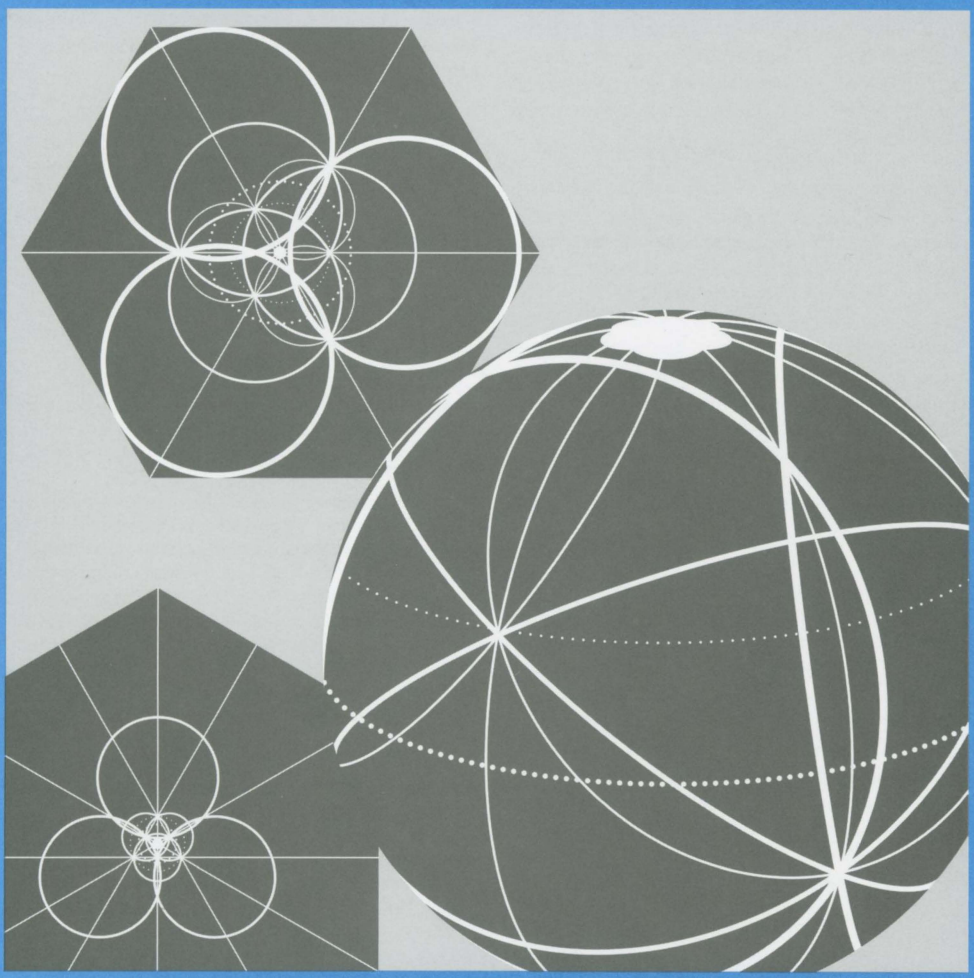
Vol. 76, No. 1, February 2003

MATHEMATICS MAGAZINE



Circle Trios

- Dr. David Harold Blackwell, African American Pioneer
- A Tale of Three Circles

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Submit new manuscripts to Frank A. Farris, Editor, Mathematics Magazine, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0373. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

Cover image: *Circle Trios*, by Charles Delman. In the upper left pattern, three trios of congruent circles are arranged symmetrically about a central point, P. Stereographic projection takes the trios to the sphere, and another stereographic projection gives the lower left image, where P has now gone to infinity. The innermost family, where each circle passes through P, is Euclidean: two stereographic projections turn the circles to straight lines. The next family, where P is inside all three circles, is elliptic: when projected onto the sphere, these circles become great circles. The outermost family, where P is outside all three circles, is hyperbolic: these are hyperbolic lines in the Poincaré disc bounded by the larger dotted circle. This image gives a hint for the question posed in "A Tale of Three Circles": When a triangular shape is created from arcs of circles, what is the sum of the interior angles?

AUTHORS

Nkechi Agwu is an Associate Professor at the Borough of Manhattan Community College (BMCC), City University of New York (CUNY) where she is the Chair of the Faculty Council Curriculum Committee. She received a Ph.D. in 1995 in Mathematics Education with a minor in Gender and Multicultural Studies from Syracuse University and a Masters in 1989 in Mathematics from the University of Connecticut. From 1996–2001, she was a participant of the MAA Institute in the History of Mathematics and Its Uses in Teaching (IHMT). Her experience at this institute facilitated her PSC-CUNY 29 research study, *Using Biography to Develop Mathematical Power, Encourage Diversity and Teach the History of Mathematics*, which led to this biography of Dr. Blackwell. Dr. Agwu is a Faculty for the 21st Century (F21) member of Project Kaleidoscope (PKAL), a national advocacy organization dedicated to exploring and defining what works in undergraduate science, technology, engineering, and mathematics (STEM), and disseminating effective practices through workshops, institutes, publications, and real virtual networks.

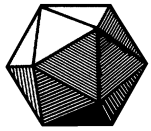
Luella Smith is a 2000 Business Management honors graduate of the Borough of Manhattan Community College, City University of New York, who is currently an Associate at JP Morgan Stanley.

Aissatou Barry is a current student at the Borough of Manhattan Community College, City University of New York, where she is majoring in Accounting.

Charles I. Delman received his bachelor's degree from Harvard and his Ph.D. from Cornell. His main mathematical interests are low-dimensional topology, billiard dynamics, and geometry. This article grew out of discussions with co-author Gregory Galperin about an undergraduate course in classical geometry, which both of them frequently teach. This subject is studied mainly by prospective secondary teachers, but many others would benefit from knowing more about it. Major nonmathematical interests are music and the visual arts. He loves wild places and may often be found backpacking with his children, Anna and Ben, and his partner, Barbara.

Gregory Galperin received his master's and Ph.D. degrees from Moscow State University, Russia, under the supervision of A.N. Kolmogorov. His mathematical interests include dynamical systems, billiard dynamics, differential geometry, automata theory, and combinatorial geometry. This article is a consequence of the one-line solution to the initial Galperin problem on three semicircles with collinear centers, the key words of which are *Poincare upper halfplane*. Galperin enjoys inventing unusual problems for mathematical competitions. He is a member of the USAMO committee, a member of the editorial board of a new mathematical Russian journal "Mathematical Enlightenment," an author of the books "Moscow Mathematical Olympiads" and "Mathematical Billiards." His nonmathematical interests include table tennis, drawing, and listening to good music.

Vol. 76, No. 1, February 2003



MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

Subscription correspondence and notice of change of address should be sent to the Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to Dave Riska (*driska@maa.org*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

Periodicals postage paid at Washington, D.C. and additional mailing offices.

Postmaster: Send address changes to Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

Printed in the United States of America

Dr. David Harold Blackwell, African American Pioneer

NKECHI AGWU
LUELLA SMITH
AISSATOU BARRY
Borough of Manhattan Community College (BMCC)
City University of New York (CUNY)
New York, NY 10007
nagwu@bmcc.cuny.edu

Dr. David Harold Blackwell 1919–
"Find something that you like.
It is more important than how much money you make." [**1**]

Dr. David Blackwell is an African American educational pioneer and eminent scholar in the fields of mathematics and statistics, whose contributions to our society extend beyond these fields. This paper highlights his significant contributions and the personal, educational, and professional experiences that groomed and nurtured him for leadership as a civic scientist. We hope this account of Dr. Blackwell's life will enhance the literature on African American achievers, and motivate students majoring in, or considering careers in mathematics and statistics, particularly those from underrepresented groups.

The education of David Blackwell

Early childhood It is April 24, 1919, an era of heightened segregation and racial discrimination in the United States. Welcome to Centralia, Illinois, a small town community on the "Mason-Dixon line," with a population of about 12,000 people, and very few African American families [**13**]. Witness the birth of David Harold Blackwell. He was to be the eldest of four children born to Grover Blackwell, a hostler for Illinois Central Railroad and Mabel Johnson Blackwell, a full-time homemaker. His

two younger brothers, J. W. and Joseph, and his younger sister Elizabeth would follow soon after.

During his early childhood, David had a grandfather and an uncle living in Ohio who were influential to his cognitive development. His grandfather, whom he had never met, was a school teacher and later a storekeeper. He endowed David with a large library of books. From this library, David read and enjoyed many books, including his first algebra book. His uncle had been home schooled by his grandfather, because of worries about the effects of racism on his son at school. He impressed David with his ability to add three columns of numbers expeditiously, in a one-step process.

David attended Centralia public schools for the first ten years of his schooling, from 1925 to 1935. His parents enrolled him in integrated schools in his southern Illinois locality, which also had racially segregated schools: whites only schools and blacks only schools. However, being at an integrated school, David was unaware of, or unaffected by, issues of racial discrimination. He attributes this to the fact that his parents shielded their children as much as possible from the effects of racism, and to the fact that he experienced few encounters where race was an issue.

High school education In high school, David developed a strong interest in games such as checkers and in geometry, but was not particularly interested in algebra and trigonometry. He pondered over questions of whether the player with the first move in these games had a higher probability of winning. He states this about geometry:

> Until a year after I finished calculus, it was the only course I had that made me see that mathematics is really beautiful and full of ideas. I still remember the concept of a helping line. You have a proposition that looks quite mysterious. Someone draws a line and suddenly it becomes obvious. That's beautiful stuff. I remember the proposition that the exterior angle of a triangle is the sum of the remote angles. When you draw that helping line it is completely clear. [**3**, p. 20]

Fortunately for David, he had teachers who nurtured his mathematical interests. His geometry teacher got him to love mathematics by helping him to see the beauty of the subject. A teacher named Mr. Huck formed a mathematics club where he would challenge students with problems from the School Science and Mathematics journal. Whenever a student came up with a good possible solution Mr. Huck would send the solution to the journal under the student's name. David's solutions got published once in the journal and he was identified three times there as having correct solutions to problems, which gave him great joy. This was something that further motivated his interest in mathematics. Consequently, long before he was admitted to college, David had decided to major in mathematics. He states, "I really fell in love with mathematics.... It became clear that it was not simply a few things that I liked. The whole subject was just beautiful." [**3**, p. 21]

Undergraduate education David graduated from high school in 1935, at the age of 16. He promptly enrolled at the University of Illinois in Champaign-Urbana, a campus with no black faculty at that time. His intention was to earn a Bachelor's degree and become an elementary school teacher. This decision was motivated by the scarcity of jobs at that time and the fact that a good friend of his father, with a strong influence on the school board in a southern Illinois town, had promised to get him hired upon graduation. However, because his decision to become an elementary school teacher was based primarily on the need for employment after graduation rather than a keen interest, he kept postponing his education courses. After a time, they were no longer necessary, due to a change in his career decision.

David's career goal to become an elementary school teacher changed in his combination junior/senior year when he took a course in elementary analysis. This course really sparked his interest in advanced level mathematics. It motivated him to consider a career that would require graduate level education in mathematics. He now set his sights on teaching at the college or high school level. He began to pursue activities that would facilitate his career goals and groom him for leadership, such as serving as president for the mathematics club at this university. His parents, who were not college educated folks, left it to him to make the hard core decisions about his college education and career. However, they supported him in every way they could and encouraged him to work hard to achieve his goals.

At the end of his freshman year, David learned that his father was borrowing money to finance his college education. A young man with strength of character, he decided to spare his father this financial ordeal by taking responsibility for supporting his college education by working as a dishwasher, a waiter, and a cleaner for equipment in the college entomology lab. In spite of having to work his way through college, David facilitated his college education by taking summer classes and passing proficiency exams, which allowed him to skip courses. Thus, in 1938, he graduated with a Bachelor's degree in mathematics within three years of admission to college.

Graduate education David continued on for graduate study from 1938 to 1941, at Champaign-Urbana, working to pay for his education as usual. In his last two years of graduate study, while he was a doctoral student, he was awarded fellowships from the university. David has mixed feelings about the motivations of the university officials in offering him these fellowships. He says this about the issue:

> One of my fellow graduate students told me that I was going to get a fellowship. I said, "How do you know?" He said, "You're good enough to be supported, either with a fellowship or a teaching assistantship, and they're certainly not going to put you in the classroom." That was funny to me because fellowships were the highest awards; they gave one the same amount of money and one didn't have to work for it. I have no doubt, looking back on it now, that race did enter into it. [**3**, p. 21]

In 1939, David earned a master's degree in mathematics, proceeding on for doctoral studies with some trepidation. He was confident that he could handle the mathematics course work and read research papers. However, he was unsure about whether he would be successful in writing a thesis. Being a determined young man, he took on this challenge, bearing in mind that he had the option of high school teaching in the event that he was not successful in completing the doctoral program.

David's thesis advisor was Joseph Doob. He was a probability and statistics professor at Champaign-Urbana, renowned for his contributions to martingale theory. David states, "Joseph Doob had the most important mathematical influence on me. I studied his work carefully and learnt a lot from it. I admired him and tried to emulate him." [**1**] This statement captures the significance of documenting the contributions and biographies of pioneers, innovators, and leaders in any field of study.

Ironically, David had never met Doob prior to approaching him to become his thesis advisor. His decision to appeal to Doob was based on the recommendation of a peer mentor, Don Kibbey, a teaching assistant in whom he placed a great deal of confidence; Doob was Don Kibbey's dissertation advisor at that time. He was also the dissertation advisor to Paul Halmos, a mathematician who contributed immensely to the development of measure theory and who was a significant peer mentor to David in this area while they were both students of Doob.

In 1941, at the age of 22, within five years of graduation from high school, David earned a doctorate in mathematics. He is the seventh African American to earn a Ph.D. in this field. His dissertation is titled, "Some Properties of Markoff's Chains." It led to his first set of publications: "Idempotent Markoff Chains," "The Existence of Anormal Chains," and "Finite Non-homogeneous Chains." [**4, 5, 6**] David credits the main idea in his thesis to his advisor Joseph Doob. In doing so, he shows how important a thesis advisor is in helping students to identify appropriate research questions.

Post-doctoral education Upon completion of his doctoral program, David was awarded a Rosenwald Post-doctoral Fellowship for a year at the Institute of Advanced Study (IAS) at Princeton University. His exposure at the IAS was the beginning of a stellar career as a renowned mathematician, statistician, and educator.

His acceptance at the IAS was not devoid of the hurdles of racism. At that time, it was customary for Princeton to appoint IAS members as visiting fellows. However, when the administrators at Princeton, particularly the president, realized that David was a black man they profusely objected to his acceptance at the IAS. Princeton had never admitted a black student nor hired a black faculty member, and the administrators wanted to maintain the status quo. Upon the insistence and threats of the IAS director, the administrators of Princeton later withdrew their objections to David's acceptance at the IAS. At the time David accepted the Rosenwald Fellowship, he was unaware of the racial controversy that took place between the IAS director and the president of Princeton. He discovered the exact details several years later during the prime of his professional career. Thus, he was shielded from the marring effects of racism on his acceptance of the Rosenwald Fellowship and his stay at the IAS.

At the IAS, there were two mathematicians in particular who influenced David's post-doctoral education, Samuel Wilks and John von Neumann, a renowned Hungarian American mathematician credited with initiating the development of game theory. David developed a keen interest in statistics by auditing Samuel Wilks' course. Wilks was a mathematician renowned for his work in developing the field of mathematical statistics. He was a founding member of the Institute of Mathematical Statistics, an international professional and scholarly society devoted to the development, dissemination, and application of statistics and probability. (Much later, in 1955, David would serve as president of this organization.)

Another important mathematician who took an interest in David was John von Neumann. He encouraged David to meet with him to discuss his thesis. David avoided this meeting for several months because he did not think that the great John von Neumann was genuinely interested, or had the time to listen to him discuss his thesis. This turned out to be a flawed assumption, for von Neumann was indeed interested in mentoring students. When David and von Neumann finally met to discuss his thesis, von Neumann spent about 10 minutes listening to David's explanation about his thesis and asking him related questions. Afterwards, he took the liberty to explain to David other simpler techniques that he could have used for his thesis problem. The time David spent with von Neumann discussing his thesis, seeing firsthand that he was willing to mentor students, certainly impressed young David. Throughout his professional career, even at the height of success, we see him mentoring students and other young professionals.

Professional career: scholar, teacher, and administrator

We only have to examine the humble beginnings of David's professional career to understand some of the negative consequences of racism, and other forms of discrimi-

nation, on society. David was a young African American pioneer with genius, integrity, and strength of character, whose work was of interest to world-class mathematicians of this period. Yet when he completed his post-doctoral education, the only universities he applied to for a faculty position were Historically Black Colleges and Universities (HBCUs), because he could envision himself nowhere else. He states:

> It never occurred to me to think about teaching in a major university since it wasn't in my horizon at all—I just assumed that I would get a job teaching in one of the black colleges. There were 105 black colleges at that time, and I wrote 105 letters of application.... I eventually got three offers, but accepted the first one I got. From Southern University. [**10**]

From 1942 to 1943, David was an instructor at Southern University in Baton Rouge, Louisiana. In 1943, he accepted an instructor position for a year, at another HBCU, Clark College, in Atlanta, Georgia. In 1944, at the end of his term at Clark College, David still envisioned himself as a faculty member at an HBCU. He accepted a tenure-track position as an assistant professor at Howard University, Washington D.C., the premier HBCU at that time, where he was one of the Mathematics Department's four faculty members. At Howard he was a generalist, teaching all mathematics courses right up to the master's degree level, which was the highest degree program in the department.

David stayed at Howard for ten years, from 1944 to 1954, rising through the ranks from Assistant Professor to Associate Professor in 1946, and finally to the position of Professor and Chairman of the Mathematics Department in 1947. In spite of the heavy teaching loads of at least 12 hours per week, and heavy administrative duties at these HBCUs, he had over 20 publications by the time he left Howard. He had also earned a strong reputation as an excellent teacher and innovative scholar in probability, statistics, and game theory.

Interestingly, although David enjoyed his work as a mathematics faculty member at Howard, it was not Howard but the larger mathematics community and professional networking that was the springboard for his professional success. He says, "I was teaching at Howard and the mathematics environment was not really very stimulating, so I had to look around beyond the university for whatever was going on in Washington that was interesting mathematically." [**10**] However, Howard should be given some credit. The administrators understood the importance of professional meetings and supported David financially and otherwise to allow him to attend them. This illustrates how important it is for students and young professionals to attend professional meetings and participate in professional organizations.

David credits Abe Girshick for initiating his professional success in statistics. He attended a meeting sponsored by the Washington Chapter of the American Statistical Association. There he listened to an interesting lecture by Girshick on sequential analysis. The lecture involved a discussion of Wald's equation, a concept David found to be unbelievable. Thus, after the meeting, David constructed a counterexample to this equation, which he mailed to Girshick. His counterexample turned out to be wrong. However, it resulted in an invitation from Girshick to David to meet with him in his office to discuss it. This meeting was the beginning of a wonderful relationship for both men and several years of collaboration, which culminated years later in a classic mathematics book, *Theory of Games and Statistical Decisions* [**9**]. It also resulted in several publications by David, including his favorites: "On an equation of Wald" [**7**] (a proof with much weaker constraints of the equation he found to be unbelievable) and "Bayes and minimax solutions of sequential decision problems." [**8**]

According to David, Abe Girshick was his most influential mentor in the field of statistics. He took time off to work in collaboration with Girshick at the RAND Corporation and Stanford University, California, while he was still a faculty member at Howard. The RAND Corporation began as Project RAND, started by the Air Force in 1946 to conduct long range studies in intercontinental warfare by means other than ground armies. David worked as a mathematician at the RAND Corporation in Santa Monica, California, from 1948 to 1950 during the summer periods, and as a Visiting Professor at Stanford, from 1950 to 1951. These were the most significant times for him. His work during this period resulted in breakthroughs that set the stage for world recognition.

David's work in game theory blossomed at the RAND Corporation while he was collaborating with Abe Girshick and other colleagues. World War II had promoted an interest in the theory of games depicting duels. The theory of duels deals with two-person, zero-sum games.

Imagine two persons initially standing $2n$ paces apart, each with a gun loaded with a single bullet. They are advancing towards each other. At every step forward each person has to decide whether to shoot or hold fire without any prior knowledge of what the other person's decision will be. A strategy certainly involves how many of the possible n steps have been taken already, knowledge of one's own shooting ability, and some guess about one's opponent. Firing too soon means the shooter might miss; firing too late might mean the shooter may have been shot. To simplify the theory, we assume that the game always ends with one person having been shot.

David explored different variations on the basic theory of duels. For instance, if the intention is to kill one's opponent, then the optimal number of steps before firing may be different than it would be if all one wants is to stay alive. It also might make a difference if both guns have silencers, so one might not know that the opponent has fired and missed. His work in the theory of duelling led to significant developments in game theory and earned him a reputation as a pioneer in this area. He developed a game theoretic proof of the Kuratowski Reduction Theorem, which was groundbreaking in that it connected the fields of topology and game theory, an achievement that gives him great pleasure.

David did not explore beyond two-person, zero-sum games. He attributes his reluctance to do so to the extreme complexity of other types of games and to the fact that the best mathematical response for certain games may have a negative social, psychological, or economical response. This had to do with the *sure thing principle*, which was formulated by Jimmie Savage, one of David's mentors at the RAND Corporation. One way of stating it is this: Suppose you have to choose between two alternatives, A and B, and you think that the outcome depends on some unknown situation X or Y. If knowing that X was the case would lead you to choose A over B and if knowing that Y was the case would still lead you to choose A over B, then, even if you do not know whether X or Y was the case, you should still choose A over B. It was thought that the arms race arising from the Cold War showed the sure thing principle at work.

Suppose that the U.S. and the Soviet Union both operate on the sure thing principle. They have to choose between arming (alternative A) or disarming (alternative B) without knowing whether the other nation is going to arm (situation X) or disarm (situation Y). The sure thing principle indicates that the best mathematical strategy is for both nations to continue arming themselves in order to stay ahead or at par with the opposing nation. This leads to the depletion of valuable resources that each nation could have spent on other important areas of development. This is like the well-known prisoner's dilemma, because both nations are actually losing when they use the sure thing principle. The winning strategy would be for both nations to disarm, a situation that is unlikely to happen due to mistrust between the two nations who both fear

that the other will double-cross them if they cooperate. David says, "I started working on this particular game where the sure thing principle led to behavior that was not best. So, I stopped working on it." [**1**] Here we see David as a moral scientist.

David's work with the RAND Corporation led him to an avid study of the works of Thomas Bayes. By a stroke of fate, an economist at the RAND Corporation asked David's mathematical opinion on how to apportion the Air Force research budget over a period of five years between immediate developmental and long-range research. The appropriate proportion is dependent on the probability of a major war within the budget period. If this probability is high, then the budget emphasis will be on immediate developmental research, and if it is low, the emphasis will shift to long-range research.

David gave a mathematically correct but unhelpful answer. He indicated that, in this situation, we are dealing with a unique event and not a sequence of repeated events, so the probability of occurrence of a major war within the five-year period is either 0 or 1, and is unknown until the five-year period has elapsed. The economist remarked, in a manner that intrigued David, that this was a common answer of statisticians. It caused him to ponder the problem, and to discuss it with Jimmie Savage on his visit to the RAND Corporation several weeks later. His discussion with Savage left David with a completely new approach to statistical inference—the Bayesian approach.

The Bayesian approach to statistical inference considers probability as the right way to deal with all degrees of uncertainty, and not just the extremes of impossibility and certainty, where the probability is 0 or 1. As a basic example, consider this: Even though we cannot observe the same five-year period repeatedly and deduce the probability of a war, we still may be able to make inferences about this probability and base decisions on our estimates of it. In more sophisticated applications, statisticians develop utility functions based on underlying probability distributions; decision-makers attempt to maximize utility.

Since David developed an appreciation for the Bayesian approach, all his statistical works have incorporated it. Thus, he credits Jimmie Savage as the second most influential person in terms of his statistical thinking.

The years at Berkeley In 1954, David accepted a visiting position for one year at the University of California, Berkeley. In the following year, he accepted a position as a full Professor at this university, and remained there until his retirement in 1988. It is noteworthy to point out that in 1942, much to David's surprise at the time, he was interviewed for a faculty position at Berkeley. However, he was not surprised or disappointed when he was not offered the position. The reason given by the university for not hiring him was that they had decided to appoint a woman, due to the war and the draft. Nevertheless, destiny prevailed. David finally ended up at Berkeley 12 years later, during the period of civil rights gains. African Americans were now beginning to enjoy more career opportunities and better employment practices.

Shortly after David's arrival at Berkeley, the Mathematics Department there was divided to make its Statistics Laboratory, headed by Jerzy Neyman, into a separate department of its own. For four years, from 1957 to 1961, David was the chair of the Department of Statistics, succeeding Neyman, the person who had interviewed him in 1942 for a possible faculty position at Berkeley. Neyman turned out to be a good friend. He had a personal influence on David through his warmth, generosity, and integrity.

David enjoyed his stint as chair of the department, but he admits that he did not miss the responsibilities of that position. He sees the primary goal of administrative leadership as creating an environment where the workers are happy. He states, "When I was department chairman, I soon discovered that my job was not to do what was right but to make people happy." [**3**, p. 30] The success of his leadership at Berkeley

shows that it was a winning strategy to build coalitions in which people enjoy working together.

David also provided leadership at Berkeley in other administrative capacities. He was the Assistant Dean of the College of Letters and Science from 1964 to 1968 at a time of serious strife at the university. He was also the Director of the University of California Study Center for the United Kingdom and Ireland, from 1973 to 1975.

While at Berkeley, David continued his scholarly work on the mathematics of competition and cooperation. Interestingly, although he accomplished many innovations while still a faculty member at Howard, he did not gain world recognition until he was a faculty member at Berkeley. Also interesting is the fact that his scholarship was not motivated by doing research for its own sake, but by attempting to understand the problems that intrigued him.

A caring teacher Surprisingly, even at Berkeley while David was at the peak of his research productivity, he taught probability and statistics courses at all levels, from elementary to graduate courses. He states, "There is beauty in mathematics at all levels of sophistication and all levels of abstraction." [**3**, p. 26] This statement highlights a very important quality that characterizes talented teachers: They are able to convey the beauty of their subject regardless of the level of mastery of the students.

David is very modest about his ability as a teacher. He sees himself as a good teacher for certain students, but not necessarily for all of them; he recognizes that there are some styles of teaching where he may not excel. He states, "People have different learning styles, abstract, concrete, visual, hearing, spatial, and so on. So it is necessary for teachers to reflect these learning styles in their teaching if they would like their students to appreciate the beauty of what they are teaching." [**1**] Many students evidently found David caring and approachable, since he served as the dissertation advisor to at least 53 students at Berkeley, a very high number.

David is a dynamic scholar and teacher who feels most comfortable when he is around students or those willing to learn and share. He is ever willing to jump to the blackboard to illustrate examples. From his conversations with colleagues and others he has granted interviews, his excitement with mathematics surfaces when he begins to ponder its beauty, how he fell in love with geometry, or how much pleasure it gave him to be challenged by a difficult proof of a theorem.

World recognition: the leader and civic scientist

David's world recognition as an eminent scholar, educator, and leader in our society is illustrated through his numerous awards, honors, and positions of leadership in professional organizations. His honorary Doctor of Science degrees alone illustrate his widespread recognition. He has received 12 honorary Doctor of Science degrees: from the University of Illinois in 1966; Michigan State University in 1969; Southern Illinois University in 1971; Carnegie-Mellon University in 1980; the National University of Lesotho in 1987; Amherst College and Harvard University in 1988; Howard University, Yale University, and the University of Warwick in 1990; Syracuse University in 1991; and the University of Southern California in 1992.

Equally amazing are his extensive leadership roles and honors in the profession, which speak to his dynamism as a civic scientist, a role in which David emerged full-fledged after he left Howard University. In 1954, he gave the invited address in probability at the International Congress of Mathematicians in Amsterdam. This address is credited with spurring Berkeley to offer him a visiting professorship. From 1959 to 1960, he was a visiting lecturer for the Mathematical Association of Amer-

ica in a program to enhance undergraduate mathematics education. In 1965, he was elected to the National Academy of Science. In 1968, he was elected to the American Academy of Arts and Sciences. By this time, David had published at least 60 books and papers.

From 1968 to 1971, David served as the vice president of the American Mathematical Society. From 1972 to 1973, he was chairman of the Faculty Research Lecture Committee. In 1973, he was president of the International Association for Statistics in the Physical Sciences. In 1974, he was the W. W. Rouse Ball Lecturer at the University of Cambridge in the United Kingdom. From 1975 to 1978, he was president of the Bernoulli Society for Mathematical Statistics and Probability. From 1975 to 1977, he was vice president of the International Statistical Institute. In 1976, he was elected Honorary Fellow of the Royal Statistical Society. In 1977, he gave the Wald Lecture for the Institute of Mathematical Statistics. In 1978, he was vice president of the American Statistical Association. Additionally, David has given the Rietz Lecture for this Institute of Mathematical Statistics and he has served on the Board of Directors of the American Association for the Advancement of Science.

In fact, the Wald and Rietz Lectures of the Institute of Mathematical Statistics were instrumental in establishing his reputation as an effective and charismatic lecturer. Noteworthy is the fact that David was among a select few chosen to be filmed by the American Mathematical Society and the Mathematical Association of America, lecturing on mathematical topics accessible to undergraduate students.

