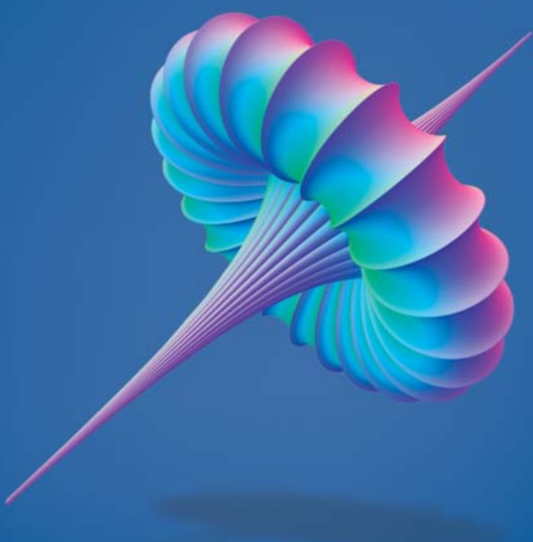
THE GREAT COURSES®

Topic
Science & Mathematics

Subtopic
Mathematics

Change and Motion: Calculus Made Clear, 2nd Edition

Professor Michael Starbird
The University of Texas at Austin



Transcript Book

PUBLISHED BY:

THE GREAT COURSES
Corporate Headquarters
4840 Westfields Boulevard, Suite 500
Chantilly, Virginia 20151-2299
Phone: 1-800-832-2412
Fax: 703-378-3819
www.thegreatcourses.com

Printed in the United States of America

Michael Starbird, Ph.D.

University Distinguished Teaching Professor of Mathematics
The University of Texas at Austin

Professor Michael Starbird is Professor of Mathematics and a University Distinguished Teaching Professor at The University of Texas at Austin. He received his B.A. degree from Pomona College in 1970 and his Ph.D. in mathematics from the University of Wisconsin, Madison, in 1974. That same year, he joined the faculty of the Department of Mathematics of The University of Texas at Austin, where he has stayed, except for leaves as a Visiting Member of the Institute for Advanced Study in Princeton, New Jersey; a Visiting Associate Professor at the University of California, San Diego; and a member of the technical staff at the Jet Propulsion Laboratory in Pasadena, California.

Professor Starbird served as Associate Dean in the College of Natural Sciences at The University of Texas at Austin from 1989 to 1997. He is a member of the Academy of Distinguished Teachers at UT. He has won many teaching awards, including the Mathematical Association of America's Deborah and Franklin Tepper Haimo Award for Distinguished College or University Teaching of Mathematics, which is awarded to three professors annually from among the 27,000 members of the MAA; a Minnie Stevens Piper Professorship, which is awarded each year to 10 professors from any subject at any college or university in the state of Texas; the inaugural award of the Dad's Association Centennial Teaching Fellowship; the Excellence Award from the Eyes of Texas, twice; the President's Associates Teaching Excellence Award; the Jean Holloway Award for Teaching Excellence, which is the oldest teaching award at UT and is presented to one professor each year; the Chad Oliver Plan II Teaching Award, which is student-selected and awarded each year to one professor in the Plan II liberal arts honors program; and the Friar Society Centennial Teaching Fellowship, which is awarded to one professor at UT annually and includes the largest monetary teaching prize given at UT. Also, in 1989, Professor Starbird was the Recreational Sports Super Racquets Champion.

The professor's mathematical research is in the field of topology. He recently served as a member-at-large of the Council of the American Mathematical Society and on the national education committees of both the American Mathematical Society and the Mathematical Association of America.

Professor Starbird is interested in bringing authentic understanding of significant ideas in mathematics to people who are not necessarily mathematically oriented. He has developed and taught an acclaimed class that presents higher-level mathematics to liberal arts students. He wrote, with co-author Edward B. Burger, *The Heart of Mathematics: An invitation to effective thinking*, which won a 2001 Robert W. Hamilton Book Award. Professors Burger and Starbird have also written a book that brings intriguing mathematical ideas to the public, entitled *Coincidences, Chaos, and All That Math Jazz: Making Light of Weighty Ideas*, published by W.W. Norton, 2005. Professor Starbird has produced three previous courses for The Teaching Company, the first edition of *Change and Motion: Calculus Made Clear*; *Meaning from Data: Statistics Made Clear*; and with collaborator Edward Burger, *The Joy of Thinking: The Beauty and Power of Classical Mathematical Ideas*. Professor Starbird loves to see real people find the intrigue and fascination that mathematics can bring. ■

Table of Contents

INTRODUCTION

LECTURE GUIDES

Table of Contents

Table of Contents

Acknowledgments

I want to thank Alex Pekker for his excellent help with every aspect of this second edition of the calculus course. Alex collaborated with me substantially on the design of the whole course, on the examples and flow of the individual lectures, on the design of the graphics, and on the written materials. Thanks also to Professor Katherine Socha for her work on the first edition of this course and for her help during the post-production process of the second edition. Thanks to Alisha Reay, Pam Greer, Lucinda Robb, Noreen Nelson, and others from The Teaching Company not only for their excellent professional work during the production of this series of lectures but also for creating a supportive and enjoyable atmosphere in which to work. Thanks to my wife, Roberta Starbird, for her design and construction of several of the props. Finally, thanks to Roberta and my children, Talley and Bryn, for their special encouragement. ■

Change and Motion: Calculus Made Clear, 2nd Edition

Scope:

Twenty-five hundred years ago, the Greek philosopher Zeno watched an arrow speeding toward its target and framed one of the most productive paradoxes in the history of human thought. He posed the paradox of motion: Namely, at every moment, the arrow is in only one place, yet it moves. This paradox evokes questions about the infinite divisibility of position and time. Two millennia later, Zeno's paradox was resolved with the invention of calculus, one of the triumphs of the human intellect.

Calculus has been one of the most influential ideas in human history. Its impact on our daily lives is incalculable, even with calculus. Economics, population growth, traffic flow, money matters, electricity, baseball, cosmology, and many other topics are modeled and explained using the ideas and the language of calculus. Calculus is also a fascinating intellectual adventure that allows us to see our world differently.

The deep concepts of calculus can be understood without the technical background traditionally required in calculus courses. Indeed, frequently, the technicalities in calculus courses completely submerge the striking, salient insights that compose the true significance of the subject. The concepts and insights at the heart of calculus are absolutely meaningful, understandable, and accessible to all intelligent people—regardless of the level or age of their previous mathematical experience.

Calculus is the exploration of two ideas, both of which arise from a clear, commonsensical analysis of our everyday experience of motion: the *derivative* and the *integral*. After an introduction, the course begins with a discussion of a car driving down a road. As we discuss velocity and position, these two foundational concepts of calculus arise naturally, and their relationship to each other becomes clear and convincing. Calculus directly describes and deals with motion. But the ideas developed there also present us with a dynamic view of the world based on a clear analysis of change. That perspective lets us view even such static objects as circles in a dynamic way—growing by accretion of infinitely thin layers. The pervasive nature of change makes calculus extremely widely applicable.

The course proceeds by exploring the rich variations and applications of the two fundamental ideas of calculus. After the introduction in the setting of motion, we proceed to develop the concepts of calculus from several points of view. We see the ideas geometrically and graphically. We interpret calculus ideas in terms of familiar formulas for areas and volumes. We see how the ideas developed in the simple setting of a car moving in a straight line can be extended to apply to motion in space. Among the many variations of the concepts of calculus, we see how calculus describes the contours of mountains and other three-dimensional objects. Finally, we explore the use of calculus in describing the physical, biological, and even architectural worlds.

One of the bases for the power of calculus lies in the fact that many questions in many subjects are equivalent when viewed at the appropriate level of abstraction. That is, the mathematical structures that one creates to study and model motion are identical, mathematically, to the structures that model phenomena from biology to economics, from traffic flow to cosmology. By looking at the mathematics itself, we strip away the extraneous features of the questions and focus on the underlying relationships and structures that govern the behavior of the system in question. Calculus is the mathematical structure that lies at the core of a world of seemingly unrelated issues.

It is in the language of calculus that scientists describe what we know of physical reality and how we express that knowledge. The language of calculus contains its share of mathematical symbols and terminology. However, we will see that every calculus idea and symbol can be understood in English, not requiring "mathese." We will not eschew formulas altogether, but we will make clear that every equation is an English sentence that has a meaning in English, and we will deal with that meaning in English. Indeed, one of the principal goals of this series of lectures is to have viewers understand the concepts of calculus as meaningful ideas, not as the manipulation of meaningless symbols.

Our daily experience of life at the beginning of the third millennium contrasts markedly with life in the 17^{th} century. Most of the differences emerged from technical advances that rely on calculus. We live differently now because we can manipulate and control nature better than we could 300 years ago. That practical, predictive understanding of the physical processes of nature is largely enabled by the power and perspective of calculus. Calculus not only provides specific tools that solve practical problems, but it also entails an intellectual perspective on how we analyze the world.

Calculus is all around us and is a landmark achievement of humans that can be enjoyed and appreciated by all. ■

Two Ideas, Vast Implications

Lecture 1

It will be my great pleasure to guide you on a journey through a world of calculus during these 24 lectures. The tour begins at our doorsteps and takes us to the stars. Calculus is all around us every day of our lives.

Calculus is the exploration of two ideas that arise from a clear, commonsensical analysis of everyday experience. But explorations of these ideas—the **derivative** and the **integral**—help us construct the very foundation of what we know of physical reality and how we express that knowledge. Many questions in many subjects are equivalent when viewed at the appropriate level of abstraction. That is, the mathematical structures that one creates to study and model motion are identical, mathematically, to the structures that model aspects of economics, population growth, traffic flow, fluid flow, electricity, baseball, planetary motion, and countless other topics. By looking at the mathematics itself, we strip away the extraneous features of the questions and focus on the underlying relationships and structures that govern and describe our world. Calculus has been one of the most effective conceptual tools in human history.

Calculus is all around us. When we're driving down a road and see where we are and how fast we are going … that's calculus. When we throw a baseball and see where it lands … that's calculus. When we see the planets and how they orbit around the Sun … that's calculus. When we lament the decline in the population of the spotted owl … that's calculus. When we analyze the stock market … that's calculus.

Calculus is an idea of enormous importance and historical impact. Calculus has been extremely effective in allowing people to bend nature to human purpose. In the 20th century, calculus has also become an essential tool for understanding social and biological sciences: It occurs every day in the description of economic trends, population growth, and medical treatments. The physical world and how it works are described using calculus—its terms, its notation, its perspective. To understand the history of the last 300 years, we must understand calculus. The technological developments in recent centuries are the story of this time, and many of those developments depend on calculus. Why is calculus so effective? Because it resolves some basic issues associated with change and motion.

Twenty-five hundred years ago, the philosopher Zeno pointed out the paradoxical nature of motion. Zeno's paradoxes confront us with questions about motion. Calculus resolves these ancient conundrums. Two of Zeno's paradoxes of motion involve an arrow in flight. The first is the *arrow paradox*: If at every moment, the arrow is at a particular point, then at every moment, it is at rest at a point. The second is the *dichotomy paradox*: To reach its target, an arrow must first fly halfway, then half the remaining distance, then half the remaining distance, and so on, forever. Because it must move an infinite number of times, it will never reach the target. Looking at familiar occurrences afresh provokes insights and questions. Zeno's paradoxes have been extremely fruitful. Zeno's paradoxes bring up questions about infinity and instantaneous motion.

