometric functions into one exponential. With only slight modifications, all the ideas of interpolation, least squares, and orthogonality continue as before.

Simplicio: While I don't mind the rule for addition, I don't see the justification for multiplication.

Virginia: Obviously, the rule for multiplication is motivated by the distributive law. For if $z_1 = a + bi$ and $z_2 = c + di$, then we can simply assume the distributive law holds, multiply out the product and gather terms. In particular,

$$\begin{aligned} z_1 * z_2 &= (a + bi) * (c + di) \\ &= ac + adi + bci + bdi^2 \\ &= ac + adi + bci - bd \\ &= (ac - bd) + (ad + bc)i. \end{aligned}$$

The real part of the number is $ac - bd$, while the imaginary part is $ad + bc$. The negative sign appears because $i^2 = -1$.

Galileo: Correct.

Simplicio: Are we done with all this Algebra?

Galileo: How about if we formulate the algebraic rules into a proposition?

Proposition 19.8.3. *If z_1, z_2, z_3 are complex numbers, then the following rules hold.*

1. $z_1 + 0 = z_1$ *(additive identity property)*
2. $z_1 * 1 = z_1$ *(multiplicative identity property)*
3. $z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3$, *(associative law for addition)*
4. $z_2 + z_1 = z_1 + z_2$, *(commutative law for addition)*
5. $z_1(z_2 z_3) = (z_1 z_2)z_3$, *(associative law for multiplication)*
6. $z_1 z_2 = z_2 z_1$, *(commutative law for multiplication)*
7. $z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3$. *(distributive law)*

Simplicio: How about an example?

Galileo: Sure.

Example 19.8.1. *If* $z_1 = \frac{1+\sqrt{3}i}{2}$, *and* $z_2 = \frac{1-\sqrt{3}i}{2}$, *then by the distributive law*

$$\begin{aligned} z_1 * z_2 &= (\frac{1+\sqrt{3}i}{2}) * (\frac{1-\sqrt{3}i}{2}) \\ &= \frac{1}{4} + \frac{3}{4} + (-\frac{1}{2}\frac{\sqrt{3}}{2} + \frac{1}{2}\frac{\sqrt{3}}{2})i \\ &= \ = 1 + 0i = 1. \end{aligned}$$

This example leads to the Geometry of the complex numbers.

Simplicio: Geometry?

Galileo: If we think of the quantity $i = \sqrt{-1}$ as a place holder for a coordinate, then the complex number $z = \frac{1+\sqrt{3}i}{2}$ can be written as the vector $z = (\frac{1}{2}, \frac{\sqrt{3}}{2})$. Thus, the proposition given above indicates that we can add, subtract, multiply, and divide two 2-dimensional vectors $z_1 = (a, b)$ and $z_2 = (c, d)$.

Simplicio: And the multiplication rule is

$$z_1 * z_2 = (a, b) * (c, d) = (ac - bd, ad + bc).$$

Galileo: Now that we have addition and multiplication out of the way, we can turn to the idea of the *modulus* of a complex number. This concept is defined by the rule:

Definition 19.8.4. *If* $z = a + bi = (a, b)$, *then the modulus is defined by* $\|z\| = \sqrt{a^2 + b^2}$.

Simplicio: But wait a minute. Haven't you just computed the length of the vector (a, b)? Is modulus another word for length?

Galileo: Correct. We could just easily have called it the $2-$norm. For complex numbers, the words modulus, length, absolute value, and $2-$norm are different terms to describe the same concept. They all have the same meaning. However, as soon as we are talking about length, we are talking about Geometry.

Virginia: And it all began with Pythagoras.

Galileo: The next geometric concept is embedded in the computation of the conjugate of a complex number. This computation can be used whenever we compute the modulus. We can visualize the conjugate as a "flip" of a complex number across the $x-$axis.

Definition 19.8.5. *If $z = a + bi = (a, b)$, then the conjugate is defined by the rule $\overline{z} = a - bi = (a, -b)$.*

The first application of the conjugate is to give us a second definition of the modulus of a complex number.

Proposition 19.8.6. *If $z = a + bi = (a, b)$, then $\|z\| = \sqrt{\overline{z}z}$.*

Proof. Simplicio: But this formula is obvious. All you have to do is make the computation. □

Virginia: It is also convenient.

Galileo: It is more than convenient. In Geometry we are also interested in whether or not two lines or vectors are orthogonal. In general, we would like to compute the angle between two vectors. Right?

Simplicio: Sure.

Galileo: How did we compute angles before?

Simplicio: We computed $\cos(\theta)$ using the dot product and norm.

Virginia: More generally, we encapsulated these computations in the idea of inner product.

Galileo: Ok, so to decide whether or not two $2-$dimensional vectors $z_1 = (a, b)$ and $z_2 = (c, d)$ are orthogonal we check whether or not $< (a, b), (c, d) >= ac + bd$ equals zero.

Virginia:

Simplicio:

where $x \in [0, 2\pi]$ and $z = a + bi$ is a point on the unit circle in the complex plane. In particular, the length of $z = \sqrt{a^2 + b^2} = 1$.

Exercise Set 19.8.

19.9 The Discrete Fourier Transform: The Complex Case

Galileo: To begin our discussion of the discrete complex Fourier transform, let us consider the "polynomial" $p_1(z) = c_0 + c_1 z + c_{-1} z^{-1}$ where $z = a + bi$ is a point on the unit circle in the complex plane. (i.e. $a^2 + b^2 = 1$.) In our discussions, we will use the letter i to denote the square root of -1. In particular, $i^2 = -1$. As was the case with previous discussions of interpolation, we have the setting:

Given the data $(z_0, y_0), (z_1, y_1), (z_2, y_2)$,

Find the constants c_0, c_1, c_{-1} so that $p_1(z_0) = y_0, p_1(z_1) = y_1, p_1(z_2) = y_2$.

This problem leads to the matrix equation:

$$\begin{pmatrix} \frac{1}{z_0} & 1 & z_0 \\ \frac{1}{z_1} & 1 & z_1 \\ \frac{1}{z_2} & 1 & z_2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_{-1} \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

As it turns out, a "smart" choice of the points is
$z_0 = 1, z_1 = \frac{-1+\sqrt{3}}{2}, z_2 = \frac{-1-\sqrt{3}}{2}$, which leads to the matrix equation

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & \frac{-1+\sqrt{3}i}{2} & \frac{-1-\sqrt{3}i}{2} \\ 1 & \frac{-1-\sqrt{3}i}{2} & \frac{-1+\sqrt{3}i}{2} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_{-1} \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

Does this equation look familiar?

Simplicio: Sure, but I still don't like those imaginary numbers in there.

Galileo: To ease the pain, how about if think begin by thinking about the geometry associated with Euler's formula.

Virginia: You mean where the variable x denotes an angle between zero and 2π and $e^{ix} = \cos(x) + \sin(x)$ represents the corresponding point on the unit circle.

To begin the discussion, let us consider the function $T_1(x) = \frac{a_0}{2} + a_1 \cos(x) + b_1 \sin(x)$ and the "polynomial" $p_1(z) = c_{-1} z^{-1} + c_0 + c_1 z$, where $x \in [0, 2\pi]$ and $z = a + bi$ is a point on the unit circle in the complex plane. In particular, the length of

$z = \sqrt{a^2 + b^2} = 1$. This setting is virtually identical to the one we gave for polynomial interpolation, where we were given $n+1$ data points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ and were expected to find a polynomial $p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$ with the property that $p_n(x_k) = y_k$ for all $k = 0, 1, 2, \ldots, n$.

Simplicio: Exactly the same? I am suspicious here.

Galileo: Well OK, if the setting were exactly the same, then I would be repeating myself. I certainly wouldn't want to bore you. The difference this time is that we now allow the variable x to be a complex number z.

Virginia: Do we still get to make "smart choices" for the points $x_0, x_1, \ldots, x_n$?

Galileo: Absolutely. However, since these numbers will be complex, we will denote them by the letters z_k. Also, the notation will be a bit easier if we assume we have n points and are interpolating with a polynomial of the form $p_{n-1}(z) = c_0 + c_1 z + \cdots + c_{n-1} z^{n-1}$.

Simplicio: Why did you change the coefficients from a_k to c_k?

Galileo: While it is part of our culture to use the coefficients a_k and b_k in the definition of the function $T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)]$, there is a close connection between these coefficients and the coefficients c_k in the "polynomial" $p_n(z) = \sum_{k=-n}^{n}$.