The year 1979 was a wonderful one for David. He was awarded the John von Neumann Theory Prize by TIMS/ORSA, which today has become INFORMS, the Institute for Operations Research and Management Sciences. This was a significant honor given that John von Neumann was one of his earliest professional mentors. The purpose of this prize is to recognize a scholar (or more than one, in cases of joint work) who has made fundamental contributions to theory in operations research and management sciences. Although recent work is not overlooked, the award is usually given for work that has stood the test of time. The criteria for the prize are broad, and include significance, innovation, depth, and scientific excellence. In addition to a cash award and medallion, the citation reads:

> The John von Neumann Theory Prize for 1979 is awarded to David Blackwell for his outstanding work in developing the theory of Markovian decision processes, and, more generally, for his many contributions in probability theory, mathematical statistics, and game theory that have strengthened the methodology of operations research and management sciences. In the area of Markovian decision processes Blackwell, in a remarkable series of papers published between 1961 and 1966, put the theory of dynamic programming on a rigorous mathematical footing. He introduced new techniques of analysis and established conditions for the existence of optimal and stationary optimal policies. Particularly noteworthy are his studies of the effect of varying the discount rate and his introduction of the important concepts of positive and negative dynamic programs. Virtually all of the subsequent developments in this field are based on these fundamental papers. In other areas, Blackwell's early work with Arrow and Girshick helped lay the foundations for sequential analysis, and his subsequent book with Girshick systematized the whole field of decision theory, to the great benefit of a generation of mathematical statisticians. The famous Rao-Blackwell theorem on statistical estimation led to a practical method for improving estimates, now known as "Rao-Blackwellization." An elegant and important form of the renewal theorem is due to Blackwell, as is a beautiful characterization of the information content of an experiment. In game theory, he initiated the study of duels (with

Girshick) and later made several deep contributions to our understanding of sequential games and the role of information therein. [12]

David, the trailblazer, did not relax after receiving the John von Neumann Theory Prize. He continued scaling the frontiers of twentieth century developments in mathematics and statistics as a leader. In 1986, he was awarded the R. A. Fisher Award from the Committee of Presidents of Statistical Societies. Upon his retirement in 1988, David received the Berkeley Citation. This is one of the highest honors given to a faculty member at Berkeley, for exemplary service to the university and outstanding achievement in one's field. David received this citation for his work in game theory, Bayesian inference, and information theory, for authoring the classic book, *Theory of Games and Statistical Decisions* [9], and for induction into the American Academy of Arts and Sciences and the National Academy of Sciences. By the time of his retirement, he had well over 90 books and papers published on dynamic programming, game theory, measure theory, probability theory, set theory, and mathematical statistics.

A tribute to David's immense contributions is the long list of lecture series and publications in his honor. The book, *Statistics, Probability and Game Theory, Papers in Honor of David Blackwell* [11], is a compilation of 26 papers edited by T. S. Ferguson, L. S. Shapley, and J. B. MacQueen. These papers treat topics related to his significant contributions in probability, statistics, gambling, game theory, Markov decision processes, set theory, and logic. The editors say this about the man honored by the volume: "It is the mark of an outstanding scientist to be influential in a variety of fields."

Another honor in this category is the Mathematical Sciences Research Institute (MSRI) conference and prize in honor of David Blackwell and Richard A. Tapia, distinguished mathematical scientists who have inspired more than a generation of African American and Hispanic American students and professionals in the mathematical sciences. The prize is awarded every second year to a mathematical scientist who has contributed significantly to his or her field of expertise, and who has served as a role model for mathematical scientists and students from underrepresented minority groups or contributed in significant ways to the addressing of the problem of the underrepresentation of minorities in mathematics.

Yet another honor in this category is the David Blackwell Lecture of the National Association of Mathematicians (NAM). This lecture is given annually at the MathFest, the popular summer meeting of the MAA. Its goal is to highlight the contributions of minorities in the mathematical community and to stimulate their professional growth.

Family life and personal tidbits

On December 27, 1944, David married a wonderful woman by the name of Ann Madison. He says, "The best thing I ever did in life was to get married to my wife" [1]. Thus, it is poignantly clear that Ann played a very supportive role in her husband's successes, and in ensuring the stability and enhancement of their family.

David and Ann have eight children, three sons and five daughters, Ann, Julia, David, Ruth, Grover, Vera, Hugo, and Sara. Notably, none of their children have exhibited any interest in mathematics, nor in a related field, an issue that is viewed positively by David. In response to a question about his children, he says this: "No, they have no particular mathematical interests at all. And I'm rather glad of that. This may sound immodest, but they probably wouldn't be as good at it as I am. People would inevitably make comparisons." [1]

On David's off time, when he's not in a classroom filled with students, or writing fascinating papers on mathematical or statistical topics, or engaging in other profes-

sional commitments, you can find him with his wife on their 40 acre property in Northern California, listening to music and enjoying themselves. He might say his dream is to sit beneath a tree and sip a martini, but in reality he is more active, and finds himself planting trees or doing yard work. Many of his good friends are professional colleagues with whom he works and collaborates to advance the fields of mathematics and statistics.

There was a time when David's family home did not have telephone service. One of his children had made quite an expensive long distance telephone call, so David and his wife decided to have the telephone disconnected for a month. During that period, he realized the advantage of not having a telephone—peace and tranquility. One month led to three months, but eventually, the advantages of the telephone won over its cons.

David enjoys playing on the computer. However, these are not trivial games that he plays. He says, "I have a little computer at home, and it is a lot of fun just to play with. In fact I'd say that I play with this computer here in my office at least as much as I do serious work with it." **[10]** He admits to attempting to use his computer to set up a program to take the square root of a positive definite matrix, to minimize functions with five variables, and to look at curves. Perhaps this kind of play is at odds with the image of the mathematician who sits down to prove a theorem, but many of us enjoy this kind of fun.

An example to remember

We learn a great deal from examples. If we know that someone has gone through a situation similar to ours, it helps us to analyze our situation in a more confident manner and to make better decisions. If we want students to make well-informed decisions concerning their educational and professional lives, we must provide them with examples of pioneers, innovators, and leaders, both in all fields of study and in all parts of our nation's history. The information about that successful person motivates students, and gives them the courage to tread similar paths. In the words of a student at Borough of Manhattan Community College (BMCC), City University of New York (CUNY), who was acquainted with Dr. David Harold Blackwell only through reading about him for a research project in an Introduction to Statistics class:

> David Blackwell's life has influenced me with the struggles that he has had to endure. He started out just wanting to teach elementary school, but his love of mathematics and his natural talent for mathematics has taken him so much further. This is an inspiration to me for I too love what I do and wish to go further. He has shown me to persevere in the face of adversity. I am happy to have learned so much about this truly incredible man. **[2]**

In addition, it is important to appreciate the contributions and accomplishments of persons from underrepresented groups in any field of study, in order to promote justice, equity, and diversity. This is an avenue for teaching cultural sensitivity and cooperation with people of different cultures, and a way to motivate students from these groups to similar or greater heights of success.

Dr. David Harold Blackwell is one of the world's most accomplished thinkers in the fields of mathematics and statistics. Of great significance is that he is one of the African American masters in these fields. He is a dynamic educator with a reputation as "one of the finest lecturers in the field." **[10]** He is a civic scientist and leader whose life history will certainly motivate others to follow his example. His example can also motivate us to develop the necessary mentoring programs and practices to open the

doors of opportunity for all students, especially for students from underrepresented groups, in the fields of mathematics and statistics.

By examining the conditions under which this mathematician rose to success, we can learn a lot about leadership, humility, strength of character, and passion for one's field. We can learn that mentoring, professional development, and active participation in professional meetings and organizations are vital opportunities. Providing them helps us to nurture students and encourage them to consider careers in mathematics and scientific fields and to groom young professionals in these disciplines. We can also learn about the social and psychological consequences of any form of discrimination on society.

Dr. David Harold Blackwell was not overly concerned about financial status when he decided to major in a career in mathematics. He had cultivated an appreciation for the subject and had a passion for examining and understanding issues that intrigued him. This passion led him to make groundbreaking innovations in the fields of mathematics and statistics. Our society has benefited from the vast contributions of this most renowned African American thinker. Unbelievable for a man that thought his story in life was to be an elementary teacher.

Acknowledgments. The authors would like to thank:

- Dr. David Harold Blackwell for providing biographical materials, including his picture for use in this paper, for the interviews he granted during the 1998–99 academic year to the first author, her BMCC Women's Research Program interns (Deanna Bernard, Kaur Gursimiran, and Luella Smith), and her MAT 150—Introduction to Statistics students, and for his editorial review of this paper.
- The BMCC Women's Research Program for supporting the second and third authors of this paper throughout this research endeavor, and for supporting Kaur Gursimiran and Deanna Bernard, research interns who helped to compile biographical information on Blackwell.
- The Professional Staff Congress, City University of New York (PSC-CUNY) Research Program for supporting the first author in this research endeavor with a PSC-CUNY 29 Research Award, entitled, *Using Biography to Develop Mathematical Power, Encourage Diversity and Teach the History of Mathematics.*

REFERENCES

1. Nkechi Agwu, et. al., *Interviews with David Blackwell*, Women Research Project, Borough of Manhattan Community College, 1998–99.
2. N. Agwu, et. al., *Biographies of Statisticians, MAT 150—Introduction to Statistics Research Project*, Borough of Manhattan Community College, 1998–1999.
3. D. J. Albers and G. L. Alexanderson, eds., *Mathematical People*, Birkhauser, Boston, 1985.
4. D. H. Blackwell, Idempotent Markoff chains, *Ann. of Math. (2)* **43**:3 (1942), 560–567.
5. D. H. Blackwell, The existence of anormal chains, *Bull. Amer. Math. Soc.* **51** (1945), 465–468.
6. D. H. Blackwell, Finite non-homogeneous chains, *Ann. of Math. (2)* **46**:4 (1945), 594–599.
7. D. H. Blackwell, On an equation of Wald, *Annals of Mathematical Statistics* **17**:1 (1946), 84–87.
8. D. H. Blackwell, K. J. Arrow, and M. A. Girshick, Bayes and minimax solutions of sequential decision problems, *Econometrica* **17** (1949), 213–244.
9. D. H. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*, Wiley and Sons, New York, 1954.
10. M. H. DeGroot, A Conversation with David Blackwell, *Statistical Science* **1**:1 (1986), 40–53.
11. T. S. Ferguson, L. S. Shapley, and J. B. MacQueen, eds., *Statistics, Probability and Game Theory, Papers in Honor of David Blackwell*, Institute of Mathematical Statistics Lecture Notes—Monograph Series **30**, Hayward, California, 1996.
12. Institute for Operations Research and Management Sciences (INFORMS), John von Neumann Theory Prize Winners, 1975–2002, http://www.informs.org/Prizes/vonNeumannDetails.html#1979.
13. S. W. Williams, Mathematicians of the African Diaspora, http://www.math.buffalo.edu/mad/PEEPS/blackwell_david.html, 2002.

A Tale of Three Circles

CHARLES I. DELMAN
GREGORY GALPERIN
Eastern Illinois University
Charleston, IL 61920
cfcid@eiu.edu, cfgg@eiu.edu

Everyone knows that the sum of the angles of a triangle formed by three lines in the plane is 180°, but is this still true for *curvilinear* triangles formed by the arcs of three circles in the plane? We invite the reader to experiment enough to see that the angle sum indeed depends on the triangle, and that no general pattern is obvious. We give a complete analysis of the situation, showing along the way, we hope, what insights can be gained by approaching the problem from several points of view and at several levels of abstraction.

We begin with an elementary solution using only the most basic concepts of Euclidean geometry. While it is direct and very short, this solution is not complete, since it works only in a special case. The key to another special case turns out to be a model of hyperbolic geometry, leading us to suspect that the various manifestations of the problem lie on a continuum of models of geometries with varying curvature. This larger geometric framework reveals many beautiful unifying themes and provides a single method of proof that completely solves the original problem. Finally, we describe a very simple formulation of the solution, whose proof relies on transformations of the plane, a fitting ending we think, since a transformation may be regarded as a change in one's point of view. The background developed earlier informs our understanding of this new perspective, and allows us to give a purely geometric description of the transformations needed.

For the reader who is unfamiliar with the classical noneuclidean geometries, in which the notions of *line* and *distance* are given new interpretations, we provide an overview that is almost entirely self-contained. Such a reader will be introduced to such things as *angle excess*, *stereographic projection*, and even a sphere of imaginary radius. For the reader who is familiar with the three classical geometries, we offer some new ways of looking at them, which we are confident will reveal some surprises.

Three problems

Consider three circles in the plane intersecting transversely (that is, with no circles tangent to each other) at a common point, P, as in FIGURE 1. What is the sum of the measures of the angles of triangle of circular arcs ABC? Answer: 180°! (The picture gives a hint, but we will spell out the solution shortly.)

Next consider FIGURE 2, showing three circles whose centers are collinear. Now two (obviously congruent) curvilinear triangles are formed. What can be said about the sum of the angle measures in this case? Answer: This time, the sum is less than 180°!

Finally, consider three circles that intersect in the pattern of a generic Venn diagram, as in FIGURE 3. The boundary of the common intersection of their interiors is a convex curvilinear triangle; the sum of its angles is greater than 180°, because the straight-sided triangle with the same vertices lies inside it. What about the other six curvilinear

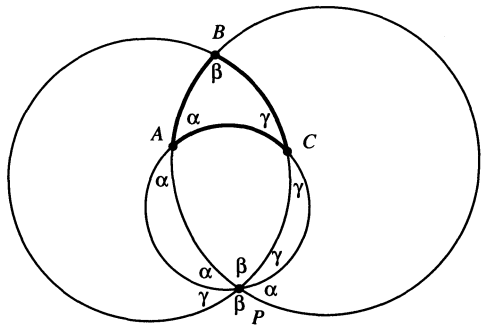


Figure 1 Three circles through a common point

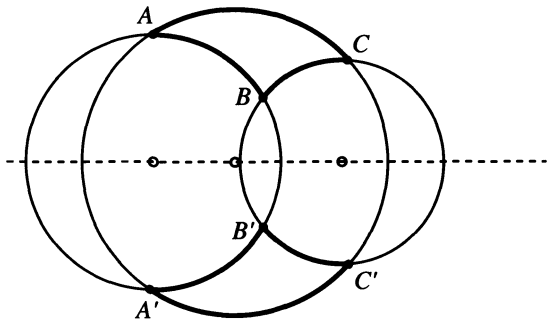


Figure 2 Three circles with collinear centers

triangles formed by these circles? Answer: As the reader might guess, although it is far from obvious from the diagram, the sum in this case is also always greater than 180°!

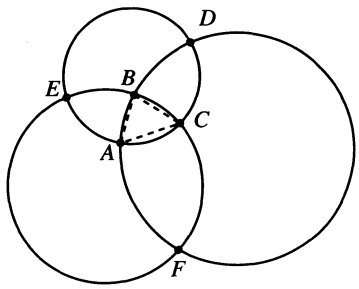


Figure 3 Three circles in Venn diagram position

The solution to the first problem admits an elementary proof. Consider triangle ABC in FIGURE 1. Note that by symmetry, the angles labeled with the letter α have the same measure, as do those labeled by β and by γ. We then see that $2(\alpha + \beta + \gamma) = 360°$, hence $\alpha + \beta + \gamma = 180°$.

The solution to the second problem also admits a simple proof, although a more advanced geometric idea is needed. Namely, the half-plane lying above the line through the three centers may be considered as the upper half-plane model of hyperbolic geometry. In this different sort of geometry, semicircles with centers on the boundary line take the place of lines in Euclidean geometry, and angles between these (hyperbolic) lines are computed using the Euclidean angles between the semicircles. A well-known

result in hyperbolic geometry states that the sum of the angle measures of any hyperbolic triangle is always less than 180°.

For the third problem, we will also show that the circles are *lines* in a model of geometry, this time spherical geometry, in which the angle sum of any triangle is more than 180°. More generally, we consider any configuration of three circles that intersect in pairs and give a simple criterion for deciding into which geometry they fall.

The general theorem

Three circles, provided they intersect in the way we have described, determine three lines by their points of pairwise intersection. It may surprise you to learn that these three lines are either concurrent or parallel. The position of the point common to these three lines, as described in the following lemma, is the key to determining which geometry will answer the question about the angle sum.

LEMMA. Let c_1, c_2, and c_3 be three circles in the plane, with any two of them intersecting in two distinct points. Let l_{12}, l_{13}, and l_{23} be the lines determined by these pairwise intersections. Then the lines $\{l_{ij}\}$ are either concurrent or parallel (in which case we consider them to be *concurrent at infinity*). Furthermore, there are exactly three possibilities for the location of the point of concurrency, P:

1. P lies on all three circles (FIGURE 4a);
2. P lies outside all three circles (FIGURE 4b), possibly at infinity; or,
3. P lies inside all three circles (FIGURE 4c).

Finally, if P lies outside all three circles, but not at infinity, the six tangents from P to the three circles all have the same length. The circle centered at P with this common length as radius is perpendicular to all three of the original circles. If P lies at infinity, the line through the centers of all three circles plays this role.

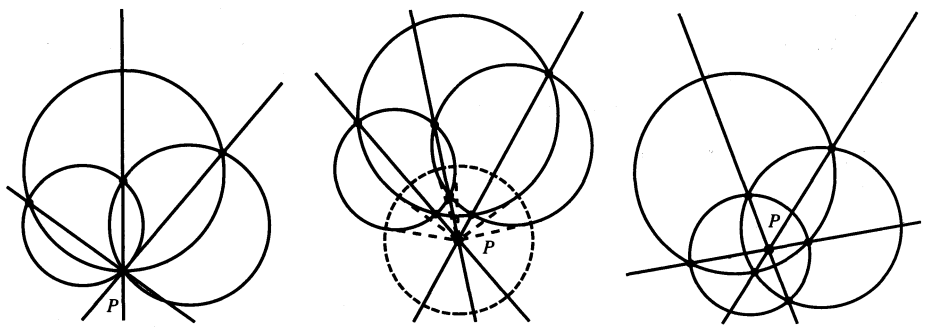


Figure 4 The three possibilities for the location of point P

Proof. Suppose a line through P intersects a circle c in two points A and B (which need not be distinct). The *power of P with respect to c* is defined as the signed product $(PA)(PB)$, where PA denotes the directed distance from P to A. We invite the reader to prove, using similar triangles, that the power does not depend on the line chosen. (A proof may be found in Coxeter and Greitzer [**2**, Theorem 2.11].)

The power of P with respect to c is positive if P is outside c, negative if P is inside c, and 0 if P lies on c. Viewing PA and PB as vectors, we can equivalently define

the power as the scalar product $PA \cdot PB$; since the vectors are parallel, the cosine of the angle between them is ± 1 according to whether they point in the same or opposite directions.

If two circles intersect at points A and B, we deduce (by considering line $\overleftrightarrow{PA}$, say) that a point P has the same power with respect to both circles if and only if it lies on line $\overleftrightarrow{AB}$. Suppose first that l_{12} and l_{23} are not parallel and let P be their point of intersection. Then P has the same power with respect to all three circles; hence, P lies on l_{13} and the lines are concurrent. If, on the other hand, two of the lines are parallel, an argument by contradiction shows that all three must be.

If the lines intersect, then, as we have just observed, the power of P with respect to all three circles is the same. Denoting this common power by $\mathcal{P}$, we have:

Case 1. If $\mathcal{P} = 0$, then P lies on all three circles.

Case 2. If $\mathcal{P} > 0$, then P lies outside all three circles.

Case 3. If $\mathcal{P} < 0$, then P lies inside all three circles.

Finally, if P lies outside, then the length of a tangent from P to any of the circles is $\sqrt{\mathcal{P}}$. The circle with this radius and center P is orthogonal to all three circles. ■

As an interesting digression, we note that for a general pair of circles, which need not intersect, a point has the same power with respect to both if and only if it lies on a particular line, called their *radical axis*, orthogonal to the segment joining their centers. In the course of computing the equation of this line in Cartesian coordinates, Coxeter and Greitzer [**2**, Section 2.2] also prove the beautiful fact that, if the equation of a circle is put in the standard form $F(x, y) = (x - a)^2 + (y - b)^2 - r^2 = 0$, then the power of any point (x, y) with respect to that circle is just $F(x, y)$! This result is not unexpected, since the power is constant on circles concentric with the given one.

We are now prepared to state the theorem. *Triangle* means a curvilinear triangle that is not subdivided by any arcs of the three circles. In case one, only the triangle that does not include P as a vertex is considered.

THEOREM. Let c_1, c_2, and c_3 be three circles in the plane, with each pair intersecting in two distinct points. Let l_{12}, l_{13}, and l_{23} be the lines determined by these pairwise intersections, with common point P. Then the sum of the angles of any triangle formed by the three circles is determined according to the three cases of the lemma:

1. if P lies on all three circles, the sum is equal to 180°;
2. if P lies outside all three circles, the sum is less than 180°; and,
3. if P lies inside all three circles, the sum is greater than 180°.

The proof of the theorem requires some knowledge of the classical noneuclidean geometries, to which we now turn our attention. The reader who is already familiar with these is invited to skip ahead to the section in which the theorem is proved as follows: For any of the possible configurations, we give a conformal map that takes our three circles to the lines of one of the three classical geometries; since the map is conformal, the angle sum is preserved, and thus found to be equal to, less than, or greater than 180° accordingly.

A quick tour of three geometries

A geometry is an abstract mathematical system in which the undefined notions of *point* and *line* are assumed to behave in accordance with certain axioms. (*Euclidean*

and non-Euclidean Geometries, by Marvin Greenberg [**4**], presents a lively account of the historical development of the classical geometries. Edwin Moise's text, *Elementary Geometry from an Advanced Standpoint* [**7**], is another very comprehensive reference.) In practice, we visualize a geometry by working with a model, which is described more concretely. Objects in the model represent points and lines in such a way that the axioms of the system are fulfilled.

Although there is much to say about the classical geometries, we continue our tale of three circles and focus on the sum of angles in a triangle.

The Cartesian plane, consisting of pairs of real numbers, is a model for Euclidean geometry, provided we adopt the usual concept of distance. Each pair of numbers represents a point, while lines are the solution sets of linear equations. Using algebra to study lines, points, and circles is called the *analytic method*. That the angle measures in a triangle sum to two right angles can be derived either from Euclid's axioms or an analytic approach.

The Cartesian plane and its physical approximations, such as tabletops, in which parallel lines remain equidistant and meet a common transversal line at the same angle on the same side, have historically been described as *flat* (as opposed to a sphere or other *curved* surface); thus, Euclidean geometry is said to be *flat*, or have *curvature zero*.

Any surface on which a distance function, or *metric*, has been defined is a model of some geometry. (These notions may be extended to higher dimensions as well.) The lines on this surface, called *geodesics*, are the curves that locally realize the shortest distances between their points.

Spherical geometry A sphere sitting in $\mathbb{R}^3$, where we know how to measure distances, inherits a metric that measures distances along the surface. Here, the lines are the *great circles*—those circles cut by planes through the sphere's center. The word *local* in the definition of geodesic is important; you must choose the short way around. The geodesic nature of great circles can be understood intuitively by noting that the distance along an arc is proportional to the central angle it subtends, and that central angles obey a sort of triangle inequality: where three planes meet, the sum of any two of the face angles is greater than the third. Thus, a path made up of very small (think *infinitesimal*) arcs will be shortest if all the arcs lie in the same plane through the center. A rigorous proof, as suggested by the preceding discussion, requires integration and other concepts from calculus. (*Geometry from a Differential Viewpoint* by John McCleary [**6**] is an accessible introduction to the application of calculus to geometry.)

Of particular importance to us is the sum of the angles of a spherical (geodesic) triangle. Our calculations will be nicer if we measure angles in radians, which we do from now on. Some experimentation, which we invite the reader to do, suggests that the angle sum of a spherical triangle always exceeds π and decreases with the triangle's area, approaching, but not reaching, π as this area approaches zero (and approaching, but not reaching, 5π as the triangle fills up the whole sphere). This observation suggests that we focus on the amount by which the angle sum of a spherical triangle exceeds π, which we call its *angle excess*. We now prove that a triangle's excess is always positive by precisely examining its relationship with the triangle's area.

Several facts will lead us to the correct relationship. First, notice that the angle excess is additive: if a triangle is subdivided into two smaller triangles, the excesses of the component triangles add up to the excess of the whole, as the reader can calculate. Second, congruent triangles clearly have the same angle excess. These are the essential properties of area: the areas of the parts of a subdivided region add up to the area of the whole, and congruent regions have equal area! Moreover, a sphere, like the plane, is *homogeneous*: any triangle can be rigidly moved, by rotations and reflections of the

sphere, to any other position without changing distances or angles (or area). Together, these facts suggest that, for a spherical triangle, angle excess is a constant multiple of area.

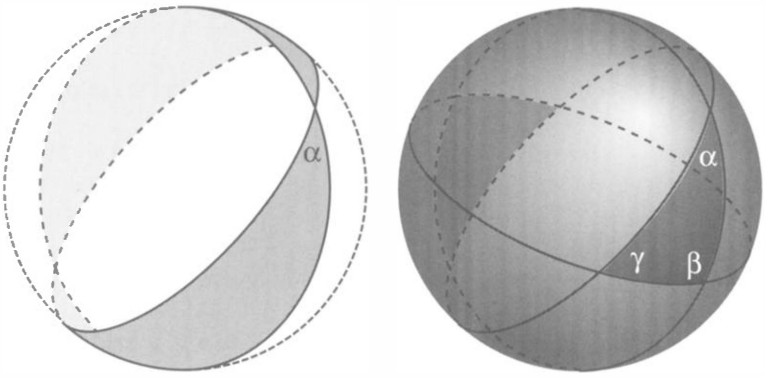


Figure 5 The sector swept out by a spherical angle

To prove that angle excess is proportional to area, observe that the angle between two great circles is proportional to the area of the sector they bound. (Note the essential role of homogeneity here!) On a sphere of radius R, the area of the sector swept out by an angle α is $(\alpha/\pi)(4\pi R^2) = 4R^2\alpha$. (See FIGURE 5a. For simplicity, we have used the same symbol to represent both the angle and its measure.) Let Δ represent the area of a triangle with angles α, β, and γ. The sectors swept out by α, β, and γ cover the sphere redundantly; the triangle and its antipodal image are each covered three times, while the remainder of the sphere is covered exactly once (FIGURE 5b). Thus, with a little algebra, we obtain the formula

$$\alpha + \beta + \gamma - \pi = \frac{\Delta}{R^2}.$$

In particular, on a sphere of unit radius, the angle excess is exactly equal to the area.

Hyperbolic geometry A sphere is said to have *constant positive curvature*, and it is easy to imagine that a small sphere has large curvature, while a large sphere has small curvature. What would it mean for a surface to have constant *negative* curvature? This question will lead us to a model where the sum of the angles in a triangle is always less than two right angles.

To see how we might describe a surface of negative curvature, we start with some formal manipulations on the equation $x^2 + y^2 + z^2 = R^2$. As the constant R^2 increases toward $+\infty$, the curvature of the sphere described by the equation diminishes toward zero. We may view the Euclidean plane as a sphere of infinite radius. What happens when the constant term passes infinity and reappears on the negative end of the number line?

We will call a surface satisfying an equation of the form $x^2 + y^2 + z^2 = -R^2 = (iR)^2$, which has been aptly described as a "sphere of imaginary radius," a *pseudo-sphere*. Of course, the equation will have no solutions unless we expand the domain of our variables from the real to the complex numbers, and its "radius" makes no sense unless we expand our notion of distance to include imaginary numbers.

Distance and angles in Cartesian space are measured via the *dot product*, familiar from multivariable calculus. This product is an example of a symmetric, bilinear form, a type of operation that plays a large role in many branches of mathematics. In addition,

the dot product is positive definite: for any vector $\mathbf{v}$, $\mathbf{v} \cdot \mathbf{v} \geq 0$, and $\mathbf{v} \cdot \mathbf{v} = 0$ if and only if $\mathbf{v} = 0$. The length of a vector $\mathbf{v}$, denoted $|\mathbf{v}|$, is defined as $\sqrt{\mathbf{v} \cdot \mathbf{v}}$ and the angle θ between two vectors is determined by the formula $\cos(\theta) = \mathbf{v} \cdot \mathbf{w}/(|\mathbf{v}||\mathbf{w}|)$.

If we allow the coordinates of our vectors to be any complex numbers, the dot product remains a symmetric, bilinear form, although it is no longer positive definite. A form with these properties is called a *pseudo-metric*. It is this form that we choose to measure distances and angles in complex Cartesian space, $\mathbb{C}^n$, allowing these quantities to have complex values.

Returning to the equation $x^2 + y^2 + z^2 = -R^2$, we consider, to obtain a two-dimensional surface, only those solutions for which x and y are real. The remaining variable, z, is then forced to be purely imaginary, so we let $z = it$. To picture the pseudosphere, we map the solution set of our equation to a surface in $\mathbb{R}^3$, using the map $(x, y, it) \mapsto (x, y, t)$. This image is a two-sheeted hyperboloid, the solution set of the equation $x^2 + y^2 - t^2 = -R^2$. The sheets are analogous to the two hemispheres of an ordinary sphere, the origin $(0, 0, 0)$ may be thought of as the center, and the points $(0, 0, \pm R)$ as the North and South poles.

In order to measure distances and angles, we must remember that our space is an image of a subspace of $\mathbb{C}^3$, where the actual distances and angles are measured using the dot product. Since a vector (x, y, t) actually represents the vector (x, y, it), its actual length is $\sqrt{x^2 + y^2 + (it)^2} = \sqrt{x^2 + y^2 - t^2}$. Similarly, the angle between two vectors (x_1, y_1, t_1) and (x_2, y_2, t_2) must be based on the form $x_1x_2 + y_1y_2 - t_1t_2$. Some readers may recognize this as a Lorentz metric, the three-dimensional version of the pseudo-metric of relativistic space-time. (For a readable and physically motivated, but advanced, introduction to pseudo-metrics, see *Semi-Riemannian Geometry*, by Barret O'Neill [**8**].) The apparent distances and angles in our picture are distorted from their actual values; for example, the true length of every radial vector from the origin to the surface is the imaginary number iR!

To measure distances and angles on the pseudosphere itself, we apply the pseudo-metric to its *tangent vectors*. For any point P in $\mathbb{R}^3$, the *tangent space at* P is the copy of $\mathbb{R}^3$ consisting of all vectors emanating from P. The tangent line at P to a curve through P on the surface is a one-dimensional subspace of this tangent space. The tangent plane to a surface at P is the two-dimensional space composed of all these lines.

It is useful (and intuitive) to write coordinates and other quantities related to tangent vectors in terms of *differential* expressions. If f is a differentiable, real-valued function on $\mathbb{R}^3$, students know how to compute ∇f, and use it to find directional derivatives by taking dot products. We prefer another vocabulary: the *differential* of f is the function that takes a tangent vector $\mathbf{v}_P$, at a point P, to the real number $\nabla f(P) \cdot \mathbf{v}_P$. This gives the best linear approximation to the change in the value of f that results from starting at P and travelling with a displacement $\mathbf{v}_P$.

In particular, the Cartesian coordinates, x, y, and t in our model may be viewed as functions of $P \in \mathbb{R}^3$, and dx, dy, and dt, are the differentials of these functions. If we suppress the vector argument of these differentials, we can use dx, dy, and dt as the first, second, and third coordinates of a general tangent vector with respect to a parallel coordinate system based at P. (In FIGURE 6a, the linear approximations dx, dy, and dt happen to be exact, since orthogonal projection onto an axis is a linear function.) Thus we may conveniently write the (Lorentz) length of a tangent vector as the differential expression $\sqrt{dx^2 + dy^2 - dt^2}$. This gives us the metric we need in order to talk about the lines of this geometry, which are the geodesics of this metric.