In the 20th century, calculus has also become an essential tool for understanding social and biological sciences: It occurs every day in the description of economic trends, population growth, and medical treatments.

In this course, we emphasize the ideas of calculus more than the mechanical side. But I must add that one of the reasons that calculus has been of such importance for these last 300 years is that it *can* be used in a mechanical way. It can be used by people who don't understand it. That's part of its power. Perhaps we think calculus is hard because the word *calculus* comes from the Greek word for *stones* (stones were used for reckoning in ancient times). Calculus does have a fearsome reputation for being very hard, and part of the goal of this course is to help you see calculus in a different light.

In describing his college entrance examinations in his autobiography, Sir Winston Churchill says, "Further dim chambers lighted by sullen, sulphurous fires were reputed to contain a dragon called the 'Differential Calculus.'

But this monster was beyond the bounds appointed by the Civil Service Commissioners who regulated this stage of Pilgrim's heavy journey." We will attempt to douse the dragon's fearsome fires.

Another reason calculus is considered so forbidding is the size of calculus textbooks. To students, a calculus book has 1,200 different pages. But to a professor, it has two ideas and lots of examples, applications, and variations.

Fortunately, the two fundamental ideas of calculus, called the derivative and the integral, come from everyday observations. Calculus does not require complicated notation or vocabulary. It can be understood in English. We will describe and define simply and understandably those two fundamental ideas in Lectures 2 and 3. Both ideas will come about from analyzing a car moving down a straight road and just thinking very clearly about that scenario. The viewer is not expected to have any sense whatsoever of the meanings of these ideas now. In fact, I hope these technical terms inspire, if anything, only a foreboding sense of impending terror. That sense will make the discovery that these ideas are commonsensical and even joyful, instead of terrifying, all the sweeter. The derivative deals with how fast things are changing (instantaneous change). The integral provides a dynamic view of the static world, showing fixed objects growing by accretion (the accumulation of small pieces). We can even view apparently static things dynamically. For example, we can view the area of a square or the volume of a cube dynamically by thinking of it as growing rather than just being at its final size. The derivative and the integral are connected by the **Fundamental Theorem of Calculus**, which we will discuss in Lecture 4. Both of the fundamental ideas of calculus arise from a straightforward discussion of a car driving down a road, but both are applicable in many other settings.

The history of calculus spans two and a half millennia. Pythagoras invented the Pythagorean Theorem in the 6th century B.C., and as we know, Zeno posed his paradoxes of motion in the 5th century B.C. In the 4th century B.C., Eudoxus developed the method of **exhaustion**, similar to the integral, to study volumes of objects. In about 300 B.C., Euclid invented the axiomatic method of geometry. In 225 B.C., Archimedes used calculus-like methods to find areas and volumes of geometric objects. For many centuries, other mathematicians developed ideas that were important prerequisites for the full development of calculus. In around 1600, Johannes Kepler and Galileo Galilei worked making mathematical formulas that described planetary motion. In 1629, **Pierre de Fermat** developed methods for finding *maxima* of values, a precursor to the idea of the derivative. In the 1630s, Bonaventura Cavalieri developed the "Method of Indivisibles," and later, **René Descartes** established the Cartesian Coordinate System, a connection between algebra and geometry. The two mathematicians whose names are associated with the invention or discovery of calculus are Isaac Newton and Gottfried Wilhelm von Leibniz. They independently developed calculus in the 1660s and 1670s.

From the time of the invention of calculus, other people contributed variations on the idea and developed applications of calculus in many areas of life. Johann and Jakob Bernoulli were two of eight Bernoullis who were involved in developing calculus. Leonhard Euler developed extensions of calculus, especially infinite series. Joseph Louis Lagrange worked on calculus variations, and Pierre Simon de Laplace worked on **partial differential equations** and applied calculus to **probability** theory. **Jean Baptiste Joseph Fourier** invented ways to approximate certain kinds of dependencies, and **Augustin Louis Cauchy** developed ideas about **infinite series** and tried to formalize the idea of **limit**. In the 1800s, **Georg Friedrich Bernhard Riemann** developed the modern definition of the integral, one of the two ideas of calculus. In the middle of the 1850s—about 185 years after Newton and Leibniz invented calculus—Karl Weierstrass formulated the rigorous definition of limit that we know today.

partial differential equation: An equation involving a function of several variables and its partial derivatives.

probability: The quantitative study of uncertainty.

limit: The result of an infinite process that converges to a single answer. Example: the sequence of numbers $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$ converges to the number 0.

Here is an overview of the lectures. In Lectures 2, 3, and 4, we will introduce the basic ideas of calculus in the context of a moving car and discuss the connection between those ideas. Then, we have a series of lectures describing the meaning of the derivative graphically, algebraically, and in many applications. Following that, we have a similar series of lectures showing the integral from graphical, algebraic, and application points of view. The last half of the course demonstrates the richness of these two ideas by showing examples of their extensions, variations, and applications.

The purpose of these lectures is to explain clearly the concepts of calculus and to convince the viewer that calculus can be understood from simple scenarios. Calculus is so effective because it deals with change and motion and allows us to view our world as dynamic rather than static. Calculus provides a tool for measuring change, whether it is change in position, change in temperature, change in demand, or change in population. Calculus is intrinsically intriguing and beautiful, as well as important. Calculus is a crowning intellectual achievement of humanity that all intelligent people can appreciate, understand, and enjoy. ■

Names to Know

Cauchy, Augustin Louis (1789–1857). Prolific French mathematician and engineer. He was professor in the Ecole Polytechnique and professor of mathematical physics at Turin. He worked in number theory, algebra, astronomy, mechanics, optics, and elasticity theory and made great contributions to analysis (particularly the study of infinite series and of complex variable theory) and the calculus of variations. He improved the foundations of calculus by refining the definitions of limit and continuity.

Descartes, René (1596–1650). French mathematician and philosopher. He served in various military campaigns and tutored Princess Elizabeth (daughter of Frederick V) and Queen Christina of Sweden. Descartes developed crucial theoretical links between algebra and geometry and his own method of constructing tangents to curves. He made substantial contributions to the development of analytic geometry.

Fermat, Pierre de (1601–1665). French lawyer and judge in Toulouse; enormously talented amateur mathematician. Fermat worked in number theory, geometry, analysis, and algebra and was the first developer of analytic geometry, including the discovery of equations of lines, circles, ellipses, parabolas, and hyperbolas. He wrote *Introduction to Plane and Solid Loci* and formulated the famed *Fermat's Last Theorem* as a note in the margin of his copy of Bachet's *Diophantus*. He developed a procedure for finding maxima and minima of functions through infinitesimal analysis, essentially by the limit definition of derivative, and applied this technique to many problems, including analyzing the refraction of light.

Fourier, Jean Baptiste Joseph (1768–1830). French mathematical physicist and professor in the Ecole Polytechnique. Fourier accompanied Napoleon on his campaign to Egypt, was appointed secretary of Napoleon's Institute of Egypt, and served as prefect of Grenoble. He carried out extensive studies in heat propagation, which form the foundation of modern partial differential equations with boundary conditions. He developed the **Fourier Series**, which represents functions by infinite (trigonometric) series.

Riemann, Georg Friedrich Bernhard (1826–1866). German mathematician; professor of mathematics at Göttingen. He made great contributions to analysis, geometry, and number theory and both extended the theory of representing a function by its Fourier series and established the foundations of complex variable theory. Riemann developed the concept and theory of the Riemann integral (as taught in standard college calculus courses) and pioneered the study of the theory of functions of a real variable. He gave the most famous job talk in the history of mathematics, in which he provided a mathematical generalization of all known geometries, a field now called Riemannian geometry.

Important Terms

derivative: Mathematical description of (instantaneous) rate of change of a function. Characterized geometrically as the slope of the tangent line to the graph of the function. The derivative of a function $f(x)$ is written $f'(x)$ or $\frac{d}{dx}(f(x))$ and is formally defined as $\lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$.

exhaustion (the Greek method of exhaustion): A geometric technique by which formulas for areas of different shapes could be verified through finer and finer approximations.

Fourier Series: An infinite series of sines and cosines, typically used to approximate a function.

Fundamental Theorem of Calculus: The most important theorem in calculus. Demonstrates the reciprocal relationship between the derivative and the integral.

infinite series (also called an infinite sum): The sum of an infinite collection of numbers. Such a series can sum to a finite number or "diverge to infinity"; for example, $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\ldots$ sums to 1, but $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots$ does not sum to a finite number.

integral: Denoted $\int_a^b v(t)dt$. If we think of function $v(t)$ as measuring the velocity of a moving car at each time t, then the integral is a number that is equal to the distance traveled, because the integral is obtained by dividing the time from to b into small increments and approximating the distance traveled by assuming that the car went at a steady speed during each of those small increments of time. By taking increasingly smaller increments of time, approximations converge to a single answer, the integral. This sounds complicated, but the naturality of it is the topic of Lecture 3. The integral is also equal to the area under the graph of $v(t)$ and above the t-axis. The integral is related to the derivative (as an inverse procedure) via the Fundamental Theorem of Calculus.

parabola: A conic section defined as the set of all points equidistant between a point and a line.

paradox: Two compelling arguments about the same situation that lead to two opposite views. Zeno's paradoxes of motion give logical reasons why motion cannot occur. On the other hand, we experience motion. The opposite conclusions deduced from Zeno's logic versus our experience compose the paradox.

Suggested Reading

Any standard calculus textbook.

Cajori, Florian. "History of Zeno's Arguments on Motion," *The American Mathematical Monthly*, Vol. 22, Nos. 1–9 (1915).

Churchill, Winston Spencer. *My Early Life: A Roving Commission*.

Questions to Consider

1. Find things or ideas in your world that you usually view as complete and fixed and think about them dynamically. That is, view their current state as the result of a growing or changing process.

2. Explore the idea of function by describing dependent relationships between varying quantities that you see in everyday life. For example, in what way is the amount of money in a savings account dependent on interest rate and time? Or, for a less quantitative example, how is happiness a function of intellectual stimulation, exercise, rest, and other variables?

3. Think of a scenario from your daily life that interests you. Keep it in mind as you progress through the lectures—can calculus be applied to understand and analyze it?

Two Ideas, Vast Implications

Lecture 1—Transcript

Welcome to calculus. It will be my great pleasure to guide you on a journey through a world of calculus during these 24 lectures. The tour begins at our doorsteps and takes us to the stars. Calculus is all around us every day of our lives. When we're driving down the road and we see where we are at every moment and we figure out how fast we're going, that's calculus. When we throw a baseball and see where it lands, that's calculus. When we see the planets and see how they orbit around the sun—calculus. When NASA sends a rocket ship to explore the solar system, and they make all these computations about the trajectories and how much fuel to burn in which direction, all of these involve calculus. When we turn on our TVs—calculus.