In particular, if $z = e^{ix}$, where $x \in [0, 2\pi]$, then by Euler's formula $p_n(z) = T_n(x)$ as long as we choose $a_k = c_k$ and $b_k = ic_k$ for all $k = 0, 1, \ldots, n$.

integer n, we will be given let $x_0 = z_0 = 1$, $x_1 = z_1 = \omega = e^{\frac{2\pi i}{n}}$, where $i = \sqrt{-1}$, and $x_k = z_k = \omega^k$ for $k = 1, 2, \ldots, n-1$.

An example of the type of problem we are solving is: Given data points y_0 and y_1, find a polynomial of the form $p_2(z) = c_0 + c_1 z$ such that $p_2(1) = y_0$ and $p_2(\omega) = y_1$. In this simple setting, $\omega = -1$ and we need to solve the matrix equation

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix}.$$

The definition of the Fourier matrix F_2 becomes

$$F_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

For three points, the problem we are solving becomes: Given data points y_0, y_1, and y_2, find a polynomial of the form $p_2(z) = c_0 + c_1 z + c_2 z^2$ such that $p_2(1) = y_0, p_2(\omega) = y_1$, and $p_2(\omega^2) = y_2$, where ω is the cube root of unity defined by $\omega = e^{\frac{2\pi i}{3}}$. In particular, $\omega^3 = 1$.

The matrix equation that must be solved is

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega^4 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

The definition of the Fourier matrix F_3 is given by

$$F_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega^4 \end{pmatrix}.$$

If $n = 4$, then

$$\begin{aligned} \omega &= e^{\frac{2\pi i}{4}} = i, \\ \omega^2 &= -1, \\ \omega^3 &= -\omega = -i, \text{ and} \\ \omega^4 &= 1. \end{aligned}$$

The corresponding Fourier matrix becomes

$$F_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & (\omega^2)^2 & (\omega^3)^2 \\ 1 & \omega^3 & (\omega^2)^3 & (\omega^3)^3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}.$$

Note that the F_2, F_3, and F_4 matrices all have the structure of a Vandermonde matrix. Since each one arose as part of a solution to a problem in polynomial interpolation, this observation is not an accident.

We are now in a position to show that each Fourier matrix F_n has two important properties. First, every pair of columns are orthogonal. Second, each column has length $\sqrt{n}$. Thus, once again we can efficiently compute the Fourier coefficients by simply multiplying both sides of the equation by a matrix A^*, which has the property that A^*A is diagonal. The purpose of the next discussion is to give a careful definition of this new matrix.

We begin with a definition of the Fourier matrix.

Definition 19.9.1. *If n is a positive integer and $\omega = e^{\frac{2\pi i}{n}}$, then the Fourier matrix F_n is defined by the rule:*

$$F_n = \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \dots & \omega^{n-1} \\ 1 & \omega^2 & (\omega^2)^2 & (\omega^3)^2 & \dots & (\omega^{n-1})^2 \\ 1 & \omega^3 & (\omega^2)^3 & (\omega^3)^3 & \dots & (\omega^{n-1})^3 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \omega^{n-1} & (\omega^2)^{n-1} & (\omega^3)^{n-1} & \dots & (\omega^{n-1})^{n-1} \end{pmatrix}.$$

To make the discussion of orthogonality more precise, we need to extend the definition from the domain of vectors in $\Re^n$ to the complex $n-$dimensional space C^n. First, recall the following definitions.

Definition 19.9.2. *If $z = a+bi \in C$, then the conjugate of z is denoted by $\overline{z} = a-bi$.*

Example 19.9.1. *If $z = 3 + 4i$, then $\overline{z} = 3 - 4i$.*

Since an equivalent way to represent a complex number $z = a + bi$ is as a point $z = (a, b)$, we can graph any complex number in the plane. Note that the graph of the complex conjugate $\overline{z} = (a, -b)$ is on the opposite side of the $x-$axis (or line $y = 0$.) from z.

Example 19.9.2. *If*

$$z = \begin{pmatrix} 2+i \\ 3 \\ 5-2i \end{pmatrix}$$

Definition 19.9.3. *If*

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{24} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & a_{34} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & a_{m4} & \dots & a_{mn} \end{pmatrix},$$

then the conjugate of A is the matrix

$$\overline{A} = \begin{pmatrix} \overline{a_{11}} & \overline{a_{12}} & \overline{a_{13}} & \overline{a_{14}} & \dots & \overline{a_{1n}} \\ \overline{a_{21}} & \overline{a_{22}} & \overline{a_{23}} & \overline{a_{24}} & \dots & \overline{a_{2n}} \\ \overline{a_{31}} & \overline{a_{32}} & \overline{a_{33}} & \overline{a_{34}} & \dots & \overline{a_{3n}} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ \overline{a_{m1}} & \overline{a_{m2}} & \overline{a_{m3}} & \overline{a_{m4}} & \dots & \overline{a_{mn}} \end{pmatrix}.$$

Definition 19.9.4. *If $A \in C^{m\times n}$, then the adjoint of A is defined by $A^* = \overline{A^t}$.*

Definition 19.9.5. *If z_1 and z_2 are complex column vectors in C^n, then the inner product is defined by $< z_1, z_2 >= z_1^* z_2$.*

Example 19.9.3. *If*

$$z_1 = \begin{pmatrix} 2+i \\ 3 \\ 5-2i \end{pmatrix} \text{ and } z_2 = \begin{pmatrix} 7+i \\ 11-5i \\ 13-8i \end{pmatrix}$$

then

$$\begin{aligned} < z_1, z_2 > &= \overline{(2+i)}(7+i) + \overline{3}(11-5i) + \overline{(5-2i)}(13-8i) \\ &= (2-i)(7+i) + 3(11-5i) + (5+2i)(13-8i) \\ &= 129 - 34i. \end{aligned}$$

Definition 19.9.6. *If z_1 and z_2 are complex column vectors in C^n, then they are orthogonal if $< z_1, z_2 >= 0$.*

Definition 19.9.7. *If z is a complex column vector in C^n, then the 2-norm of z is given by $|z| = \sqrt{< z, z >}$.*

Note that the 2-norm of a complex vector simply represents its length.

Proposition 19.9.8. *If n is a positive integer and $\omega = e^{\frac{2\pi i}{n}}$, then $1 + \omega + \omega^2 + \ldots + \omega^{n-1} = 0$.*

Proof. Since $\omega^n = 1$, we know by the formula for the geometric series that $1 + \omega + \omega^2 + \ldots + \omega^{n-1} = \frac{1-\omega^n}{1-\omega} = 0$. □

Proposition 19.9.9 (Complex Fourier: Orthogonality and Equal Lengths). *The columns of the Fourier matrix F_n are pairwise orthogonal and the matrix $D = F_n^* F_n$ is a diagonal matrix with each entry on the diagonal equal to the integer n. In particular, the 2-norm of each column of F_n is $\sqrt{n}$.*

Proof. This proposition follows immediately from the assumption that $\omega^n = 1$, the fact that $\overline{\omega}\omega = 1$, and the previous proposition. □

Proposition 19.9.10. *The inverse of the Fourier matrix F_n is the matrix $\frac{1}{n}F_n^*$.*

Proof. This fact follows immediately from the previous proposition. □

Theorem 19.9.11 (Complex Fourier: Coefficient Formulas). *If $y_0, y_1, \ldots, y_{n-1}$ is a given set of data and $\omega = e^{\frac{2\pi i}{n}}$, then the coefficients of the polynomial $p_{n-1}(z) = c_0 + c_1 z + \ldots + c_{n-1} z^{n-1}$ with the property that $p_{n-1}(\omega^k) = y_k$ for all $k = 0, 1, \ldots, n-1$ are $c_k = \frac{1}{n}\sum_{j=0}^{n-1} y_j \overline{\omega}^{jk}$, for $k = 0, 1, \ldots, n-1$.*

Proof. The matrix equation that must be solved is $F_n c = y$, where $c = (c_0, c_1, \ldots, c_{n-1})^t$ and $y = (y_0, y_1, \ldots, y_{n-1})^t$. Since the columns of F_n are pairwise orthogonal and all have 2-norm equal to $\sqrt{n}$, $c_k = \frac{1}{n} < \omega_k, y >$, where ω_k is the k^{th} column of F_n. Since $< \omega_k, y >= \sum_{j=0}^{n-1} y_j \overline{\omega}^{jk}$, we are done. □

Theorem 19.9.12 (Complex Parseval/Pythagoras). *If $y_0, y_1, \ldots, y_{n-1}$ is a given set of data in C, $\omega = e^{\frac{2\pi i}{n}}$, $p_{n-1}(z) = c_0 + c_1 z + \ldots + c_{n-1} z^{n-1}$ with $c_k \in C$ for all $k = 0, 1, \ldots, n-1$ is a polynomial with the property that $p_{n-1}(\omega^k) = y_k$ for all $k = 0, 1, \ldots, n-1$, then*

$$\sum_{j=0}^{n-1} y_j^2 = \frac{1}{n}(|c_0|^2 + |c_1|^2 + \ldots + |c_{n-1}|^2).$$

Exercise Set 19.9.