A beautiful fact is that this differential expression is real and positive, when applied to vectors tangent to the pseudosphere. Since the length of a path is computed by

integrating the lengths of its tangent vectors, paths between points on the pseudosphere have real, positive length, and it makes sense to talk about geodesics as paths of locally minimal length.

To see that $dx^2 + dy^2 - dt^2 > 0$, it helps to use cylindrical coordinates. As an orthonormal basis for the tangent space to $\mathbb{R}^3$ at the point with cylindrical coordinates (r, θ, t), take a radial unit vector, a circumferential unit vector in the counter-clockwise direction, and a vertical unit vector; with respect to this basis, the coordinates of a tangent vector are $(dr, rd\theta, dt)$. (The factor of r in the circumferential coordinate results from the fact that changing the central angle by a small amount changes the distance travelled by r times that amount. See FIGURE 6b.) The expression $dx^2 + dy^2$, the squared length of the vector's horizontal component, is equal to $dr^2 + r^2 d\theta^2$. For a vector tangent to the hyperboloid, $dt^2 < dr^2$, since the radial slope of the hyperboloid is less than that of the cone to which it is asymptotic. (Algebraically, it follows from the equation $r^2 - t^2 = -R^2$ that $r^2 < t^2$, and by differentiating we calculate that $dt^2 = (r^2/t^2)\, dr^2$.) Thus $dx^2 + dy^2 - dt^2 = r^2 d\theta^2 + (dr^2 - dt^2) > 0$.

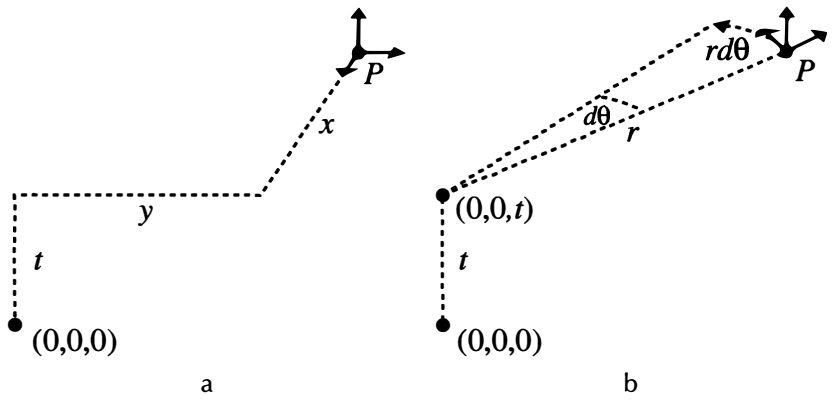


Figure 6 Comparison of two bases for the tangent space at P

The geodesics turn out to be the intersections of the pseudosphere with planes through the origin, analogous to those on an ordinary sphere, although it is harder to see why this is so, and we won't prove it here. The pseudosphere is also homogeneous, and the angle sum of a triangle satisfies a completely analogous formula:

$$\alpha + \beta + \gamma - \pi = \frac{\Delta}{-R^2}.$$

It follows that this sum is always less than π. Of course, there is a different pseudosphere for each value of R. By analogy with the sphere, one might guess that a small value of R yields a pseudosphere of large negative curvature, while a large value of R gives a pseudosphere that is so little curved as to be nearly flat. However great or small the curvature, the sum of the angles in a triangle is still less than two right angles. (A book by B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov [3] gives a thorough discussion of both the sphere and pseudosphere.)

Each sheet of the pseudosphere is a model of a geometric object called a *hyperbolic plane*. It would be nice to be able to see this plane looking more like a plane, without having to work with an object as complicated as the Lorentz metric. There is a planar map of the pseudosphere that shows angles accurately, obtained by a method called *stereographic projection*. A similar map is also available for the ordinary sphere. Any

map that preserves angle measure is called *conformal*. Since our tale of three circles involves angle measure, conformal maps will be a powerful tool.

Stereographic projection

The sphere To obtain a conformal mapping of the ordinary sphere onto the plane, project from any point of the sphere onto the plane tangent to its antipodal point. Any such map, or its inverse map from the plane to the sphere, is called a *stereographic projection*. Every circle on the sphere (not just the great circles) projects stereographically to a circle or line (which may be thought of as a circle through ∞) in the plane, and it maps to a line if and only if it passes through the point of projection (which maps to ∞). Conversely, every line or circle in the plane is the image of a circle on the sphere. (See FIGURE 7, and note there that the center of a circle on the plane is not, in general, the projection of the center of the corresponding circle on the sphere, but rather the apex of the cone tangent to it.) These valuable properties may be proven by elementary arguments. (The arguments are outlined and nicely illustrated in Hilbert and Cohn-Vossen's classic book, *Geometry and the Imagination* [**5**, pp. 248–251].)

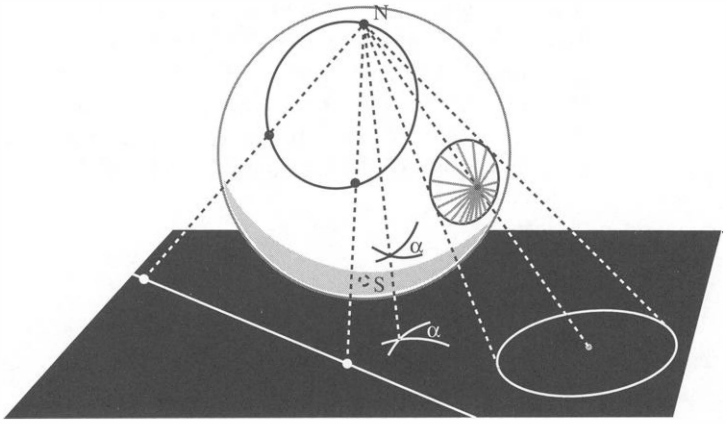


Figure 7 Stereographic projection preserves angles and takes circles to circles or lines

To see that stereographic projection from the sphere to the plane is conformal, consider FIGURE 8a, which shows the cross-section of the sphere cut off by a plane through the point of projection, N, its antipodal point, S, and another point P on the sphere; $\pi(P)$ is the projected image of P, and A is the point where the cross-section of the plane tangent to the sphere at P intersects the tangent plane at S.

An angle with vertex at P in the plane tangent to the sphere is cut by two planes intersecting along line $\overleftrightarrow{NP}$. Since $\angle\pi(P)PA \cong \angle P\pi(P)A$, these planes cut the same angle at $\pi(P)$ in the horizontal plane through S, by symmetry. To see this, imagine an angle formed by two stiff planes of paper; if you snip at the same angle to the spine in either direction, the angle you make is the same. (The projected image of the angle is its reflection in the plane through A that is perpendicular to $\overleftrightarrow{NP\pi(P)}$.)

FIGURE 8b illustrates the local behavior of stereographic projection. If Q is a nearby point on the tangent plane through P, then $\triangle NQP$ is approximately similar to $\triangle N\pi(P)\pi(Q)$. Therefore, stereographic projection is linearly approximated at each point P by a dilation (uniform scaling). The dilation factor varies with the latitude of P, increasing without bound as P approaches N, with a minimum value of 1 at S.

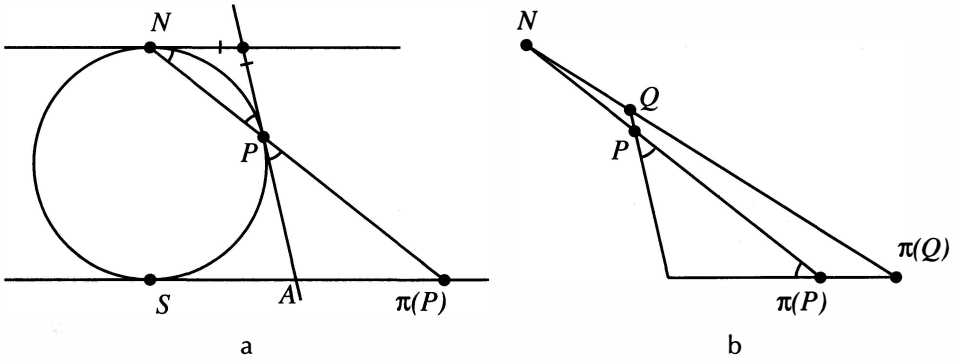


Figure 8 Stereographic projection is conformal and locally approximated by a dilation

The pseudosphere To obtain a conformal map of the pseudosphere onto the plane, project each point of the hyperboloid model onto the horizontal plane through the South pole, $(0, 0, -R)$ (that is, the plane $t = -R$), via the line connecting it to the North pole, $(0, 0, R)$, as in FIGURE 9. This projection maps the southern hemisphere onto the interior of the disk of radius $2R$ centered at the origin, while the northern hemisphere, except for the North pole, goes onto the disk's exterior. The disk's interior is known as a *Poincaré Disk* model of a hyperbolic plane, and its boundary is called the *circle at infinity*. It can be shown that each geodesic maps to a circle or line orthogonal to the circle at infinity (with the points of intersection removed); the geodesics through the poles go to lines. (See B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov [3] for a proof.)

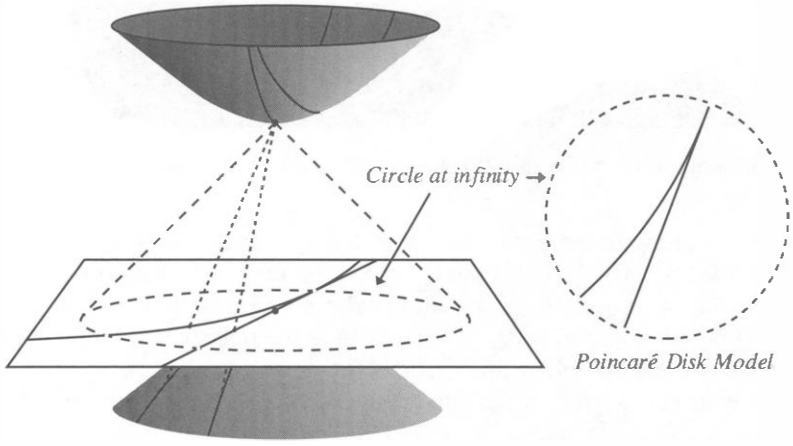


Figure 9 Projection of the hyperboloid model of the pseudosphere onto a plane

The conformality of the projection of the hyperboloid model cannot be demonstrated by elementary geometric arguments because the angles on the hyperboloid are measured using a different bilinear form than the one used to measure angles in the plane. So to prove that angle measure is preserved, we must resort to calculation to compare the angles between tangent vectors to curves, before and after projection. The reader who wishes to just believe us and avoid the technicalities involved may skip the

calculation below without any loss of continuity. For those who wish to venture in, the proof provides a nice application of calculus techniques to geometry.

Proof that projection to the Poincaré disk model is conformal We restrict our attention to the southern hemisphere; the proof for the northern hemisphere is similar. Let π denote the projection map, extended to the region $t < R$, and viewed as a map onto $\mathbb{R}^2$ by ignoring the last coordinate $(-R)$ in the image. Suppose curves α and β intersect at $P = (x, y, t)$. The derivative of π at P, $D\pi(P)$, carries vectors tangent to α and β at P to vectors tangent to their projected images at $\pi(P)$ (by the chain rule).

Let $\langle \mathbf{v}_P, \mathbf{w}_P \rangle_L$ denote the Lorentz product of tangent vectors $\mathbf{v}_P$ and $\mathbf{w}_P$ at P, that is,

$$\langle (x_1, y_1, t_1), (x_2, y_2, t_2) \rangle_L = x_1x_2 + y_1y_2 - t_1t_2.$$

We sketch a proof that, if $\mathbf{v}_P$ and $\mathbf{w}_P$ are tangent to the hyperboloid, then $D\pi(P)(\mathbf{v}_P) \cdot D\pi(P)(\mathbf{w}_P) = [4R^2/(R-t)^2]\langle \mathbf{v}_P, \mathbf{w}_P \rangle_L$. In other words, the dot product of the images is just scaled by a constant factor from the Lorentz product of the preimages. It follows that the angle, θ, between $\mathbf{v}_P$ and $\mathbf{w}_P$, which is determined by

$$\cos\theta = \frac{\langle \mathbf{v}_P, \mathbf{w}_P \rangle_L}{\langle \mathbf{v}_P, \mathbf{v}_P \rangle_L^{\frac{1}{2}} \langle \mathbf{w}_P, \mathbf{w}_P \rangle_L^{\frac{1}{2}}}$$

is equal to the angle, ψ, between $D\pi(P)(\mathbf{v}_P)$ and $D\pi(P)(\mathbf{w}_P)$, which is determined by

$$\cos\psi = \frac{D\pi(P)(\mathbf{v}_P) \cdot D\pi(P)(\mathbf{w}_P))}{[D\pi(P)(\mathbf{v}_P) \cdot D\pi(P)(\mathbf{v}_P)]^{\frac{1}{2}} [D\pi(P)(\mathbf{w}_P) \cdot D\pi(P)(\mathbf{w}_P)]^{\frac{1}{2}}},$$

since the scaling factor cancels out. In contrast to the sphere, the scaling factor *decreases* as P moves away from S (t decreases), with a *maximum* value of 1 at S.

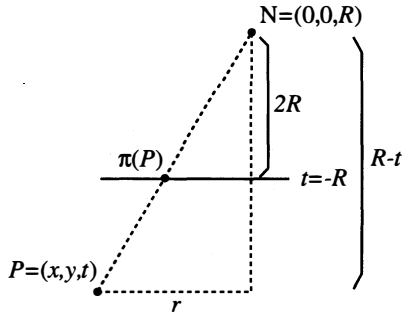


Figure 10 Effect of the projection π on the radial coordinate

Referring to similar triangles, as in FIGURE 10, we see that the image under π of the point P with cylindrical coordinates (r, θ, t) is the point whose polar coordinates are $(\frac{2Rr}{R-t}, \theta)$. The derivative at P of the conversion from Cartesian to cylindrical coordinates takes $(dr, rd\theta, dt)$ to $(dr, d\theta, dt)$. If $\phi : \mathbb{R}^3 \to \mathbb{R}^2$ is the map $(r, \theta, t) \mapsto (\frac{2Rr}{R-t}, \theta)$, its derivative (or Jacobian, if you prefer) at (r, θ, t) is

$$D\phi(r, \theta, t) = \begin{pmatrix} \frac{2R}{R-t} & 0 & \frac{2Rr}{(R-t)^2} \\ 0 & 1 & 0 \end{pmatrix}.$$

Applying $D\phi(r, \theta, t)$ to $(dr, d\theta, dt)$, followed by the derivative at $\phi(r, \theta, t) = (\frac{2Rr}{R-t}, \theta)$ of the conversion from polar back to Cartesian coordinates, we calculate (thanks to the ever-valuable chain rule) that $D\pi(P)(dr, rd\theta, dt) = \frac{2R}{R-t}(dr + \frac{rdt}{R-t}, rd\theta)$.

For any point P on the hyperboloid, $r^2 - t^2 = -R^2$, and for any vector tangent to the hyperboloid at P, $dt = (r/t)\,dr$. After some simplification using these substitutions, we find that $D\pi(P)(dr, rd\theta, (r/t)\,dr) = \frac{2R}{R-t}(-(R/t)dr, rd\theta)$. The reader can now check that for two tangent vectors at P, $D\pi(P)(v_P) \cdot D\pi(P)(w_P) = [4R^2/(R-t)^2]\langle v_P, w_P\rangle_L$, as stated.

Sequences of stereographic projections Stereographic projection is an indispensable tool for transforming geometric models without changing the angles between geodesics. In particular, it provides us with another celebrated model of the hyperbolic plane called the *Poincaré half-plane*. To obtain it, first project the Poincaré disk of radius two from a horizontal plane onto the southern hemisphere of the unit sphere (and its complement, including the point at ∞, onto the other hemisphere). Then project onto any vertical plane tangent to the equator. The image of the circle at infinity under this sequence of projections is called the *line at infinity*. Each half-plane is an image of a hyperbolic plane, as in FIGURE 11.

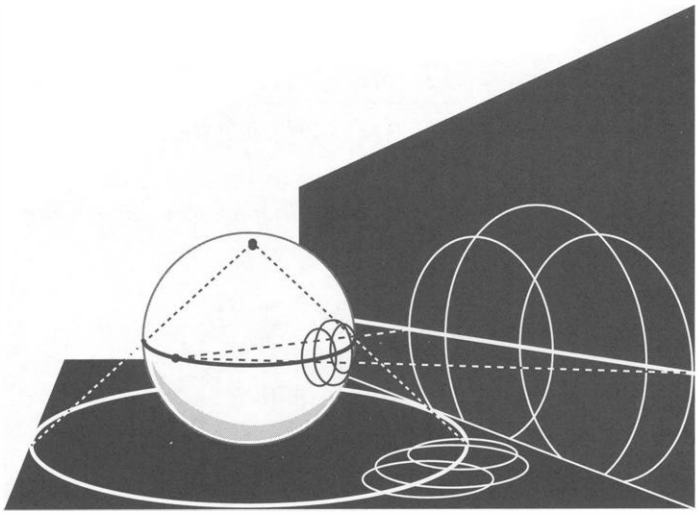


Figure 11 Conformal transformations between the Poincaré disk, hemisphere, and Poincaré half-plane models of hyperbolic geometry

By a sequence of two stereographic projections from different points, we also obtain a conformal map of the Euclidean plane in which the images of the geodesics are either lines or circles. See FIGURE 12.

In summary, we now have maps of the plane, sphere, and pseudosphere in which the geodesics are represented by lines and circles. Just as a map of the earth must be distorted in order to print it on the flat pages of an atlas, our maps distort the true distances between points in the geometric objects they depict. The art of map-making revolves around choosing a projection whose particular type of distortion allows the map to be useful for its intended purpose. For example, the famous Mercator projection of the earth's surface is useful to navigators because it accurately depicts the compass

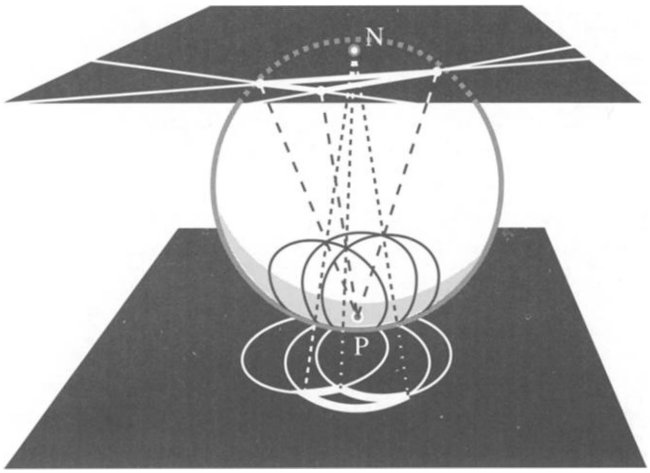


Figure 12 Euclidean geometry on the sphere

bearing between any two points. (McCleary [**6**, chapter 8^{bis}] gives a nice discussion of map projections.)

Since we are interested in the properties of angle measure, we have chosen to view our circles through conformal maps, which render the true angles between smooth curves. The question remains: Given a trio of circles that intersect in pairs, how can we interpret the configuration as a geodesics on a map of one of the geometries we have studied? Is such an interpretation even possible?

Choosing a geometry: a proof of the theorem

Recall that P is the intersection of the three lines determined by the pairwise intersections of the three circles. We will show that, depending on the location of P relative to the circles, there is a conformal map that takes the circles to lines of a standard geometric model—a model for which we know a great deal about angle sums.

Case 1: P lies on all three circles. In this case consider a sphere tangent to the plane with South pole at P, and a second plane tangent to this sphere at the North pole, as in FIGURE 12. The images of our three circles on the sphere are circles through the South pole. If we project again, this time from the South pole of the sphere onto the second plane, these circles are taken to straight lines. Thus, c_1, c_2, and c_3 are geodesics in a conformal planar model of Euclidean geometry, and the sum of the angles of the triangle they form is 180°.

Case 2: P lies outside all three circles. We already studied the special case where the three lines determined by pairwise intersections of the circles are parallel; this happens when the centers of all three circles are collinear, and that line was seen to be the boundary line of a hyperbolic plane. Having dispensed with that case, we assume that the lines are concurrent at a point P lying outside of all three circles. In this case, the Lemma presents us with a circle, d, that is orthogonal to all three circles. Thus, each arc of the circles lying inside or outside d is a geodesic in a model of a hyperbolic plane, and thus the sum of the angles of any triangle they form is less than 180°.

Case 3: P lies inside all three circles. Consider the family of spheres (of varying size) tangent to the plane with South pole at P. Consider how the area of the stereo-

graphic projection of the disk bounded by c_1 onto each of these spheres compares to the surface area of the sphere: if the sphere is very small, the projected region will have area more than half that of the sphere (in fact, it will include the entire Southern hemisphere); if the sphere is very large, the projected area will be less than half the surface. This is just a matter of having c_1 lie inside or outside the preimage of the equator on the plane. (See FIGURE 13.) Allowing the sphere to vary continuously, we see that there is a unique sphere, $\mathcal{S}$, such that the projected area on $\mathcal{S}$ is exactly half the surface area. (For an alternative argument, see the third remark below.) Consequently, the image on $\mathcal{S}$ of c_1 is a great circle.

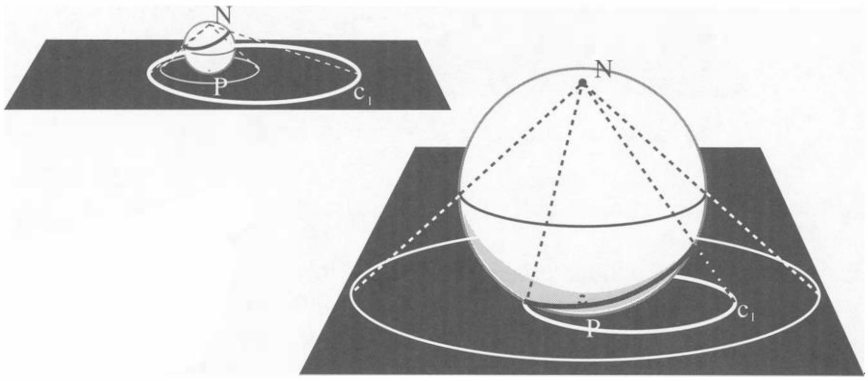


Figure 13 The family of spheres tangent to the plane at P

We claim that the images on $\mathcal{S}$ of c_2 and c_3 are also great circles. It suffices to prove this for c_2, as the same argument works for c_3. To do so, observe that the image of l_{12} under stereographic projection is a *meridian*, that is, a great circle passing through the North and South poles of the sphere. Since the image of c_1 is also a great circle, the images of the points of intersection of c_1 with l_{12} are antipodal. Since the points of $c_1 \cap l_{12}$ also lie on c_2, it follows that the image of c_2 is a great circle. FIGURE 14 illustrates this.

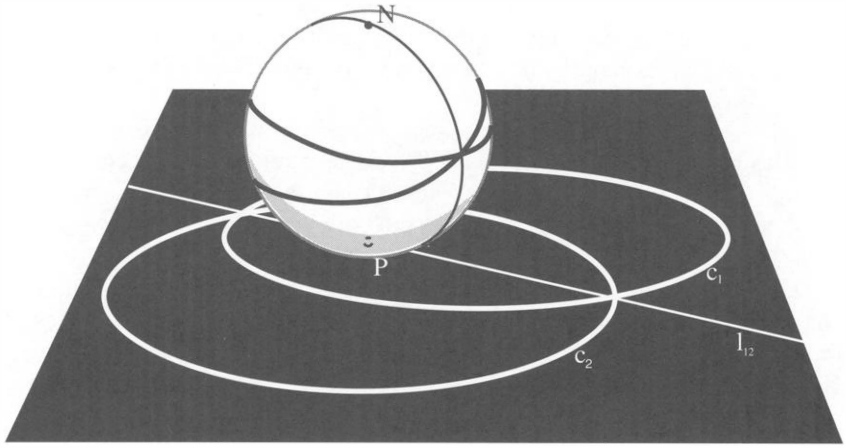


Figure 14 The images of the circles on the sphere $\mathcal{S}$ are great circles

We have thus shown that the images on $\mathcal{S}$ of the three circles are great circles; that is, they are spherical geodesics. Thus c_1, c_2, and c_3 are geodesics in a conformal planar model of spherical geometry, and the sum of the angles of any triangle formed by them is greater than 180°. ■

REMARK. That Euclidean geometry occurs only in the instance that point P lies exactly on the circles is an illustration of the fact that our familiar flat geometry is just a single point in the spectrum of geometries, from the sphere of large radius R, to the flat plane, where R is effectively infinite, to the pseudosphere, whose radius is imaginary in the model we have shown.

REMARK. Using the sequence of stereographic projections described earlier, we may transform the general picture for Case 2 into the half-plane picture.

REMARK. Although we like the continuity argument in Case 3 of the Theorem, it may be avoided as follows. Let $s = \sqrt{-\mathcal{P}(P)}$. The sphere $\mathcal{S}$ in the above proof is the one that has radius $s/2$. To show this, consider, for each of the circles c_1, c_2, and c_3, the chord through P that is perpendicular to the radius through P. Each of these chords has length $2s$ and midpoint P. Let e be the circle centered at P with radius s. This circle projects stereographically to the equator of $\mathcal{S}$, and c_1, c_2, and c_3, which intersect e at diametrically opposite points, also project to great circles. This alternative argument emphasizes the parallels between Cases 2 and 3.

REMARK. The theorem remains true if *circle* is understood in its more general sense to include straight lines as well.

We invite the reader to extend the theorem to the limiting cases in which two or more of the circles are tangent. The possibilities are illustrated in FIGURE 15. Case I

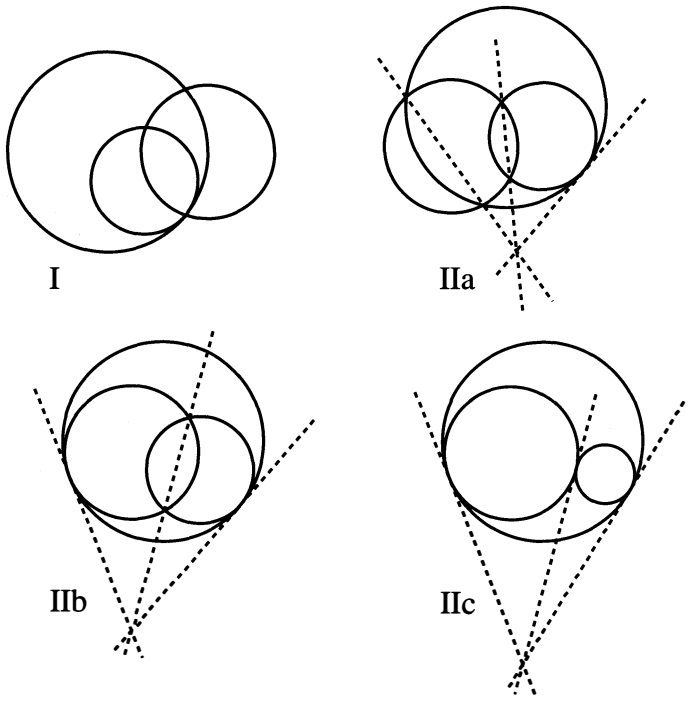


Figure 15 The possibilities for two circles to be tangent

of the figure leads to Euclidean geometry, when viewed with the right map: One vertex of the triangle is thrown out to infinity by the map that takes the circles to straight lines; hence, two sides of the triangle become parallel lines. The possibilities in Case II are hyperbolic, with one or more vertices of the triangle lying on the circle (or line) at infinity. What is the measure of an angle whose vertex lies on the circle at infinity? Note that none of the limiting cases is spherical.

Changing your point of view: a transformational approach

We are indebted to Keith Burns for pointing out the following very elegant formulation and proof of the Theorem. His proof uses the group of Möbius transformations of the extended complex plane (that is, the plane, regarded as the field of complex numbers, together with a point at ∞). Möbius transformations are invertible, bicontinuous, and conformal, and take generalized circles (with lines regarded as circles through ∞) to circles. Moreover, given any two ordered sets consisting of three points each, there is a (unique) Möbius transformation taking each point of one set to the corresponding point of the other. (The excellent, very readable text on geometry from a Kleinian point of view by Brannan, Esplen, and Gray contains a thorough discussion of the Möbius group and its geometric properties [**1**, Chapter 5].)

In particular, and this is all we will need, there is a Möbius transformation taking any given point to ∞. As we have seen, a transformation with the required properties can be constructed by composing a pair of stereographic projections.

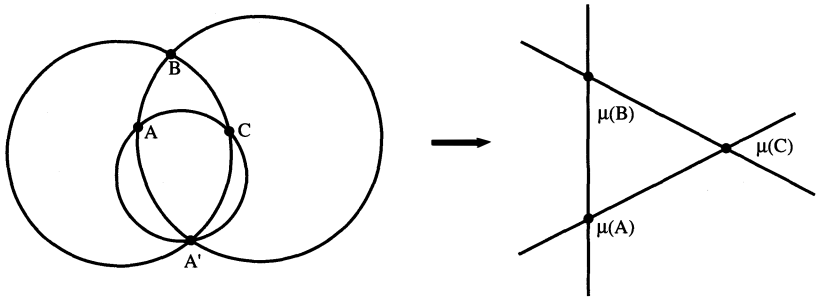
Consider a curvilinear triangle ABC formed by the three circles c_1, c_2, and c_3. Without loss of generality, assume that A lies on c_1 and c_2, and let A' be the other point of intersection of c_1 and c_2. We then have three possibilities for the positions of A and A' with respect to the third circle, c_3, which correspond to the three cases of the Lemma and Theorem: A' lies on c_3; A' lies on the opposite side of c_3 from A; or A' lies on the same side of c_3 as A. By simply changing our point of view, placing A' at ∞, we can discern by inspection how the angle sum of triangle ABC compares to 180°. Although this approach does not demonstrate the underlying global geometry in which the circles are geodesics, it does have the advantage of being delightfully simple and direct.