But, calculus is not restricted only to physical issues. When we lament the decline in the population of the spotted owl, that's calculus. When we analyze the stock market and we look at economic trends, that's calculus. Calculus comprises a collection of ideas that have had tremendous historical impact. And the reason is that calculus is enormously effective in allowing people to bend nature to human purpose. Much of the scientific description of our world is based on calculus; descriptions of motion, certainly; of electricity and magnetism; of sound waves or waterways—all of these involve calculus. But in addition to that, calculus is an essential tool for understanding social and biological sciences. It occurs every day when we describe economic trends, when we talk about population growth or decline, or medical treatments; all of the description are couched in terms of calculus. That is, its vocabulary, its notation, but most important, more important than any of those, are its ideas, its perspectives.

So, to understand the history of the last 300 years, we certainly must understand calculus. What is it that makes us experience life differently now from the way people understood it and experienced it 300 years ago? Well, most of the difference comes from technological developments, and most of those technological developments are the story of calculus; that is, the implications of calculus that allowed us to develop the technology that changed the way we experienced life. It would be extremely difficult to exaggerate the impact of calculus on our life experience.

So, that brings us to the question of why? Why is calculus so effective? And the answer is calculus resolves some basic issues associated with change and motion, just basic every day motion. If you think about every day motion, it turns out that that idea, just thinking about things moving, is a concept that is much trickier than we might at first realize. In fact, this was pointed out 2,500 years ago. The philosopher Zeno, in 450 B.C., proposed several paradoxes about motion, and let me first just say a word or two about what a paradox is. A paradox is an example where you look at some issue from two different points of view, and the two different points of view are both extremely persuasive. For each one you say, "That's the way to look at the thing." And then you discover that they are, in direct opposition to each other. So, when you have a paradox, it is pointing the way to a fundamental issue that needs to be resolved.

Well, Zeno's paradoxes were paradoxes about motion, and he forced us to confront some questions about basic motion that were not resolved for more than 2,000 years, and they were resolved by calculus. So, let me tell you about a couple of Zeno's paradoxes of motion. Two of them involve an arrow in flight. So, here was the idea of Zeno's paradox.

The first paradox was the following: you have an arrow shooting across the room, zooming across the room, and Zeno asked the following question, he said, "Okay. We see that the arrow is moving across the room, but let's ask the question, at one instant is it moving? At one instant it's in one place. So, in what sense is it, at that very instant, moving? This was a real puzzle. This was a real puzzle about motion. And, if you think about it, motion is one of these things that becomes more difficult to understand the more you think about it because of the nature of Zeno's paradox of the arrow. At every moment it is not moving; it's in one place. That was a big challenge for people historically, and it was not resolved until calculus.

Another example of Zeno's paradox was his paradox called the dichotomy paradox, and this was another paradox involving an arrow. This time somebody is standing over there and shoots an arrow right at me, and the arrow is coming at me. But, if I'm Zeno, I'm very calm. I'm calm because I think the following thoughts, I say, "Well, the arrow has to come halfway toward me, and then it has to come half of the remaining distance, and after that it has to come half of the remaining distance, and so on half and half and half." Therefore it has to accomplish infinitely many things; infinitely many traversals of distances. Consequently, it will never get to me and I can be perfectly sanguine in my future healthy by thinking about this paradox.

Well, of course, we know in both of these cases that reality is different. That is to say, we know the arrow is moving; we know the arrow will eventually strike its target. So, the question is, how can we resolve these issues that Zeno pointed out: instantaneous motion and the idea that there have to be infinite many traversals in order for the arrow to reach its target. Well, when we look at familiar issues, familiar occurrences, in new ways, we turn out to find all sorts of fruitful avenues for investigation. And in the case of Zeno's paradoxes, Zeno's paradoxes led to the calculus, among other things, which had tremendous impact, as we said, in many areas.

Zeno's paradoxes actually bring up issues about infinity, the case of infinitely many traversals before the arrow gets to the target and the question of infinity about motion. You see, the trouble with motion is that motion really refers to the basic idea that something in motion is at one place at one moment and in another place at another moment. Well, when you talk about instantaneous motion, you only have one moment to discuss. So, it's a question about dealing with zero divided by zero. The arrow has gone zero distance in zero amount of time, and that issue was a terrific conundrum that was not resolved until, in fact, not only—it was resolved by calculus, but it required further elaboration, even after calculus was invented, to really come to grips with an understanding of that. In fact, that took several hundred years after, or 200 years, after calculus was invented to really pin down that idea.

Well, in this course we're going to emphasize the ideas of calculus, the concepts of calculus, more than the mechanical side of calculus. But, I have to tell you that one of the main reasons that calculus has been of such importance during these last 300 years is that it can be used in a mechanical way. It can be used by people who don't really understand calculus. That's part of the power of calculus. And in fact, the word itself, the word *calculus*, refers to calculating; that is, calculating in this sort of mechanical way. And, by mechanical way, what I mean is this: We all learn how to multiply in elementary school, where we took two numbers and did something and we moved it down and we added things, and, well, most people, when they learn that mechanical way to do multiplication, they're learning a technique, but don't really understand why that technique leads to the answer that they're getting. So, it's in the same way that calculus has this calculating, mechanical quality to it, that a person can learn the mechanistic strategies by which calculus derives its implications without actually knowing what those, why they work, and what's involved in the conceptual background. And, I have to tell you, that my own students at The University of Texas, in my class, when they emerge from my class, they probably have that view of calculus, too, because when you take a calculus class, a lot of the time that you spend in the class is mastering these mechanical methods of doing the actual computations.

Well, in this set of lectures we're going to emphasize the ideas of calculus much more than the mechanical side. Actually, I think it's the mechanical side that gives calculus its reputation for being sort of a fearsome and maybe difficult and forbidding topic, but I have a theory about why calculus is often viewed as a very hard subject. Maybe it's because—and this is just a theory; I may be wrong—but I think it comes from the word. You see, calculus comes from the Greek word *calculi*, which is the Greek word meaning stones. See, stones were used for counting things in ancient Greek time. So, well, stones are hard, you see, and so that's why calculus—well, okay, maybe not. Okay, anyway. But, regardless of the reason, calculus certainly does have the reputation for being extremely hard and sort of fierce, and part of the goal of this course is to, of course, make it seen in a totally different light. Calculus can be our friend.

But, you know, I teach mathematics and I teach calculus and so I often hear about the fright and the pain associated with mathematics in general, or calculus in particular, like when I'm in an airplane and I sit next to somebody and I divulge that I teach mathematics or I teach calculus. Well, they first tend to move away; that's the first thing that happens. They've sort of been afraid of it. But, one of—a particularly delightful example of this representation of the fear and loathing of mathematics and calculus in general came about in Sir Winston Churchill's autobiography, where he was writing about his early life in his autobiography called *My Early Life, A Roving Commission.* And in Chapter Three of this, he talked about taking examinations which he had to pass in order to enter the military academy. He had trouble passing these examinations—which, of course, is always astounded me that Winston Churchill, of all people, would have trouble passing examinations. But, that's just the way it was. So, in this Chapter Three he talks about mathematics and how he had to learn mathematics to pass this examination. So, he says the following: "All my life, from time to time, I've had to get up disagreeable subjects at short notice; but I consider my triumph, moral and technical, was in learning mathematics in six months." So that was, first of all, was pretty good. He goes on to describe his recollections of these days. "When I look back upon those care-laded months, their prominent features rise from the abyss of memory. We were arrived in an *Alice in Wonderland* world at the portals of which stood a quadratic equation." Here's where he gets to calculus. He said "Further dim chambers, lighted by sullen, sulphurous fires, were reputed to contain a dragon called the 'Differential Calculus.' But this monster was beyond the bounds appointed by the Civil Service Commissioners who regulated this stage of pilgrim's heavy journey." Isn't this great? So, calculus has this reputation of being a real monster; a dragon with a fiery breath. But, anyway, we will attempt to quench these fearsome fires. By the way, I must say, reading this little excerpt from Churchill—Churchill, of course, is amazing, having written 52 volumes of history; he won the Nobel Prize in literature; so, it is delightful to read about Churchill, and particularly this book about his early life.

But I think there are other reasons why calculus is viewed as a difficult subject and sort of forbidding, and I have to admit that I think mathematicians will not win the prize for using psychology that calms and sooths the soul. I think we're not going to win that prize because I think another reason that calculus is so forbidding is the simple size of calculus textbooks. Now, I brought a sample here—this is a calculus textbook—and, first of all, it does have a lot of benefits, not only mental, but also physical because people—you learn to heft heavy weights and so you get strong in that way, too. But, this is a typical calculus textbook; it has more than 1,300 pages. And, of course, to a student—and they're very big pages and sort of forbidding—that, to a student, these 1,300 pages are all different, but to a professor it has two ideas and 1,298 pages of examples, applications, and variations. The range and the richness of the implications of those two ideas show the power of those two insights. The exploitation of those two fundamental ideas leads to this whole world of calculus. Fortunately for us, and for this course, those two fundamental concepts of calculus—which, by the way, are called the derivative and the integral—both of those ideas come from every day observations. Calculus does not require a complicated vocabulary or notation to understand it really authentically. It can be understood in English; and, for the most part, we will be talking, in this course, in English about the ideas of calculus.

Okay, so calculus is the study of these two basic ideas: the *derivative* and the *integral*. Now, we're going to describe them and we're going to clearly say them in an understandable way in Lectures Two and Three respectively; derivative Lecture Two and integral Lecture Three. These ideas both will come about by analyzing a very simple situation of a car driving down a straight road; and then thinking very clearly about the implication of that idea and how it is that we need to think about such questions as Zeno's paradox of instantaneous motion. What are we going to mean by that? Now, let me make absolutely clear, that I'm not expecting any viewer of this course to have any sense whatsoever about the meaning of these ideas. In fact, in a way, I sort of hope the technical terms "derivative" and "integral" inspire in you, maybe, something that's foreboding, maybe a sense of impending terror, because that way when you find out that, in fact, these ideas are accessible and enjoyable, it will be all the sweeter. So, you see, this is good.

Well, the derivative deals with instantaneous change—how fast things move and change; and the integral gives us a dynamic view of the world, even of the static world. It talks about the accumulation of little things that accumulate into a big thing, and that's what I love about calculus. That it can take even the—it gives us a dynamic view of the world. That even things such as the area of a square or the volume of a cube, all of these things that seem so static can be viewed as growing into the size of the cube or growing into the size of a ball, and looking at that perspective allows us to see the world in a different way.

Well, the derivative and the integral turn out to be connected to each other, and they're connected by the fundamental theorem of calculus; and, we'll talk about that in Lecture Four. But, as I said before, both of these ideas arise from the discussion of a car moving down a road, but the same ideas that we get from analyzing this car moving down the road are applicable in these many, many other settings. That's the strength and the power of mathematics in general, and of these two ideas, derivative and integral, in particular.

Now, we're not going to, in this course, we not going to take the shortcut of over-simplifying. Einstein had a quote on that; it went something like, "I want things presented to me as simply as possible, but not simpler." And in this course, we're going to present things that really give you the authentic understanding of the ideas that make up calculus.

But the two ideas that come from calculus that really make up the foundations of calculus, did not appear by magic. The history of calculus spans two and a half millennia. From the time of Zeno, it was more than 2,000 years that elapsed before calculus became fully developed; and during that time there was a steady progress of mathematical ideas that needed to be created before we can understood calculus the way we know it now. So, I thought that we might take a few minutes here and go through and look at some of the people who participated in this long development and to hear a little about them.