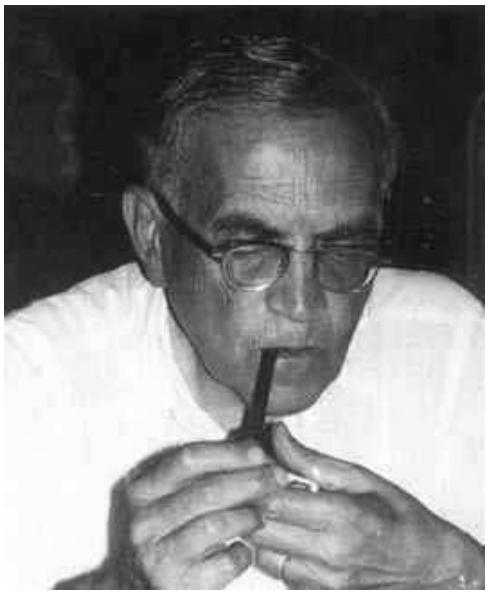
1. If
$$z = \begin{pmatrix} 2+i \\ 3 \\ 5-2i \end{pmatrix},$$
then compute the 2-norm of the vector z.

2. If $\omega = e^{\frac{2\pi i}{3}}$ and $y_0 = 2, y_1 = 3, y_2 = 5$ are given points, then find constants c_0, c_1, c_2 such that the polynomial $p_2(z) = c_0 + c_1 z + c_2 z^2$ has the property that $p_2(1) = y_0, p_2(\omega) = y_1$, and $p_2(\omega^2) = y_2$.

3. If $\omega = e^{\frac{2\pi i}{3}}$ and y_0, y_1, y_2 are given points, then find constants c_0, c_1, c_2 such that the polynomial $p_2(z) = c_0 + c_1 z + c_2 z^2$ has the property that $p_2(1) = y_0, p_2(\omega) = y_1$, and $p_2(\omega^2) = y_2$.

Chapter 20

Cubic Spline Interpolation



Isaac Schoenberg

Galileo: The idea behind this next discussion is to show that polynomials can be of great use as long as you make an effort to control them. These ideas were first developed by Romanian born Isaac Schoenberg (1903-1990), who is recognized as the inventor of splines. While he was more interested in their use in theoretical mathematics, they now play a fundamental role in numerous applications including

data fitting, computer graphics, and computer-aided design. The primary reason spline techniques are used in so many different real-world applications is that they are stable.

Simplicio: I am not sure of the meaning of the word stable when used in this context.

Galileo: Hopefully you remember our discussion of the Runge example, where the smooth function $f(x) = \frac{1}{1+25x^2}$ was approximated by interpolating polynomials on the interval $[-1, 1]$. As the degree of the polynomial was increased, the accuracy of the approximation became worse.

Simplicio: Oh yes, the approximation was particularly poor at the endpoints.

Galileo: The good news is that with splines that type of problem will never occur.

Simplicio: Sounds good.

Galileo: While a multitude of different kinds of splines have been devised, we will consider only four different types: B-splines, clamped, free, and periodic. Note that some researchers refer to clamped splines as "complete splines" and some people call free splines "natural."

While each of the different types of splines have their uses, periodic splines are particularly useful because they can be used to construct digital contours in the plane passing through a given finite set of points. These contours will be smooth and thus not contain any sharp corners.

Simplicio: Where do we begin?

Galileo: For several reasons the class of piecewise linear functions provide a natural entry point into the discussion of splines. The first reason is a pedagogical issue. Namely, the idea of a spline is most accessible if the construction of a piecewise approximation is well understood. The second reason is that the error formula for piecewise linear approximation is not only useful by itself, but also provides a key piece of information used in the proof of the convergence formula for clamped splines.

The theory for clamped splines turns out to be special in a number of ways. First, the clamped spline is characterized as the smooth interpolating function with the "fewest number of oscillations." This idea can be formulated mathematically as

an integral of the second derivative squared. This integral can be thought of as a measure of the "energy" of the function. The less energy in the function, the fewer oscillations. The clamped cubic spline is the smooth interpolant, which minimizes this energy function. Remarkably, integration by parts and a modern version of the Pythagorean Theorem are used in the proof of this minimization theorem.

Simplicio: Why should I care about the integral of the square of the second derivative?

Galileo: While the minimization theorem provides enjoyable reading for a mathematician, an engineer is more likely to be interested by the high convergence rates provided by clamped cubic splines. For most situations, the convergence rate is 4^{th}-order for the interpolants and 2^{nd}-order for the second derivatives. As a demonstration of the power of the method, we will apply splines to the function $f(x) = \frac{1}{1+25x^2}$ and show that not only does the sequence of splines converge to the function, but the sequence of second derivatives converge as well.

Simplicio: What are the differences between these splines?

Galileo: To avoid technical difficulties, we will limit our discussion to the setting where the partition has equally spaced points. When we make this assumption, the interpolant can be written as a linear combination of functions which are formed as translated and scaled versions of a single standard spline function. In all four cases, the constants can be found by solving a system of equations, where the coefficient matrix has a special easy to understand form. The coefficient matrix for the B-spline interpolant has $1's$ on the diagonal and $\frac{1}{4}'s$ on the super and sub-diagonals. Every other entry in the matrix is zero. The matrices for the other three types of splines are minor variations of this one.

The B-spline interpolation technique is the easiest to explain because no discussion of the endpoints is required. The other three types of splines are the same as B-splines except that additional restrictions are placed on the endpoints of the interval. For the clamped spline, the first derivatives of the interpolant $S_C(x)$ are forced to be equal to preset values at the two endpoints. Thus, $S'_C(a) = y'_0$ and $S'_C(b) = y'_n$, where y'_0 and y'_n are given values. For the free spline, the second derivatives of the interpolant

$S_F(x)$ are set equal to zero at the two endpoints. Thus, $S_F'(a) = 0$ and $S_F'(b) = 0$. For periodic splines, the interpolant $S_P(x)$ is forced to have the property that derivatives agree at the endpoints. Thus, $S_P(a) = S_P(b), S_P'(a) = S_P'(b)$, and $S_P''(a) = S_P''(b)$.

Simplicio: While this all sounds interesting, I am not sure I am a believer yet.

Galileo: It took Schoenberg 20 years to get people to pay attention to what he was doing. However, with the advent of the computer in the early 1960's interest skyrocketed because engineers found them useful in a multitude of applications.

20.1 Piecewise Linear Interpolation

Galileo: Even though the focus of this section is on splines, we begin with a discussion of linear interpolation. While we could have presented this material earlier, it provides an excellent introduction into the ideas and convergence theorems we will encounter for clamped splines.

We begin our discussion with a brief review of some notation and a brief introduction to some new notation.

Let $P = \{a = x_0 < x_1 < \ldots x_n = b\}$ denote a fixed partition of $[a, b]$.

Definition 20.1.1. *If $P = \{a = x_0 < x_1 < \ldots x_n = b\}$ is a partition, then the mesh of P is defined by $\|P\| = \max\{x_{i+1} - x_i : i = 0, 1, \ldots, n-1\}$.*

If P is a partition of $[a, b]$, then let $C^P[a, b]$ denote the set of all continuous functions on $[a, b]$ which are linear on each segment $[x_i, x_{i+1}]$. This collection of functions will be referred to as the piecewise linear functions.

Definition 20.1.2. *A Piecewise Linear Bump or Chapeau function is defined by*

$$B_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & x \in [x_i, x_{i+1}] \\ 0 & \textit{otherwise.} \end{cases}$$

Note that the functions $B_i(x)$ are continuous on $[a,b]$. In fact, these functions are zero from $-\infty$ to x_{i-1}, a straight line with positive slope from x_{i-1} to x_i, a straight line with negative slope from x_i to x_{i+1}, and zero from x_{i+1} to ∞.