THEOREM. (ALTERNATIVE FORMULATION) Let c_1, c_2, and c_3 be three circles in the plane, with each pair intersecting in two distinct points. Then exactly one of the following three conditions holds and determines the sum of the angles of any triangle formed by the three circles:

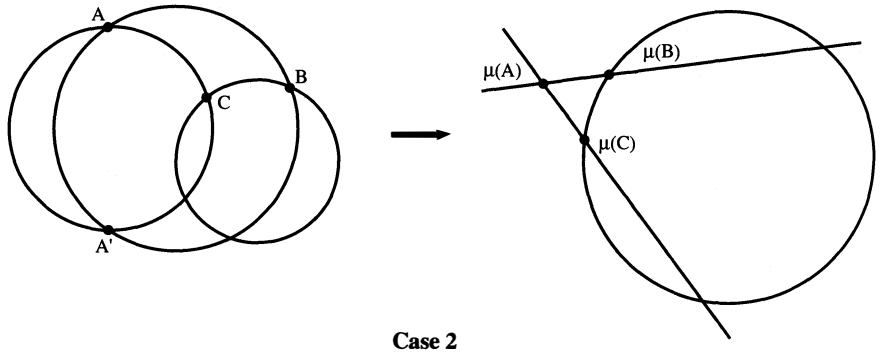
1. the intersection of each pair of circles contains a point of the third circle, in which case the sum of the angles is 180°;
2. the intersection of each pair of circles lies entirely inside or outside the third circle, in which case the sum of the angles is less than 180°; or,
3. the intersection of each pair of circles contains one point inside and one point outside the third circle, in which case the sum of the angles is greater than 180°.

Proof. Let A, B, and C be the vertices of a triangle formed by c_1, c_2, and c_3, with $\{A, A'\} = c_1 \cap c_2$, as in the paragraph preceding the statement of the theorem. The angle sum of the triangle does not depend on which pair of circles we consider, so it suffices to show that this sum is determined by the positions of A and A' relative to c_3.

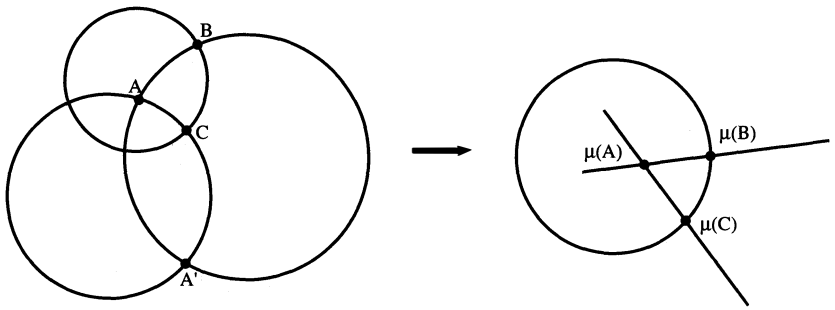
Apply a Möbius transformation, μ, that takes A' to ∞. Under this transformation, the images of c_1 and c_2, which pass through A', are lines. If A' lies on c_3, then the image of c_3 is also a line, hence the angle sum of the image of triangle ABC is 180°.



Case 1



Case 2



Case 3

Figure 16 The effect of a Möbius transformation, μ, taking A' to ∞

If A and A' lie on the same side of c_3, then the image of c_3 is a circle with the image of A *outside* it, since ∞ is outside; hence the angle sum of the image of triangle ABC is less than 180°. If A and A' lie on opposite sides of c_3, then the image of c_3 is a circle with the image of A *inside* it; hence the angle sum of the image of triangle ABC is greater than 180°. The three possibilities are illustrated in FIGURE 16. ■

Conclusion

Each of the classical geometries, Euclidean, spherical, and hyperbolic, has a variety of conformal representations on the plane, obtained by stereographic projection. Rec-

ognizing these representations enabled us to classify the generalized triangles whose sides are either segments or circular arcs as belonging to one of three types of geometries, allowing us to make the correct conclusion about the sum of the angles. Finally, by using a group of conformal transformations of the extended plane, we were able to refine our solution to be very simple, although perhaps less informative. Our success is just one example of the value of studying noneuclidean geometries and transformation groups.

Acknowledgments. In addition to Keith Burns, we would like to thank Duane Broline, Leo Comerford, Bob Foote, Yuri Ionin, and Rosemary Schmalz for their helpful comments. We would also like to thank an early referee for pointing out the argument in the third remark.

REFERENCES

1. D. Brannan, M. Esplen, and J. Gray, *Geometry*, Cambridge University Press, Cambridge, 1999.
2. H. S. M. Coxeter and S. L. Greitzer, *Geometry Revisited*, Random House, New York, 1967.
3. B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern Geometry—Methods and Applications. Part I. The geometry of surfaces, transformation groups, and fields*, Graduate Texts in Mathematics # 93, Springer-Verlag, New York, 1992.
4. M. J. Greenberg, *Euclidean and non-Euclidean Geometries*, Freeman, New York, 1993.
5. D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelsea, New York, 1952.
6. J. McCleary, *Geometry from a Differentiable Viewpoint*, Cambridge University Press, Cambridge, 1994.
7. E. Moise, *Elementary Geometry from an Advanced Standpoint*, Addison-Wesley, Reading, 1990.
8. B. O'Neill, *Semi-Riemannian Geometry, with Applications to Relativity*, Academic Press, New York, 1983.

Pete, Repete, and One Bagel

Pete: Do you want a bagel?

Repete: Oh, no, I couldn't eat that much.

Pete: I could halve the bagel.

Repete: Yes, you should have it.

Pete: No, if we halve the bagel, we could each have half a bagel.

Repete: Oh, OK. But that brings up a tricky question.

Pete: Yes?

Repete: Before you halve a bagel, a bagel will have a hole. After you halve the bagel, does half a bagel have half a bagel hole, or it is a whole hole?

Pete: A whole!

Repete: But if half a bagel is to have a whole hole, then when you halve a whole bagel, you don't, in fact, halve a bagel hole, since in each half you have a whole hole, which is two holes.

Pete: Mysterious! So when you halve a bagel, you get to have a bagel half and a bagel hole.

Repete: Right! This means that when you have a whole bagel, you get half the bagel holes you get when you halve a whole bagel!

Pete: Holy Cow!

—GLENN APPLEBY
SANTA CLARA UNIVERSITY
SANTA CLARA, CA 95053

NOTES

Power Distribution in Four-Player Weighted Voting Systems

JOHN TOLLE
Carnegie Mellon University
Pittsburgh, PA 15213-3890
tolle@qwes.math.cmu.edu

The Hometown Muckraker is a small newspaper with a few writers and layout personnel, and an editorial staff of four. When major policy decisions require that the editorial staff vote, the Editor-in-Chief gets 3 votes; the Managing Editor gets 2 votes; and the News Editor and Feature Editor each get 1 vote, for a total of 7. A majority (4 votes) is needed to pass a motion.

Lately, however, the Managing Editor has begun to feel that she has no more say in policy decisions than the News Editor or the Feature Editor has. In a private meeting with the Editor-in-Chief, she was told, "That's ridiculous; you have twice as many votes as they do, and therefore you have twice as much say." Is the Editor-in-Chief right? Or is it in fact true that the News Editor and the Feature Editor have just as much influence as the Managing Editor?

In this note we show that not only is the Editor-in-Chief wrong, but there is no way to remedy the problem, at least in the sense of restructuring the voting system so that there is the following hierarchy of influence, or *say* on policy matters:

(i) The Editor-in-Chief has the most influence, but less than half of the *total influence.*

(ii) The Managing Editor has less say than the Editor-in-Chief but more influence than the other two editors.

(iii) The News and Feature Editors have equal influence, but less than their senior editors.

Now consider a second scenario: Pennsylvania has decided to secede from the United States and invites its neighbors to do the same. There are three takers: Ohio, New York, and West Virginia. The four states decide to keep the existing geographical boundaries in place and form the Republic of PAIN (Pennsylvania and its Neighbors).

An electoral college system is instituted to elect the head of the executive branch of government, who is called the Big Cheese. An integer number of electoral votes for each state is chosen which is roughly proportional to population. (How to fairly apportion the total number of votes among the four states is another problem entirely; potential solutions to that problem have been much studied, and we shall assume that the framers of the Republic of PAIN are oblivious to the complexity of that problem and have apportioned the electoral votes in a manner satisfactory to them. Balinski and Young [**1**] provide a comprehensive treatment of the apportionment problem.) A candidate must receive a strict majority of the electoral votes in order to become the

Big Cheese. These votes are distributed as follows:

New York	38
Pennsylvania	25
Ohio	23
West Virginia	4

Since there are a total of 90 electoral votes, 46 are required for a victory.

A provision is in place to reapportion the electoral votes based on population changes. However, the population of each state continues to grow at a (universally) constant percentage rate, so that the four states maintain their ratios to the total population. However, after a few Big Cheese elections, West Virginia legislators propose a restructuring of the electoral college system, because they have begun to believe that their citizens actually have no influence whatsoever in Big Cheese elections. Unfortunately, their claim falls on deaf ears, because the other three states argue that the small number of electoral votes West Virginia casts is perfectly appropriate for its size and that these four votes *do* matter. Who is right?

Perhaps in this second scenario it is easier to see the validity of the objection: Any two of Ohio, New York, or Pennsylvania can agree on a Big Cheese candidate, and that candidate will win. Conversely, no candidate can win the election without carrying two of these states. As we will argue more carefully in the next section, it is in fact true that West Virginia's votes *don't* count; the key point here is that West Virginia can never cast the *deciding* votes. But if it is inappropriate for the Republic of PAIN to allocate votes in proportion to population, then how should the voting system be restructured to give West Virginia some measure of influence befitting its size? Or will it turn out, as in the previous example, that the desired objective cannot be achieved?

The Banzhaf Power Index The tool which we shall use to measure influence in each of our two hypothetical examples is called the *Banzhaf power index*, developed by attorney John Banzhaf [**2**] in the 1960s to argue that among the six districts of Nassau County, New York, represented by the Board of Supervisors, only the three largest districts wielded any real influence. (See also chapter 2 of the book by Tannenbaum and Arnold [**4**].) Informally speaking, to measure a voting party's influence, we ask, "How likely is it (indeed, does it occur at all) that this party can cast deciding votes?"

Following Tannenbaum and Arnold [**4**], we assume a finite number of voting parties called *players*, denoted $P_1, P_2, \ldots, P_n$. Player P_i casts a positive integer number of votes, v_i. The number of votes q needed to pass a motion shall be called the *quota*, and we assume

$$\frac{v_1 + \cdots + v_n}{2} < q \leq v_1 + \cdots + v_n.$$

We shall call this a *weighted voting system* (WVS) of size n, and represent it by

$$[q; v_1, v_2, \ldots, v_n]$$

and assume $v_1 \geq v_2 \geq \cdots \geq v_n$.

We shall also stipulate that $v_i < q$ holds for each i; otherwise it would be possible for some P_i to vote alone and pass a motion. But in addition, we assume that for all i,

$$\sum_{j \neq i} v_j \geq q.$$

If this were not true for some i, then P_i would have the power to prevent any motion from passing; the votes of all the other players would not exceed the quota. We call such a condition *veto power*.

A *coalition* in a WVS is a subset of the players. A *winning coalition* is a coalition for which the combined votes of the players exceed the quota. Otherwise, we have a *losing coalition.* We call a player in a winning coalition *critical* if the removal of that player results in a losing coalition. The presence of a critical player in a winning coalition will be called a *critical instance*; if a winning coalition has k critical players, we shall say that there are k critical instances corresponding to that coalition.

We are now ready to define the *Banzhaf power index* for a player P_i as the ratio of the number of instances in which P_i is critical to the total number of critical instances. We shall denote this ratio by $B(P_i)$. Note that since each $B(P_i)$ is a percentage, we have

$$B(P_1) + \cdots + B(P_n) = 1.$$

The aim of the Banzhaf power index is to measure the percentage of the total amount of power each player in the WVS possesses. The underlying philosophy is that power is characterized by the ability to cast deciding votes, and if there is a player who can never do this, then that player has no effective power.

There are other measures of power in weighted voting systems, such as the Shapley-Shubik power index, which in fact predated the Banzhaf index by a decade. The Shapley-Shubik model, however, is better suited to situations for which the order in which the players vote is a factor. In that case, one considers all $n!$ permutations of the n-player coalition; associated to each one is a unique player whose votes bring the cumulative total up to (or past) the quota, if all players vote in favor of a motion. That player is considered critical. The Shapley-Shubik power index for P_i is then the total number of instances in which P_i is critical, divided by $n!$.

The Banzhaf and Shapley-Shubik power distributions for a given WVS can sometimes agree, but they can also be dramatically different. (Chapter 9 of Taylor's book [**5**] provides an example, and also other models of power.) We have chosen the Banzhaf model here because we have no reason to distinguish the various permutations of a given winning coalition. Our tacit assumption is that all players vote at once, and each possesses no knowledge of how the others are voting.

With the aid of Banzhaf's power index, let us verify that West Virginia is powerless in the Republic of PAIN example, represented by the WVS [46; 38, 25, 23, 4]. The players here are P_1, P_2, P_3, and P_4, corresponding to New York, Pennsylvania, Ohio, and West Virginia, respectively. The winning coalitions, along with their *weights* (total numbers of votes cast), are as follows:

Coalition	Weight
$\{P_1, P_2\}$	63
$\{P_1, P_3\}$	61
$\{P_2, P_3\}$	48
$\{P_1, P_2, P_3\}$	71
$\{P_1, P_2, P_4\}$	67
$\{P_1, P_3, P_4\}$	65
$\{P_2, P_3, P_4\}$	52
$\{P_1, P_2, P_3, P_4\}$	90

Observe that no winning coalition to which West Virginia (P_4) belongs has a weight exceeding the quota by less than 4 votes. For this reason, West Virginia is not a critical player in any winning coalition, so that $B(P_4) = 0$. This is already an interesting finding, but let us look closer, and note that in each 2-player winning coalition, each player

is critical. This gives 6 critical instances so far. Then when we examine $\{P_1, P_2, P_3\}$, we see that the removal of any one player from this coalition results in a 2-player winning coalition; even without arithmetic we can see this by referring to the list of 2-player winning coalitions. Hence $\{P_1, P_2, P_3\}$ yields no critical instances. For the other 3-player winning coalitions, however, we find

- P_1 and P_2 are critical in $\{P_1, P_2, P_4\}$
- P_1 and P_3 are critical in $\{P_1, P_3, P_4\}$
- P_2 and P_3 are critical in $\{P_2, P_3, P_4\}$

This gives another 6 critical instances. The 4-player winning coalition has no critical players (so that veto power is absent in this WVS), and we have a total of 12 critical instances, with each of P_1, P_2, and P_3 critical in 4 instances. Hence we find $B(P_1) = B(P_2) = B(P_3) = 1/3$.

The surprising revelation behind our analysis, then, is that not only does West Virginia have no power, but the other three states have equal power. So despite the proportional representation apparently afforded by the electoral college system, the allocation of votes may as well be as follows: 0 to West Virginia and 1 to each of the other states, with 2 electoral votes needed to become the Big Cheese. (According to our stipulation that each v_i should be positive, we would not consider [2; 1, 1, 1, 0] to be an acceptable WVS. However, any WVS of the form $[2m; m, m, m, 1]$, with $m \geq 2$, would yield the same distribution of power as we find in the Republic of PAIN example.) The outcomes of all elections would be the same.

No wonder, then, that West Virginia proposes a change. The good news is that change is possible; the bad news, as the following result reveals, is that the options are quite limited.

THEOREM. *In any 4-player WVS with no veto power, there are only five possible power distributions:*

(a) $B(P_i) = 1/4$ *for every* i.
(b) $B(P_4) = 0$ *and* $B(P_i) = 1/3$ *for* $i \neq 4$.
(c) $B(P_1) = 1/2$ *and* $B(P_i) = 1/6$ *for* $i \neq 1$.
(d) $B(P_1) = B(P_2) = 1/3$ *and* $B(P_3) = B(P_4) = 1/6$.
(e) $B(P_1) = 5/12$, $B(P_2) = B(P_3) = 1/4$, *and* $B(P_4) = 1/12$.

Moreover, if $v_3 = v_4$, *then alternatives (b) and (e) are not possible.*

The reader can check that alternative (c) holds in the Hometown Muckraker example, so that indeed, the Managing Editor has no more power in the Banzhaf model than the News and Feature Editors have. Furthermore, it is impossible for the editorial staff to effect a hierarchy of power that is consistent with rank.

In the Republic of PAIN example, alternative (b) holds, and it would seem that the only sensible way to restructure the WVS so that all four states have some power, in amounts commensurate with their populations, is to effect alternative (e). The WVS [11; 7, 5, 4, 2] does this, and if Ohio should object to having only twice as many electoral votes as West Virginia, even though its population is over six times larger, then perhaps the WVS [13; 8, 6, 6, 1] would be acceptable; the power distribution still corresponds to alternative (e). Another example is [90; 80, 65, 20, 5], in which Pennsylvania casts over three times as many votes as Ohio, a condition which might be psychologically disturbing to Ohio residents. Yet the power index for the two states is the same. The point is that, given two players P_i and P_j, the ratio v_i/v_j is not a reliable indicator of comparative power.

To further reinforce the point, consider the 3-player WVS [13; 12, 11, 2]. All 2-player coalitions are winning coalitions, and in all three cases, both players are critical. But in the 3-player coalition, none of the players is critical. Thus the total number of critical instances is 6, and each player is critical in two instances. Hence each player has power index 1/3. This is true despite the fact that P_1 casts six times as many votes as P_3.

We include a fairly uninspiring proof of the theorem; one objective of this note is to stimulate interest in finding a slicker way to analyze weighted voting systems of a given size. We have treated size 4 because the brute force argument we give is not too cumbersome. One pleasing aspect of the proof is the revelation that one need only consider the various combinations of winning coalitions, and the proof relies only on the absence of veto power and not on the actual value of the quota q. Nor must one ever consider by how many votes a coalition wins; one need only know whether the coalition wins or loses. The proof also reveals that no matter the size of a WVS, only finitely many power distributions are possible (a fact which may be readily evident to a combinatorialist), so that in general, it is not possible to construct a WVS with a prescribed power distribution.

Proof of the Theorem We show that only five power distributions are possible in a 4-player WVS by exhausting the possible compositions of the set of winning coalitions. We use three important consequences of the absence of veto power:

1. The 4-player coalition is a winning coalition that does not yield any critical instances (since if one player were critical, then that player would have veto power). Therefore all critical instances occur among the 2- and 3-player winning coalitions.
2. All of the 3-player coalitions must be winners; if not, the missing player has veto power, since the 4-player coalition is certainly a winner. Hence the various cases are distinguished completely by which 2-player coalitions are winners. (We must point out, however, that changing the collection of 2-player winning coalitions will affect instances of criticality in 3-player winning coalitions, so the 3-player coalitions must still be examined.)
3. In any 2-player winning coalition, both players must be critical, since we do not permit single-player winning coalitions.

Our work is further simplified by observing that critical instances occurring in 3-player winning coalitions can always be detected by referring to the list of winning 2-player coalitions.

Since we need only consider cases distinguished by which 2-player coalitions win, let us start by observing the following rankings, in descending order of total weight, for 2-player winning coalitions.

- $\{P_1, P_2\}, \{P_1, P_3\}, \{P_1, P_4\}, \{P_2, P_4\}, \{P_3, P_4\}$
- $\{P_1, P_2\}, \{P_1, P_3\}, \{P_2, P_3\}, \{P_2, P_4\}, \{P_3, P_4\}$

Note that what is uncertain is how $\{P_1, P_4\}$ and $\{P_2, P_3\}$ compare. However, if one of these coalitions wins, then the other loses. In fact, if any 2-player coalition wins, then its complement loses; for example, if $\{P_1, P_2\}$ wins, then $\{P_3, P_4\}$ must lose, for otherwise we would have

$$v_1 + v_2 \geq q \quad \text{and} \quad v_3 + v_4 \geq q,$$

so that the strict majority requirement

$$q > \frac{v_1 + v_2 + v_3 + v_4}{2}$$

cannot hold for the quota.

It follows that there cannot be more than three 2-player winning coalitions in a 4-player WVS and that the coalitions $\{P_2, P_4\}$ and $\{P_3, P_4\}$ always lose.

We now consider the various cases. First, if there are no 2-player winning coalitions, then every player is critical in each of the four 3-player winning coalitions. Hence the total number of critical instances is 12, and each player is critical three times, so that $B(P_i) = 1/4$ for every i, which is alternative (a).

Next suppose that there is only one 2-player winning coalition; then it must be $\{P_1, P_2\}$, so that P_1 and P_2 are critical in this coalition and also in the coalitions $\{P_1, P_2, P_3\}$ and $\{P_1, P_2, P_4\}$. On the other hand, P_3 and P_4 are not critical in these last two coalitions but are critical in the other two 3-player coalitions. Also, P_1 is critical in $\{P_1, P_3, P_4\}$, and P_2 is critical in $\{P_2, P_3, P_4\}$. Hence we have 4 critical instances for each of P_1 and P_2 and 2 critical instances for each of P_3 and P_4, so that alternative (d) results.

If there are two winning 2-player coalitions, then they must be $\{P_1, P_2\}$ and $\{P_1, P_3\}$, and we note that this can only occur if $v_3 > v_4$ (for if $v_3 = v_4$, then $\{P_1, P_3\}$ winning would imply that $\{P_1, P_4\}$ wins as well). P_1 is critical in each of the three 3-player coalitions containing P_1, yielding 5 critical instances for P_1. Then P_2 is critical in $\{P_1, P_2, P_4\}$ and $\{P_2, P_3, P_4\}$; P_3 is critical in $\{P_1, P_3, P_4\}$ and $\{P_2, P_3, P_4\}$; and the only 3-player coalition for which P_4 is critical is $\{P_2, P_3, P_4\}$. Again we have 12 critical instances in total, and we find that alternative (e) holds in this case.

There are now two cases involving three winning 2-player coalitions. We may have $\{P_1, P_2\}$, $\{P_1, P_3\}$, and $\{P_1, P_4\}$ winning, yielding 6 critical instances so far. Then we observe that only P_1 is critical in the coalitions $\{P_1, P_2, P_3\}$, $\{P_1, P_2, P_4\}$, and $\{P_1, P_3, P_4\}$, while all three players are critical in $\{P_2, P_3, P_4\}$. The result is alternative (c).

With $v_3 > v_4$ we may now encounter the case in which $\{P_1, P_2\}$, $\{P_1, P_3\}$, and $\{P_2, P_3\}$ win. Note that here, P_4 is never a critical player, because removing P_4 from any 3-player coalition to which it belongs leaves one of the three winning 2-player coalitions. Hence $B(P_4) = 0$. Furthermore, removing any player from $\{P_1, P_2, P_3\}$ still leaves a winning coalition, but in the other three 3-player coalitions containing P_4, the other two players are critical. Here, as in every case, we have a total of 12 critical instances, with each of P_1, P_2, and P_3 critical 4 times. Hence we find $B(P_1) = B(P_2) = B(P_3) = 1/3$, so that alternative (b) holds. The proof is complete. ■

Next steps We pose the following questions for investigation:

- For weighted voting systems of size n, is there a formula in terms of n for the number of feasible power distributions? (In case $n = 3$, a simple check reveals that all three players must have equal power, so that only one power distribution is possible in the absence of veto power.)
- With a complete enumeration of the power distributions feasible for weighted voting systems of size n, can one efficiently generate a complete list of feasible power distributions for size $n + 1$ weighted voting systems?
- If a certain power distribution is desired, can one efficiently construct a WVS with the feasible power distribution that comes closest to the ideal (by some measure)? We might expect that the solution to this problem is dependent on the choice of norm used to measure deviation from the ideal power distribution. Such norm-dependence

does arise in the apportionment problem mentioned above. One incarnation of that problem is to fairly mete out congressional representatives to states in direct proportion to the state's population. The ideal share for each state, however, is in general not an integer, so given a positive integer partition of the total number of representatives, one wishes to measure the overall deviation from the ideal. An article by Ernst [3] gives a thorough discussion.

- If a player is to be removed from an n-player WVS, can an $(n-1)$-player WVS be efficiently constructed so as to preserve, as nearly as possible, the existing distribution of power among the remaining players?

One final observation is that in the case of a 4-player WVS, a *strict hierarchy* of power is impossible; that is, there are always at least two players with the same Banzhaf power index. The least value of n for which we can have

$$B(P_1) > B(P_2) > \cdots > B(P_n)$$

turns out to be $n = 5$. An example, due to the author, Christopher Carter Gay, and Jabari Harris, is the WVS [9; 5, 4, 3, 2, 1], which yields Banzhaf power distribution

$$\left(\frac{9}{25}, \frac{7}{25}, \frac{1}{5}, \frac{3}{25}, \frac{1}{25}\right).$$

(The reader can check that the WVS [8; 5, 4, 3, 2, 1] does not induce the above Banzhaf power distribution.) We have also found, in unpublished work, that this is the *only* power distribution, of the 35 possibilities in the 5-player case, with strict hierarchy of power. What we have not yet found is an efficient way to analyze the 5-player problem in order to establish the number of power distributions feasible or generate the complete list of these power distributions.

The most naive attempt to effect strict hierarchy of power with $n = 6$, the WVS [11; 6, 5, 4, 3, 2, 1], fails to deliver, for this WVS yields power distribution

$$\left(\frac{9}{28}, \frac{1}{4}, \frac{5}{28}, \frac{3}{28}, \frac{3}{28}, \frac{1}{28}\right).$$

However, [15; 9, 7, 4, 3, 2, 1] yields power distribution

$$\left(\frac{5}{12}, \frac{3}{16}, \frac{1}{6}, \frac{1}{8}, \frac{1}{16}, \frac{1}{24}\right).$$

Acknowledgments. The author thanks the referees for excellent suggestions and Vic Mizel for his interest.

REFERENCES

1. M. L. Balinski and H. P. Young, *Fair Representation: Meeting the Ideal of One Man, One Vote*, Yale University Press, 1982.
2. J. F. Banzhaf III, Weighted voting doesn't work: a mathematical analysis, *Rutgers Law Review* **19** (1965), 317–343.
3. L. R. Ernst, Apportionment methods for the House of Representatives and the court challenges, *Management Science* **40** (1994), 1207–1227.
4. P. Tannenbaum and R. Arnold, *Excursions in Modern Mathematics*, 4th ed., Prentice-Hall, 2001.
5. A. D. Taylor, *Mathematics and Politics: Strategy, Voting, Power and Proof*, Springer-Verlag, New York, 1995.

Self-Similar Structure in Hilbert's Space-Filling Curve

MARK MCCLURE
University of North Carolina at Asheville
Asheville, North Carolina 28801
mcmcclur@bulldog.unca.edu

Hilbert's space-filling curve is a continuous function that maps the unit interval onto the unit square. The construction of such curves in the 1890s surprised mathematicians of the time and led, in part, to the development of dimension theory. In this note, we discuss how modern notions of self-similarity illuminate the structure of this curve. In particular, we show that Hilbert's curve has a basic self-similar structure and can be generated using what is called an *iterated function system*, or *IFS*. Furthermore, its coordinate functions display a generalized type of self-similarity called digraph self-affinity and may be described using an appropriately generalized iterated function system.

The notions of self-similarity used here are described in the text by Edgar [**2**, Ch. 4]. The definition of a digraph IFS was originally formulated in a research paper by Mauldin and Williams [**4**], although similar ideas have appeared elsewhere. The author has published a *Mathematica* package implementing the digraph IFS scheme [**6**]. A broad introduction to space-filling curves may be found in the book by Sagan [**7**].

Iterated function systems The curve K shown in FIGURE 1 is called the *Koch curve* and is an example of a *self-similar set*. In the figure, K_1 is the image of K scaled by the factor $1/3$ about the left endpoint of the curve and K_4 is the image of K scaled by the factor $1/3$ about the right endpoint of the curve. The portions K_2 and K_3 have been rotated by $60°$ and $-60°$ respectively and shifted, in addition to being scaled. The four copies of K illustrating the decomposition have been slightly separated.

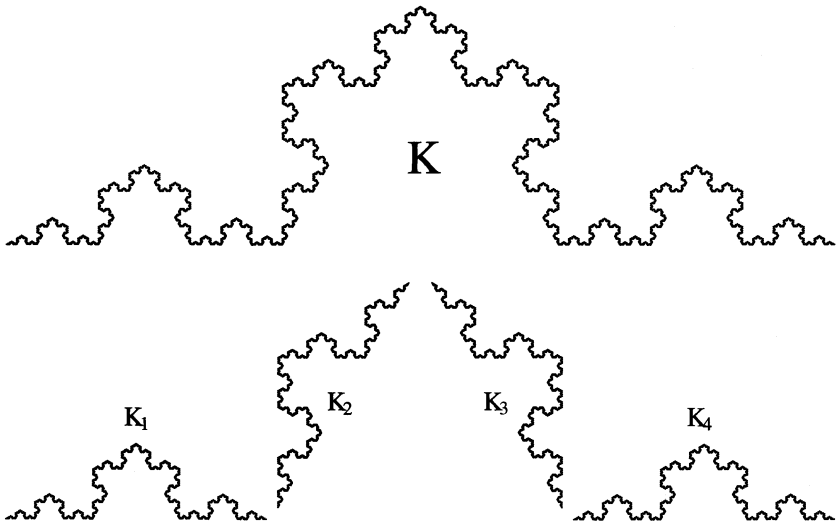


Figure 1 The Koch curve

Any self-similar set may be described using an iterated function system, or IFS. Indeed, we will define self-similar sets using this concept, so we develop it first. A *contraction* of $\mathbb{R}^2$ is a function $f : \mathbb{R}^2 \to \mathbb{R}^2$ such that for all $\vec{x}, \vec{y} \in \mathbb{R}^2$ and some $r \in (0, 1)$, we have $|f(\vec{x}) - f(\vec{y})| \leq r|\vec{x} - \vec{y}|$. If for all $\vec{x}, \vec{y} \in \mathbb{R}^2$ we have $|f(\vec{x}) - f(\vec{y})| = r|\vec{x} - \vec{y}|$, then f is called a *similarity*. An *iterated function system* is a finite collection of contractions $\{f_i\}_{i=1}^m$. A surprising fact about an IFS, and an important reason that we study them, is that there is always a unique nonempty, closed, bounded subset E of $\mathbb{R}^2$ such that

$$E = \bigcup_{i=1}^{m} f_i(E),$$

that is, E consists exactly of contractions of itself. The set E is called the *invariant set* of the IFS. If all contractions of the IFS are similarities, then the invariant set is also called *self-similar*.

Matrices provide a convenient notation to describe many iterated function systems. A function of the form $f(\vec{x}) = A\vec{x} + \vec{b}$, where A is a matrix and $\vec{b}$ is a translation vector, is called an *affinity*. If all the functions of the IFS are affinities, then the invariant set is called *self-affine*. A similarity is a special type of affinity. To build an affinity, we start with a rotation about the origin through angle θ, which can be represented using a matrix of the form

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Dividing the matrix through by $r > 1$ turns this into a contraction; adding a translation vector turns it into a similarity. For example, the following functions make up the IFS for the Koch curve.

$$f_1(\vec{x}) = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \vec{x}$$

$$f_2(\vec{x}) = \begin{pmatrix} \frac{1}{6} & -\frac{1}{2\sqrt{3}} \\ \frac{1}{2\sqrt{3}} & \frac{1}{6} \end{pmatrix} \vec{x} + \begin{pmatrix} \frac{1}{3} \\ 0 \end{pmatrix}$$

$$f_3(\vec{x}) = \begin{pmatrix} \frac{1}{6} & \frac{1}{2\sqrt{3}} \\ -\frac{1}{2\sqrt{3}} & \frac{1}{6} \end{pmatrix} \vec{x} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2\sqrt{3}} \end{pmatrix}$$

$$f_4(\vec{x}) = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \vec{x} + \begin{pmatrix} \frac{2}{3} \\ 0 \end{pmatrix}$$

Note that each f_i maps K onto K_i in FIGURE 1.