So we should actually begin a little bit before Zeno, in the 6th century, Pythagoras started his famous school of mathematics and, of course, he proved the Pythagorean Theorem. Then Zeno came along in the middle of the 5th century B.C. In the middle of the 4th century B.C., there was a Greek mathematician by the name of Eudoxus, who developed a method of exhaustion that was very similar to the integral, one of the basic ideas of calculus. In fact, we'll see an example of Eudoxus's method in Lecture Number Nine.

In about the year 300 B.C., Euclid invented the axiomatic method of geometry and wrote maybe the most famous mathematics book ever written, *The Elements.* In 225 B.C., Archimedes used a method of exhaustion to develop formulas for areas and volumes of geometrical figures, and we'll see a great example of that that, in fact, involves his lever, in Lecture Number Nine.

Well, let's skip ahead a couple thousand years and get to the year 1600. There were some contributions in between, but it really got going again in the year 1600 when we have Johannes Kepler and Galileo Galilei working on motion, that planetary motion and describing the motion of objects just in the world, and making mathematical formulas that would describe that motion.

In 1629, Pierre de Fermat developed methods for finding maxima of values; and, in fact, that idea was very close to the idea of the derivative and we'll see it applied in our lecture about optimization. So, he was very close to being an inventor of calculus, but not quite.

In the 1630s, Bonaventura Cavalieri developed a method that he called the "method of indivisibles," and we'll see an example of that in a later lecture. René Descartes was famous for his Cartesian coordinate system, made a connection between algebra and geometry.

In 1669, Isaac Barrow, who was Newton's teacher, gave up his chair in mathematics for his famous student, Newton, and I thought this was such a nice thing to do until I learned that actually he had gotten a better job and went into a different line of work.

But finally, after all this, we come to the two people whose names are most closely associated with the development of calculus, and these are Isaac Newton and Gottfried Wilhelm von Leibniz. These are two people who are most associated with calculus, and we might say they invented calculus or discovered calculus; but, actually, all of the developments of calculus were definitely incremental. And the fact that both of them independently invented calculus at nearly the same time shows that the ideas were in the air. In fact, Isaac Barrow, Newton's teacher, remember, he actually expressed the fundamental theorem of calculus in one some of his lectures, but didn't appreciate the significance of calculus. He wrote it down, but didn't appreciate the significance. It took Newton and Leibniz to really systematize calculus, which was what we come to think of as the core of it.

Well, from the time that the calculus was actually invented in the late 17th century, it still took many years for calculus to be really understood; and many people developed variations to the idea and applied it to many applications in life, from physics, economics, probability—all sorts of things, biology. Johannes and Jakob Bernoulli were two of eight Bernoullis that were involved in developing calculus and applying it in Europe. Leonhard Euler developed a tremendous amount of mathematics, including many applications and extensions of calculus, especially infinite series. Joseph Louis Lagrange worked on the calculus variations. Pierre Simon de Laplace worked on partial differential equations and applied calculus to probability theory. John Baptiste Joseph Fourier invented ways to approximate certain kinds of dependencies, certain kinds of descriptions of things, in terms of what are called circular functions, like sines and cosines. And, by the way, this idea of taking a path, such as the path of a planet, and describing it by a collection of circles, really came from the old idea of using epicycles to try to describe the elliptical orbits of planets. So, he by adding a lot of these circular motions, you can approximate other kinds of curves. Augustin Louis Cauchy developed ideas about infinite series and tried to formalize the idea of limit. We'll find that the idea of a limit comes up in both the derivative and the integral, and it is one of the challenges to calculus that neither Newton nor Leibniz really were able to master. It took another 180 years before people really could pin down that idea of limit, which is when things, numbers, are converging toward one number. But we'll talk about that much more in future lectures.

In the 1800s, Georg Friedrich Bernhard Riemann developed the modern definition of the integral, which is one of those two ideas of calculus; and, in the middle of the 1850s, Karl Weierstrass formulated, finally, the rigorous definition of limit that occurs in this calculus book and we use today. Notice, again, that this 1850 is 185 years since Newton first developed calculus in 1665.

So, it took a long time to understand calculus, even after it was invented. So, if you don't understand calculus, you can take some solace in the fact that it took humanity 2,000 years to develop it and, even after it was developed, it took several hundred years to understand. It takes everybody a long time to understand it. And, in fact, essentially nobody understands calculus the first time they take it. I didn't understand calculus the first time I took it and among my friends who teach mathematics, they didn't understand calculus the first time they took it. Almost everybody—in fact, for most of us who teach calculus, the time we understood calculus was the time we taught it, and I recommend this as a method for learning anything.

Well, this then, gives a sense of the history of calculus. I'd like to take the remaining part of this lecture to just tell you the structure of the upcoming lectures, the structure of the course.

So, we'll begin in Lectures Two, Three, and Four comprise a collection of lectures that really present the fundamental ideas of calculus. In Lecture Two we introduce the idea of the derivative and say what the definition is and what it means. Then, in Lecture Three, we do the same thing to the integral and say what the definition of the integral is, how it comes about. And, then, in Lecture Four, we introduce the fundamental theorem of calculus, which connects the two. And, if I have done my job correctly, you will find that Lecture Four is sort of a letdown in that it is obvious that the derivative and the integral are two sides of the same coin; that they are related to each other. But, that was an insight that took some time, historically, to clarify.

Well, after we get through Lecture Four, then we proceed to investigate each of these fundamental ideas in more detail. The next lectures are a sequence of lectures about the derivative, and in the collection of lectures about the derivative, we're going to look at the derivative and see that, in fact, although we had presented the idea of the derivative in terms of motion, of a car moving on a straight line, we'll see that that same abstract procedure that's the derivative also tells us other things. It tells about—it can be interpreted in terms of graphs of functions, which we'll talk about and introduce in later lectures. But it says the same procedure that tells us the instantaneous velocity of a car also tells us the slope of a tangent line of a curve, and that kind of connection is an interesting one that gives the power to the ideas of the derivative.

Similarly, we're going to take the derivative and look at it from the point of view of its algebraic manifestations. One of the properties that makes the derivative so potent, and the integral, is that you can do it algebraically, and that's this mechanical side that students view as—the most common part that they deal with most is learning how to manipulate the algebra. So, we'll see the derivative then physically, that is, with the car moving; and then graphically; and then algebraically; and, then, we'll see it applied to different application areas such as volume, formulas for the volumes of objects—all of these things have manifestations about the derivative.

After that, we turn to the integral and we have a similar sequence of lectures that present the integral in these terms. That is to say, graphically, algebraically, and in terms of their applications to many different areas. So by taking these fundamental ideas and viewing them in different ways, that will show the richness of these themes. Then, the last half of this course will demonstrate the richness of these two ideas by showing lots of examples of their extensions, their variations, and their applications.

Well the purpose of these lectures is to explain clearly the concepts of calculus and to convince you that calculus can be understood from simple scenarios. Calculus is so effective because it deals with change and motion, and it allows us to view our world as a dynamic rather than just a static place. Calculus provides a tool for measuring change, whether it's a change in position, change in temperature, change in demand, or change in population. But, in addition to that, I like to think that calculus is intrinsically intriguing and beautiful, as well as just being merely important. So, calculus is a crowning intellectual achievement of humanity that all intelligent people can appreciate, enjoy, and understand. I look forward to exploring calculus with you during the next 23 lectures. Bye for now.

Stop Sign Crime—
The First Idea of Calculus—The Derivative
Lecture 2

In this lecture, the word for the day is *derivative*. Today the whole purpose of this lecture is to introduce the idea of derivative and to explain what it means.

Change is a fundamental feature of our world: temperature, pressure, the stock market, the population—all change. But the most basic example of change is motion—a change in position with respect to time. We will start with a simple example of motion as our vehicle for developing an effective way to analyze change. Specifically, suppose we run a stop sign, but in preparation for potential citations, we have a camera take a picture of our car neatly lined up with the stop sign at the exact instant that we were there. We show this photograph to the officer and ask to have the ticket dismissed, presenting the photograph as evidence. The officer responds by analyzing our motion in a persuasive way that illustrates the first of the two fundamental ideas of calculus—the *derivative*. We get the ticket but can take some solace in resolving one of **Zeno**'s paradoxes.

Calculus has two fundamental ideas (called the *derivative* and the *integral*)—one centered on a method for analyzing change; the other, on a method of combining pieces to get the whole. Both of the fundamental concepts of calculus arise from analyzing simple situations, such as a car moving down a straight road. This lecture presents an everyday scenario that leads to one of the two ideas of calculus—the derivative.

The following stop-sign scenario is a modern-day enactment of one of Zeno's paradoxes of motion. Let us suppose we have a car driving on a road, and there is a mileage marker at every point along the road. Such a simple scenario can be represented in a **graph**. The horizontal axis is the time axis. The vertical axis tells us the position of the car at each moment of time. For the sake of arithmetic simplicity, we will talk about measuring the **velocity** (speed) of the car in miles per minute. Therefore, the vertical axis of our graph is in miles and the horizontal axis is in minutes.

Suppose Zeno is driving this car, and he goes through a stop sign without slowing down. Soon thereafter, he is pulled over by Officers **Newton** and **Leibniz**. (The corny names will make memorable how the roles in this drama relate to Zeno's paradox and the invention of calculus.) The driver, Zeno, protests by showing a still picture of his car exactly at the stop sign at the exact moment, 1 minute after the hour, when he is supposed to have been running the stop sign. On this street, time is measured by stating the minutes only. Zeno claims that there could be no violation because the car was in one place at that moment.

Officers Newton and Leibniz produce additional evidence. The officers produce a still photo of the car at 2 minutes after the hour clearly showing the car 1 mile beyond the stop sign. Zeno argues, "So what? At 1 minute, I was stopped at the stop sign." Newton: "But you must admit your average velocity between 1 and 2 minutes was 1 mile per minute." To compute the average velocity during any interval of time, you need to know the position of the car at the beginning, the position at the end, and the amount of time that passed. The average velocity is change in position divided by change in time, that is, how far you went divided by how long it took. The average velocity does not rely on what happened between those two moments—just on where the car is at the beginning and at the end. Again, Zeno says, "So what?" Officers Newton and Leibniz produce an infinite amount of additional evidence, all incriminating. They note that Zeno was at the 1.1 mile marker at 1.1 minute, at the 1.01 mile marker at 1.01 minute, and so on, all of it proving that Zeno's velocity was 1 mile per minute. The idea of instantaneous velocity is the result of an infinite amount of data. All the evidence is that, on every even incredibly tiny interval of time, the average velocity was 1 mile per minute. The cumulative effect of all the evidence—an infinite number of intervals—leads to the idea of instantaneous velocity.

> **"Motion is the fulfillment of what exists potentially insofar as it exists potentially."**
>
> **—Aristotle**

Knowing the position of a car at every moment allows us to compute the velocity at every moment. This can be illustrated with a car whose velocity is increasing. Let's now consider the example where at every time, measured in minutes and denoted by the letter t, we are at mileage marker t^2 miles. For example, at time 1, the car is at position 1, but at time 2, it is at position 4 (2^2), and at time 0.5, it is at position 0.25 (0.5^2). If we know where the car is at every time during an hour, we can tell how fast it was going at any selected moment by doing the infinite process of finding instantaneous velocity.