Proposition 20.1.3. *If a function $f(x)$ is defined on an interval $[a,b]$ and $P = \{a = x_0 < x_1 < \ldots x_n = b\}$ is a partition of $[a,b]$, then the piecewise linear function $I_f(x) = \sum_{k=0}^{n-1} f(x_k) \cdot B_k(x)$ has the property that $I_f(x_k) = f(x_k)$ for all $k = 0, 1, \ldots, n$.*

Proof. Note that $B_i(x_j) = \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$ □

Note that the function $I_f(x) = \sum_{k=0}^{n-1} f(x_k) \cdot B_k(x)$ is called the piecewise linear approximation of $f(x)$. Since any piecewise linear function $\phi(x)$ in the collection $C^P[a,b]$ can be written in the form $\phi(x) = \sum_{k=0}^{n-1} \phi(x_k) \cdot B_k(x)$, the set of functions $\{B_k : k = 0, 1, \ldots, n-1\}$ form a basis for the set of all functions in $C^P[a,b]$. This set of functions is comparable to the functions $L_k(x)$, which were used in the Lagrange method for polynomial interpolation. In particular, note that the Chapeau function $B_k(x_j)$ is equal to zero at all points x_j, where $j \neq k$.

Definition 20.1.4. *If a function $f(x)$ is defined on an interval $[a,b]$, then the $\infty-$norm (or sup norm) of $f(x)$ is defined by the rule $\|f\|_\infty = max\{|f(x)| : x \in [a,b]\}$.*

More intuitively, $\|f\|_\infty$ is the maximum value of $|f(x)|$ on the interval $[a,b]$.

Proposition 20.1.5. *If $f(x) \in C^2[a,b]$ and $f(a) = f(b) = 0$, then $\|f\|_\infty \leq \frac{1}{8}\|f''\|_\infty$.*

Proof. By the Lagrange error formula there is a first degree polynomial $p_1(x)$ such that for every $x \in [a,b]$ there is a z such that

$$f(x) = p_1(x) + \frac{f''(z)}{2}(x-a)(x-b).$$

Since $f(a) = f(b) = 0$, $p_1(x) = 0$ for all $x \in [a,b]$. Thus, there is a point $z \in [a,b]$ such that

$$f(x) = \frac{f''(z)}{2}(x-a)(x-b).$$

But the extreme (i.e. minimum) value of parabola $(x-a)\cdot(x-b)$ occurs at the point $\frac{a+b}{2}$ so that for all $x \in [a,b]$

$$\begin{aligned} |f(x)| &\leq \frac{\|f''\|_\infty}{2}\left(\frac{b-a}{2}\right)^2 \\ &\leq \frac{\|f''\|_\infty}{8}\cdot(b-a)^2. \end{aligned}$$

Therefore, $\|f\|_\infty \leq \frac{\|f''\|_\infty}{8}\cdot(b-a)^2$. □

The next corollary is a precise statement that as the partition is refined to have a smaller mesh size, the piecewise linear interpolants will converge to the given function. Even more important is the fact that the convergence rate is quadratic.

Corollary 20.1.6. *Let $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ be a partition of $[a,b]$. If $f \in C^2[a,b]$, then the interpolating function $I_f(x) = \sum_{k=0}^{n} f(x_k)\cdot B_k(x)$ has the property that*

$$\|f - I_f\|_\infty \leq \frac{\|P\|^2}{8}\cdot\|f''\|_\infty.$$

Proof. The proof of this corollary follows immediately from the application of the previous proposition applied to each interval $[x_k, x_{k+1}]$. □

While the next corollary is an immediate consequence of the previous one, it will be used as one of the key steps in the proof of convergence for the clamped cubic splines.

Theorem 20.1.7 (Error Theorem For Piecewise Linear Approximation). *Let $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ be a partition of $[a,b]$. If $f \in C^4[a,b]$, then the interpolating function $I_{f''}(x) = \sum_{k=0}^{n} f''(x_k)\cdot B_k(x)$ has the property that $\|f'' - I_{f''}\|_\infty \leq \frac{\|P\|^2}{8}\cdot\|f^{(4)}\|_\infty$.*

Proof. Simply replace the function $f(x)$ by the function $f(x)'' - I_{f''}(x)$ in the previous proposition. □

Exercise Set 20.1.

1. Given that the $\cos(23) = 0.92050485345244$ and $\cos(24) = 0.91354545764260$, what is the piecewise linear approximation of $\cos(23.56)$? Use your calculator to check that $\cos(23.56) = 0.91664200257852$. How does the difference between these two numbers compare with the estimate provided by the Error Theorem For Piecewise Linear Approximation?

2. If $f(x) = \cos(x)$ for $x \in [-\pi, \pi]$ and $tol = \frac{1}{10^5}$, then how many equally spaced points will be required to guarantee that the piecewise linear approximation $I_f(x)$ will approximate $\cos(x)$ with error less than $\frac{1}{10^5}$ for all $x \in [-\pi, \pi]$?

3. If $f(x) = \frac{1}{1+25x^2}$ for $x \in [-1, 1]$ and $tol = \frac{1}{10^5}$, then how many equally spaced points will be required to guarantee that the piecewise linear approximation $I_f(x)$ will approximate $f(x)$ with error less than $\frac{1}{10^5}$ for all $x \in [-1, 1]$?

20.2 Cubic B-Spline Interpolation

Galileo: The most straight forward path to understanding splines is through the study of standard "bumps." The first standard bump is the piecewise linear Chapeau function from the previous section.

Definition 20.2.1. *The standard piecewise linear bump is defined by the following rules:*

$$B(x) = \begin{cases} 0 & x \leq -1 \\ x+1 & x \in [-1, 0] \\ 1-x & x \in [0, 1] \\ 0 & x \geq 1. \end{cases}$$

A graph of this function is displayed in Figure 20.1.

Galileo: Note that if the points in a partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ are equally spaced, then the functions $B_i(x)$ defined in the previous section can be defined by the formulas $B_i(x) = B(\frac{x-x_i}{h})$, where $h = \frac{b-a}{n}$. In other words, the function

$B_i(x)$ is nothing more than a translation by x_i and a stretch by h of the Standard Chapeau function $B(x)$. Thus, any continuous function $f(x)$ can be approximated by linear combinations of translations and stretches of $B(x)$.

Since numerous applications (e.g. computer graphics) require the use of smooth curves rather than curves with sharp corners, these functions are not always appropriate. However, the same concepts can be translated into the domain of smooth approximations. The only technical difficulty is to create a smooth bump. The next definition provides the formulas needed for the standard cubic spline bump. The graph of this function is presented in Figure 20.2.

Definition 20.2.2. *The standard spline bump is a piecewise cubic polynomial defined by the following rules:*

$$S(x) = \begin{cases} 0 & x \leq -2 \\ \frac{1}{4}[(2-x)^3 - 4(1-x)^3 - 6x^3 + 4(1+x)^3] & x \in [-2,-1] \\ \frac{1}{4}[(2-x)^3 - 4(1-x)^3 - 6x^3] & x \in [-1,0] \\ \frac{1}{4}[(2-x)^3 - 4(1-x)^3] & x \in [0,1] \\ \frac{1}{4}(2-x)^3 & x \in [1,2] \\ 0 & x \geq 2. \end{cases}$$

Proposition 20.2.3. *The standard spline bump $S(x)$ has the property that it is in*

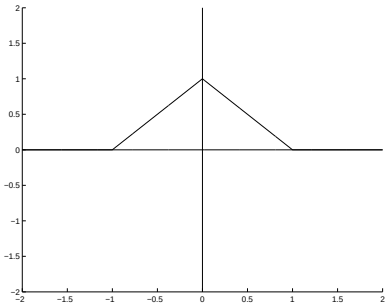


Figure 20.1: The Graph of the Standard Chapeau PWL Bump $B(x)$

$C^2(-\infty,\infty)$. *In particular,* $S(x), S'(x)$, *and* $S''(x)$ *are all continuous for all* $x \in (-\infty,\infty)$. *Moreover,*

1. $S(0) = 1, S(\pm 1) = \frac{1}{4}$, *and* $S(\pm 2) = 0$,
2. $S'(-1) = \frac{3}{4}, S'(1) = -\frac{3}{4}, S'(-2) = S'(0) = S'(2) = 0$, *and*
3. $S''(-2) = S''(2) = 0, S''(-1) = S''(1) = \frac{3}{2}$, *and* $S''(0) = -3$.

Proof. Since $S(x)$ is a cubic polynomial at all points $x \in (-\infty,\infty)$ except where two polynomials join. Thus, we only need to check continuity at the five points $x = -2, -1, 0, 1$, and 2. However, since $S(\pm 2) = 0, S(\pm 1) = \frac{1}{4}$, and $S(0) = 1$ whether computed by the formula on the left side or right side of the possible trouble spot, the function is continuous.