Hilbert's curve FIGURE 2 depicts the action of an IFS with four transformations on the unit square at the corner of the first quadrant. This IFS is also described in section 6.5 of the text by Barnsley [**1**]. In FIGURE 2a, we see the unit square together with a path through four subsquares. In FIGURE 2b we see the image of FIGURE 2a under each of the four functions of the IFS. If we drop the arrows and connect the terminal point of one path to the initial point of the subsequent path, we obtain the bold path shown in FIGURE 2c. The ordering in which these paths are connected is determined by the initial path in FIGURE 2a. If we iterate this procedure two more times we obtain the fourth level approximation shown in FIGURE 2d. These paths

represent approximations to a continuous function h mapping the unit interval onto the unit square. The function h is called *Hilbert's space-filling curve*. Careful proofs of its basic properties may be found in the book by Sagan [7, Ch. 2].

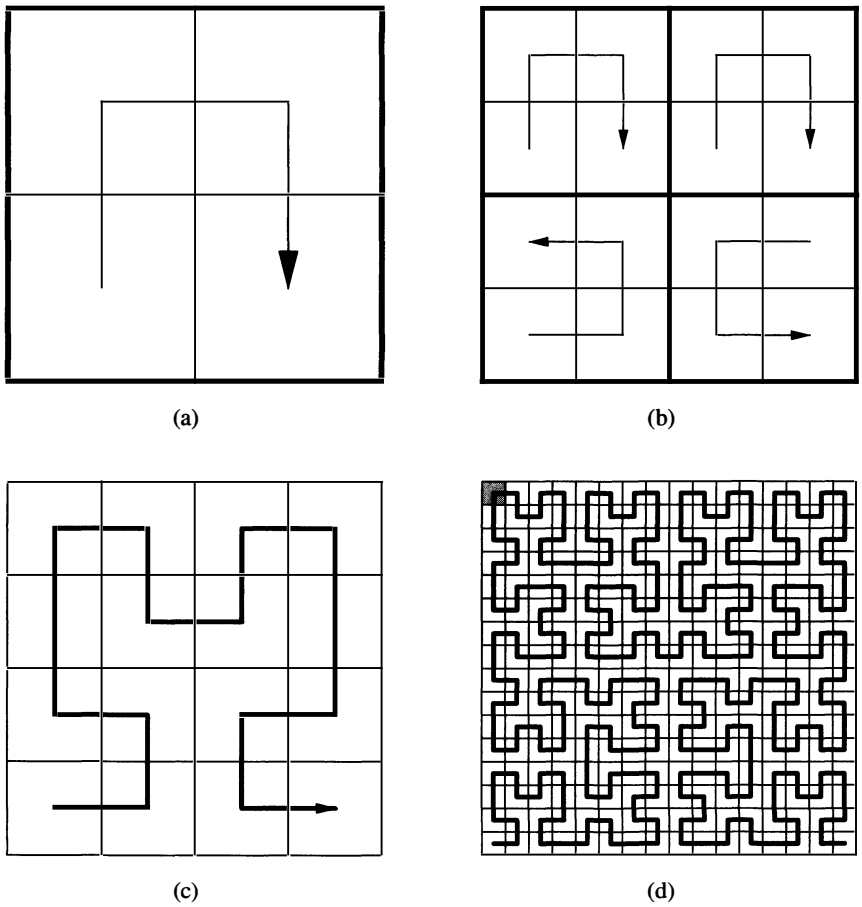


Figure 2 Approximations to Hilbert's space-filling curve

In the previous example, it was the invariant set that was interesting, namely the Koch curve. Here, that set is the entire unit square. Instead, the object of interest is the function h, and it is important to keep in mind that h is a continuous function. This function may be rather difficult to picture as a curve, since it maps the interval *onto* the unit square. However, some estimates may help: Given $t \in [0, 1]$, suppose that $(i-1)/4^n \le t \le i/4^n$; then $h(t)$ lies in the ith closed subsquare determined by following the nth approximating curve, counting the squares as you progress along the curve. If $t = i/4^n$ for some i, then $h(t)$ lies on the border of two adjacent subsquares. For example, since $85/4^4 < 1/3 < 86/4^4$, $h(1/3)$ lies in the 86th subsquare determined by the fourth level approximation. This subsquare is shaded a light gray in the upper left corner of FIGURE 2d.

Digraph iterated function systems In order to understand the coordinate functions of h, we need to introduce the notion of a *directed graph iterated function system* or *digraph IFS*. Consider the two curves A and B shown in FIGURE 3. The curve A is composed of one copy of itself, scaled by the factor $1/2$, and two copies of B, rotated

and scaled by the factor 1/2. The curve B is composed of one copy of itself, scaled by the factor 1/2 and one copy of A, reflected and scaled by the factor 1/2. The sets A and B form a pair of *digraph self-similar sets.*

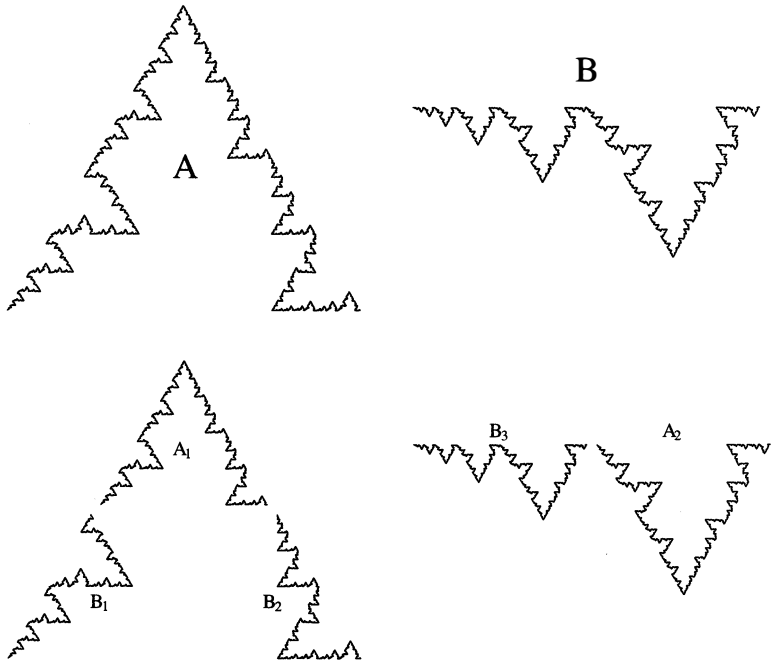


Figure 3 Digraph self-similar curves

Any collection of digraph self-similar sets can be described using a *digraph IFS*, which consists of a directed multigraph G together with a contraction f_e from $\mathbb{R}^2$ to $\mathbb{R}^2$ associated with each edge e of G. A *directed multigraph* consists of a finite set V of vertices and a finite set E of directed edges between vertices, possibly having multiple edges between vertices and even edges connecting a vertex to itself. Given two vertices, u and v, we denote the set of all edges from u to v by E_{uv}. Given a digraph IFS, there is a unique collection of nonempty, closed, bounded sets K_v, one for each $v \in V$, such that for every $u \in V$

$$K_u = \bigcup_{v \in V,\ e \in E_{uv}} f_e(K_v).$$

The set $\{K_u : u \in V\}$ is called the *invariant list* of the digraph IFS and its members are the *invariant sets* of the digraph IFS. If all the functions in the digraph IFS are similarities, then the invariant sets are also called digraph self-similar sets. If all the functions in the digraph IFS are affinities, then the invariant sets are called *digraph self-affine sets.*

The digraph IFS for the curves A and B is shown in FIGURE 4. The labels on the edges correspond to similarities mapping one set to part of another (perhaps the same) set. For example the label a_2 corresponds to the similarity mapping A to the portion of B labeled A_2 and is given by

$$a_2(\vec{x}) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix} \vec{x} + \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

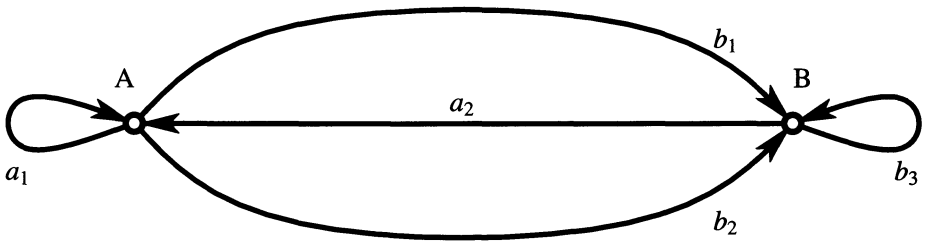


Figure 4 Digraph IFS for the curves

The coordinate functions of *h* If we write Hilbert's space-filling curve in the form $h(t) = (x(t), y(t))$, then it turns out that the graphs of the coordinate functions $x(t)$ and $y(t)$ form a pair of digraph self-affine sets. To show this, we will create a digraph IFS with invariant sets X and Y. We then show that these sets coincide with the graphs of $x(t)$ and $y(t)$. Define matrices A and B:

$$A = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \qquad B = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}.$$

Note that the linear mappings defined by A and B both contract by the factor $1/4$ in the horizontal direction and by the factor $1/2$ in the vertical direction; these are affinities, not similarities. The transformation defined by B has the additional effect of reflecting about the horizontal axis. Now, let $\vec{x} \in \mathbb{R}^2$ denote a column vector and define affine functions as follows.

$$\begin{aligned}
a_{xx}(\vec{x}) &= A\vec{x} + \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix} & a_{yy}(\vec{x}) &= A\vec{x} + \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \end{pmatrix} \\
b_{xx}(\vec{x}) &= A\vec{x} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} & b_{yy}(\vec{x}) &= A\vec{x} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \\
c_{xy}(\vec{x}) &= A\vec{x} & c_{yx}(\vec{x}) &= A\vec{x} \\
d_{xy}(\vec{x}) &= B\vec{x} + \begin{pmatrix} \frac{3}{4} \\ 1 \end{pmatrix} & d_{yx}(\vec{x}) &= B\vec{x} + \begin{pmatrix} \frac{3}{4} \\ \frac{1}{2} \end{pmatrix}
\end{aligned}$$

These are the affine functions used to define the digraph IFS shown in FIGURE 5. The sets X and Y that form the invariant list of this digraph IFS are shown in FIGURE 6.

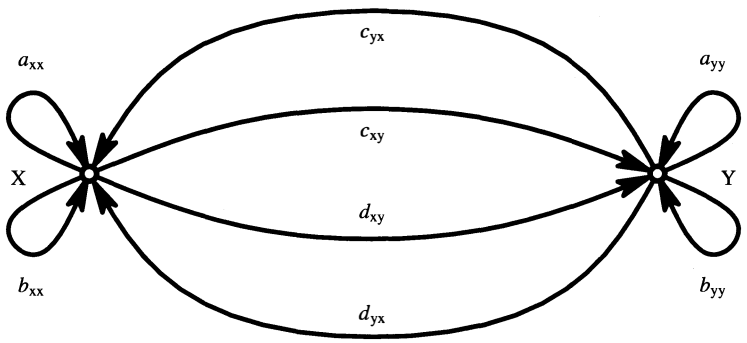


Figure 5 The digraph for X and Y

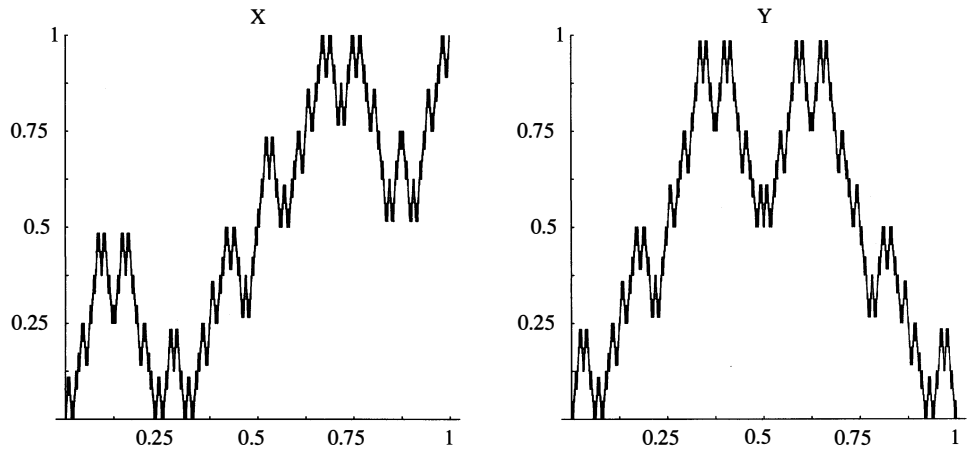


Figure 6 The sets X and Y

The function c_{xy}, for example, maps the set Y onto the portion of X lying over the interval $[0, 1/4]$.

We claim that the sets X and Y are the graphs of $x(t)$ and $y(t)$. To show this, we need only show that the graphs form an invariant list for the digraph IFS, since such a list is known to be unique. This may be deduced from the self-similar structure in Hilbert's construction. Note that each stage in the construction of Hilbert's curve (see FIGURE 2) may be obtained by piecing together four images similar to the previous stage scaled by the factor $1/2$. For example, the portion of Hilbert's curve in the upper left quarter of the unit square is similar to the whole curve but scaled by the factor $1/2$. More precisely, $h : [0, 1] \to [0, 1] \times [0, 1]$ scales to $h : [1/4, 1/2] \to [0, 1/2] \times [1/2, 1]$. Thus the graph of $x(t)$ scales from $[0, 1] \times [0, 1]$ onto $[1/4, 1/2] \times [0, 1/2]$, which is accomplished by a_{xx}. The other affinities may be derived in a similar manner. Note that the roles of x and y switch on the lower left and lower right quarters, due to the rotation in the similarities mapping the unit square onto those quarters.

Box-counting dimension The digraph IFS scheme is useful not only for generating images, but also for computing dimensions. The box-counting dimension of the graph of a function $f : \mathbb{R} \to \mathbb{R}$ is a number in $[0, 1]$ that measures how complicated or rough a graph is. It takes a very complicated graph to approach dimension 2, while a smooth curve has dimension 1. An exposition of the box-counting dimension may be found in Falconer [**3**, Ch. 3]. We will show that the box-counting dimensions of the sets X and Y are both $3/2$, indicating that these functions are fairly complicated. An older and deeper notion of dimension is called *Hausdorff dimension.* An earlier paper of the author [**5**] shows that the Hausdorff dimensions of X and Y are also $3/2$.

To define the box-counting dimension of a bounded set $S \subset \mathbb{R}^2$ we first consider covers of S by small squares. For $\varepsilon > 0$, the *ε-mesh* for $\mathbb{R}^2$ is the grid of squares of side length ε with the origin at one corner and sides parallel to the coordinate axes. For a bounded set $S \subset \mathbb{R}^2$, define

$$N_\varepsilon(S) = \text{number of } \varepsilon\text{-mesh squares that intersect } S.$$

As ε shrinks toward 0, we would expect $N_\varepsilon(S)$ to grow larger. The rate at which $N_\varepsilon(S)$ grows reflects the dimension of S. For example, if $\mathcal{I}$ is the unit interval and $\mathcal{Q}$ is the unit square, then $N_\varepsilon(\mathcal{I})$ grows as $1/\varepsilon$ while $N_\varepsilon(\mathcal{Q})$ grows as $1/\varepsilon^2$. The exponent of ε

indicates the dimension of the set. Thus we define the box-counting dimension by

$$\dim_b(S) = \lim_{\varepsilon \to 0^+} \frac{\log(N_\varepsilon(S))}{\log(1/\varepsilon)},$$

provided this limit exists. An important simplifying property of $\dim_b$ (proved in Falconer's book [**3**, p. 41]) is this: if we take the limit along some sequence $\{c^n\}_{n=1}^{\infty}$ where $c \in (0, 1)$, we still obtain the same value.

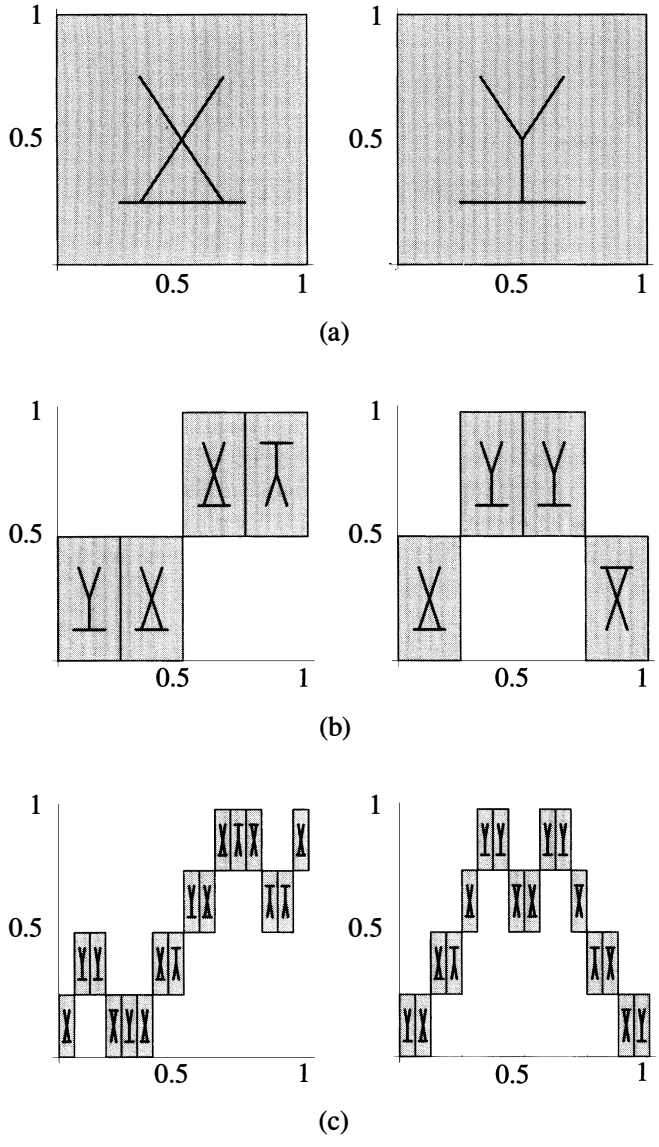


Figure 7 Approximations to X and Y using the digraph IFS

To compute the box-counting dimensions of X and Y, we use the rectangular covers generated by the digraph IFS illustrated in FIGURE 7. These covers are generated as follows. There are two versions of the unit square shown in FIGURE 7a, one labeled X and one labeled Y. We see how these parts fit together under the action of the digraph

IFS after one iteration in FIGURE 7b and after two iterations in FIGURE 7c. The underscores allow us to observe the orientation of the rectangles. For example, the lower left hand rectangle labeled Y in the approximation to X of FIGURE 7b is the image of the unit square labeled Y under the affine function c_{xy}. This process is then iterated. A proof by induction shows that the rectangles of each level of the approximation are contained in the rectangles of the previous approximation. Thus these approximations in fact cover the invariant sets X and Y. It can also be proved by induction that after n iterations, each cover consists of 4^n rectangles of width 4^{-n} and height 2^{-n}. Furthermore, inside each of these rectangles is an affine image of either X or Y with height 2^{-n}. Each of the rectangles may be subdivided into a column of 2^n squares of side length 4^{-n}. Thus

$$N_{4^{-n}}(X) = N_{4^{-n}}(Y) = 2^n \cdot 4^n = 8^n$$

and

$$\dim_b(X) = \dim_b(Y) = \lim_{n\to\infty} \frac{\log(8^n)}{\log(1/4^{-n})} = \lim_{n\to\infty} \frac{\log(2^{3n})}{\log(2^{2n})} = \frac{3}{2}.$$

Directions for undergraduate research There are many more space-filling curves defined in Sagan's book [7]. Applying the ideas in this paper to those curves is a source of potential undergraduate research projects. Here are a few ideas.

Use a digraph IFS scheme to describe those curves on a case-by-case basis. Better yet, perhaps some general principle could be identified. Can this principle be used to define some new space-filling curves?

Are there any space-filling curves for which the digraph IFS characterization fails? Perhaps some modification of one of the curves described in Sagan's book will work. Perhaps an entirely new construction will be needed.

What are the possible dimensions of the graphs of the coordinate functions? The dimension should certainly be between 1 and 2. Loosely speaking, the larger the dimension the rougher the graphs of the coordinate functions. Are there space-filling curves whose coordinate functions have dimension 1 or dimension 2? Must the coordinate functions have graphs of the same dimension?

REFERENCES

1. Michael Barnsley, *Fractals Everywhere*, Academic Press, San Diego, 1988.
2. Gerald Edgar, *Measure, Topology, and Fractal Geometry*, Springer-Verlag, New York, 1990.
3. Kenneth Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley and Sons, West Sussex, UK, 1990.
4. R. D. Mauldin and S. C. Williams, Hausdorff dimension in graph directed constructions, *Trans. Amer. Math. Soc.* **309**:2 (1988), 811–829.
5. Mark McClure, The Hausdorff dimension of Hilbert's coordinate functions, *The Real Anal. Ex.* **24**:2 (1998/99), 875–883.
6. Mark McClure, Directed-graph iterated function systems, *Mathematica in Education and Research* **9**:2 (2000).
7. Hans Sagan, *Space-Filling Curves*, Springer-Verlag, New York, 1994.

A Dynamical Systems Proof of Fermat's Little Theorem

KEVIN IGA
Pepperdine University
Malibu, CA 90263
kiga@pepvax.pepperdine.edu

At first glance, the subject of dynamical systems seems unrelated to number theory. Number theory deals with integers: more specifically, integer solutions to equations, primes, reducing modulo n, and so on. In dynamical systems, we study how quantities change in time when governed by such things as differential equations or recurrence relations.

But in this note we will prove an important and foundational result in number theory using techniques from dynamical systems. Most of the proof is geometric, as is common in dynamical systems, rather than algebraic, which is common in number theory. The result, known as Fermat's little theorem, says this:

THEOREM. (FERMAT) Let p be a prime number, and a any integer. Then

$$a^p \equiv a \pmod{p}.$$

This theorem, not to be confused with its more famous brother, Fermat's Last Theorem, is useful not only when you want to compute what 6^{9397} is modulo 13 (though it is useful for that kind of problem); it is also foundational to the RSA code in modern cryptology (Churchhouse's book on codes [2] gives a readable account of this). Fermat's little theorem and its generalizations make frequent appearances in any course in number theory and abstract algebra.

There are many elementary proofs of this result: Almost every number theory textbook (like Niven and Zuckerman's book *An Introduction to the Theory of Numbers* [5]) and abstract algebra textbook (like Artin's *Algebra* [1] and Herstein's *Topics in Algebra* [4]) states, proves, and applies this result.

Usually, the theorem is proved using algebraic ideas—some of them very abstract. In this note we will prove it with a simple argument from dynamical systems on the unit interval [0, 1]. The basic idea is to count points of minimum period p of a particular dynamical system, and note that this number must be divisible by p.

A referee pointed out another dynamical-systems proof of this result, by Hausner [3]. Although that proof yields other impressive results beyond Fermat's little theorem, it is much less visual.

Proof. For every integer n, we define the function $T_n : [0, 1] \to [0, 1]$ as follows:

$$T_n(x) = \begin{cases} \{nx\}, & x \neq 1 \\ 1, & x = 1 \end{cases}$$

Here the braces denote the fractional part, so that $T_n(x)$ is in the interval [0, 1] . We can picture T_n in two different but important ways. First, we can draw a graph, as in FIGURE 1, which shows a graph of $T_3(x)$. Second, we can think of T_n as something that happens to points on the interval [0, 1]. That is, we imagine that the point $x \in [0, 1]$ gets moved to the point $T_n(x)$. You may wish to draw your own sketch of $T_3(x)$ as a transformation of our interval, but one will be provided soon.

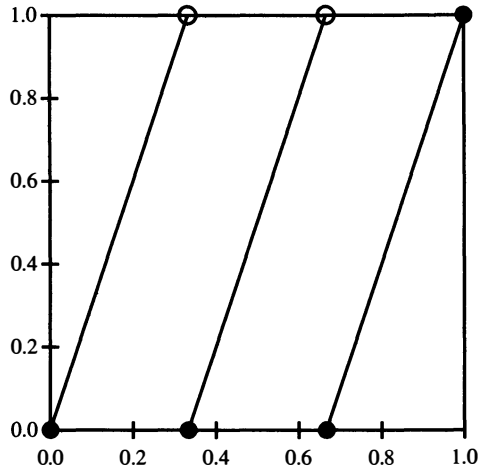


Figure 1 The graph of T_3

The important thing is that some points get left where they were (these are called *fixed points*, and occur when $T_n(x) = x$), and some points get moved somewhere else. We now imagine that this operation of applying T_n happens again and again, so that if we imagine a frog on the interval at point x at time 0, then at time 1, the frog jumps to $T_n(x)$; at time 2, it is at $T_n(T_n(x))$, and so on. Some points in [0, 1] will be periodic, in the sense that a frog starting at that point will repeat its path over and over again, so that $T_n(T_n(\ldots(T_n(x)))) = x$ for some number of iterations of T_n. We call a point k-periodic if it comes back after k times, and call k its minimal period if it doesn't come back for any earlier time. So the 1-periodic points are fixed points. Note that there might be points that never return to their starting position and so are not periodic at all (it turns out there are many of these for the function T_n, but that is not important in this proof).

We now determine two crucial properties of the family of functions T_n:

LEMMA 1. *Let n be an integer greater than 1. The function $T_n(x)$ has n fixed points in* [0, 1].

Proof. Recall that a fixed point of T_n is a point where $T_n(x) = x$. On a graph of T_n, then, this happens when the graph crosses the line $y = x$, because those are the points where the x-value and the y-value are equal.

FIGURE 1 shows a graph of T_3. Notice that the line $y = x$ crosses this graph in exactly 3 places, once for each linear segment in the graph. More generally, if n is a positive integer greater than 1, then there will be n linear segments, each of slope n, and which proceed in y-value from 0 to 1. In particular, each crosses the line $y = x$ exactly once. So there are n fixed points of T_n in [0, 1]. ■

Now for those of you who prefer a more formal, nonvisual argument for the same result:

Proof. We will show that the fixed points have the form $x = k/(n-1)$ where k is any integer in the range $\{0, 1, \ldots, n-1\}$. When $0 \le k < n-1$,

$$T_n\left(\frac{k}{n-1}\right) = \left\{\frac{nk}{n-1}\right\} = \left\{\frac{nk}{n-1} - k\right\} = \left\{\frac{k}{n-1}\right\}.$$

In this case, $0 \leq k < n-1$, so $0 \leq k/(n-1) < 1$, and so $\{k/(n-1)\} = k/(n-1)$. Therefore, when $0 \leq k < n-1$,

$$T_n\left(\frac{k}{n-1}\right) = \frac{k}{n-1}$$

and $k/(n-1)$ is a fixed point of T_n in $[0, 1]$. When $k = n-1$, then $k/(n-1) = 1$, and $T_n(1) = 1$, so $k/(n-1)$ is a fixed point of T_n in $[0, 1]$ in this case also.

Conversely, suppose x is a fixed point of T_n in $[0, 1]$. Then $T_n(x) = x$. Using the formula for T_n, we see that $nx = k + x$ for some integer k, so that $(n-1)x = k$, and therefore $x = \frac{k}{n-1}$. We have listed the set of fixed points as $\{\frac{0}{n-1}, \frac{1}{n-1}, \ldots, \frac{n-1}{n-1}\}$, showing that there are n fixed points of T_n in $[0, 1]$. ■

LEMMA 2. *Let a and b be positive integers. Then for all $x \in [0, 1]$,*

$$T_a(T_b(x)) = T_{ab}(x).$$

Proof. Again, this proof can be done geometrically, noticing that each of the line segments that constitute the graph of T_b has as its image the interval $[0, 1]$, so the graph of T_a will be replicated inside each of these corresponding intervals, as shown in FIGURE 2. ■

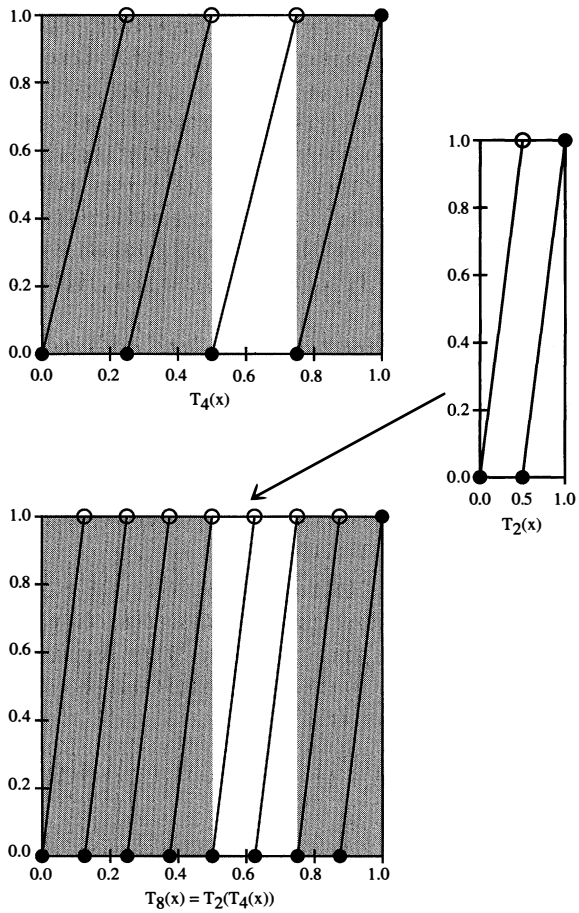


Figure 2 Composing $T_2(T_4(x))$

The algebraic, nonvisual proof of the same result goes as follows:

Proof. The lemma is straightforward to prove for $x = 1$. Now consider $x < 1$. Then $T_b(x) = \{bx\}$, so that for some integer m, we have $T_b(x) = bx + m$. So

$$T_a(T_b(x)) = \{a(bx + m)\} = \{abx + am\} = \{abx\} = T_{ab}(x). \quad \blacksquare$$

Now recall the statement of the theorem. We have p a prime, and a any integer, and we want to prove that

$$a^p \equiv a \pmod{p}.$$

Using these values of a and p, we consider the p-periodic points of T_a. These p-periodic points are the fixed points of T_a iterated p times, which is T_{a^p}. This has a^p fixed points. Of these, exactly a are fixed points of T_a. Since p is prime, the rest of them have minimal period p under T_a. This means that there are $a^p - a$ points that have minimal period p.