The idea of instantaneous velocity is the result of an infinite amount of data.

Let's apply that infinite process to this moving car at several different times t. First, if time $t = 2$ minutes, the position of the car is $2^2 = 4$. Then, if time $t = 1$ minute, the position of the car is $1^2 = 1$. Subtracting 1 from 4 and dividing by 1 minute, we have 3 miles per minute. In other words, by looking at where the car was 1 minute after the 1-minute mark, we find that the average velocity was 3 miles per minute. However, when we look at shorter intervals of time, we get a different story, as shown in the chart below. We will find the positions of the car at various nearby times, such as 1.1, 1.01, 1.001, and 0.99 minutes, and compute our average velocity between time 1 and those times.

Position: $p(t) = t^2$

Initial Time	Final Time	Average Velocity (mi/min)
1	2	3
1	1.1	2.1
1	1.01	2.01
1	1.001	2.001
0.999	1	1.999

Instantaneous velocity at $t = 1$ is 2 mi/min

The whole collection of average velocities leads us to conclude that the instantaneous velocity is 2 miles per minute. This process of taking smaller and smaller intervals of time to arrive at the instantaneous velocity is called a *limit*; the instantaneous velocity is the limit of the average velocities as the intervals of time get smaller and smaller. The infinite process used to find the velocity at each time is the *derivative*. The derivative gives the instantaneous velocity at any given moment using this infinite process, if we know the position of the car at each moment.

If we look at the same question for other times, 0.7 minutes, for example, we have the following results:

Position: $p(t) = t^2$

Initial Time	Final Time	Average Velocity (mi/min)
0.7	1.7	2.4
0.7	0.8	1.5
0.7	0.71	1.41
0.7	0.701	1.401
0.699	0.7	1.399

Instantaneous velocity at $t = 0.7$ is 1.4 mi/min

If we look at the same question for other times, such as 1.4, 2, or 3 minutes, we have similar results for the instantaneous velocity.

We have been computing instantaneous velocities at various times for a car that is moving in such a manner that at every time t minutes, the car is at mileage marker t^2 miles. If we look at a chart of all our examples of instantaneous velocity, we see a pattern that indicates that at every time t, the velocity is $2t$ miles per minute.

Position: $p(t) = t^2$

Time (min)	Instantaneous Velocity (mi/min)
0.7	1.4
1	2
1.4	2.8
2	4
3	6

Instantaneous velocity: $v(t) = 2t$

The first idea of calculus, the derivative, quantifies the idea of instantaneous velocity. We have taken a simple, everyday scenario (a moving car) and developed a simple, though infinite, process that made sense of the intuitive idea of motion or velocity at an instant. We now have an equation that tells us the instantaneous velocity at each moment of time.

The derivative of a **function** $p(t)$ at time t is obtained by computing $\frac{p(t+\Delta t)-p(t)}{\Delta t}$, where Δt is a small increment of time, and then seeing what number those values approach as Δt becomes increasingly small. The single number to which those values approach is called the limit: $\frac{p(t+\Delta t)-p(t)}{\Delta t}$ as Δt approaches 0. ■

Names to Know

Leibniz, Gottfried Wilhelm von (1646–1716). German diplomat, logician, politician, philosopher, linguist, and mathematician; president of the Berlin Academy. Leibniz is regarded, with Newton, as a co-inventor of calculus. He was the first to publish a theory of calculus. Leibniz's notation is used currently. He made substantial contributions to formal logic, leading to the establishment of symbolic logic as a field of study. He discovered an infinite series formula for $\frac{\pi}{4}$. He was accused of plagiarism by British partisans of Newton, and his supporters counterclaimed that Newton was the plagiarist. Now, he is acknowledged to have independently discovered calculus.

Newton, Sir Isaac (1642–1727). Great English mathematician and scientist; Lucasian professor of mathematics at Cambridge. Newton was the first discoverer of differential and integral calculus. He formulated the law of universal gravitation and his three laws of motion, upon which classical physics is based. In 1687, he published his results in *Philosophiae Naturalis Principia Mathematica*. He formulated the theory of colors (in optics) and proved the binomial theorem. He is possibly the greatest genius of all time. Newton was a Member of Parliament (Cambridge), long-time president of the Royal Society, and Master of the Mint. The controversy with Leibniz over attribution of the discovery of calculus poisoned relations between British and Continental scientists, leading to the isolation of British mathematicians for much of the 18th century.

Zeno of Elea (c. 495–430 B.C.). Ancient Greek dialectician and logician. He is noted for his four paradoxes of motion. He was a student of Parmenides, whose school of philosophy rivaled that of the Pythagoreans.

Important Terms

function: Mathematical description of dependency. A rule or correspondence that provides exactly one output value for each input value. Often written algebraically as $f(x)$ = formula involving x (for example, $f(x) = x^2$).

graph: A geometric representation of a function, showing correspondences via pairs of points (input, output) drawn on a standard Cartesian (x-y) plane.

velocity: Average velocity is total distance divided by the time it took to traverse that distance; units are length/time (for example, miles per hour). Instantaneous velocity is the speed at one moment of time, approximated by average velocity for smaller and smaller time intervals; units are also length/time. The instantaneous velocity is the derivative of the position function for a moving object.

Suggested Reading

Any standard calculus textbook, section introducing the derivative.

Questions to Consider

1. Suppose a rocket is traveling along a road, and at each time t minutes its position is t^3 miles from where it started. Use the derivative process (and a calculator) to guess the instantaneous velocity of the rocket at various times, such as 1 minute, 2 minutes, and 3 minutes after starting.

2. Would it make sense to view position and time as discrete quantities having a little width to them? Then, a moving object could pause for an instant of time at one "point" before moving on. Zeno's fourth paradox of motion treats this situation when he considers three objects on parallel tracks, one moving right, one moving left, and one fixed. The relative velocities of these moving objects present a challenge. What is it?

Stop Sign Crime—
The First Idea of Calculus—The Derivative
Lecture 2—Transcript

Welcome back to *Change and Motion: Calculus Made Clear***.** In this lecture, the word for the day is *derivative*. Today the whole purpose of this lecture is to introduce the idea of derivative and to explain what it means. So, it's a very clear goal for the lecture. Remember, calculus is based on two ideas; it's the exploration of two ideas. One of them is the derivative, which we'll meet today, and the other one is the integral, which we'll meet in lecture number three. And both of them come from the same scenario of a car driving down a straight road, and so, our challenge today is to analyze this very simple situation of a car moving down a straight road; and, from it, deriving the definition and the concept of the derivative.

So, let's just get started right away with the following scenario: Suppose that we have a car that's driving along a road. And so, here, we have a car, you see, and it's driving a long a straight road like this; at every moment of time the car is in some place along the road, and this road is amazingly well marked. You know how there are mileage markers on some roads? Well, this is a road that has a mileage marker at every single point on the road, so you know exactly where you are at every single moment. So, at every moment of time, the car is at some position on the road, you see? So, this is a basic idea; and, one thing that I wanted to say before we develop the scenario of the derivative is the idea of how can we represent a simple scenario such as a car moving on a road in a way that is useful to describe. Well, one way is to look at a graph that captures the idea of a car moving in a road. The way that a graph works is very simple. We have a horizontal axis, which is the time axis, and then we have a vertical axis that tells us the position of the car at each moment of time. So, if we have a car that's moving, for example, just like this on our road, and it's sort of a steady kind of a way, then that would correspond to the car's position changing, and it could be recorded that at each moment of time it is in a certain position. And, so, every dot on this graph represents the position of a car moving on a road.

Now, I do want to say one thing about this: That, in all of the lectures today and tomorrow and the next day, we're going to be talking about this car moving on a road, and for purposes of arithmetic simplicity, we will be talking about measuring the speed of the car in miles per minute rather than miles per hour, and the reason is that 1 mile per minute is a smaller number than 60 miles per hour. So, it's just computationally easier. So, in our graph here, the vertical axis will be in miles and the horizontal axis will be in minutes.

Okay, now, here is the scenario that we're going to discuss first in our situation of a car moving down the road. Here's the car; and here is, as you see, a stop sign. Now, here's the way the car goes; it just goes like this. I guess the stop sign should really be like this. Now, you notice something about that car moving down the road; the car moved down the road—notice, it didn't stop. It just went straight through the stop sign. Now this, by the way, is going to illustrate Zeno's paradox, his arrow paradox, but in the modern terms of a car moving down the road.

Now so here's what happened in our scenario: This car moved down the road, it just saw the stop sign, and then just drove right on through. That was what the car did. A little later, the car, as it was driving down the road, it heard a siren in the background and the policemen pulled up and stopped the car and said, "You just went through that stop sign. What do you think you're doing? You didn't even slow down? You just went right through the stop sign." And the driver of the car, though, it turns out that the driver of the car—by the way, the driver of the car's name is Zeno; that's the driver of the car—was prepared for this possibility for being pulled over, and said, "No, no, officer, officers, that's not true. You say that I went through the stop sign; but, in fact, I have a photograph that proves that I was just at the stop sign at the moment that you say I was moving through it;" and so Zeno had a picture of the car right at the stop sign. See? A still photograph of the car right at the stop sign which he had prepared in advance to be taken, and he took that and it was sent electronically to him in this car, and he said, "Here is this photograph of the car at the line of the stop sign. What do you mean by claiming that I was moving at that time? I wasn't moving. Look at this still picture. Any jury in its right mind could not possibly convict me of going through that stop sign when I have this evidence that I was in one place at one time."

Well, at that point, Zeno looks at the little nametags of the officers, and one of the officer's names is Newton and the other officer's named Leibniz. Now, Zeno, if he had lived a few thousand years later, would be a little bit nervous at this point because it turned out that Newton and Leibniz, these officers, say, “Ah, but we have some additional evidence about this case; we have additional evidence that’s going to illustrate why you were moving at the time you claim you were stopped at the stop sign; why you were moving at the stop sign.” So, they say, “Look. You have this photograph of you right at the stop sign at a given time, but we have a photograph of you one minute later one mile down the road. See? And, so, we can compute how fast you were going during that minute, and here’s the way we do it.” You see, it’s very easy to compute how fast a car is going between two moments of time because you simply look at the position that the car was in at the last time and the position the car was in at the first time, subtract the two to get the distance traveled—so, if this is the position of the car at the first time, and this was at the last time, then that difference in distance is the distance traveled; and, then, dividing by the amount of time it took, in this case one minute, would give us the average velocity in that the car was traveling during those two instances of time. And that’s the nature of motion. The nature of motion—the most fundamental aspect of motion is that the car is at one place at one instant and it’s at another place at another instant.