The first derivative of $S(x)$ is given by the rules:

$$S'(x) = \begin{cases} 0 & x \le -2 \\ \frac{1}{4}[-3(2-x)^2 + 12(1-x)^2 - 18x^2 + 12(1+x)^2] & x \in [-2,-1] \\ \frac{1}{4}[-3(2-x)^2 + 12(1-x)^2 - 18x^2] & x \in [-1,0] \\ \frac{1}{4}[-3(2-x)^2 + 12(1-x)^2] & x \in [0,1] \\ -\frac{3}{4}(2-x)^2 & x \in [1,2] \\ 0 & x \ge 2. \end{cases}$$

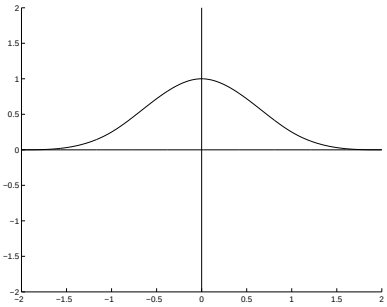


Figure 20.2: The Graph of the Standard Spline Bump $S(x)$

Again, we only need to check the five possible trouble spots, where the quadratic polynomials are joined. However, $S'(-1) = \frac{3}{4}, S'(-2) = S'(0) = S'(2) = 0$, and $S'(1) = -\frac{3}{4}$. Thus, the function $S'(x)$ is continuous for each $x \in (-\infty, \infty)$.

The second derivative of $S(x)$ is given by the rules:

$$S''(x) = \begin{cases} 0 & x \leq -2 \\ \frac{1}{4}[6(2-x) - 24(1-x) - 36x + 24(1+x)] & x \in [-2,-1] \\ \frac{1}{4}[6(2-x) - 24(1-x) - 36x] & x \in [-1,0] \\ \frac{1}{4}[6(2-x) - 24(1-x)] & x \in [0,1] \\ \frac{6}{4}(2-x) & x \in [1,2] \\ 0 & x \geq 2. \end{cases}$$

For the second derivative $S''(x)$, the calculations are $S''(-2) = S''(2) = 0$, $S''(-1) = S''(1) = \frac{3}{2}$, and $S''(0) = -3$.

Thus, the function $S''(x)$ is continuous for each $x \in (-\infty, \infty)$. □

Simplicio: How would anyone think up those weird formulas?

Galileo: In any research project, one of the key ingredients is to ask the right questions. The best questions are straightforward to understand, but whose answers provide insight beyond the stated question. To answer your immediate question, the function $S(x)$ equals the convolution of $B(x)$ with itself.

Simplicio: The word convolution means nothing to me.

Galileo: The convolution of two functions is a fancy word for the integration of their product–in a very particular way. Not only does this idea provide solutions to a number of questions in differential equations, but also occurs whenever filtering is discussed in signal processing and image processing. I have placed the topic on the agenda for a meeting in the not-to-distant future. In any case, we now give the formal definition.

Definition 20.2.4. *If $f(x)$ and $g(x)$ are continuous functions on $(-\infty, \infty)$ with the property that $\int_{-\infty}^{\infty} f(x)^2\, dx < \infty$ and $\int_{-\infty}^{\infty} g(x)^2\, dx < \infty$, then the convolution of $f(x)$*

and $g(x)$ is given by the formula

$$f * g(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\ dt.$$

Galileo: As it turns out, the bump $S(x)$ is the convolution of $B(x)$ with itself. If we also know that the operation of convolution tends to make a function smoother, then we are closer to the fact that the function $S(x)$ has continuous first and second derivatives.

Simplicio: I bet the computation is messy.

Galileo: Maybe so, but the computation can be visualized as simply dragging one copy of $B(x)$ across another. When the bumps are disjoint, the integrals are zero. As they begin to intersect, we are integrating the product of two straight lines so the answer is a cubic polynomial.

Simplicio: And if we would like to construct a piecewise linear 5^{th} degree polynomial bump, then we simply convolve the functions $B(x)$ and $S(x)$ to create a function which has continuous first, second, third, and fourth derivatives. Is that not correct?

Galileo: You have the picture.

Simplicio: But is there method a with easier formulas?

Galileo: Actually, some researchers use the piecewise 6^{th} degree polynomial:

$$C(x) = \begin{cases} (x-1)^3(x+1)^3 & \text{if } x \in [-1,1] \\ 0 & \text{if } |x| \geq 1 \end{cases}$$

Galileo: However, the more popular method is the one we described. Since the first and second derivatives of $S(x)$ play an important role in both the theory and application of splines, we present their graphs in Figures 20.3 and 20.4.

Galileo: We now turn to the problem of constructing the $B-$spline from building blocks provided by the spline bump $S(x)$. Given a partition of equally spaced points $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, the second step is to translate and stretch the standard bump $n+1$ times so that the k^{th} bump, $S_k(x)$, has center x_k and equals zero outside the interval $[x_{k-2}, x_{k+2}]$.

Given data points (x_k, y_k) $k = 0, 1, 2, 3, \ldots, n$, where $x_{k+1} - x_k = h$ for all $k = 0, 1, 2, \ldots, n$, the goal now is to find constants c_k, where $k = 0, 1, 2, \ldots, n$ so that $S_B(x) = \sum_{k=0}^{n} c_k S(\frac{x-x_k}{h})$ has the property that it interpolates the data. In particular, we insist that $S_B(x_k) = y_k$ for all $k = 0, 1, \ldots, n$.

If we let $S_k(x) = S(\frac{x-x_k}{h})$, then we can write $S_B(x) = \sum_{k=0}^{n} c_k S_k(x)$. As was the case for both polynomial and Fourier interpolation the constants c_k can be found by solving the matrix equation $\mathbf{S}_B\mathbf{c} = \mathbf{y}$, where

$$\mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}, \text{ and}$$

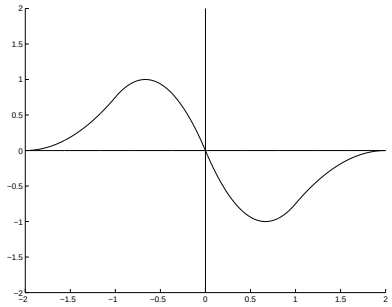


Figure 20.3: The Graph of $S'(x)$

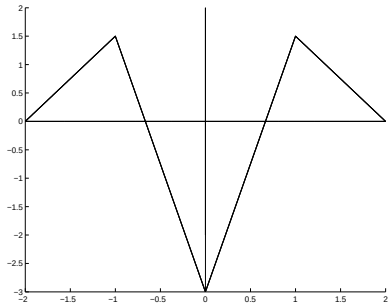


Figure 20.4: The Graph of $S''(x)$

$$\mathbf{S}_B = \begin{pmatrix} 1 & \frac{1}{4} & 0 & 0 & 0 & \dots & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & 1 & \frac{1}{4} & 0 \\ \vdots & & & & \frac{1}{4} & 1 & \frac{1}{4} \\ 0 & \dots & \dots & \dots & 0 & \frac{1}{4} & 1 \end{pmatrix}.$$

The constants 1 and $\frac{1}{4}$ in the $(n+1) \times (n+1)$ matrix $\mathbf{S}_B$ are forced by the relationships $S_k(x_k) = S(0) = 1$ and $S_k(x_{k+1}) = S_k(x_{k-1}) = S(\pm 1) = \frac{1}{4}$, respectively. The zero entries in the matrix follow from the fact that $S_k(x) = 0$ for all x such that $|x - x_k| \geq 2h$.

The beauty of the matrix $\mathbf{S}_B$ is that it is tridiagonal, diagonally dominant, symmetric, and well-conditioned [1].

Exercise Set 20.2.

1. Given the data $(0,2),(1,2),(2,2),(3,2),(4,2)$, set up the matrix equation that must be solved to compute the constants for the B-spline interpolation function $S_B(x)$. Use computer software to compute the constants c_0, c_1, c_2, c_3, c_4. Use these constants to compute $S_B(x)$ for $x = 1, 5, 7$.

2. Compute the LU factorization of the 5×5 spline matrix $\mathbf{S}_B$. How would you use this factorization to write efficient code to solve the matrix equation $\mathbf{S}_B\mathbf{c} = \mathbf{y}$?

20.3 Clamped Cubic Spline Interpolation

Galileo: We now turn to the problem of clamped cubic spline interpolation. In this application, we again have a partition of equally spaced points $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, where $x_{k+1} - x_k = h$ for all $k = 0, 1, 2, \dots, n-1$. As before, we are also given data points (x_k, y_k) for $k = 0, 1, 2, \dots, n$. The difference this time consists

of two new pieces of information, y_0' and y_n', constraining the value of the derivative at the two endpoints. In particular, we require that the interpolating function $S_C(x)$ has the property that $S_C'(a) = S_C'(x_0) = y_0'$ and $S_C'(b) = S_C'(x_n) = y_n'$. Since we have two new constraints, we must have two new free variables. The trick is to simply add two new bumps.