Since each point with minimal period p lies in an orbit of size p, there are $(a^p - a)/p$ orbits of size p. Since this is an integer, we see that p divides $a^p - a$, and $a^p \equiv a \pmod{p}$. $\blacksquare$

Three benefits of this approach come to mind: First, the approach may motivate students who are interested in number theory to think about dynamical systems, and vice versa. Second, this is amenable to a very geometric explanation that may help students who are visual learners. Third, this generalizes differently when p is replaced by a composite number. For example, the 15-periodic points have minimal period 1, 3, 5, and 15. Now from the discussion above T_a has a^{15} 15-periodic points, of which $a^3 - a$ have minimal period 3, $a^5 - a$ have minimal period 5, and a have minimal period 1 (are fixed points). Therefore, the number of points with minimal period 15 is $a^{15} - (a^3 - a) - (a^5 - a) - a$, that is, $a^{15} - a^5 - a^3 + a$. So as before, 15 divides $a^{15} - a^5 - a^3 + a$, or if you prefer, $a^{15} \equiv a^5 + a^3 - a \pmod{15}$. More generally, the inclusion-exclusion principle (stated in many books on combinatorics, for instance Van Lint and Wilson's book on combinatorics [**6**]) can be applied to get similar interesting congruences when 15 is replaced by other numbers.

This is not the usual generalization of Fermat's little theorem, which can be found in the abstract algebra, number theory, and cryptology books mentioned in the introduction [**1, 2, 4, 5**]. The usual generalization of Fermat's little theorem for modulo 15 would give $a^8 \equiv a \pmod{15}$.

Acknowledgments. I'd like to thank the referees for the helpful suggestions, and for pointing out Hausner's dynamical systems proof [**3**].

REFERENCES

1. M. Artin, *Algebra*, Prentice Hall, Upper Saddle River, NJ, 1991.
2. R. Churchhouse, *Codes and Ciphers: Julius Caesar, the Enigma, and the Internet*, Cambridge University Press, Cambridge, 2002.
3. M. Hausner, Applications of a simple counting technique, *Amer. Math. Monthly* **90** (1983), 127–129.
4. I. Herstein, *Topics in Algebra*, John Wiley & Sons, New York, 1975.
5. I. Niven and H. S. Zuckerman, *An Introduction to the Theory of Numbers*, John Wiley & Sons, New York, 1960.
6. J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge University Press, Cambridge, 1992.

Using Tangent Lines to Define Means

BRIAN C. DIETEL
RUSSELL A. GORDON
Whitman College
Walla Walla, WA 99362
gordon@whitman.edu
dietelbc@whitman.edu

Let $(a, f(a))$ and $(b, f(b))$ be two distinct points on the graph of a differentiable function f. Suppose that the tangent lines of f at these two points intersect, and call the point of intersection (c, d). Verifying the following facts is elementary.

1. If $f(x) = x^2$, then $c = (a + b)/2$, the arithmetic mean of a and b.
2. If $f(x) = \sqrt{x}$, then $c = \sqrt{ab}$, the geometric mean of a and b.
3. If $f(x) = 1/x$, then $c = 2ab/(a + b)$, the harmonic mean of a and b.

We found these results to be quite intriguing and wondered if other means could be generated in this way by using different functions. As it turns out, this idea can be applied to a wide class of functions and the locations of the corresponding points c exhibit a surprising degree of simplicity and elegance. We will provide the details in this paper under the assumption that the reader has only a limited knowledge of the vast field of means. At the end of the paper, we will relate the means defined here to other types of means that have been considered in the literature.

Given two distinct real numbers a and b, a *mean* $M(a, b)$ is a number that lies between a and b. A mean is *symmetric* if $M(a, b) = M(b, a)$ for all a and b. In this paper, we will only consider symmetric means defined for positive numbers. The most familiar example of a mean is the *arithmetic mean* (or average) of two numbers, but there are many more examples. For instance, the *geometric mean* of two positive numbers a and b is $\sqrt{ab}$. This number is the side length of a square whose area is equal to that of a rectangle with lengths a and b.

The value between a and b, often called c, whose existence is asserted by the Mean Value Theorem from differential calculus, also provides a way to define a mean of two numbers. As with the point of intersection of the tangent lines, a mean defined in this way depends on the function f. We will mention a connection between these two types of means.

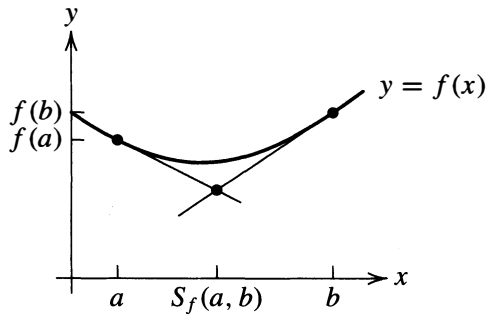
Means generated by tangent lines

We now turn to the method for constructing means introduced in the opening paragraph. Let f be a twice differentiable function defined on $(0, \infty)$ such that f'' is continuous. In order to guarantee that every pair of distinct tangent lines intersect, we must assume that f'' is nonzero on $(0, \infty)$. (One way to see this involves a simple application of Rolle's Theorem.) Given distinct positive numbers a and b, define $S_f(a, b)$ to be the x-coordinate of the point of intersection of the lines tangent to the graph of $y = f(x)$ at $(a, f(a))$ and $(b, f(b))$, as shown in the figure. The number $S_f(a, b)$ is the value of x that satisfies

$$f'(a)(x - a) + f(a) = f'(b)(x - b) + f(b).$$

Solving for this value yields

$$S_f(a, b) = \frac{bf'(b) - af'(a) - \big(f(b) - f(a)\big)}{f'(b) - f'(a)} = \frac{\big(xf'(x) - f(x)\big)\big|_a^b}{f'(x)\big|_a^b} = \frac{\int_a^b xf''(x)\,dx}{\int_a^b f''(x)\,dx}.$$



The integral representation for this value is quite remarkable. It shows that the number $S_f(a, b)$ is the x-coordinate of the centroid of the region bounded by the graph of $y = f''(x)$ and the x-axis on the interval $[a, b]$. Given the formula for $S_f(a, b)$, there is no loss of generality if we assume that f'' is positive on $(0, \infty)$. If $0 < a < b$, then

$$a\int_a^b f''(x)\,dx < \int_a^b xf''(x)\,dx < b\int_a^b f''(x)\,dx,$$

which indicates that $S_f(a, b)$ is between a and b. This shows that $S_f(a, b)$ is a mean, and this mean is clearly symmetric. Before looking at specific means generated in this way, we consider some general properties of $S_f(a, b)$.

General properties of S_f means

In what follows, all functions used to define a mean $S_f(a, b)$ are assumed to have continuous nonzero second derivatives on $(0, \infty)$. Let a be a fixed positive number and consider the function $S_f(a, x)$, where $x > 0$. It makes intuitive sense that $S_f(a, a)$ should be defined to be a, and this is consistent with the fact that (use L'Hôpital's Rule and the Fundamental Theorem of Calculus)

$$\lim_{x\to a} S_f(a, x) = \lim_{x\to a} \frac{\int_a^x tf''(t)\,dt}{\int_a^x f''(t)\,dt} = \lim_{x\to a} \frac{xf''(x)}{f''(x)} = a.$$

Hence, the function $S_f(a, x)$ is continuous at a. Since the functions defined by integrals in the formula for $S_f(a, x)$ are continuous functions, it follows that $S_f(a, x)$ is continuous on $(0, \infty)$. More properties of this function are given in Theorem 2.

THEOREM 1. If there exist constants $\alpha \neq 0$, β, and γ such that $g(x) = \alpha f(x) + \beta x + \gamma$ for all $x > 0$, then $S_f(a, b) = S_g(a, b)$ for all positive numbers a and b.

Proof. This result is a trivial consequence of the definition of $S_f(a, b)$. ∎

THEOREM 2. If $a > 0$, then the function $S_f(a, x)$ is strictly increasing on $(0, \infty)$ and has a continuous positive derivative on $(0, \infty)$.

Proof. Without loss of generality, we may assume that $f(a) = 0$, $f'(a) = 0$, and $f''(x) > 0$ for all $x > 0$ (replace $f(x)$ with $\pm(f(x) - f'(a)(x - a) - f(a))$, where the sign is chosen appropriately, and use Theorem 1). These conditions on f guarantee that $f(x) > 0$ for all positive numbers $x \neq a$. Note that

$$S_f(a,x) = \frac{xf'(x) - f(x)}{f'(x)} \quad \text{and} \quad S_f'(a,x) = \frac{f(x)f''(x)}{(f'(x))^2}$$

for all $x \neq a$. (Under the assumptions on f, the function $S_f(a,x)$ is the same function that appears in Newton's method for approximating roots.) It follows that $S_f'(a,x)$ is continuous and positive on the intervals $(0,a)$ and (a,∞); this fact guarantees that $S_f(a,x)$ is strictly increasing on $(0,\infty)$. To show that $S_f'(a,x)$ is defined and continuous at a, first use L'Hôpital's Rule and the continuity of f'' to compute

$$\lim_{x\to a} \frac{f(x)}{(f'(x))^2} = \lim_{x\to a} \frac{f'(x)}{2f'(x)f''(x)} = \frac{1}{2f''(a)}$$

and thus conclude that $\lim_{x\to a} S_f'(a,x) = 1/2$. Therefore, using the Mean Value Theorem, the function $S_f(a,x)$ is differentiable at a and $S_f'(a,x)$ is continuous on $(0,\infty)$. ■

It is interesting to note that the value $S_f'(a,a) = 1/2$ is independent of both a and f. The linear approximation to $S_f(a,x)$ for x near a is thus

$$S_f(a,x) \approx a + \frac{1}{2}(x-a) = \frac{a+x}{2}.$$

Therefore, the mean $S_f(a,x)$ is locally arithmetic, that is, its values are close to the arithmetic mean when x is near a.

As an exercise, the reader is encouraged to find conditions on f so that

$$\lim_{x\to\infty} S_f(a,x)$$

is finite and to find the limit in some specific cases for which it is finite. It is also possible to consider $\lim_{x\to 0^+} S_f(a,x)$ and find conditions on f so that the limit is not zero.

Using the fact that $S_f(a,x)$ is a differentiable function, we can prove the converse of Theorem 1.

THEOREM 3. If the means $S_f(a,b)$ and $S_g(a,b)$ are equal for all positive numbers a and b, then there exist constants $\alpha \neq 0$, β, and γ such that $g(x) = \alpha f(x) + \beta x + \gamma$ for all $x > 0$.

Proof. The idea behind this proof is due to Groetsch [**1**, p. 56]. Suppose that f and g generate the same mean and let $a > 0$. Without loss of generality, we may assume that both f'' and g'' are positive on $(0,\infty)$. An elementary computation reveals that

$$S_f'(a,x) = \frac{f''(x)}{\int_a^x f''(t)\,dt}\big(x - S_f(a,x)\big)$$

for all $x \neq a$; a similar equation is valid for $S_g'(a,x)$. Since $S_f'(a,x) = S_g'(a,x)$ for all $x > 0$ and $x - S_f(a,x) \neq 0$ for $x \neq a$, we find that

$$\frac{f''(x)}{f'(x) - f'(a)} = \frac{g''(x)}{g'(x) - g'(a)}$$

for all $x \neq a$. It follows that

$$\frac{d}{dx} \ln\left(\frac{g'(x) - g'(a)}{f'(x) - f'(a)}\right) = 0$$

for all $x \neq a$. Note that the quantity inside the logarithm function is positive on $(0, \infty)$, and continuous there, provided it is assigned the value $g''(a)/f''(a)$ at a. Hence,

$$\frac{g'(x) - g'(a)}{f'(x) - f'(a)} = C$$

for some positive constant C. Solving for $g'(x)$ and antidifferentiating gives the desired result. ∎

A mean $M(a, b)$ is *homogeneous* if $M(ka, kb) = kM(a, b)$ for all positive numbers a, b, and k. Most of the familiar means have this property. As we now show, the mean $S_f(a, b)$ is homogeneous if and only if f has a rather simple form. A function f is said to be multiplicative if $f(xy) = f(x)f(y)$ for all x and y in the domain of f. It is not difficult to show that a nonzero differentiable function f defined on $(0, \infty)$ is multiplicative if and only if $f(x) = x^r$ for some constant r. (Differentiate the equation $f(xy) = f(x)f(y)$ with respect to y, then let $y = 1$ to obtain the differential equation $xf'(x) = f'(1)f(x)$.)

THEOREM 4. The mean $S_f(a, b)$ is homogeneous if and only if $f''(x) = Cx^r$ for some $C \neq 0$ and real number r.

Proof. Suppose that $S_f(a, b)$ is a homogeneous mean and, without loss of generality, assume that $f''(1) = 1$. Let k be a positive number and define a function g by $g(x) = f(kx)$. Then

$$S_g(a, b) = \frac{\int_a^b xk^2 f''(kx)\,dx}{\int_a^b k^2 f''(kx)\,dx} = \frac{\int_{ka}^{kb} tf''(t)\,dt}{k\int_{ka}^{kb} f''(t)\,dt} = \frac{1}{k} S_f(ka, kb) = S_f(a, b).$$

By Theorem 3, we find that $g''(x) = g''(1)f''(x)$ for all $x > 0$. This equation can be written as $f''(kx) = f''(k)f''(x)$, and it is valid for all positive numbers k and x. This shows that the function f'' is multiplicative, that is, $f''(x) = x^r$ for some real number r.

The proof of the converse is purely mechanical. Assume that f'' has the given form, find the formula for the mean $S_f(a, b)$ (see the next paragraph), and show that it is homogeneous. The simple details will be omitted. ∎

Specific examples of S_f means

Since the more familiar means arise from power functions, we will focus on these and let the reader explore what happens for other types of functions. Let $S_r(a, b)$ represent the mean generated by the function $f(x) = x^r$. A routine computation shows that

$$S_r(a, b) = \frac{r - 1}{r} \cdot \frac{b^r - a^r}{b^{r-1} - a^{r-1}},$$

for values of r other than 0 and 1. It is not difficult to show that $\lim_{r\to 0} S_r(a, b)$ corresponds to the mean generated by the function $f(x) = \ln x$ and that $\lim_{r\to 1} S_r(a, b)$ corresponds to the mean generated by the function $f(x) = x \ln x$. We will use these values as the definitions for $S_0(a, b)$ and $S_1(a, b)$, respectively. The means generated by some particular values of r are recorded below.

$$S_2(a,b) = \frac{a+b}{2} = A(a,b) \qquad \text{arithmetic mean}$$

$$S_{1/2}(a,b) = \sqrt{ab} = G(a,b) \qquad \text{geometric mean}$$

$$S_{-1}(a,b) = \frac{2ab}{a+b} = H(a,b) \qquad \text{harmonic mean}$$

$$S_{3/2}(a,b) = \frac{a+\sqrt{ab}+b}{3} = He(a,b) \qquad \text{Heronian mean}$$

$$S_1(a,b) = \frac{b-a}{\ln b - \ln a} = L(a,b) \qquad \text{logarithmic mean}$$

$$S_0(a,b) = \frac{\ln b - \ln a}{\frac{1}{a} - \frac{1}{b}} = \frac{(G(a,b))^2}{L(a,b)}$$

$$S_3(a,b) = \frac{a^2+ab+b^2}{3} \div \frac{b+a}{2} = \frac{He(a^2,b^2)}{A(a,b)}$$

For fixed positive numbers a and b, it is elementary to verify that $S_r(a,b)$ is a continuous function of r on $(-\infty, \infty)$ and that

$$\lim_{r\to-\infty} S_r(a,b) = a \quad \text{and} \quad \lim_{r\to\infty} S_r(a,b) = b.$$

The reader with some knowledge of means may notice that the means $S_r(a,b)$ appear to increase with r. This is indeed the case and will be confirmed at the end of the paper.

A second way to obtain a mean

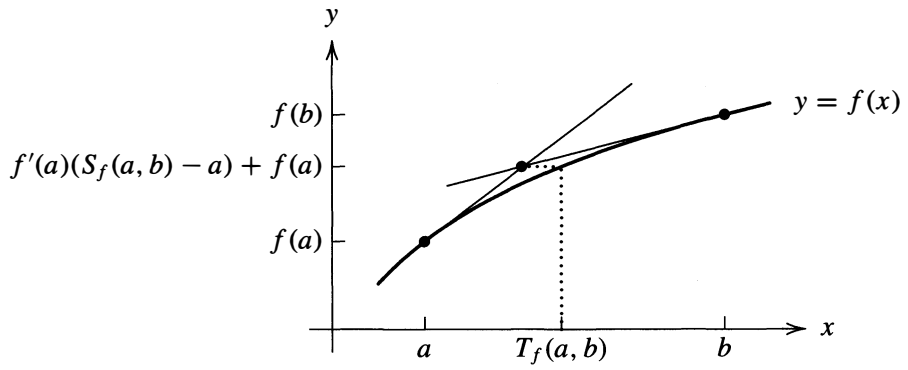
Another way to use tangent lines to define a mean involves the y-coordinate of the point of intersection of the tangent lines. Suppose that, in addition to the conditions already imposed on f, the function f' is nonzero on $(0, \infty)$. In this case, the function f has a twice differentiable inverse f^{-1}. Let a and b be distinct positive numbers and once again let (c, d) be the point of intersection of the tangent lines of f at $(a, f(a))$ and $(b, f(b))$. If f' is positive on $(0, \infty)$, then

$$d = f'(a)(c-a) + f(a) > f(a) \quad \text{and} \quad d = f'(b)(c-b) + f(b) < f(b);$$

the inequalities are reversed if f' is negative on $(0, \infty)$. In either case, the number d is between $f(a)$ and $f(b)$. Since f is continuous and one-to-one on $[a, b]$, there exists a unique number $t \in (a, b)$ such that $f(t) = d$. We will denote this mean, depicted in the figure, by $T_f(a,b)$, that is,

$$T_f(a,b) = f^{-1}\Big(f'(a)\big(S_f(a,b) - a\big) + f(a)\Big).$$

Since this definition is rather awkward and involves the mean $S_f(a,b)$, it is helpful to find another expression for this quantity. Consider the tangent lines of f^{-1} at the points $(f(a), a)$ and $(f(b), b)$. Since $(f^{-1})'(f(x)) = 1/f'(x)$ for all $x > 0$, the equations of these tangent lines are



$$y = \frac{1}{f'(a)}\big(x - f(a)\big) + a, \quad \text{or} \quad x = f'(a)(y - a) + f(a),$$

and

$$y = \frac{1}{f'(b)}\big(x - f(b)\big) + b, \quad \text{or} \quad x = f'(b)(y - b) + f(b).$$

These equations show that if (c, d) is the point of intersection of the tangent lines to $y = f(x)$ at $(a, f(a))$ and $(b, f(b))$, then (d, c) is the point of intersection of the tangent lines to $y = f^{-1}(x)$ at $(f(a), a)$ and $(f(b), b)$. Therefore, the mean $T_f(a, b)$ satisfies

$$f\big(T_f(a, b)\big) = \frac{\int_{f(a)}^{f(b)} x(f^{-1})''(x)\,dx}{\int_{f(a)}^{f(b)} (f^{-1})''(x)\,dx} = \frac{\int_a^b f(x) \cdot \frac{f''(x)}{(f'(x))^2}\,dx}{\int_a^b \frac{f''(x)}{(f'(x))^2}\,dx}.$$

The second equality follows from the relationship between $(f^{-1})''$ and the derivatives of f. This formulation for the T_f mean is useful when it is difficult or impossible to find a formula for f^{-1}.

General properties of T_f means

Let a be a fixed positive number and consider the function $T_f(a, x)$, where $x > 0$. As with the mean $S_f(a, x)$, this function is defined for all positive $x \neq a$ and becomes continuous on $(0, \infty)$ if $T_f(a, a)$ is defined to be a. The order relationship between the means $S_f(a, b)$ and $T_f(a, b)$ is given in the following theorem.

THEOREM 5. Let a and b be distinct positive numbers.

a) If the product $f'f''$ is positive on $(0, \infty)$, then $T_f(a, b) < S_f(a, b)$.

b) If the product $f'f''$ is negative on $(0, \infty)$, then $S_f(a, b) < T_f(a, b)$.

Proof. We will consider one of four possible cases; the proofs of the other cases are similar. Suppose that f' is positive and f'' is negative (as in the above graph). Since f'' is negative, the lines tangent to the graph of $y = f(x)$ lie above the curve. Therefore, the point of intersection $(S_f(a, b), f(T_f(a, b)))$ lies above the curve, that is, $f(S_f(a, b)) < f(T_f(a, b))$. Since f is an increasing function, it follows that $S_f(a, b) < T_f(a, b)$. ■

The next three theorems are the analogues of Theorems 1, 2, and 4.

THEOREM 6. If there exist constants $\alpha \neq 0$ and β such that $g(x) = \alpha f(x) + \beta$ for all $x > 0$, then $T_f(a, b) = T_g(a, b)$ for all positive numbers a and b.

Proof. Let a and b be positive numbers. By Theorem 1, we know that $S_f(a, b) = S_g(a, b)$. Since

$$\begin{aligned} g\big(T_g(a,b)\big) &= g'(a)\big(S_g(a,x) - a\big) + g(a) \\ &= \alpha f'(a)\big(S_f(a,x) - a\big) + \alpha f(a) + \beta \\ &= \alpha f\big(T_f(a,b)\big) + \beta = g\big(T_f(a,b)\big) \end{aligned}$$

and g is one-to-one, it follows that $T_g(a, b) = T_f(a, b)$. ∎

THEOREM 7. If $a > 0$, then the function $T_f(a, x)$ is strictly increasing on $(0, \infty)$ and has a continuous positive derivative on $(0, \infty)$.

Proof. By Theorem 6, we may assume that $f(a) = 0$ and $f'(a) = 1$ (replace the function $f(x)$ with $(f(x) - f(a))/f'(a)$). Under these conditions,

$$T_f(a, x) = f^{-1}\big(S_f(a, x) - a\big)$$

for all $x > 0$. Since $T_f(a, x)$ is the composition of two continuously differentiable functions, it is continuously differentiable on $(0, \infty)$ and

$$T_f'(a, x) = \frac{S_f'(a, x)}{f'\big(f^{-1}\big(S_f(a, x) - a\big)\big)} = \frac{S_f'(a, x)}{f'(T_f(a, x))}$$

for all $x > 0$. Note that $T_f'(a, a) = 1/2$. Since the functions $S_f'(a, x)$ and f' are positive on $(0, \infty)$, the function $T_f'(a, x)$ is positive on $(0, \infty)$ and $T_f(a, x)$ is strictly increasing on $(0, \infty)$. ∎

THEOREM 8. If $f'(x) = Cx^r$ for all $x > 0$, where C and r are nonzero constants, then the mean $T_f(a, b)$ is homogeneous.

Proof. As with $S_f(a, b)$, direct computation of $T_f(a, b)$ (see the next paragraph) shows that it is homogeneous. ∎

Specific examples of T_f means

If $T_r(a, b)$ represents the mean generated by $f(x) = x^r$, then it is easy to verify that

$$T_r(a, b) = \left((1 - r) \cdot \frac{b - a}{b^{1-r} - a^{1-r}}\right)^{1/r}$$

for values of r other than 0 and 1. Note that $T_r(a, b)$ is the point guaranteed to exist by the Mean Value Theorem for the function $f(x) = x^{1-r}$ on the interval $[a, b]$. It is not difficult to show that $\lim_{r \to 0} T_r(a, b)$ corresponds to the mean generated by the

function $f(x) = \ln x$ and that $\lim_{r\to 1} T_r(a, b)$ gives the logarithmic mean of a and b. (As with the S_r means, the function $f(x) = x \ln x$ would seem to correspond to this case. However, its derivative does not have constant sign on $(0, \infty)$. It does however have constant sign for $x > 1$, but for $b > a > 1$ it generates a mean different than the logarithmic mean.) We will use these values as the definitions for $T_0(a, b)$ and $T_1(a, b)$, respectively. The means generated by some particular values of r are recorded below.

$$T_2(a,b) = \sqrt{ab} = G(a,b) \qquad \text{geometric mean}$$

$$T_{-1}(a,b) = \frac{a+b}{2} = A(a,b) \qquad \text{arithmetic mean}$$

$$T_{1/2}(a,b) = \left(\frac{\sqrt{a}+\sqrt{b}}{2}\right)^2 = P_{1/2}(a,b) \qquad \text{power mean, } P_r(a,b) = \left(\frac{a^r+b^r}{2}\right)^{1/r}$$

$$T_1(a,b) = \frac{b-a}{\ln b - \ln a} = L(a,b) \qquad \text{logarithmic mean}$$

$$T_0(a,b) = \frac{1}{e}\cdot\left(\frac{b^b}{a^a}\right)^{1/(b-a)} = I(a,b) \qquad \text{identric mean}$$

$$T_3(a,b) = \left(\frac{2ab}{a+b}\cdot ab\right)^{1/3} = \big(H(a,b)\big(G(a,b)\big)^2\big)^{1/3}$$

For fixed positive numbers a and b, it is elementary to verify that $T_r(a, b)$ is a continuous function of r on $(-\infty, \infty)$. In addition, applications of L'Hôpital's Rule reveal that $\lim_{r\to-\infty} T_r(a, b) = b$ and $\lim_{r\to\infty} T_r(a, b) = a$. A proof that the means $T_r(a, b)$ decrease as r increases will be discussed at the end of the paper.

Solving the equations $T_r(1, 2) = H(1, 2)$ and $T_r(2, 3) = H(2, 3)$ for r and obtaining different values shows that the harmonic mean is not a T_r mean. The reader is invited to investigate means generated by other types of functions. For instance, the mean generated by $f(x) = e^x$ is $T_f(a, b) = a + b - c$, where c is the point guaranteed to exist by the Mean Value Theorem for e^x on $[a, b]$.

In looking over the results for the $S_f(a, b)$ means and the $T_f(a, b)$ means, one notices a lack of converses for Theorems 6 and 8. Although we are inclined to believe that the converse of Theorem 6 is valid, we have not been able to prove it. The following result gives a partial converse.

THEOREM 9. If $T_f(a, b) = T_g(a, b)$ and $S_f(a, b) = S_g(a, b)$ for all positive numbers a and b, then there exist constants $\alpha \neq 0$ and β such that $g = \alpha f + \beta$.

Proof. Fix $a > 0$ and let $T(x) = T_f(a, x)$ and $S(x) = S_f(a, x)$. Define functions f_1 and g_1 by

$$f_1(x) = \frac{f(x) - f(a)}{f'(a)} \quad \text{and} \quad g_1(x) = \frac{g(x) - g(a)}{g'(a)}.$$

Note that $f_1(a) = 0 = g_1(a)$ and $f_1'(a) = 1 = g_1'(a)$. By Theorems 1 and 6, we know that $T_{f_1}(a, x) = T(x) = T_{g_1}(a, x)$ and $S_{f_1}(a, x) = S(x) = S_{g_1}(a, x)$. By definition of the T means,

$$f_1(T(x)) = S(x) - a \quad \text{and} \quad g_1(T(x)) = S(x) - a$$

for all $x > 0$. Since T is one-to-one and $f_1 \circ T = g_1 \circ T$, it follows that $f_1 = g_1$.

Consequently,

$$g(x) = \frac{g'(a)}{f'(a)}\big(f(x) - f(a)\big) + g(a)$$

for all $x > 0$. This completes the proof. ■

COROLLARY 10. *If $c \neq 0$ and $g(x) = f(x) + cx$, then the means $T_f(a, b)$ and $T_g(a, b)$ are different.*

Proof. Suppose that $T_f(a, b) = T_g(a, b)$ for all positive numbers a and b. Since $S_f(a, b) = S_g(a, b)$ for all positive numbers a and b, the previous theorem guarantees the existence of constants $\alpha \neq 0$ and β such that $g = \alpha f + \beta$. Using the hypothesis for the form of g, this equation can be written as $(1 - \alpha)f(x) = \beta - cx$. If $\alpha = 1$, then $\beta = cx$, a contradiction to the fact that β is constant, and if $\alpha \neq 1$, then f is linear, a contradiction to the fact that f'' is nonzero. ■

As a final comment on the converses of Theorems 6 and 8, we note that the converse of Theorem 8 follows from the converse of Theorem 6. To prove this, assume that the converse of Theorem 6 is valid and that the mean $T_f(a, b)$ is homogeneous. Without loss of generality, we may assume that $f'(1) = 1$. Let k be a positive number and define a function g by $g(x) = f(kx)$. Then

$$g\big(T_g(a, b)\big) = \frac{\int_a^b \frac{gg''}{(g')^2}}{\int_a^b \frac{g''}{(g')^2}} = \frac{\int_{ka}^{kb} \frac{ff''}{(f')^2}}{\int_{ka}^{kb} \frac{f''}{(f')^2}} = f\big(T_f(ka, kb)\big) = f\big(kT_f(a, b)\big) = g\big(T_f(a, b)\big).$$

If the converse of Theorem 6 were true, we would conclude that $g'(x) = g'(1)f'(x)$ for all $x > 0$. This equation can be written as $f'(kx) = f'(k)f'(x)$, and is valid for all positive numbers k and x. This shows that the function f' is multiplicative, that is, $f'(x) = x^r$ for some real number r. Since f'' must be nonzero, r must be nonzero.

Relationship to other means

To prove that the means $S_r(a, b)$ increase with r and the means $T_r(a, b)$ decrease with r, we will relate these means to those found in the literature. The means $S_r(a, b)$ and $T_r(a, b)$ are special cases of means considered by Stolarsky [7]. Stolarsky defines the generalized logarithmic mean as (we have modified the notation)

$$L_r(a, b) = \left(\frac{1}{r} \cdot \frac{b^r - a^r}{b - a}\right)^{1/(r-1)}.$$

It then follows that $T_r(a, b) = L_{1-r}(a, b)$. In his proof of monotonicity, Stolarsky considers the mean

$$E(r, s : a, b) = \left(\frac{s}{r} \cdot \frac{b^r - a^r}{b^s - a^s}\right)^{1/(r-s)},$$

where we have adopted the notation developed by Leach and Sholander [4]. Both of these papers [4, 7] give proofs that the means $E(r, s : a, b)$ increase with increase in either r or s. (The proofs do not involve any ideas beyond elementary real analysis, but the details are a bit tedious and require some effort on the part of the reader.) It follows that the mean $S_r(a, b) = E(r, r - 1 : a, b)$ increases as r increases, and that

the mean $T_r(a,b) = E(1, 1-r : a, b)$ decreases as r increases. The two values are the same when $r = 1$; the common value is the logarithmic mean.