Now, let’s be specific about our times so we can make this computation exact. Our time is being measured in minutes, and, if you wish, you may think of these as minutes after an hour. The car was at the stop sign at precisely 1 minute, your watch was reading exactly 1 minute; that’s where the car was. One minute later, the car was 1 mile down the road. So, we have the photograph of the car at 1 minute past the hour, and here’s the photograph of the car at 2 minutes past the hour. The position of the car means where the car was against the mileage marker running along the side of the road. Now, these markers, by the way, don’t correspond.

$$\text{Average speed} = \frac{p(2) - p(1)}{2 - 1} = \frac{2 - 1}{1} = 1\text{mi/min}$$

This is a ruler; don’t look at the ruler for the mileage markers. We’re going to say—I’ll put the ruler down—that at the time that the car was at the stop sign, the stop sign was at the 1-mile position on this straight road. And, then, 1 minute later, it was at the 2-mile position on the straight road. In other words, you look out your window and that's what you see, the 2-mile marker. So, here we can compute the average speed, or the average velocity, during that time was the position of time 2 minus the position of time 1, divided by the elapsed time, which was—the time 2 minutes was what time the person ended up 1 mile down the road; and then minus 1 minute, that was the time that the person was at the stop sign; and this division, $\frac{2-1}{1}$, is 1 mile per minute. And that's the average time, the average velocity, of the car during that minute.

$$\text{Average speed} = \frac{p(1.1) - p(1)}{1.1 - 1} = \frac{1.1 - 1}{1} = 1\text{mi/min}$$

Well, of course, Zeno says, "So what? So what? What do you mean to say that I was going through the stop sign because all of the evidence that you have is where I was at the time of the stop sign and where I was a minute later. But how do you know that I wasn't stopped at the stop sign and then I just speeded up to get a mile down the road in the minute? It doesn't tell me anything about what I was doing at the instant that you claim I went through the stop sign."

Well, the two officers, Leibniz and Newton, say, "We have additional evidence. We had other photographs taken, and the other photographs we had taken were photographs that took place at exactly 1.1 minute—the time was 1.1 minute—we took a photograph, and where was the car? It was at the position 1.1 mile marker." Well, and so what does that say? That says that the position at time 1.1 was 1.1; that's where the car was at time 1.1. At time 1, the car was at position 1. The difference between those two is .1. So, the car traveled .1 mile in the elapsed time. The elapsed time was 1.1, that was the time that we measured at the second position minus the initial time, 1; elapsed time of .1; dividing the two gives us a value of 1 mile per minute.

Zeno says, "Look. That still doesn't prove it. That just says where I was at the beginning and the end of this interval; it doesn't tell me how fast I was going at the instant you claimed I was moving through the stop sign."

So, Newton and Leibniz say, "Wait a minute; we have more evidence. Look at time 1.01; where were you? You were at the position 1.01; the elapsed time was .01 minutes; dividing it out, we see that it was the same story—you were going an average of 1 mile per minute when measured by these two moments of time that were very, very close to each other." And now then they say, "Look. We have more evidence. We have two pictures—one at time 1.001; where were you? You were at position 1.001. So, you traveled a total of .001 miles in .001 minutes; which, once again, is telling the story of 1 mile per minute." And in fact, Newton and Leibniz have infinite amount of information to give to the jury. They have the information of every single instant of time of where the car was before and after the time message number 1, and each time it told the same story. The story was that the car was going at 1 mile per minute. Even taking a photograph of where the car was immediately before the 1-minute time, at the .999 minute on your watch, the car was at position .999 on the road. Since all of that evidence, an infinite amount of evidence, all is telling the same story, we conclude that the instantaneous velocity at the 1-minute mark was exactly 1 mile per minute.

Okay, by the way, you may say this was a long and tedious story, and I have to tell you, this is a bit of an expository challenge because the whole concept of the derivative is, by its very nature, repetitive and similar. We're doing the same thing, subtracting the position at one moment minus the position at another moment and dividing by the elapsed time, and we're doing it over and over and over again, infinitely many times, and we're seeing that it's telling the same story; and that's what we're going to define as the instantaneous velocity of the car at time 1 because all of these measurements are telling us that the car was going at average speeds, average velocities, of 1 mile per hour, and therefore that is going to be declared to be the instantaneous velocity at that time.

$$\text{Average speed} = \frac{p(1.001)-p(1)}{1.001-1} = \frac{1.001-1}{.001} = 1 \text{ mi/min}$$

Notice, in order to talk about the instantaneous velocity, we needed information about the position of the car at all times nearby, and that was the thing that allowed us to make a decision about the instantaneous velocity at that moment. So, it was not just the information about where the car was at that moment, as Zeno would like it to have been for his defense; that photograph of him correctly at the stop sign at that time, that's not sufficient. In fact, you needed to know where the car was at all times nearby.

Now, you may say, well, okay, this is the first thing you'd think of—maybe not; I don't know if you think that way or not. But, I wanted to read to you some statement from Aristotle, who thought about Zeno's paradoxes and tried to shed light on them, and here is what Aristotle had to say about motion because instantaneous motion was a real conceptual challenge to people for many millennia. So, here's what he said. Aristotle was trying to give a definition of motion; he said the following: "Motion is the fulfillment of what exists potentially insofar as it exists potentially." And he goes on to say, here's a further remark on motion, "We can define motion as the fulfillment of the moveable qua moveable."

Now, as you know, I am a teacher, and I sometimes get papers from students that say things like that, and I always wonder did they feel that they were actually saying something meaningful? Maybe Aristotle felt he was shedding light on this issue, but it certainly is difficult to get anything from it. But, the point is, that we have had to do an infinite process to come up with a single number.

Now, there was a defect to the example that I just gave, and the defect was that in all instances, for every single pair of instances of time, we got the same answer. So, all the evidence was leading to the conclusion that the car was going at 1 mile per minute; every single time it said 1 mile per minute. Let's do another example.

Here's an example where the car is moving according to another scenario. So, here we have our car; it's moving on our, once again, on our well-traveled road, but this time the car, instead of going at a steady velocity, the car is speeding up as it goes. It's speeding up according to a very well-defined notion, and that is the following: That at every moment of time, you look at your watch and you say that the car is going to be at the position that's the square of the time that's on your watch. So at time 1 it is at position 1, as before; but at time 2, it is at 2×2 at position 4; and at time .5, it's at position .25. Follow me? So, that it's speeding up as it goes; speeding up as it goes. So that is the position of the car. [This is represented by the equation $p(t) = p^2$.]

Let's see if we can undertake exactly the same question about what the instantaneous velocity of the car is at time 1. Let's undertake the same thing. Now I know it's laborious, but it's important because this is one of the two basic ideas of calculus. So, here we go.

$$\text{Average speed} = \frac{p(1.1) - p(1)}{1.1 - 1} = \frac{1.21 - 1}{.1} = 2.1 \text{ mi/min}$$

Suppose that you looked at the position of the car at time 2. Well, the position is given by taking the time and squaring it, multiplying it by itself, so that's 4. So, the position at time 1 is given by squaring that, so it's at position 1, so the difference is 4 minus 1, that is, the car traveled 3 miles during that 1 minute of elapsed time, between time 2 and time 1; dividing 3 by 1, we get 3 miles per minute. So, by looking at where the car was 1 minute after the 1-minute mark, and we're undertaking to understand what the instantaneous speed was at time 1 minute—well, by looking at the evidence given by saying where the car was at time 2 minutes, we see that the average velocity during that time was 3 miles per minute. But, in this case, when we look at shorter intervals of time, we get a different story.

Now, let's look at the shorter interval of time that would happen if we looked at where the car was at the 1.1 minute mark. Well, the position at 1.1 is obtained by taking 1.1 and multiplying it by itself. We see that we're at the position 1.21 on our well-marked highway. The distance traveled between 1.21 and 1, subtracting, is that the car traveled .21 miles during the elapsed time of .1 minutes; dividing, we see that we get the car was traveling during that interval of time, that .1 interval of time, 1/10 of a minute interval of time, it was traveling at the average velocity of 2.1 miles per minute. You see, that's a different story than we had before. Before it said 3 miles per minute; now 2.1 miles per minute was the average velocity during that shorter interval of time. And we get a sense, by the way, that the shorter intervals of time are giving us a better indication of what the car was doing at the exact moment in which we're interested, because during that shorter interval of time, the car is not changing its velocity as much; doesn't have time to do so.

$$\text{Average speed} = \frac{p(2) - p(1)}{2 - 1} = \frac{4 - 1}{1} = 3 \text{ mi/min}$$

Let's continue this laborious process by looking at the 1.01 moment of time. This is 1/100 of a minute after the 1-minute mark. Where was the car? Well, if we multiplied 1.01 by itself, we get 1.0201. That's the position on this well-marked road. We subtract 1 from it; we get .0201 is the elapsed distance traveled during the 1/100 of a minute that we're talking. It's very precise. And we see, now, that its velocity during that very short interval of time was 2.01 miles per minute. Now, this is getting a little interesting, that we're getting different answers. As we take different, shorter intervals of time we're getting different answers, but those different answers are tending toward one number, and this is going to be one of the main characteristics of calculus and of doing the derivative in particular. So, this is—we're going to do this yet one more time. And, in fact, we're going to do it infinitely many times, except, of course, we won't actually do it on camera infinitely many times, but we will do it infinitely many times.

$$\text{Average speed} = \frac{p(1.01) - p(1)}{1.01 - 1} = \frac{1.0201 - 1}{.01} = 2.01 \text{ mi/mi}$$

Let's start with, once again, 1.001. So, this is 1/1000 of a minute after the time that we're interested in. We see how far the car went during that interval of time. It went .002001 miles, and it took 1/1000 of a minute to do so; when we divide, we get that the average velocity during that 1/1000 of a minute was 2.001 miles per minute. So, you can see that the values of these are—evidence that's taken by increasingly shorter intervals of time—are giving us a sense of the velocity of the car, getting closer and closer to 2 miles per minute.

$$\text{Average speed} = \frac{p(1.001) - p(1)}{1.001 - 1} = \frac{1.002001 - 1}{.001} = 2.001/\text{mi/min}$$

Let's just do one more example, just to see it; namely, if we take the car slightly before the 1-minute mark, at .999; if we do the multiplication; do the division; we see that the indicated average velocity during that 1/1000 of a minute was 1.999 miles per minute.

$$\text{Average speed} = \frac{p(1) - p(0.999)}{1 - 0.999} = \frac{1 - 0.998001}{0.001} = 1.999 \text{ mi/min}$$

So, what is the evidence telling us? The evidence is giving us a sense that the velocity of the car is converging; that all of this different evidence, each piece of evidence, taken from two instants of time, seeing how far the car went during that short interval of time, and then dividing by the elapsed time—all of that evidence, each one gives us a number, and those numbers are getting closer and closer to the single number 2. That process is called a limiting process. We're taking a limit of values that are getting closer and closer to 2, and then we say that at the instance of 1, that the value is 2. That's the concept of the derivative. The derivative—for every instant of time we have to do an infinite process; and when we do the infinite process, we get a single number, namely, the number to which all of that evidence converges.

Well, this is, as you can see, this is a very laborious kind of a process. Even for this car moving at this time where its position at every moment of time is given by the square of the time on our watch, we see that it took us a long time, and it was sort of boring, to figure out how fast the car was going at the 1-minute mark. But, we could do this at every single moment of time. That is, this long, laborious process could be undertaken at different points of time.