Simplicio: But, where are you going to add them?

Galileo: Simply add one at each end of the interval $[a, b]$. Since the points in P are assumed to be equally spaced, we simply add the points x_{-1} and x_{n+1} to the partition so that $x_0 - x_{-1} = h$ and $x_{n+1} - x_n = h$. We now have to solve $n+3$ equations and $n+3$ unknowns for $c_{-1}, c_0, c_1, \ldots, c_n, c_{n+1}$. The new interpolant is defined by the linear combination $S_C(x) = \sum_{k=-1}^{n+1} c_k S(\frac{x-x_k}{h})$.

The two new constraints are forced by the equations $S_C'(x_0) = y_0'$ and $S_C'(x_n) = y_n'$. But if $x = x_0$, then

$$S_C'(x_0) = c_{-1}S_{-1}'(x_0) + c_0 S_0'(x_0) + c_1 S_1'(x_0) = c_{-1}\frac{3}{4h} + 0 + c_1\frac{-3}{4h} = y_0'.$$

If $x = x_n$, then

$$S_C'(x_n) = c_{n-1}S_{n-1}'(x_n) + c_n S_n'(x_n) + c_{n+1}S_{n+1}'(x_n) = c_{n-1}\frac{3}{4h} + 0 + c_{n+1}\frac{-3}{4h} = y_n'.$$

Solving the first equation for c_{-1} and the second for c_{n+1}, we find that $c_{-1} = \frac{4h}{3}y_0' + c_1$ and $c_{n+1} = \frac{4h}{3}y_n' + c_{n-1}$. Thus, the two new equations become:

$$c_0 + \frac{1}{2}c_1 = y_0 - \frac{h}{3}y_0'$$

and

$$\frac{1}{2}c_{n-1} + c_n = y_n - \frac{h}{3}y_n'.$$

Thus, the modified system of equations we need to solve becomes $\mathbf{S}_C\mathbf{c} = \mathbf{yy}$, where

$$\mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix}, \ \mathbf{yy} = \begin{pmatrix} y_0 - \frac{h}{3}y_0' \\ y_1 \\ \vdots \\ y_n - \frac{h}{3}y_n' \end{pmatrix}, \text{ and}$$

$$\mathbf{S}_C = \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 & \dots & \dots & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & 1 & \frac{1}{4} & 0 \\ \vdots & & & & \frac{1}{4} & 1 & \frac{1}{4} \\ 0 & \dots & \dots & \dots & 0 & \frac{1}{2} & 1 \end{pmatrix}.$$

Again, the constants 1 and $\frac{1}{4}$ in the $(n+1) \times (n+1)$ matrix $\mathbf{S}_C$ are forced by the relationships $S_k(x_k) = S(0) = 1$ and $S_k(x_{k+1}) = S_k(x_{k-1}) = S(\pm 1) = \frac{1}{4}$, respectively. The zero entries in the matrix come from the fact that $S_k(x) = 0$ for all x such that $|x - x_k| \geq 2h$.

As before, the matrix $\mathbf{S}_C$ is tridiagonal, diagonally dominant, almost symmetric, and well-conditioned. Again, the LU-factorization can be used to solve the matrix equation $\mathbf{S}_C\mathbf{c} = \mathbf{yy}$.

Exercise Set 20.3.

1. Given the data $(0,2), (1,2), (2,2), (3,2), (4,2), y_0' = 3$ and $y_4' = 7$, set up the matrix equation that must be solved to compute the constants for the clamped cubic spline interpolation function $S_C(x)$. Use computer software to compute the constants c_0, c_1, c_2, c_3, c_4. Use these constants to compute $S_C(x)$ for $x = 1, 5, 7$.

2. Compute the LU factorization of the 5×5 clamped spline matrix $\mathbf{S}_C$. How would you use this factorization to write efficient code to solve the matrix equation $\mathbf{S}_C\mathbf{c} = \mathbf{yy}$. How does this factorization compare with the factorization of the 5×5 matrix S_B?

20.4 Natural Cubic Spline Interpolation

Galileo: We now turn to the problem of natural cubic spline interpolation. Sometimes this type of spline is referred to as a free spline. In this application, we will again

assume we have been given data points (x_k, y_k) for $k = 0, 1, 2, \ldots, n$, where the partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ has equally spaced points and $x_{k+1} - x_k = h$ for all $k = 0, 1, 2, \ldots, n-1$. While the spirit is the same as clamped splines, the strategy this time is to simply define the endpoint conditions by the rules: $y_0'' = 0$ and $y_n'' = 0$.

Simplicio: Why would you make this assumption?

Galileo: You may not have any information on the first derivatives and yet you may want to temper the behavior at the endpoints.

Simplicio: May I guess that you simply add two new bumps, which provide two new free variables?

Galileo: Exactly! If we let $S_N(x) = \sum_{k=-1}^{n+1} c_k S(\frac{x-x_k}{h})$, then we can create two new constraints: $S_N''(x_0) = y_0'' = 0$ and $S_N''(x_n) = y_n'' = 0$. These constraints provide us with two new endpoint conditions.

First, if $x = x_0$, then

$$S_N''(x_0) = c_{-1}S_{-1}''(x_0) + c_0 S_0''(x_0) + c_1 S_1''(x_0) = c_{-1}\frac{3}{2h^2} - c_0\frac{3}{h^2} + c_1\frac{3}{2h^2} = y_0'' = 0.$$

Second, it $x = x_n$, then

$$S_N'(x_n) = c_{n-1}S_{n-1}''(x_n) + c_n S_n''(x_n) + c_{n+1}S_{n+1}''(x_n) = c_{n-1}\frac{3}{2h^2} - c_n\frac{3}{h^2} + c_{n+1}\frac{3}{2h^2} = y_n'' = 0.$$

These equations simplify to the following:

$$\begin{aligned} c_{-1} - 2c_0 + c_1 &= 0 \\ c_{n-1} - 2c_n + c_{n+1} &= 0. \end{aligned}$$

Solving for the variables c_{-1} and c_{n+1}, we immediately see that $c_{-1} = 2c_0 - c_1$ and $c_{n+1} = 2c_n - c_{n-1}$.

Thus, the matrix equation becomes: $\mathbf{S}_N\mathbf{c} = \mathbf{y}$, where

$$\mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}, \text{ and}$$

$$\mathbf{S}_N = \begin{pmatrix} \frac{3}{2} & 0 & 0 & 0 & \dots & \dots & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & 1 & \frac{1}{4} & 0 \\ \vdots & & & & \frac{1}{4} & 1 & \frac{1}{4} \\ 0 & \dots & \dots & \dots & 0 & 0 & \frac{3}{2} \end{pmatrix}.$$

Simplicio: Let me finish your thoughts by saying that the matrix $\mathbf{S}_N$ is tridiagonal, diagonally dominant, almost symmetric, and well-conditioned. Again, the LU-factorization can be used to solve the matrix equation $\mathbf{S}_N\mathbf{c} = \mathbf{y}$.

Exercise Set 20.4.

1. Given the data $(0, 2), (1, 2), (2, 2), (3, 2), (4, 2)$, set up the matrix equation that must be solved to compute the constants for the natural spline interpolation function $S_N(x)$. Use computer software to compute the constants c_0, c_1, c_2, c_3, c_4. Use these constants to compute the value of $S_N(x)$ for $x = 1, 5, 7$.

2. Compute the LU factorization of the 5×5 natural spline matrix $\mathbf{S}_N$. How does this factorization compare with the factorizations for the 5×5 matrices $\mathbf{S}_B$ and $\mathbf{S}_C$? How would you use this factorization to write efficient code to solve the matrix equation $\mathbf{S}_N\mathbf{c} = \mathbf{y}$?

20.5 Periodic Cubic Spline Interpolation

Galileo: We now turn to the problem of periodic cubic spline interpolation for equally spaced points. If you understood what we did before, this discussion will only take a

minute.

Simplicio: I think I have time in my schedule for this item.