Acknowledgments and further reading

One of the authors, Brian Dietel, is an undergraduate student at Whitman College, and the work presented here is the result of a grant provided by the Perry Summer Research Scholarship Program. We first started working on this material after reading the paper by Stenlund [**6**]. Our results offer another proof of the main result in that paper. After obtaining the results presented here, we discovered the article by Horwitz [**3**], in which he considers the point of intersection of Taylor polynomials of odd degree and proves a number of interesting results about the corresponding means. Of course, tangent lines are the special case in which the degree of the Taylor polynomial is one. Horwitz's results offer generalizations of some of the results discussed in this paper as well as different (but more advanced) proofs. We highly recommend this paper to those who have a good background in analysis. For other ways to use a function to define a mean, see the paper by Mays [**5**] and the references found there. The classic work by Hardy, Littlewood, and Pólya [**2**] is also a good source of information about means.

REFERENCES

1. C. W. Groetsch, *Inverse Problems*, Mathematical Association of America, 1999.
2. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge, 1952.
3. A. Horwitz, Means and Taylor polynomials, *J. Math. Anal. Appl.* **149** (1990), 220–235.
4. E. B. Leach and M. C. Sholander, Extended mean values, *Amer. Math. Monthly* **85** (1978), 84–90.
5. M. E. Mays, Functions which parametrize means, *Amer. Math. Monthly* **90** (1983), 677–683.
6. M. Stenlund, On the tangent lines of a parabola, *College Math. J.* **32** (2001), 194–196.
7. K. B. Stolarsky, Generalizations of the logarithmic mean, this MAGAZINE, **48** (1975), 87–92.

Characterization of Polynomials Using Divided Differences

ELIAS DEEBA
PLAMEN SIMEONOV
University of Houston–Downtown
Houston, TX 77002
deebae@dt.uh.edu,simeonov@dt.uh.edu

A standard exercise in calculus books [**9**] is to prove that if a function f is continuous on a closed interval I and differentiable inside it, and if for any points $a, b \in I$

$$\frac{f(b) - f(a)}{b - a} = f'\left(\frac{a+b}{2}\right),$$

then $f(x)$ is a quadratic function over that interval. Geometrically, if the slope of the secant line over x-values a and b is equal to the slope of the tangent line at $x = (a+b)/2$ for every a and b, the graph of $f(x)$ is a parabola.

A natural generalization of this property replaces (or extends) the average value of the function at two points with the *divided difference* of f at $n+1$ points. The divided difference of the function f at the points $\{x_j\}_{j=0}^n$ is defined recursively by

$$f[x_i] := f(x_i), \quad i \in \{0, \ldots, n\}$$

and then

$$f[x_s, \ldots, x_t] := \frac{f[x_{s+1}, \ldots, x_t] - f[x_s, \ldots, x_{t-1}]}{x_t - x_s}, \qquad 0 \le s < t,$$

where the points $\{x_j\}_{j=0}^n$ are assumed to be distinct. In particular,

$$f[x_0, x_1] := \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

and

$$f[x_0, \ldots, x_n] := \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}.$$

Then a reasonable generalization of the above result is: If f is a continuous function on $[a, b]$, its nth derivative is continuous in (a, b), and

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!} f^{(n)}\left(\frac{x_0 + x_1 + \cdots + x_n}{n+1}\right) \tag{1}$$

for every set of $n+1$ distinct points in $[a, b]$, then f is a polynomial of degree at most $n+1$.

In general, the divided difference of an arbitrary function is a complicated object. However, for polynomials there are explicit formulas for the divided differences of any order. Therefore it is of interest to characterize those situations that will determine whether a given function is a polynomial.

We will prove the following generalization of (1):

THEOREM. *Let $n \ge 1$ be fixed. Let f be a bounded function on the compact subsets of* **R**. *Assume that the divided difference of f at any $n+1$ distinct points $\{x_j\}_{j=0}^n$ satisfies the functional equation*

$$f[x_0, x_1, \ldots, x_n] = H(x_0 + x_1 + \cdots + x_n). \tag{2}$$

Then f is a polynomial of degree at most $n+1$. Furthermore, $H(t) = a_{n+1}t + a_n$, where a_{n+1} and a_n are the leading coefficients of f.

Polynomials and divided differences have been treated by many authors. The case $n = 1$ was studied by Aczel [**2**]. Bailey [**3**] considered the case $n = 2$ and proved that f is a cubic polynomial. Andersen [**5**] gave a complete characterization based on (2) and proved that f is a polynomial of degree at most $n+1$. We establish Andersen's result by providing a new proof that extends the argument employed by Bailey [**3**] for the case $n = 2$. Different proofs of this theorem can be found in many sources [**4, 5, 6, 7**]. What is remarkable about the result is the seeming weakness of the assumptions made on the functions f and H (which are not even assumed to be continuous), and the very strong conclusions that follow from them. Furthermore, our proof will use rather elementary methods.

In the proof of the theorem we will use the following well-known properties (Proofs can be found in de Boor's book [1], though they are not too hard to establish—the first by induction and the second by an application of Taylor's Theorem.):

PROPERTY 1. The divided difference at $n+1$ distinct points has the representation

$$f[x_0, x_1, \ldots, x_n] = \sum_{j=0}^{n} \frac{f(x_j)}{p'_{n+1}(x_j)}, \tag{3}$$

where $p_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$.

PROPERTY 2. If f has an nth derivative, then

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}, \tag{4}$$

where $\xi \in (\min\{x_0, \ldots, x_n\}, \max\{x_0, \ldots, x_n\})$. In particular, if $f \in \mathcal{P}_{n-1}$, the set of polynomials of degree at most $n-1$, then $f[x_0, \ldots, x_n] = 0$.

Proof of the theorem. The first step is to show that f is continuous. Using equation (3), for distinct numbers $\{x_j\}_{j=0}^{n}$ we can write the divided difference $f[x_0, x_1, \ldots, x_n]$ in the form

$$\begin{aligned} f[x_0, x_1, \ldots, x_n] &= \sum_{j=0}^{n} \frac{f(x_j)}{p'_{n+1}(x_j)} = \frac{f(x_0)}{p'_{n+1}(x_0)} + \frac{f(x_1)}{p'_{n+1}(x_1)} + \sum_{j=2}^{n} \frac{f(x_j)}{p'_{n+1}(x_j)} \\ &= \frac{f(x_1)}{p'_{n+1}(x_1)} - \frac{f(x_0)}{p'_{n+1}(x_1)} + \frac{f(x_0)}{p'_{n+1}(x_1)} + \frac{f(x_0)}{p'_{n+1}(x_0)} + \sum_{j=2}^{n} \frac{f(x_j)}{p'_{n+1}(x_j)} \\ &= \left(\frac{f(x_1) - f(x_0)}{x_1 - x_0}\right)\left(\frac{x_1 - x_0}{p'_{n+1}(x_1)}\right) \\ &\quad + f(x_0)\left(\frac{1}{p'_{n+1}(x_0)} + \frac{1}{p'_{n+1}(x_1)}\right) + \sum_{j=2}^{n} \frac{f(x_j)}{p'_{n+1}(x_j)}. \end{aligned}$$

But then by (2) we have

$$\begin{aligned} &\left(\frac{f(x_1) - f(x_0)}{x_1 - x_0}\right)\left(\frac{x_1 - x_0}{p'_{n+1}(x_1)}\right) + f(x_0)\left(\frac{1}{p'_{n+1}(x_0)} + \frac{1}{p'_{n+1}(x_1)}\right) \\ &\quad + \sum_{j=2}^{n} \frac{f(x_j)}{p'_{n+1}(x_j)} = H(x_0 + x_1 + \cdots + x_n). \end{aligned} \tag{5}$$

Fix the numbers $x_0, x_2, \ldots, x_n$ and let $x_1 \to x_0$ in (5). As $x_1 \to x_0$ all terms in (5) remain bounded except possibly the first term. The second factor in the first term,

$$\frac{x_1 - x_0}{p'_{n+1}(x_1)} = \frac{1}{(x_1 - x_2)\cdots(x_1 - x_n)},$$

tends to a nonzero constant. Therefore, $(f(x_1) - f(x_0))/(x_1 - x_0)$ remains bounded as $x_1 \to x_0$, which is possible only if $f(x_1) \to f(x_0)$ as $x_1 \to x_0$. Thus, f is a continuous function. This in turn implies that H is continuous.

Next, we show that f has a continuous first derivative. Let again x_1 tend to x_0 in (5). The right-hand side of (5) tends to $H(2x_0 + x_2 + \cdots + x_n)$. The second term on the left-hand side of (5) tends to $f(x_0)Q'(x_0)$, where $Q(x) = 1/\prod_{j=2}^{n}(x - x_j)$ (this last claim is proved by combining the fractions and taking a limit). The sum on the left-hand side also has a finite limit. Therefore, the first term on the left-hand side of (5) has a finite limit, that is, $f'(x_0)$ exists. Furthermore, since the points $x_0, x_2, \ldots, x_n$ are distinct, all terms in the equation obtained from (5) after taking the limit $x_1 \to x_0$, except possibly the first term, are continuous at x_0 for fixed $\{x_j\}_{j=2}^{n}$. Then that first term $f'(x_0)/((x_0 - x_2)\cdots(x_0 - x_n))$ is also continuous, and therefore f' is continuous. This in turn implies that $H'(t) = H'(x_0 + \cdots + x_n)$ is continuous as well.

We are now in a position to show that f has derivatives of arbitrary order. From (2) and (3) we get

$$f[x_0 + h, \ldots, x_n + h] = \sum_{j=0}^{n} \frac{f(x_j + h)}{p'_{n+1}(x_j)} = H(x_0 + \cdots + x_n + (n+1)h). \tag{6}$$

We differentiate (6) with respect to h and set $h = 0$ to obtain

$$f'[x_0, \ldots, x_n] = \sum_{j=0}^{n} \frac{f'(x_j)}{p'_{n+1}(x_j)} = (n+1)H'(x_0 + \cdots + x_n). \tag{7}$$

This functional equation for f' has the same form as equation (2) for f. Using the above argument and induction we obtain

$$f^{(m)}[x_0, \ldots, x_n] = (n+1)^m H^{(m)}(x_0 + x_1 + \cdots + x_n)$$

for every m. Thus, f and H have continuous derivatives of arbitrary order.

We now select equally spaced points $x_j = t + (j - \frac{n}{2})h$, $j = 0, \ldots, n$. Then

$$p'_{n+1}(x_j) = \prod_{k=0, k\neq j}^{n} ((j-k)h) = (-1)^{n-j} j!\,(n-j)!\,h^n.$$

For these points from (2) and (3) it follows that

$$\sum_{j=0}^{n} \frac{(-1)^{n-j} f(t + (j - \frac{n}{2})h)}{j!\,(n-j)!} = h^n H((n+1)t). \tag{8}$$

We differentiate (8) with respect to h, $n+2$ times, and set $h = 0$. We get

$$\left[\sum_{j=0}^{n} \frac{(-1)^{n-j}}{j!\,(n-j)!}\left(j - \frac{n}{2}\right)^{n+2}\right] f^{(n+2)}(t) = 0. \tag{9}$$

The conclusion that $f^{(n+2)}(t) = 0$ and hence that $f \in \mathcal{P}_{n+1}$ will follow immediately from (9) provided we can show that the coefficient of $f^{(n+2)}(t)$ in (9), to be denoted by K_n, is not zero.

We will give an explicit evaluation of K_n, namely we will show that

$$K_n = n(n+1)(n+2)/24.$$

Observe that K_n is a divided difference:

$$K_n = \left(x - \frac{n}{2}\right)^{n+2}[0, 1, \ldots, n] = \sum_{s=0}^{n+2} \binom{n+2}{s} \left(-\frac{n}{2}\right)^{n+2-s} x^s[0, 1, \ldots, n]. \quad (10)$$

By (4), for $s < n$ the divided difference $x^s[0, 1, \ldots, n]$ is zero and (10) reduces to a sum of at most three nonzero terms. To evaluate these terms, for $n \geq 1$ and $l \geq 0$ we set

$$c_l(n) := x^{n+l}[0, 1, \ldots, n].$$

We have $c_l(1) = 1$, and from (4) we get $c_0(n) = 1$. For $l \geq 1$ we have

$$\begin{aligned} c_l(n) &= \sum_{j=1}^{n} \frac{(-1)^{n-j} j^{n+l-1}}{(j-1)!\,(n-j)!} = \sum_{\nu=0}^{n-1} \frac{(-1)^{n-1-\nu}(\nu+1)^{n-1+l}}{\nu!\,(n-1-\nu)!} \\ &= \sum_{\nu=0}^{n-1} \frac{(-1)^{n-1-\nu}}{\nu!\,(n-1-\nu)!} \sum_{s=0}^{n-1+l} \binom{n-1+l}{s} \nu^s \\ &= \sum_{s=0}^{n-1+l} \binom{n-1+l}{s} t^s[0, 1, \ldots, n-1] \\ &= \sum_{s=n-1}^{n-1+l} \binom{n-1+l}{s} c_{s-(n-1)}(n-1) \\ &= c_l(n-1) + \sum_{k=0}^{l-1} \binom{n-1+l}{n-1+k} c_k(n-1), \end{aligned} \quad (11)$$

where we used (4) to eliminate the terms for $s < n-1$. From (11) we get

$$c_1(n) = c_1(1) + \sum_{j=2}^{n} (c_1(j) - c_1(j-1)) = \sum_{j=1}^{n} j = n(n+1)/2. \quad (12)$$

Similarly, from (11) for $l = 2$ we get

$$c_2(n) - c_2(n-1) = (n(n+1)/2)c_0(n-1) + (n+1)c_1(n-1) = n^2(n+1)/2,$$

and then

$$\begin{aligned} c_2(n) &= c_2(1) + \sum_{j=2}^{n} (c_2(j) - c_2(j-1)) = 1 + \sum_{j=2}^{n} (j^3 + j^2)/2 \\ &= n(n+1)(n+2)(3n+1)/24. \end{aligned} \quad (13)$$

Combining (10), (12), and (13) we obtain

$$K_n = \sum_{s=n}^{n+2} \binom{n+2}{s} (-n/2)^{n+2-s} c_{s-n}(n) = n(n+1)(n+2)/24.$$

To prove the statement of the theorem concerning H we note that since $f \in \mathcal{P}_{n+1}$, by (4)

$$H(x_0 + x_1 + \cdots + x_n) = (n+1)a_{n+1}\xi + a_n$$

for some $\xi \in (\min\{x_0, \ldots, x_n\}, \max\{x_0, \ldots, x_n\})$, where a_{n+1} and a_n are the leading coefficients of f. Fixing t and letting the distinct values of $x_0, \ldots, x_n$ all tend to t we obtain $H((n+1)t) = (n+1)a_{n+1}t + a_n$, whence $H(t) = a_{n+1}t + a_n$ as desired. ■

Open problems

Let n and s be fixed positive integers, and let $\sigma_k(x_0, x_1, \ldots, x_n)$, $k = 1, 2, \ldots$ denote the elementary symmetric polynomials of the variables $x_0, x_1, \ldots, x_n$. Determine the set of functions f that are bounded on the compact subsets of $\mathbf{R}$ and satisfy a functional equation of the form

$$f[x_0, \ldots, x_n] = H(\sigma_1(x_0, \ldots, x_n), \ldots, \sigma_s(x_0, \ldots, x_n)).$$

We have solved the problem in the special case $s = 1$. The solution for this special case suggests that in the general case f is a polynomial of degree at most $n + s$. The boundedness condition is indeed necessary as the following example shows [**8**]:

$$(1/x)[x_0, \ldots, x_n] = (-1)^n/(x_0 \cdots x_n).$$

We believe that the proof of our theorem can be modified to the case $s = 2$. It would be of interest to find a solution of this open problem for the general case.

Acknowledgments. The authors would like to thank the referees for their valuable suggestions and comments.

REFERENCES

1. Carl de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.
2. J. Aczel, A mean value property of the derivative of quadratic polynomials-without mean values and derivatives, this MAGAZINE **58**:1 (1985), 42–45.
3. D. F. Bailey, A mean-value property of cubic polynomials—without mean values, this MAGAZINE **65**:2 (1992), 123–124.
4. R. O. Davies and G. Rousseau, A divided difference characterization of polynomials over a general field, *Aequationes Math.* **55** (1998), 73–78.
5. K. M. Andersen, A characterization of polynomials, this MAGAZINE **69**:2 (1996), 137–142.
6. P. L. Kanappan and P. K. Sahoo, Characterization of polynomials and divided difference, *Proc. Indian Acad. Sci. (Math. Sci.)* **105** (1995), 287–290.
7. J. Schwaiger, On a characterization of polynomials by divided differences, *Aequationes Math.* **48** (1994), 317–323.
8. P. Simeonov, Characterization of a class of functions using divided differences, *Abstract and Applied Analysis*, **5** (2001), 1–6.
9. J. Stewart, *Calculus*, Brooks/Cole Publishing Company, Pacific Grove, California, 1999.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN GELCA, Texas Tech University; ROBERT GREGORAC, Iowa State University; GERALD HEUER, Concordia College; VANIA MASCIONI, Western Washington University; PAUL ZEITZ, The University of San Francisco

Proposals

To be considered for publication, solutions should be received by July 1, 2003.

1662. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College, Bronx, NY.*

Let x_k, $1 \le k \le n$, be positive real numbers with $\sum_{k=1}^{n} x_k^{2k-1} \le n$. Prove that $\sum_{k=1}^{n}(2k-1)x_k \le n^2$.

1663. *Proposed by Michel Bataille, Rouen, France.*

Let m and n be integers such that $1 \le m < n+1$. Evaluate

$$\sum_{k=1}^{n+1}\left((k+1)\sin^{k-1}\left(\frac{2\pi m}{n+1}\right)\prod_{j=1}^{k}\left(\cot\left(\frac{\pi m}{n+1}\right)-\cot\left(\frac{\pi j}{k+1}\right)\right)\right).$$

1664. *Proposed by Tim Ferguson, student, Linganore High School, Frederick, MD, and Lenny Jones, Shippensburg University, Shippensburg, PA.*

Given a positive integer n, a sequence $\lambda_1, \lambda_2, \ldots, \lambda_k$ of positive integers is called a partition of n if $\sum_{j=1}^{k} \lambda_j = n$. Given a partition $\pi : \lambda_1, \lambda_2, \ldots, \lambda_k$ of n, let $\mathrm{LCM}(\pi) = \mathrm{LCM}(\lambda_1, \lambda_2, \ldots, \lambda_k)$, and define

$$M_n = \max\{\mathrm{LCM}(\pi) : \pi \text{ a partition of } n\}.$$

Let p be a prime such that p^a divides M_n for some integer $a \ge 3$. Prove that if q is a prime with $p < q < p^{a-1}$, then q divides M_n.

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames, IA 50011, or mailed electronically (ideally as a LaTeX file) to `ehjohnst@iastate.edu`. All communications should include the reader's name, full address, and an e-mail address and/or FAX number.

1665. *Proposed by Mihàly Bencze, Brasov, Romania.*

Let M be a point in the interior of triangle ABC and let P, Q, and R be the projections of M onto BC, CA, and AB, respectively. Prove that

$$MA^2 \sin^2 \frac{A}{2} + MB^2 \sin^2 \frac{B}{2} + MC^2 \sin^2 \frac{C}{2} \leq MP^2 + MQ^2 + MR^2.$$

1666. *Proposed by Razvan A. Satnoianu, City University, London, England.*

Let $f, g : [0, \infty) \longrightarrow [0, \infty)$ be functions with f increasing, g entire, and $g^{(n)}(0) \geq 0$ for all nonnegative integers n. Prove that for all $x, y, z \geq 0$,

$$f(x)\,(g(x) - g(y))\,(x - z) + f(y)\,(g(y) - g(z))\,(y - x) + f(z)\,(g(z) - g(x))\,(z - y) \geq 0.$$

Quickies

Answers to the Quickies are on page 73.

Q927. *Proposed by Mowaffaq Hajja, Yarmouk University, Irbid, Jordan.*

Prove that the angles of a triangle all have rational cosines if and only if the triangle is similar to one with rational sides.

Q928. *Proposed by Norman Schaumberger (Emeritus), Bronx Community College, Bronx, NY.*

Prove that if $a, b, c, d > 0$, then

$$\left(\frac{a+b+c+d}{4}\right)^{3(a+b+c+d)} \geq a^{(b+c+d)} b^{(a+c+d)} c^{(a+b+d)} d^{(a+b+c)}.$$

Solutions

Fibonacci Congruences **February 2002**

1638. *Proposed by Jody M. Lockhart and William P. Wardlaw, U. S. Naval Academy, Annapolis, MD.*

Let m be a positive integer. Show that there are infinitely many positive integers k such that m is a divisor of f_k and $f_{k+1} - 1$, where f_n denotes the nth Fibonacci number.

Solution by Nicholas C. Singer, Annandale, VA.

We prove a generalization:

> Let the integer sequence $\{x_n\}_{n=0}^{\infty}$ satisfy $x_{n+2} = ax_{n+1} + bx_n + c$ for $n \geq 0$, where a, b, and c are integers and $b \neq 0$. Let m be a positive integer relatively prime to b. Then there are infinitely many positive integers k such that m is a divisor of $x_k - x_0$ and $x_{k+1} - x_1$. (The stated problem is the case with $a = b = 1$, $c = 0$, $x_0 = 0$, and $x_1 = 1$.)

Let $x_n \equiv y_n \pmod m$ and consider the sequence $\{y_n\}$. Because there are at most m^2 possible ordered pairs (y_n, y_{n+1}), the sequence of ordered pairs eventually repeats and

hence becomes periodic. Now suppose that y_{n+1} and y_{n+2} are known for some $n > 0$. We can then determine y_k for $0 \le k \le n$. Indeed, for any $k \ge 0$, $y_{k+2} - ay_{k+1} - c \equiv by_k \pmod m$. Because b and m are relatively prime, y_k is uniquely determined, modulo m. It follows that the sequence $\{y_n\}_{n=0}^{\infty}$ is periodic. Hence there are infinitely many k with $x_k \equiv x_0 \pmod m$ and $x_{k+1} \equiv x_1 \pmod m$.

Also solved by Michel Bataille (France), Kenneth Bernstein, J. C. Binz (Switzerland), John Christopher, Con Amore Problem Group (Denmark), Chip Curtis, Daniele Donini (Italy), Robert L. Doucette, Brian D. Ginsberg, Elana C. Greenspan, Ralph P. Grimaldi, Jerrold W. Grossman, The Ithaca College Solvers, Tom Jager, Ken Korbin, Harris Kwong, Elias Lampakis (Greece), James Magliano, Paul Martin, Tim McMillan, Rolf Richberg (Germany), Heinz-Jürgen Seiffert (Germany), Skidmore College Problem Group, Lawrence Somer, Albert Stadler (Switzerland), Thai-Duong Tran, Dave Trautman, Daniel Treat, Michael Vowe (Switzerland), Li Zhou, and the proposers.

A Complex Identity — February 2002

1639. *Proposed by José Luis Díaz and Juan José Egozcue, Universitat Politécnica de Catalunya, Barcelona, Spain.*

Let $n \ge 3$ be a positive integer and let $z_1, z_2, \ldots, z_n$ be distinct, nonzero, complex numbers. Prove that

$$\sum_{k=1}^{n} \frac{1}{z_k} \left(-1 + (1 + z_k^{n-1}) \prod_{\substack{j=1 \\ j \ne k}}^{n} \frac{z_j}{z_j - z_k} \right) = 0.$$

I. *Solution by Albert Stadler, Dübendorf, Switzerland.*

Let

$$p(z) = (z - z_1)(z - z_2) \cdots (z - z_n) \quad \text{and} \quad f(z) = -\frac{p'(z)}{zp(z)} - \frac{p(0)(1 + z^{n-1})}{z^2 p(z)}.$$

Then f is a rational function with a pole of order 2 at $z = 0$ and a simple pole at each of $z_1, z_2, \ldots, z_n$. Furthermore, $f(z) = O(1/z^2)$ as $|z| \to \infty$. Thus, by Cauchy's Theorem, $\frac{1}{2\pi i} \int_{|z|=R} f(z)\, dz = 0$ for all $R > \max\{|z_1|, |z_2|, \ldots, |z_n|\}$. On the other hand, the value of the integral is the sum of the residues at each pole inside the circle. The residue at $z = 0$ is

$$\lim_{z \to 0} z \left(f(z) - \frac{1}{z^2} \lim_{w \to 0} w^2 f(w) \right) = \lim_{z \to 0} z \left(f(z) + \frac{1}{z^2} \right) = 0,$$

while the residue at z_k is

$$\lim_{z \to z_k} (z - z_k) f(z) = \frac{1}{z_k} \left(-1 + (1 + z_k^{n-1}) \prod_{\substack{j=1 \\ j \ne k}}^{n} \frac{z_j}{z_j - z_k} \right).$$

This completes the proof of the identity.

II. *Solution by Tom Jager, Calvin College, Grand Rapids, MI.*

We prove the more general result,

$$\sum_{k=1}^{n} \frac{1}{z_k} \left(-1 + g(z_k) \prod_{\substack{j=1 \\ j \ne k}}^{n} \frac{z_j}{z_j - z_k} \right) = a_1, \qquad (*)$$

for any polynomial $g(z) = 1 + a_1 z + a_2 a^2 + \cdots + a_n z^n$. Let $w_k = 1/z_k$. Then (*) is equivalent to

$$\sum_{k=1}^{n} w_k^n g\,(1/w_k) \prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{1}{w_k - w_j} = \sum_{k=1}^{n} w_k + a_1.$$

Because $w_k^n g(1/w_k) = w_k^n + a_1 w_k^{n-1} + \cdots + a_{n-1} w_k + a_n$, this is an immediate consequence of the following lemma:

If $w_1, \ldots, w_n$ are distinct complex numbers and λ is a nonnegative integer with $0 \leq \lambda \leq n$, then

$$f_\lambda(w_1, \ldots, w_n) = \sum_{k=1}^{n} w_k^\lambda \prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{1}{w_k - w_j} = \begin{cases} 0, & 0 \leq \lambda \leq n-2 \\ 1, & \lambda = n-1 \\ w_1 + \cdots + w_n, & \lambda = n \end{cases}.$$

For the proof of the lemma, let $h_\lambda(z) = f_\lambda(z, w_2, \ldots, w_n) \prod_{j=2}^{n}(z - w_j)$. Then h_λ is a polynomial in z and for $i = 2, \ldots, n$, $h_\lambda(w_i) = w_i^\lambda - w_i^\lambda = 0$. If $0 \leq \lambda \leq n-2$, then h_λ has $n-1$ distinct zeroes and had degree at most $n-2$. It follows that $h_\lambda(z) \equiv 0$ and hence that $f_\lambda(w_1, \ldots, w_n) = 0$. If $\lambda = n-1$, then h_λ is a monic polynomial of degree $n-1$. Thus $h_\lambda(z) = \prod_{j=2}^{n}(z - w_j)$ and $f_\lambda(w_1, \ldots, w_n) = 1$. Finally, if $\lambda = n$, then h_λ is of degree n and the coefficient of z^{n-1} is 0. Hence $h_\lambda(z) = (z - w^*) \prod_{j=2}^{n}(z - w_j)$ for some w^*, and the sum of the zeroes of h_λ is 0. Thus $f_\lambda(w_1, w_2, \ldots, w_n) = w_1 - w^* = w_1 + w_2 + \cdots + w_n$.

Also solved by Reza Akhlaghi, The Assumption College Problems Group, Michel Bataille (France), Jany C. Binz (Switzerland), Joseph Coster (Luxembourg), Ivko Dimitric, Daniele Donini (Italy), Robert L. Doucette, Ovidiu Furdui, Natalio H. Guersenzvaig (Argentina), John G. Heuver, Stephen Kaczkowski, Rolf Richberg (Germany), Heinz-Jürgen Seiffert (Germany), Nicholas C. Singer, Li Zhou, and the proposer.

A Trigonometric Inequality **February 2002**

1640. *Proposed by Péter Ivády, Budapest, Hungary.*

Show that for $0 < x, y < \pi/2$,

$$\frac{x \csc x + y \csc y}{2} < \sec\left(\frac{x+y}{2}\right).$$

Solution by John Spellmann, Southwest Texas State University, San Marcos, TX.

Let $0 < x \leq y < \pi/2$. We show that

$$\frac{x \csc x + y \csc y}{2} < \sec\left(\frac{x}{2}\right)\sec\left(\frac{y}{2}\right) < \sec\left(\frac{x+y}{2}\right).$$

The right inequality follows easily from the angle addition formula for the cosine. For the left inequality, let T be the triangle with vertices $(0, 0)$, $(1, 0)$, and $(1, \tan(x/4))$, and let S be the portion of the unit disk contained in T. Because the area of S is less than the area of T, it follows that

$$x < 4\tan\left(\frac{x}{4}\right) = \sin x \sec\left(\frac{x}{2}\right)\sec^2\left(\frac{x}{4}\right)$$

and hence that

$$x \csc x < \sec\left(\frac{x}{2}\right)\sec^2\left(\frac{x}{4}\right).$$

We then have

$$\begin{aligned}\frac{x\csc x + y\csc y}{2} &< \frac{\sec(x/2)\sec^2(x/4) + \sec(y/2)\sec^2(y/4)}{2}\\ &= \sec(x/2)\sec(y/2)\frac{\cos(y/2)\sec^2(x/4) + \cos(x/2)\sec^2(y/4)}{2}\\ &= \sec(x/2)\sec(y/2)\left(\frac{\cos(y/2)}{1+\cos(x/2)} + \frac{\cos(x/2)}{1+\cos(y/2)}\right)\\ &< \sec(x/2)\sec(y/2)\left(\frac{\cos(y/2)}{1+\cos(y/2)} + \frac{1}{1+\cos(y/2)}\right)\\ &= \sec(x/2)\sec(y/2).\end{aligned}$$

This completes the proof.

Also solved by Michel Bataille (France), Gerald E. Bilodeau, Con Amore Problem Group (Denmark), Daniele Donini (Italy), Robert L. Doucette, Ovidiu Furdui, Brian D. Ginsberg, Natalio H. Guersenzvaig (Argentina), Tom Jager, Stephen Kaczkowski, Charles Kicey, Elias Lampakis (Greece), Kee-Wai Lau (China), Peter W. Lindstrom, Paul Martin, Rolf Richberg (Germany), Albert Stadler (Switzerland), Michael Vowe (Switzerland) Li Zhou, and the proposer. There was one solution with no name.