So, for example, this is a chart that accumulates all of the evidence that we had seen from our previous investigations. Namely, the evidence was that the average velocity during the time 1 and time 2 was 3 miles per minute; between 1 and 1.1 it was 2.1 miles per minute; between 1 and 1.01 it was 2.01 miles per minute; between .999 and 1 to was 1.999 miles per minutes; and we, therefore, conclude that the instantaneous velocity at time $t = 1$ is 2 miles per minute. So, this summarizes our previous information.

Now, let's consider this same question for different times. In other words, instead of thinking about the time when our watch says 1 minute, let's consider when our watch says something else, like .7 minutes. Well, if we look at .7 minutes, and we do this same laborious computation, that is, we take .7, that's our initial time, and then we look at a minute later and we take the average velocity, we get 2.4; and we take .7 and we just do 1/10 of a minute afterwards, we compute the average velocity, 1.5; we take 1/100 of a minute after .7, we get an average velocity of 1.41. So, in each case we're doing this characteristic process of the position, the final position of the car, the initial position of the car, dividing by the elapsed time. That's what each of these numbers represents.

Notice that as we get closer and closer to the .7, the time at which we're interested in the instantaneous velocity, when we take just 1/1000 of a minute afterwards, or just 1/1000 of a minute before, our computations are leading us to believe that the actual instantaneous velocity is the number 1.4. So, we conclude, by doing infinitely many experiments and increasingly shorter amounts of time, that the instantaneous velocity is 1.4 miles per minute. Well, that's the computation associated with the value .7.

I'm just going to quickly show you some other charts where we did exactly the same laborious process. In the case of—at the 1.4 initial time, we do this same computation, and we conclude that the numbers to which those average velocities are tending looks like it's tending towards 2.8; 2.8 is the instantaneous velocity at the time 1.4.

When we're looking at the initial time 2, once again we can do this laborious process of taking nearby times, and we see that those values are getting near 4 miles per minute. When we look at an initial time of 3, we see that our instantaneous velocity is converging down to the number 6 miles per minute. By the way, now we're sort of beyond the idea of a car moving, since that's 360 miles per hour, but it's 6 miles per minute.

The point is that each of these laborious computations gives us a value of the instantaneous velocity; but, we see a pattern. Look at our pattern here. We did each one of these values here, the instantaneous velocity in this column, at the time .7, we deduced that that instantaneous velocity, computed from an infinite process, was 1.4 miles per minute. At the time 1, the instantaneous velocity was 2. At the time 1.4, the instantaneous velocity was 2.8. At the time 2, the instantaneous velocity was 4. At the time 3, the instantaneous velocity was 6. We see the pattern. Do you see the pattern? The pattern is that the instantaneous velocity is always exactly twice the time. So, once we observe this pattern, we see that if the car is moving at a rate so that its position at the road is always measured by t^2, then its velocity is measured by 2 × the time. And once we accumulate that into that very simple form, then we don't have to go through the laborious process at each moment of time to see what the instantaneous velocity is. We have one equation that tells us instantly the instantaneous velocity at each moment of time. And that's the power of the derivative.

> The derivative of a function $p(t)$ at time t is
>
> $$\frac{p(t+\Delta t)-p(t)}{\Delta t}$$
>
> as Δt becomes small.

So, the word of the day is derivative. And when you think of derivative, what I want for you to come to mind is this repetitious thing that we've been doing all during this half-hour. Namely, if we have a position of a car moving along a straight road, and we want to know if the instantaneous velocity at the time t, we compute its position at time $t + \Delta t$, where by tradition Δ means Δt is a small increment of time and viewed as an increasing small increment of time, but a specific small increment of time, we see where we were a little bit after or before t, we subtract where we were at time t to find the elapsed distance traveled during that very short interval of time, and then we divide by Δt—which is the elapsed time, the time between $t + \Delta t$ and the time t. The elapsed time was Δt. This characteristic fraction, just looking at this fraction, tells us the average velocity of the car between two instants of time. Notice there's nothing continuous about this; it's one instant, and then another instant, and we compute the—everybody can agree on what the average velocity is if you give just two instants because it's the net instants divided by the elapsed time, and then what we do is we do this process that's the limiting process as Δt becomes increasingly small and goes to 0. That is the process of the derivative.

So, we've now been introduced to one of the two fundamental ideas of calculus, the derivative. In the next lecture, we're going to take the same scenario of a car driving down the road, and introduce from it the idea of the integral. I look forward to seeing you then.

Another Car, Another Crime—
The Second Idea of Calculus—The Integral
Lecture 3

In this lecture, we're going to introduce the second fundamental idea of calculus, that is, the *integral*. And, once again, it's going to involve an infinite process, and it's going to involve a process that comes about as a natural result of analyzing a situation once again of a vehicle moving along the road.

The second idea of calculus helps us understand how to take information about tiny parts of a problem and combine this information to construct the whole answer. To develop this idea, we return to the scenario of a moving car. In preparation, we take our car out to a straight road, say to El Paso. We then videotape only the speedometer as the car moves. We show some friends the video of the speedometer and ask them to predict where we were at the end of an hour. The process by which they use information about velocity to compute the exact location of the car at the end of the hour is the second of the ideas of calculus—the integral.

In the last lecture, we introduced the derivative—the first of the two basic ideas of calculus. The derivative allowed us to settle one of Zeno's paradoxes of motion because it told us what we mean by the instantaneous velocity of a moving car or arrow. Note that we did not fall into the trap of trying to divide 0 by 0 to get the velocity at an instant.

The second fundamental idea of calculus arises from a scenario involving another car and another crime. In this scenario, you are kidnapped, tied up in the back of the car, and driven off on a straight road. You cannot see out of the car, but fortunately, you can see the speedometer, and you have a video camera to take time-stamped pictures of the speedometer. (There is no odometer in sight.) After 1 hour, you are dumped on the side of the road. How far have you gone?

What information can we extract from the videotape? Let's take a simple case: The car was going at a constant velocity. At a constant velocity, computing how far we have gone in a given amount of time is easy. For example, if we go 1 mile per minute for 60 minutes, we will have gone 60 miles during that hour. If we go 2 miles per minute for 20 minutes, we will have traveled 2 × 20, or 40 miles. On a graph, this constant velocity would appear as a horizontal line.

Let's take a harder case: Suppose our velocity changes. How do we compute the distance traveled? Let's look at examples where our velocity is steady for some time, then abruptly changes to another velocity, and so on. Suppose we travel at: 1 mile per minute between times 0 and 10 minutes, 2 miles per minute for the time between 10 and 20 minutes, 3 miles per minute for the time between 20 and 30 minutes, 4 miles per minute for the time between 30 and 40 minutes, 5 miles per minute for the time between 40 and 50 minutes, and 6 miles per minute for the time between 50 and 60 minutes. On a graph, this changing velocity would appear as "stair steps" going up. Our total distance traveled will be: (1 × 10) + (2 × 10) + (3 × 10) + (4 × 10) + (5 × 10) + (6 × 10) miles; that is, 210 miles. (We were speeding.) We now know the process of computing the total distance traveled if we know the velocity at each time and the velocities are steady at one velocity for a while, then jump to another velocity and so on throughout the whole hour.

Our strategy is first to underestimate the distance traveled, then to overestimate the distance traveled, then to determine that the distance traveled is somewhere between these two.

We know, however, that velocities do not jump like this and, instead, increase smoothly from one velocity to another. Dealing with **variable** velocities involves doing a little at a time and adding them up. Our strategy is first to underestimate the distance traveled, then to overestimate the distance traveled, then to determine that the distance traveled is somewhere between these two. Let's consider a car that is moving at each time t at velocity $2t$ miles per minute. That is, at 1 minute, we are traveling at a speed of 2 miles per minute; at 2 minutes, we are traveling at 4 miles per minute; and so forth. Because our car is going so fast, let's just see how far it goes during the first 3 minutes of travel. In order to discover the distance covered, we will compare our car traveling smoothly at an increasingly higher velocity to a second car that is moving in the “jerky” or “jumpy” fashion we have seen above. We will also break the 3 minutes into smaller intervals of time, for example, half-minute intervals. We then add up approximations of the distances traveled by the second “jerky” car during each short interval to approximate the total distance traveled. During each short interval of one-half minute, the car changed velocities. But if we assumed that the second “jerky” car continued at a steady velocity equal to the initial lower velocity during each little interval of time, we would get an approximation of the total distance traveled during that interval (7.5 miles). Because our car is always speeding up, the distance traveled by the second, “jerky” car will be an underestimate of the total distance our car traveled. Similarly, if we assumed that the second car went at a steady velocity equal to the fastest velocity our car actually went during each little interval of time, we would get an approximation to the total distance traveled during that 3-minute interval (10.5 miles). This approximation to the total distance traveled would be an overestimate. The correct answer would have to be somewhere between those two estimates.

The smaller the intervals, the more accurate will be the approximation to the distance traveled. Let’s try breaking the time into intervals 1/10th of a minute long. Once again, we can get an underestimate (8.7 miles) and an overestimate (9.3 miles). Notice that these over- and underestimates are closer to each other than before. As the intervals get smaller, it doesn’t matter what velocity we select from the range of velocities of the car in the intervals, because the car doesn’t change velocity much in the tiny intervals. Using increasingly smaller intervals produces increasingly better approximations.

The exact distance traveled can’t be found with any single division of the interval of time. The exact answer is obtained by looking at infinitely many increasingly improved approximations. The finer approximations get closer and closer to a single value—the limit of the approximations. This infinite process is the second fundamental idea of calculus—the integral. If we know the velocity of a car at every moment in a given interval of time, the integral tells us how far the car traveled during that interval. Remember that the derivative was the limit of the average velocities as the intervals got smaller and smaller; likewise, the integral is the limit of the approximations as the intervals get smaller and smaller.

We can use the same analysis to find out where we were at any moment. In the example above where the speedometer always reads $2t$, where are we after 1 minute, 2 minutes, 2.5 minutes, 3 minutes? In each case, we will use this infinite procedure to see how far we traveled from time 0 to these times. Let's look at the table and see if there is a pattern.

Instantaneous velocity: $v(t) = 2t$

Time (min)	Distance
0.5	0.25
1	1
1.5	2.25
2	4
3	9

Position: $p(t) = t^2$

We can make a shrewd guess as to where we were after any time t: Namely, it appears that we are at the t^2 mileage marker. Distance traveled is, thus, the square of the time interval taken, or $p(t) = t^2$. The integral can be used to find the position of a moving car at each moment, if we know the velocity at each moment. The integral process involves dividing the interval of time into small increments, seeing how far the car would have traveled if it had gone at a steady velocity during each small interval of time, and then adding up those distances to approximate the total distance traveled. Therefore, the formula to determine the distance traveled between time a and time b is: $(v(a) \times \Delta t) + (v(a+\Delta t) \times \Delta t) + (v(a+2\Delta t) \times \Delta t) + \ldots + (v(b - \Delta t) \times \Delta t$, as Δt becomes increasingly smaller. By taking smaller and smaller subdivisions and taking a limit, we arrive at the actual value of the integral. ■

Important Term

variable: The independent quantity in a functional relationship. For example, if position is a function of time, time is the variable.

Suggested Reading

Any standard calculus textbook, the section introducing the definite integral.