Galileo: To continue, when we are given a data set (x_k, y_k) for $k = 0, 1, 2, \ldots, n$, we assume the data is smoothly periodic. Thus, we not only assume $y_n = y_0$, but also that $y'_n = y'_0$ and $y''_n = y''_0$. A moments reflection makes us realize that the data can be thought of as continuing from $-\infty$ to ∞ so the interpolating function has the form $S_P(x) = \sum_{k=-\infty}^{\infty} c_k S(\frac{x-x_k}{h})$. Better yet, the solution will have the property that $c_k = c_{n+k}$ for any integer k.

Simplicio: But won't we be adding up an infinite number of numbers?

Galileo: Not really, because for any given x only a finite number of integers k exist with the property that $S(\frac{x-x_k}{h}) \neq 0$. In fact, if $S(\frac{x-x_k}{h}) \neq 0$ for some k, then $S(\frac{x-x_j}{h}) = 0$ for all $j \geq k+4$ and all $j \leq k-4$. Since $c_{-1} = c_{n-1}$ and $c_0 = c_n$, the matrix equation becomes $\mathbf{S}_P\mathbf{c} = \mathbf{y}$, where

$$\mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{pmatrix}, \text{ and}$$

$$\mathbf{S}_P = \begin{pmatrix} 1 & \frac{1}{4} & 0 & 0 & \ldots & \ldots & \frac{1}{4} \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & \ldots & \ldots & 0 \\ 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & 1 & \frac{1}{4} & 0 \\ \vdots & & & & \frac{1}{4} & 1 & \frac{1}{4} \\ \frac{1}{4} & \ldots & \ldots & \ldots & 0 & \frac{1}{4} & 1 \end{pmatrix}.$$

Simplicio: In other words, you wrapped the data around from beginning to end and threw away a bump because the data at x_0 equals the data at x_n.

Galileo: Exactly.

Exercise Set 20.5.

1. Given the data $(0,2),(1,2),(2,2),(3,2),(4,2)$, set up the matrix equation that must be solved to compute the constants for the periodic spline interpolation function $S_N(x)$. Use computer software to compute the constants c_0, c_1, c_2, c_3, c_4. Use these constants to compute the value of $S_P(x)$ for $x = 1, 5, 7$.

2. Compute the LU factorization of the 5×5 natural spline matrix $\mathbf{S}_P$. How does this factorization compare with the factorizations for the 5×5 matrices $\mathbf{S}_B$, $\mathbf{S}_C$, and $\mathbf{S}_N$? How would you use this factorization to write efficient code to solve the matrix equation $\mathbf{S}_P\mathbf{c} = \mathbf{y}$?

20.6 Orthogonality Property for Clamped Cubic Splines

Galileo:

The purpose of this section is to prove an orthogonality property for splines, which is analogous to the Pythagorean Theorem. This property is also fundamental to the stability and convergence properties that make splines useful.

Let $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ be a partition of $[a, b]$.

If $g \in C^2[a, b]$, then let g_s denote the clamped spline associated with g. In particular, $g_s(x_i) = g(x_i)$ for $i = 0, 1, 2, \ldots, n$ and $g_s'(a) = g'(a)$ and $g_s'(b) = g'(b)$.

Let $e_g(x) = g(x) - g_s(x)$.

Note that $e_g(x_i) = 0$ for all $i = 0, 1, \ldots, n$ and $e_g'(a) = e_g'(b) = 0$.

Lemma 20.6.1. *If $\phi(x)$ is a piecewise linear continuous function on $[a, b]$ which is linear on each interval $[x_i, x_{i+1}]$, then*

$$\int_a^b e_g''(x) \cdot \phi(x) dx = 0.$$

Proof. The idea behind the proof is to integrate the integral by parts. Since we are using clamped splines, the endpoint information $g_s'(a) = g'(a)$ and $g_s'(b) = g'(b)$ will ensure that the integral is zero.

□

Theorem 20.6.2 (Orthogonality Property For Clamped Cubic Splines). *If* $g \in C^2[a,b]$ *and* $e_g(x) = g(x) - g_s(x)$, *then* $\int_a^b [g''(x)]^2 dx = \int_a^b [g_s''(x)]^2 dx + \int_a^b [e_g''(x)]^2 dx$.

Proof. Since $g_s''(x)$ is piecewise linear and continuous, we observe that

$$\begin{aligned}\int_a^b [g''(x)]^2 dx &= \int_a^b [g_s''(x) + e_g''(x)]^2 dx \\ &= \int_a^b [g_s''(x)]^2 dx + 2\int_a^b g_s''(x) \cdot e_g''(x) dx \\ &\quad + \int_a^b [e_g''(x)]^2 dx \\ &= \int_a^b [g_s''(x)]^2 dx + \int_a^b [e_g''(x)]^2 dx.\end{aligned}$$

□

20.7 Minimization Property for Splines

We now present a mathematical formulation of the intuitive concept that spline interpolation provides the fit with the fewest "wiggles." This minimization property will be one of the key facts needed to prove the convergence theorem for splines.

Let $C_g^2[a,b]$ denote the set of all $\phi \in C^2[a,b]$ such that $\phi(x_i) = g(x_i)$ for all $i = 0, 1, 2, \ldots, n$ and $\phi'(a) = g'(a)$ and $\phi'(b) = g'(b)$.

Note that the set $C_g^2[a,b]$ is a convex subset of $C^2[a,b]$.

Proposition 20.7.1. *If* $\phi \in C_g^2[a,b]$, *then* $\phi_s(x) = g_s(x)$ *for all* $x \in [a,b]$.

Theorem 20.7.2. *(Minimization Property) If* $g \in C^2[a,b]$ *and any* $\phi \in C_g^2[a,b]$, *then*

$$\int_a^b [g_s''(x)]^2 dx \leq \int_a^b [\phi''(x)]^2 dx.$$

Proof. By the orthogonality property

$$\int_a^b [\phi''(x)]^2 dx = \int_a^b [\phi_s''(x)]^2 dx + \int_a^b [\phi''(x) - \phi_s''(x)]^2 dx.$$

Since $\phi \in C_g^2[a,b]$, $\phi_s \equiv g_s$ on $[a,b]$,

$$\begin{aligned}\int_a^b [\phi''(x)]^2 dx &= \int_a^b [g_s''(x)]^2 dx + \int_a^b [\phi''(x) - \phi_s''(x)]^2 dx \\ &\geq \int_a^b [g_s''(x)]^2.\end{aligned}$$

□

20.8 Convergence for Splines

The first step in the proof of the convergence theorem for splines is to attack the second derivative of a function by a linear combination of "hat" functions, which are best as measured by a least squares fit, (i.e. in the L_2 norm).

Definition 20.8.1. *If P is a partition of $[a,b]$ and $g \in C^0[a,b]$, then a function $g_{LS} \in C^P[a,b]$ is called the best piecewise linear approximation to g in the least squares sense, if*

$$\int_a^b (g(x) - g_{LS}(x))^2 dx \leq \int_a^b (g(x) - \phi(x))^2 dx$$

for all $\phi \in C^P[a,b]$.

The next proposition provides the solution to the least squares problem for the second derivative of a function. This proposition states that the second derivative of the clamped cubic spline provides the best least squares approximation to the second derivative of a given function. Note that the proof of this theorem uses the fact that the spline of the sum is the sum of the splines.

Proposition 20.8.2 (Corollary). *If $g \in C^2[a,b]$ and g_s denotes the clamped cubic spline approximation of g, then $g_s'' = (g_s)''$ is the best piecewise linear approximation of g'' in the least squares sense. In particular, $(g'')_{LS} = (g_s)''$.*

Proof. To prove this proposition we must show that if ϕ is any member of $C^P[a,b]$, then

$$\int_a^b [g''(x) - g_s'']^2 dx \leq \int_a^b [g''(x) - \phi(x)]^2 dx.$$

By the fundamental theorem of calculus, a function Φ can be found in $C^2[a,b]$ with the property that $\Phi''(x) = \phi(x)$ for all $x \in [a,b]$. (i.e. $\Phi(x)$ is the double antiderivative of $\phi(x)$.)

Let $G(x) = g(x) - \Phi(x)$.