Tetrahedrons, Spheres, and the Orthocenter **February 2002**

1641. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Show that if the midpoints of the six edges of a tetrahedron lie on a sphere, then the tetrahedron has an orthocenter.

Solution by Daniele Donini, Bertinoro, Italy.

Let A_1, A_2, A_3, and A_4 be the vertices of the tetrahedron, and let O and r be, respectively, the center and radius of the sphere. Fix an orthogonal coordinate system with origin O. Given a point P, let $\mathbf{P}$ denote the vector from O to P. By hypothesis, $\|(\mathbf{A}_i + \mathbf{A}_j)/2\| = r$ for each pair of distinct indices i, j. This condition can be rewritten as

$$4r^2 = \|\mathbf{A}_i + \mathbf{A}_j\|^2 = \mathbf{A}_i \cdot \mathbf{A}_i + 2\mathbf{A}_i \cdot \mathbf{A}_j + \mathbf{A}_j \cdot \mathbf{A}_j,$$

or equivalently

$$\mathbf{A}_i \cdot \mathbf{A}_j = 2r^2 - \tfrac{1}{2}(\mathbf{A}_i \cdot \mathbf{A}_i + \mathbf{A}_j \cdot \mathbf{A}_j). \tag{1}$$

The orthocenter of $A_1A_2A_3A_4$ is a point H such that the vectors $\mathbf{A}_i - \mathbf{A}_j$ and $\mathbf{A}_k - \mathbf{H}$ are orthogonal, that is, such that

$$(\mathbf{A}_i - \mathbf{A}_j) \cdot (\mathbf{A}_k - \mathbf{H}) = 0 \qquad \text{for any triple of distinct indices } i, j, k.$$

By (1), this condition is equivalent to

$$(\mathbf{A}_i - \mathbf{A}_j) \cdot \mathbf{H} = \tfrac{1}{2}(\mathbf{A}_j \cdot \mathbf{A}_j - \mathbf{A}_i \cdot \mathbf{A}_i) \qquad \text{for any pair of distinct indices } i, j,$$

which is in turn equivalent to the conditions

$$(\mathbf{A}_i - \mathbf{A}_1) \cdot \mathbf{H} = \tfrac{1}{2}(\mathbf{A}_1 \cdot \mathbf{A}_1 - \mathbf{A}_i \cdot \mathbf{A}_i) \qquad \text{for any index } i = 2, 3, 4. \tag{2}$$

Note that the reverse implication follows from

$$\begin{aligned}(\mathbf{A}_i - \mathbf{A}_j)\cdot\mathbf{H} &= (\mathbf{A}_i - \mathbf{A}_1)\cdot\mathbf{H} - (\mathbf{A}_j - \mathbf{A}_1)\cdot\mathbf{H}\\ &= \tfrac{1}{2}(\mathbf{A}_1\cdot\mathbf{A}_1 - \mathbf{A}_i\cdot\mathbf{A}_i) - \tfrac{1}{2}(\mathbf{A}_1\cdot\mathbf{A}_1 - \mathbf{A}_j\cdot\mathbf{A}_j)\\ &= \tfrac{1}{2}(\mathbf{A}_j\cdot\mathbf{A}_j - \mathbf{A}_i\cdot\mathbf{A}_i).\end{aligned}$$

Now consider (2) as a system of three linear equations in three unknowns; the unknowns are the three coordinates of H. Because the three vectors $\mathbf{A}_2 - \mathbf{A}_1$, $\mathbf{A}_3 - \mathbf{A}_1$, and $\mathbf{A}_4 - \mathbf{A}_1$ are linearly independent, the system has exactly one solution. The solution gives the coordinates of the orthocenter H.

Also solved by Michel Bataille (France), Robert L. Doucette, Ovidiu Furdui, H. Guggenheimer, Mowaffaq Hajja, (Jordan), John G. Heuver, Peter Y. Woo, Li Zhou, and the proposer.

Prime Time **February 2002**

1642. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College, Bronx, NY.*

Prove that for any positive integer k there is a positive integer n such that there exist n consecutive odd integers that include more than $k\lfloor\sqrt{n}\rfloor$ primes.

Solution by Rolf Richberg, Stolberg, Germany.

Assume that the assertion is false. Then there exists a natural number k such that no set $\{2a + 2v - 1 : v = 1, 2, \ldots, n\}$, with $a, n \in \mathbb{N}$, contains more than $k\lfloor\sqrt{n}\rfloor$ primes. In particular, letting $j \in \mathbb{N}$ and choosing $2a = (2j-1)^3 + 1$ and $n = 12j^2 + 1$, we find that the set

$$\begin{aligned}S_j &= \{(2j-1)^3 + 2v : v = 1, 2, \ldots, 12j^2 + 1\}\\ &= \{(2j-1)^3 + 2, (2j-1)^3 + 4, \ldots, (2j+1)^3\}\end{aligned} \tag{1}$$

contains at most $k\lfloor\sqrt{12j^2+1}\rfloor$ primes. Observing that $(2j-1)^3 + 2 > 2j^3$ and $k\lfloor\sqrt{12j^2+1}\rfloor < 4kj$, we find that for all $j \in \mathbb{N}$,

$$\sum_{\substack{p\in S_j\\ p \text{ prime}}} \frac{1}{p} < \frac{4kj}{(2j-1)^3+2} < \frac{2k}{j^2}. \tag{2}$$

From (1) we see that the sets S_j partition the set of odd integers greater than or equal to 3. Thus by (2),

$$\sum_{\substack{p\geq 3\\ p \text{ prime}}} \frac{1}{p} = \sum_{j=1}^{\infty}\left(\sum_{\substack{p\in S_j\\ p \text{ prime}}} \frac{1}{p}\right) < 2k\sum_{j=1}^{\infty}\frac{1}{j^2} < \infty,$$

which contradicts Euler's result that $\sum_{p \text{ prime}} 1/p$ diverges. This completes the proof. We note that by a similar argument we can obtain a proof for the assertion that arises by replacing $\sqrt{n}$ by n^{α}, with $0 < \alpha < 1$.

Also solved by Roy Barbara (Lebanon), Michel Bataille (France), John Christopher, Daniele Donini (Italy), Robert L. Doucette, Tom Jager, Ken Korbin, Peter W. Lindstrom, Nicholas C. Singer, Li Zhou, and the proposers.

Answers

Solutions to the Quickies from page 68.

A927. If each side of a triangle has rational length, then it follows from the Law of Cosines that each of its angles has rational cosine. Conversely, suppose that ABC is a triangle and that $\cos A$, $\cos B$, and $\cos C$ are rational. Because

$$-\cos A = \cos(B + C) = \cos B \cos C - \sin B \sin C,$$

it follows that $\sin B \sin C$ is rational, (and by a similar argument, $\sin C \sin A$ and $\sin A \sin B$ are also rational.) Therefore,

$$\sin A : \sin B : \sin C = r_1 : r_2 : r_3,$$

for some rational numbers r_1, r_2, r_3. In addition, by the Law of Sines,

$$\sin A : \sin B : \sin C = BC : CA : AB.$$

This establishes the desired result.

A928. Note that

$$\begin{aligned}
\frac{a+b+c+d}{4} &= \left(\frac{(a+b+c+d)^2}{4(a+b+c+d)}\right) \\
&= \frac{a^2+b^2+c^2+d^2+2ab+2ac+2ad+2bc+2bd+2cd}{4(a+b+c+d)} \\
&\geq \frac{\frac{1}{3}(2ab+2ac+2ad+2bc+2bd+2cd)+2ab+2ac+2ad+2bc+2bd+2cd}{4(a+b+c+d)} \\
&= \frac{2(ab+ac+ad+bc+bd+cd)}{3(a+b+c+d)} \\
&= \frac{a(b+c+d)+b(a+c+d)+c(a+b+d)+d(a+b+c)}{3(a+b+c+d)}.
\end{aligned}$$

To this last expression apply the weighted arithmetic-geometric mean inequality with weights

$$\frac{b+c+d}{3(a+b+c+d)}, \frac{a+c+d}{3(a+b+c+d)}, \frac{a+b+d}{3(a+b+c+d)}, \text{ and } \frac{a+b+c}{3(a+b+c+d)},$$

(which total to 1) to get

$$\geq a^{\frac{b+c+d}{3(a+b+c+d)}} b^{\frac{a+c+d}{3(a+b+c+d)}} c^{\frac{a+b+d}{3(a+b+c+d)}} d^{\frac{a+b+c}{3(a+b+c+d)}}.$$

This completes the proof.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles, books, and other materials are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of mathematics literature. Readers are invited to suggest items for review to the editors.

Polster, Burkard, What is the best way to lace your shoes?, *Nature* 420 (5 December 2002) 476.

Burkard Polster (Monash University, Australia) continues research of recent years on the mathematics—and here also the physics—of lacing shoes. The shoelace problem can be thought of as a restricted traveling sales problem. A *dense* lacing is one that "zigzags back and forth between the two columns of eyelets"; among such lacings, the traditional American criss-cross lacing is the shortest one. The relative distance between eyelets on one side of the shoe to the distance between the rows of eyelets determines which lacing is strongest (in terms of horizontal tension); for most shoes, the ratio is close to the boundary between criss-cross lacing and straight lacing, so either method is fine from that point of view. So why then do my son's shoes come untied every hour or so? Well, no matter how strong the lacing, the bow that most people tie results in an unstable granny knot; a simple change of habit—reversing the orientation of one of the half-knots in the bow—would give a much stronger square (reef) knot.

Pearson, Helen, Maths proves Tetris is tough, *Nature* (28 October 2002), `http://www.nature.com/nsu/021021/021021-9.html` . NP or not NP?, *The Economist* (31 October 2002), `\http://beta.economist.com/science/displayStory.cfm?story_id=1416184` . Demaine, E.D., S. Hohenberger, and D. Liben-Nowell, Tetris is hard, even to approximate, `http://www.arxiv.org/abs/cs.CC/0210020` .

In the popular computer game of Tetris, invented by Alexey Pazhitnov in 1984, the computer generates one tetromino after another and the player tries to fit them into a rectangular gameboard. Demaine et al. show that for several different goals of the game and variations of the rules, playing the game optimally is NP-complete and "highly inapproximable." Their proof is for the easier "offline" game, in which the player knows in advance the order of all of the pieces that are to appear. The authors reduce Tetris to the 3-partition problem, which is known to be NP-complete.

Grossman, Jerrold W., Patterns of collaboration in mathematical research, *SIAM News* 35 (9) (November 2002) 1, 8–9.

Grossman analyzes the co-authorship (collaboration) graph of the corpus of *Mathematical Reviews* 1940–1999. A few highlights are: Almost half of the authors are authors of just a single paper, the other authors are separated on average by a distance between 7 and 8 ("eight degrees of separation"), and the current mean number of papers per author per year is 0.5. Grossman also investigates the analogue of "Erdős number" for other mathematicians, finding that different mathematicians have different means but all have similar distributions ("people farther from the heart of the graph might take longer to get to the heart but, once there, have the same fan-out pattern").

Crilly, Rob, Scientists have secret of Christmas uncovered, *Glasgow Herald* (20 December 2002), http://www.theherald.co.uk/news/archive/20-12-19102-22-48-33.html . Carey, Tanith, Wrapper's formula for success, *The Mirror* (UK), (19 December 2002), http//www.mirror.co.uk/news/allnews/page.cfm?objectid=12466166&method=full&siteid=50143 . Wrapping it all up!, *The Daily Record* (Scotland), (20 December 2002), http://www.dailyrecord.co.uk/news/page.cfm?object=12469007&method=full&site=89488 .

This news reaches you too late for holiday presents in 2002, but there's always the next birthday occasion. Mark Chaplain (Dundee University) has formulated the minimal area to wrap a rectangular present with perfect overlap and no waste: $(2L + 2H + X)(B + 2H)$, where L, B, H are the length, breadth, and height of the box and X is the overlap of paper on the top. Presumably, the piece of paper is $(2L + 2H + X)$ by $(B + 2H)$; but something must have gotten lost in translation from the Scottish into these news articles. The wrapping algorithm is not specified, and the roles in the formula of L, B, and H are asymmetric—so the amount of paper used depends on which side up the package is wrapped. Is it true that orienting the box so that $H \leq B \leq L$ always minimizes area?

Mackenzie, Dana, The Stanford flip: A magician turned mathematician saves the casinos' shirts, *Discover* 23 (10) (October 2002) 22–23. Klarreich, Erica, Coming up trumps, *New Scientist* (20 July 2002) 42–44.

A manufacturer of casino equipment consulted Persi Diaconis and Susan Holmes (Stanford University) about its new card-shuffling machine. Their analysis showed that the "shelf shuffler," far from producing random decks, would allow a knowledgeable player to guess 9 of 52 cards correctly! Their advice: Feed the cards through twice. There are surprising connections to noncommutative geometry in physics and to genetic drift in biology.

Hull, Thomas (ed.), *Origami^3: Third International Meeting of Origami Science, Mathematics, and Education*, A K Peters, 2002; xi + 353 pp, $49 (P). ISBN 1–56881–181–0.

An ancient pastime, origami has become the focus of mathematical inquiry only in the past few years. Origami can be investigated on many levels: as a geometric construction device (more powerful than straightedge and compasses), in terms of decision procedures (will a given crease pattern fold flat? can a particular shape be made?), and as an application of mathematics (to suggest and develop new origami designs). All of these topics, plus the use of origami in mathematics education, are treated in the essays in this book.

Livio, Mario, *The Golden Ratio: The Story of Phi, the World's Most Astonishing Number*, Broadway Books, 2002; ix + 294 pp, $24.95. ISBN 0–7679–0815–5.

Astronomer Livio has written an exciting and balanced account about $\phi = (1 + \sqrt{5})/2$. Despite the hyperbole of the title, he does not see ϕ everywhere; he carefully weighs the evidence regarding claims of the appearance of ϕ in the Great Pyramid, in the Parthenon, and in the works of various artists. The narrative proceeds historically without bogging down in equations; various proofs and mathematical elaborations appear in 10 one-page appendices. This book, which ends with a discussion of the nature of mathematics (discovered or invented?), is a great example of popularization of mathematical ideas to inspire the public to see and appreciate mathematics in the world.

Nowakowski, RIchard J., *More Games of No Chance*, Cambridge University Press, 2002; xii + 535 pp, $55. ISBN 0–521–80832–4.

A successor to *Games of No Chance* (1996), this book contains papers from a conference on combinatorial games. Topics vary from tic-tac-toe on a hypercube to group theory of partisan games, from coin-moving puzzles to "Go endgames are PSPACE-hard," plus articles on the "new classics" among games. The chapter of unsolved problems and the bibliography of over 900 items bring the reader to the frontier of research. Alas, no index.

63nd Annual William Lowell Putnam Mathematical Competition

Editor's Note: Additional solutions to Putnam problems will be printing in the *American Mathematical Monthly* later in the year.

Problems

A1 Let k be a fixed positive integer. The nth derivative of $\frac{1}{x^k-1}$ has the form $\frac{P_n(x)}{(x^k-1)^{n+1}}$ where $P_n(x)$ is a polynomial. Find $P_n(1)$.

A2 Given any five points on a sphere, show that some four of them must lie on a closed hemisphere.

A3 Let $n \geq 2$ be an integer and T_n be the number of non-empty subsets S of $\{1, 2, 3, \ldots, n\}$ with the property that the average of the elements of S is an integer. Prove that $T_n - n$ is always even.

A4 In Determinant Tic-Tac-Toe, Player 1 enters a 1 in an empty 3×3 matrix. Player 0 counters with a 0 in a vacant position, and play continues in turn until the 3×3 matrix is completed with five 1's and four 0's. Player 0 wins if the determinant is 0 and player 1 wins otherwise. Assuming both players pursue optimal strategies, who will win and how?

A5 Define a sequence by $a_0 = 1$, together with the rules $a_{2n+1} = a_n$ and $a_{2n+2} = a_n + a_{n+1}$ for each integer $n \geq 0$. Prove that every positive rational number appears in the set

$$\left\{\frac{a_{n-1}}{a_n} : n \geq 1\right\} = \left\{\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{2}, \cdots\right\}.$$

A6 Fix an integer $b \geq 2$. Let $f(1) = 1$, $f(2) = 2$, and for each $n \geq 3$, define $f(n) = nf(d)$, where d is the number of base-b digits of n. For which values of b does

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

converge?

B1 Shanille O'Keal shoots free throws on a basketball court. She hits the first and misses the second, and thereafter the probability that she hits the next shot is equal to the proportion of shots she has hit so far. What is the probability she hits exactly 50 of her first 100 shots?

B2 Consider a polyhedron with at least five faces such that exactly three edges emerge from each of its vertices. Two players play the following game:

> Each player, in turn, signs his or her name on a previously unsigned face. The winner is the player who first succeeds in signing three faces that share a common vertex.

Show that the player who signs first will always win by playing as well as possible.

B3 Show that, for all integers $n > 1$,

$$\frac{1}{2ne} < \frac{1}{e} - \left(1 - \frac{1}{n}\right)^n < \frac{1}{ne}.$$

B4 An integer n, unknown to you, has been randomly chosen in the interval [1, 2002] with uniform probability. Your objective is to pick n in an **odd** number of guesses. After each incorrect guess, you are informed whether n is higher or lower, and you **must** guess an integer on your next turn among the numbers that are still feasibly correct. Show that you have a strategy so that the chance of winning is greater than 2/3.

B5 A palindrome in base b is a positive integer whose base-b digits read the same backwards and forwards; for example, 2002 is a 4-digit palindrome in base 10. Note that 200 is not a palindrome in base 10, but it is the 3-digit palindrome 242 in base 9, and 404 in base 7. Prove that there is an integer which is a 3-digit palindrome in base b for at least 2002 different values of b.

B6 Let p be a prime number. Prove that the determinant of the matrix

$$\begin{pmatrix} x & y & z \\ x^p & y^p & z^p \\ x^{p^2} & y^{p^2} & z^{p^2} \end{pmatrix}$$

is congruent modulo p to a product of polynomials of the form $ax + by + cz$, where a, b, c are integers. (We say two integer polynomials are congruent modulo p if corresponding coefficients are congruent modulo p.)

Solutions

Solution to A1 Differentiating gives $P_{n+1}(x) = (x^k - 1)P_n'(x) - (n+1)kx^{k-1}P_n(x)$ so that $P_{n+1}(1) = -k(n+1)P_n(1)$. Therefore as $P_0(1) = 1$, by induction, $P_n(1) = (-k)^n n!$

Solution to A2 Consider a great circle through any pair of points. The sphere is split by it into two closed hemispheres overlapping on this circle. Of the three remaining points, at least two must live in one of these hemispheres. That hemisphere has at least 4 points.

Solution to A3 The sets counted by T_n include $\{1\}, \{2\}, \ldots, \{n\}$, and the rest may be separated into two classes: those that contain their average and those that do not. The latter two classes are in one-to-one correspondence obtained by mapping a set A in the first class to the set A', obtained by deleting the average from A, in the second class, so $T_n \equiv n \pmod 2$.

Solution to A4 Player 0 wins. The determinant $a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$ can be made to equal 0 if each of the six products are 0. The first part of Player 0's strategy is to use the first two plays to make the two products affected by Player 1's initial play equal zero, Since there are two possible plays for each of the two products, Player 1 cannot prevent this. Since each space appears in one product with a plus sign and one with a minus sign, the remaining two product where 0 has yet to play are of opposite sign, and thus share a space. If this space is available for 0's third play, 0 plays there and wins. Otherwise, in 1's second and third turn, 1 played on the shared space, and perhaps one other space on these two products, as well. But 0 can use its third turn to block 1 from completing one of these products, and still have room to play on the remaining product on the final turn.

Solution to A5 We prove the result for A/B with $\gcd(A, B) = 1$ and $A, B \geq 1$, by using induction on $A + B$. For $A = B = 1$, $a_0/a_1 = 1/1$. If $A > B$, select n with $a_{n-1}/a_n = (A - B)/B$, so that

$$\frac{a_{2n}}{a_{2n+1}} = \frac{a_n + a_{n-1}}{a_n} = 1 + \frac{(A - B)}{B} = \frac{A}{B}.$$

If $A < B$, select n with $a_{n-1}/a_n = A/(B - A)$, so that

$$\frac{a_{2n-1}}{a_{2n}} = \frac{a_{n-1}}{a_{n-1} + a_n} = \frac{A}{B}.$$

Solution to A6 Note that

$$\sum_{b^{d-1} \leq n < b^d} \frac{1}{f(n)} = \frac{1}{f(d)} \sum_{b^{d-1} \leq n < b^d} \frac{1}{n}.$$

In fact since

$$\sum_{n=A}^{B-1} \frac{1}{n} > \int_A^B \frac{dt}{t} = \log\left(\frac{B}{A}\right),$$

we get

$$\sum_{b^{d-1} \leq n < b^d} \frac{1}{f(n)} > \frac{\log b}{f(d)}.$$

Now, summing over all $d \geq 1$, we get

$$\sum_{n \geq 1} \frac{1}{f(n)} = \sum_{d \geq 1} \sum_{b^{d-1} \leq n < b^d} \frac{1}{f(n)} > \log b \sum_{d \geq 1} \frac{1}{f(d)}.$$

Since $\log 3 > 1$, this last inequality is nonsense for $b \geq 3$, unless $\sum_{n \geq 1}(1/f(n))$ diverges.

For $b = 2$, we note that for $d \geq 2$,

$$\sum_{2^{d-1} \leq n < 2^d} \frac{1}{n} \leq \frac{2^{d-2}}{2^{d-1}} + \frac{2^{d-2}}{2^{d-1} + 2^{d-2}} = \frac{1}{2} + \frac{1}{3} = \frac{5}{6},$$

so that

$$\sum_{2^{d-1}\le n<2^d} \frac{1}{f(d)} = \frac{1}{f(d)} \sum_{2^{d-1}\le n<2^d} \frac{1}{n} \le \frac{5}{6}\cdot\frac{1}{f(2)}.$$

Let $a_1 = 4$ and inductively let $a_{k+1} = 2^{a_k-1}$. Then it follows by induction that

$$\sum_{a_k\le n\le a_{k+1}} \frac{1}{f(n)} \le \left(\frac{5}{6}\right)^k \frac{1}{f(3)},$$

so that

$$\sum_{n\ge 4} \frac{1}{f(n)} \le \frac{1}{f(3)} \sum_{k\ge 1} \left(\frac{5}{6}\right)^k = \frac{5}{f(3)},$$

and thus the infinite sum converges.

Solution to B1 Let $p(m, n)$ be the probability that Shanille O'Keal hits m and misses n out of the first $m + n$ shots. The problem stipulates that $p(m, 0) = p(0, n) = 0$ and $p(1, 1) = 1$. We now prove by induction on $m + n$ that $p(m, n) = 1/(m + n - 1)$ for all $m, n \ge 1$. We have

$$\begin{aligned} p(m,n) &= \frac{m-1}{(m-1)+n}\cdot p(m-1,n) + \frac{n-1}{m+(n-1)}\cdot p(m,n-1) \\ &= \frac{(m-1)+(n-1)}{m+n-1}\cdot\frac{1}{m+n-2} \qquad \text{(by induction when } n>1\text{)} \\ &= \frac{1}{m+n-1}. \end{aligned}$$

Solution to B2 The statement of the problem should have included the stipulation that no two faces on the polyhedron meet in multiple edges. Call the players, player A and player B. Player A signs the face with the most edges. If that face has ≥ 4 edges, no matter where B signs, A can next sign a face, adjacent to his first face, but not adjacent to the face B signed. But then there are two unsigned faces adjacent to the two faces that A has already signed and no matter which B signs, A has a winning face to sign. Thus we may assume every face has exactly 3 edges.

Consider a vertex and the three edges that emerge from it, say to vertices a, b, and c. Any two of these edges lie on a common face and, since each face has exactly 3 edges, there must be an edge between each pair of vertices amongst a, b, and c. We now have a tetrahedron and no further edges may emerge from these vertices, since they each have degree 3. Thus the solid has only four faces contradicting the hypothesis.

Solution to B3 The right-hand inequality is

$$\left(1-\frac{1}{n}\right)^n > \frac{1}{e}\left(1-\frac{1}{n}\right).$$

Replacing n by $1/x$, it suffices to show that $(1-x)^{\frac{1}{x}-1} > 1/e$ for $0 < x < 1$. Taking logs, we must show that $f(x) := (1-x)\log(1-x) + x > 0$ on $(0, 1)$. This holds since $f(0) = 0$ and $f'(x) = -\log(1-x) > 0$ on $(0, 1)$. Similarly, the left-hand inequality is equivalent to

$$f(x) := -x + x\log\left(1-\frac{x}{2}\right) - \log(1-x) > 0$$

on (0, 1). Here

$$f(0) = 0, \quad f'(0) = 0, \quad f''(x) = \frac{x(x^2 - 5x + 5)}{(x-2)^2(x-1)^2} > 0$$

for $0 < x < 1$. Hence $f(x) > 0$ on (0, 1).

Solution to B4 One strategy goes as follows: For $k = 0, 1, 2, \ldots$ in that order, either the chosen number n was $\leq 3k$ and has already been discovered or else you name $3k + 1$ as the first of two guesses. This is an odd-numbered guess, so you win if $3k + 1 = n$. Otherwise, $3k + 1$ will be pronounced too small, and you guess $3k + 3$. If this is n, you lose; but if it is pronounced too large, you win by guessing $3k + 2$, and if it is pronounced too small, proceed to the case $k + 1$. Thus you win if $n \equiv 1$ or 2 mod 3. Since $2002 \equiv 1 \bmod 3$ this accounts for $2 \cdot 667 + 1 = 1335$ integers up to 2002 and

$$\frac{1335}{2002} = \frac{2 \cdot 667 + 1}{3 \cdot 667 + 1} > \frac{2}{3}.$$

Solution to B5 Let's try for palindromes of the form $c\ 2c\ c$. Such a palindromic n is $cb^2 + 2cb + c = c(b+1)^2$, where $1 \leq c < b/2$. Choosing $n = 2^k$, $c = 2^{k-2r}$, and $b = 2^r - 1$, we only need $0 \leq k - 2r \leq r - 2$, that is, $2r \leq k \leq 3r - 2$, or $(k+2)/3 \leq r \leq k/2$. Selecting $k = 12010$, we get 2002 values of r.

Solution to B6 Denote the determinant by $D(x, y, z)$, which we consider as a polynomial in $(\mathbf{Z}/p\mathbf{Z})[x, y, z]$. For any $a, b, c \bmod p$ we have $ax^p + by^p + cz^p \equiv (ax + by + cz)^p \bmod p$, and similarly (by raising both sides of this congruence to the pth power),

$$ax^{p^2} + by^{p^2} + cz^{p^2} \equiv (ax + by + cz)^{p^2} \bmod p.$$

We use the following.

LEMMA. If not all of a, b, c are $0 \bmod p$, then $ax + by + cz$ is a factor of $D(x, y, z)$.

Assume the lemma for now. The number of distinct factors $ax + by + cz$ (that is, no two being scalar multiples of one another) is $(p^3 - 1)/(p - 1) = p^2 + p + 1$, the total degree of $D(x, y, z)$, and therefore $D(x, y, z)$ is the product of all (distinct) such factors times some constant.

Proof of the lemma: Without essential loss of generality, assume that a is not $0 \bmod p$. By multiplying through by the inverse of $a \bmod p$, we may assume that $a = 1$. Let $u = x + by + cz$. By legal column operations we have

$$D(x, y, z) = \det \begin{pmatrix} x + by + cz & y & z \\ x^p + by^p + cz^p & y^p & z^p \\ x^{p^2} + by^{p^2} + cz^{p^2} & y^{p^2} & z^{p^2} \end{pmatrix} = \det \begin{pmatrix} u & y & z \\ u^p & y^p & z^p \\ u^{p^2} & y^{p^2} & z^{p^2} \end{pmatrix}.$$

But u clearly divides this last determinant, so that $x + by + cz$ divides $D(x, y, z)$.

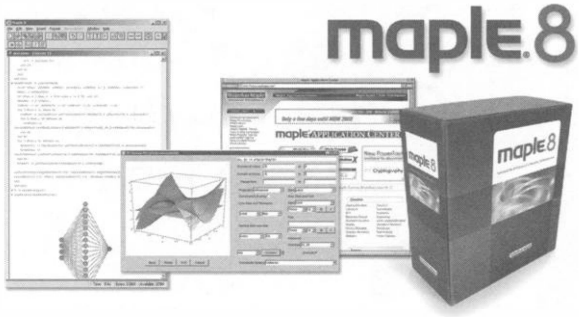
Acknowledgment. Thanks to Byron Walden for editorial assistance with these solutions.

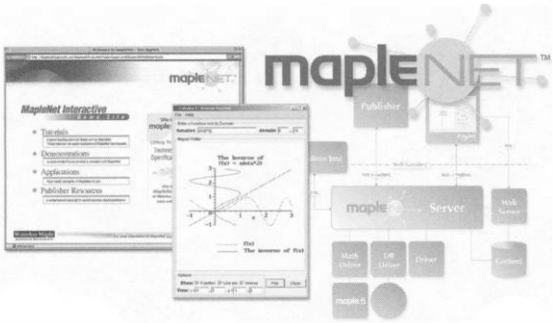
Distance Asynchronous Online On Campus

Education is no longer just about classrooms and labs. With the growing diversity and complexity of educational programs, you need a software system that lets you efficiently deliver effective learning tools to literally, the world. Maple® now offers you a choice to address the reality of today's mathematics education.

Maple® 8 – the standard

Perfect for students in mathematics, sciences, and engineering. Maple® 8 offers all the power, flexibility, and resources your technical students need to manage even the most complex mathematical concepts.





MapleNet™ – online education

A complete standards-based solution for authoring, delivering, and managing interactive learning modules through Web browsers. Derived from the legendary Maple® engine, MapleNet™ is the only comprehensive solution for distance education in mathematics.

Give your institution and your students the competitive edge.

For a **FREE 30-day Maple® 8 Trial CD** for Windows®, or to register for a **FREE MapleNet™ Online Seminar** call ***1/800 267.6583*** or e-mail ***info@maplesoft.com***.

Waterloo Maple
ADVANCING MATHEMATICS

WWW.MAPLESOFT.COM | INFO@MAPLESOFT.COM | WWW.MAPLEAPPS.COM | NORTH AMERICAN SALES 1/800 267.6583

CONTENTS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036