If $e_G(x) = G(x) - G_s(x)$, then by the orthogonality property

$$\int_a^b [G''(x)]^2 dx = \int_a^b [G_s''(x)]^2 dx + \int_a^b [e_G''(x)]^2 dx.$$

Since $e_G''(x) = G''(x) - G_s''(x)$ and $\Phi''(x) = \Phi_s''(x) = \phi(x)$ for all $x \in [a,b]$,

$$\begin{aligned} e_G''(x) &= g''(x) - \phi(x) - (g_s''(x) - \Phi_s''(x)) \\ &= g''(x) - g_s''(x). \end{aligned}$$

Thus,

$$\begin{aligned} \int_a^b [g''(x) - \phi(x)]^2 dx &= \int_a^b [G''(x)]^2 dx \\ &\geq \int_a^b [e_G''(x)]^2 dx \\ &= \int_a^b [g''(x) - g_s''(x)]^2 dx. \end{aligned}$$

Since ϕ is an arbitrary member of $C^P[a,b]$ we are done. □

Corollary 20.8.3. *Let $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ be a partition of $[a,b]$. If $g \in C^2[a,b]$, then the clamped cubic spline $g_s(x)$ has the property that*

$$\|g - g_s\|_\infty \leq \frac{\|P\|^2}{8} \cdot \|g'' - g_s''\|_\infty.$$

Proof. Simply let $G(x) = g(x) - g_s(x)$ and apply Corollary 4.4. □

Note that if the points x_i in the partition are equally spaced, then the matrix $\mathbf{A}$ has the form

$$\mathbf{A} = 2h \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & 0 & 0 & & \dots & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & & \dots & 0 \\ 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & & \vdots \\ \ddots & \ddots & \ddots & \ddots & & & \vdots \\ & & \ddots & \ddots & \ddots & \ddots & 0 \\ & & & \ddots & \ddots & \ddots & \frac{1}{4} \\ 0 & & & & 0 & \frac{1}{4} & 1 \end{pmatrix}.$$

Proposition 20.8.4. *If $g \in C^0[a,b]$ and g_{LS} is the best piecewise linear approximation to g in the least squares sense, then*

$$\|g_{LS}\|_\infty \leq 3 \cdot \|g\|_\infty.$$

Proof. The proof of this proposition is quite technical and thus omitted. □

Corollary 20.8.5. *If $g \in C^2[a,b]$, then $\|g_s''\|_\infty \leq 3 \cdot \|g''\|_\infty$.*

Proof. By the previous proposition $\|g_{LS}''\|_\infty \leq 3 \cdot \|g''\|_\infty$. By the corollary to Pythagoras, $g_{LS} = g_s''$. Therefore $\|g_s''\|_\infty \leq 3 \cdot \|g''\|_\infty$. □

The previous proposition shows that the most simple–minded interpolation is no more than twice as bad as the best.

Proposition 20.8.6. *If $g \in C^2[a,b]$, then $\|g'' - g_s''\|_\infty \leq 4 \cdot \|g'' - I_{g''}\|_\infty$.*

Proof. $\|g'' - g_s''\|_\infty \leq \|g'' - I_{g''}\|_\infty + \|I_{g''} - g_s''\|_\infty \leq \|g'' - I_{g''}\|_\infty + 3\|I_{g''} - g''\|_\infty = 4 \cdot \|g'' - I_{g''}\|_\infty$.

□

20.9 Convergence for Clamped Splines

The purpose of the next discussion is to prove the convergence theorem for the clamped cubic spline. Better yet, this theorem guarantees a $4th$-order convergence rate. In addition, the convergence theorem for the second derivatives is also proved.

The following list of 3 facts provides a summary of the key steps used to prove convergence for the clamped cubic spline.

1. $\|g - g_s\|_\infty \leq \frac{1}{8} \cdot \|g'' - g_s''\|_\infty \cdot \|P\|^2$
2. $\|g'' - g_s''\|_\infty \leq 4 \cdot \|g'' - I_{g''}\|_\infty$
3. $\|g'' - I_{g''}\|_\infty \leq \frac{1}{8}\|g^{(4)}\|_\infty \cdot \|P\|^2$

Theorem 20.9.1. *Convergence for Clamped Splines*

If $g \in C^4[a,b]$, then

$$\|g - g_s\|_\infty \leq \frac{1}{16} \cdot \|g^{(4)}\|_\infty \cdot \|P\|^4.$$

Proof. By fact 1,

$$\|g - g_s\|_\infty \leq \frac{1}{8}\|g'' - g_s''\|_\infty \cdot \|P\|^2.$$

By fact 2,

$$\|g'' - g_s''\|_\infty \leq 4 \cdot \|g'' - I_{g''}\|_\infty.$$

Therefore,

$$\|g - g_s\|_\infty \leq \frac{1}{2}\|g'' - I_{g''}\|_\infty \cdot \|P\|^2.$$

By fact 3,

$$\begin{aligned} \|g - g_s\|_\infty &\leq \frac{1}{2} \cdot \frac{1}{8}\|g^{(4)}\|_\infty \cdot \|P\|^2 \cdot \|P\|^2 \\ &= \frac{1}{16} \cdot \|g^{(4)}\|_\infty \cdot \|P\|^4. \end{aligned}$$

□

Note that the best result is by Hall in 1968 where he showed:

$$\|g - g_s\|_\infty \leq \frac{5}{384} \cdot \|P\|^4 \cdot \|g^{(4)}\|_\infty.$$

The next theorem guarantees a $2nd$-order convergence rate for the second derivative of the claimed cubic splines to converge to the second derivative of the function.

Theorem 20.9.2 (Convergence for the 2nd derivative of the clamped cubic splines.). *If $g \in C^{(4)}[a,b]$, then*

$$\|g'' - g_s''\|_\infty \leq \frac{1}{2} \cdot \|g^{(4)}\|_\infty \cdot \|P\|^2.$$

Proof. By the previous proposition $\|g'' - g_s''\|_\infty \leq 4 \cdot \|g'' - I_{g''}\|_\infty$.

By fact 2, $\|g'' - I_{g''}\|_\infty \leq \frac{1}{8}\|g^{(4)}\|_\infty \ \|P\|^2$. Therefore,

$$\|g'' - g_s''\|_\infty \leq \frac{4}{8} \cdot \|g^{(4)}\|_\infty \cdot \|P\|^2.$$

□

Note that Hall and Meyer showed in 1976 [2] that

$$\|g'' - g_s''\|_\infty \leq \frac{3}{8}\|g^{(4)}\|_\infty \cdot \|P\|^2.$$

Exercise Set 20.9.

1. If $g(x) = \cos(x)$ for $x \in [-\pi, \pi]$ and $tol = \frac{1}{10^5}$, then how many equally spaced points will be required to guarantee that the clamped cubic spline approximation $g_s(x)$ will approximate $\cos(x)$ with error less than $\frac{1}{10^5}$ for all $x \in [-\pi, \pi]$? Repeat this exercise for $g''(x)$. Compare your answer with the answer you found for the piecewise linear approximation.

2. If $g(x) = \frac{1}{1+25x^2}$ for $x \in [-1, 1]$ and $tol = \frac{1}{10^5}$, then how many equally spaced points will be required to guarantee that the piecewise linear approximation $g_s(x)$ will approximate $g(x)$ with error less than $\frac{1}{10^5}$ for all $x \in [-1, 1]$? Repeat this exercise for $g''(x)$. Compare your answer with the answer you found for the piecewise linear approximation.

Bibliography

[1] K. E. Atkinson, *An Introduction to Numerical Analysis, 2^{nd} Edition,* John Wiley & Sons, 1989.

[2] Carl de Boor, *A Practical Guide to Splines,* Applied Mathematical Sciences, volume 27, Springer-Verlag, New York, 1978.

[3] Galileo Galilei, *Dialogue on the Great World Systems in the Salusbury Translation,* Revised, Annotated, and with an Introduction by Giogio de Santillana, The University of Chicago Press, Chicago and London, 1953.

[4] D. R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid,* Vintage Books, 1989.

[5] V. J. Katz, *A History of Mathematics: An Introduction,* HarperCollins College Publishers, 1993.

[6] D. E. Knuth, *Surreal Numbers,* Addison Wesley, 1974.

[7] I. Lakatos, *Proofs and Refutations,* Cambridge University Press, 1976.

[8] J J O'Connor and E F Robertson, *Karl Pearson,* http://www-gap.dcs.st-and.ac.uk/ history/Mathematicians/Pearson.html.

[9] J J O'Connor and E F Robertson, *Abstract linear spaces,* http://www-groups.dcs.st-and.ac.uk/ history/HistTopics/Abstract_linear_spaces.html.

[10] G. Polya, *How to Solve It,* Princeton University Press, First Princeton Science Library Edition, 1988.

[11] W. Rudin, *Principles of Mathematical Analysis, 2nd Edition,* International Series in Pure and Applied Mathematics, McGraw-Hill, New York, 1964.

[12] G. F. Simmons, *Differential Equations with Applicatons and Historical Notes,* McGraw-Hill Book Company, New York, 1972.

[13] D. Sobel, *Galileo's Daughter,* Penguin Books, 2000.

[14] G. P. Tolstov, *Fourier Series,* Dover, New York, 1962.