Questions to Consider

1. Suppose the speedometer on your rocket ship reads exactly $3t^2$ miles per minute at each time t minutes. Use the integral process to compute how far you will have traveled after 1 minute, 2 minutes, 3 minutes. Do you see an expression in t that gives the same answer?

2. Suppose your velocity is always $2t$ miles per minute at each time t minutes. How fine must your divisions of the interval of time be between time 0 and time 3 in order to be certain that your summation in the integral process is definitely within 1 mile of the correct answer?

Another Car, Another Crime—
The Second Idea of Calculus—The Integral

Lecture 3—Transcript

Welcome back. In the last lecture we introduced one of the two fundamental concepts of calculus, namely, the derivative; and we saw that the derivative allowed us to actually settle one of Zeno's paradoxes in motion because it told us what we were to mean by the instantaneous velocity of a moving car or arrow. So, that was an accomplishment that we did, and notice that it never involved the mistake of dividing zero by zero. We always looked at points of time that were close to each other so that there was a difference between the increments of time so we were never dividing by zero.

In this lecture, we're going to introduce the second fundamental idea of calculus, that is, the *integral*. And, once again, it's going to involve an infinite process, and it's going to involve a process that comes about as a natural result of analyzing a situation once again of a vehicle moving along the road. And in order to keep in our theme of sticking to crime, we're going to involve this lecture with another crime. So, this is called "Another Car, Another Crime." So, this is what we're going to do.

So, suppose this time the crime is a much more heinous offense than just merely running a stop sign, as we had in the last lecture. In this case, the crime is kidnapping. So, here's what happened. You were an agent for the Mission Impossible crew, and you were kidnapped by some nefarious bad people and you were put in the back of a van. Now, you have to think of this as a van in this case. So, here you are in the back of a van, and you're put in the van, and you know where you were captured. You were put in this van. But, the van has these walls on the side of it, so you can't see out on this well-marked road; you can't see any of the markings on the side of the road. But, the car starts out and goes along, and there it is driving along, and the only thing that you can see inside this back of the van where you're thrust is through a little hole that looks into the front compartment of the van; you can see the speedometer of the van. So, you see the speedometer and that's all you see. And, by the way, you do not see the odometer. You don't see the one that tells you how far you've gone. All you see is the speedometer. And, fortunately, you have a video camera with you, and so you take this video film; you point this video camera at the speedometer and you watch this extremely entertaining vision of this speedometer going like this, you see, back and forth. So, you have this hour of film of this speedometer of this car.

Now, you're captured for an hour, and you know that you're going just on a straight road because it never turns; it's definitely a straight road. And at every moment you've got this video film of the speedometer; at the end of exactly one hour, you're thrown out the side of the road; the van drives away; and there you are at the side of the road.

Well, you've got to be rescued by your compatriots there at Mission Impossible headquarters, and you have your radio. You call them up and say, "I don't know where I am on this road, but I have the video of the speedometer." And so, you send it back electronically to Mission Impossible headquarters and they, then, look at this hour film of the velocity—namely, the speed—of the speedometer during that hour. Now, it's a pretty boring movie, but they're goal is to compute where you are on the road. Now, just think about it. You see, you have the information; knowing how fast you're going at every moment, you have the information that allows you to conclude how far you've gone on the road. And, as we figure out that strategy—the strategy by which you can take this moving picture of the speedometer and deduce from that the distance that you've traveled—that is the second idea of calculus, the integral.

Let's go ahead and do it. Now, what's the best strategy for thinking of, figuring out, difficult issues? The best strategy almost always, certainly in mathematics, is to take a very simple case. We do a simple case, and then we figure it out; we've taught ourselves something, and then we do more complicated cases until we can do more general cases.

So, let's begin with a very simple case. Here is the simple case: Suppose that the speedometer is the most boring possible movie because during the entire time of the hour in which you were sitting in the car, the speedometer never moved. It was exactly at one place the entire time. So, here's an example. Now, given this moving picture of the velocity; that is, the speedometer you see, is telling what was the velocity of that van at every moment of time, that's what the speedometer is telling us. Well, if the speedometer were just completely at the same speed the whole time, this extremely boring movie—sounds like an avant-garde thing, we should think about it. But, we could graph that velocity movie in a very simple way. Namely, at each moment of time we can just make a mark as to what the velocity was reading, that is, what the speedometer was saying. And, if it were completely a constant speed, we would just see this horizontal line, and that would be the velocity of this moving car. [The equation of the line would be $v(t) = c$.]

Now, let's suppose, for example—as before, we will be talking about miles per minute. So, let's suppose the velocity is 1 mile per minute; and it just sat there at the 1 mile per minute velocity and stayed at that one place during that whole hour. Well, how could we compute how far we went at the end of the hour, by the end of the hour?

That's very simple. We would say if we went 1 mile per minute for 60 minutes, then the total distance that we went is 60 miles. We multiply. We multiply the velocity at each time, if it's a constant velocity × the elapsed time, to get the distance that we traveled. So, if it were a constant velocity, we would be in very good shape.

Well, therefore, we've learned something. We've learned the very simplest case. But, we can do harder cases. Once we've taught ourselves simpler cases, we can do harder cases. For example, suppose that our car, the velocity, that is to say, that the speedometer that we were taking on our film, suppose that it didn't smoothly change. Suppose that it went at one speed for a certain amount of time, a certain number of minutes, and then instantly jumped to another speed for another number of minutes, and then instantly jumped to another speed for another number of minutes, and so on. So, let's be specific and look at a specific moving picture of the speedometer that we would want to analyze in order to rescue our compatriot here.

Well, here's an example. Suppose that during the first 10 minutes the speedometer read 1 mile per minute; and then it instantly jumped, during the next 10 minutes, to 2 miles per minute; and then during the next 10 minutes it moved to 3 miles per minute; and so on, 4 miles per minute, 5 miles per minute, 6 miles per minute. And, at the end the 60 minutes, we would have a graph of the velocity that was capturing what this motion picture of the speedometer showed us that would look like this. It's a step function. That is to say, it has a certain fixed value for an interval, and then another fixed value for an interval, and so on. How could we compute the distance that we traveled during that hour so that we know where to go and rescue our friend?

Well, it's easy to compute how far the van went during that first 10 minutes, because it went 1 mile per minute for 10 minutes. The product, velocity × time, is the distance traveled. So, we went 1 × 10; that's 10 miles. During the next 10 minutes we went 2 miles per minute for 10 minutes; so, the distance traveled is the product of the velocity, 2 miles per minute, × the time elapsed, 10 minutes—2 × 10 is 20 miles, and so on.

So, to compute the total distance traveled, we have to add up the distances traveled during each of those 10-minute intervals. So this is very simple, right? This is what we would do if we were tasked with the goal of finding how far that van went during the hour. Velocity × time + velocity × time + velocity × time, and so on, giving us a total of 210 miles during the hour. By the way, this car is really moving. At the last it's going 6 miles per minute, which is 360 miles an hour. But, these people don't mind breaking the law, so that's no problem.

What we have taught ourselves so far is how to compute the total distance traveled if the car had the property that the velocities went for an increment of time at a steady rate and then went at a steady rate and then went at a steady rate. But in real life cars don't do that; in real life, cars smoothly go from one velocity to another velocity. And so, if we looked at our actual picture of the speed of the car, and that is the speedometer, we might see it moving like this, gradually moving up and down or whatever it did. And, so, we've got to extend the strategy that we've so far developed to deal with the situation where the velocity is actually changing. So, let's go ahead and do this in another situation.

Suppose that in this moving picture of our speedometer, the velocity we're given to us by the time always staying at 2 × the time on our watch was the speed. In other words, it started at a stop; at time 0 it was stopped, which was probably good. And, then, as we went forward, at the 1-minute mark it was going 2 miles per minute; and then at the 2-minute mark it was going 4 miles per minute; at the 3-minute mark it was going 6 miles per minute; and so on. That would be a way that the speedometer may behave. How can we compute the distance traveled in this case?

Well, in this case we're not in a position to actually get the exact distance traveled during an interval of time; at least, it's not obvious how to do it, as it was when the velocity was constant. When the velocity was constant, it was clear what we should do. But, in this case, our velocity is varying. So, our strategy in this case is to do two things. One is, we're going to underestimate the distance traveled, and then we're going to overestimate the distance traveled, and then say that the distance actually traveled has to be somewhere between them. So, that's our general strategy. Well, how could we underestimate the velocity—I mean, the distance traveled?

So, one way to underestimate the distance traveled is to compute the distance that another car would have traveled had it always been going somewhat slower than the car that actually we're trying to compute. Well, since we know how to compute the distances traveled by cars going in this "jerky" fashion, going at a certain speed for a fixed time, then jumping up to another speed for another time. Let's consider the following: How about a car, this blue car is the car we're actually talking about, or the van, that's increasing in its speed entirely—at every moment it's increasing in time in speed. It's increasing in its velocity at every moment in time. Now, we could compare that to a car that stays still for a certain amount of time; and then, when this blue car has attained a certain velocity, then the red car begins going at that fixed velocity for an interval of time, and then it goes at the new velocity for an interval of time, and so on. So, the red car is going in this "jerky" fashion, whereas the blue car is continually increasing in its velocity. Let's go ahead and look and see graphically what this might look like.

Graphically, this is a graph of the velocity of the actual moving car. That is, we're assuming that it's velocity at every time t is $2t$. [The equation of the line is $v(t) = 2t$.] Well, these lower horizontal lines represent a movie of another related car whose velocity is always less than the given car. In the sense that, during the first interval of time—and, by the way, we're just going to think of the first 3 minutes of this motion instead of 60 minutes because we're going to think about dividing it up into a lot of pieces and

60 is too big a number. So, we're just going to think about how far this car went in the first 3 minutes of its travel. So our strategy for doing this is to say, let's divide our time interval into smaller intervals of time. In this case, we chose to divide the interval between 0 and 3 minutes into intervals of 1/2 minute. And, during the first 1/2 minute, we're imagining a car that just stays still for the first 1/2 minute; then it jumps up to go at 1 mile per minute for the next 1/2 minute. Now, notice that our car whose velocity is varying at the—prescribed by velocity at each time t is $2t$. At time 1/2 minute, the velocity of our moving car has attained 1 mile per minute. Now, our new car, the one that we're thinking about going slightly slower speed, then stays at that slower speed for 1/2 minute. Then it says, well what is the speed of our original car at time 1 minute? Well, it's going 2 miles per minute. So, our new car, the one that is moving in this "jerky" fashion, stays, then, at the speed of 2 miles per minute for the next 1/2 minute. And then it jumps up the speed that's been attained by the car at the 1-1/2 minute mark, which is 3 miles per minute, and stays at that speed for 1/2 minute, and so on.

Notice that the car that's moving in this "jerky" fashion is always going at a speed that is lower than the velocity of the given car. And, by the way, I have to catch myself, sometimes I'm saying "speed", sometimes I'm saying "velocity." Now, what's the difference? Velocity is speed with a direction. And sometimes we're going to later be talking about the car going around and turning backwards, in which case the velocity will be a negative number. And, so, I should be saying velocity each time, but as long as we're going forward it means the same thing.

Okay. So, here we have a case where the velocity is being approximated by a car that's going in this "jerky" fashion, but is always somewhat slower. Well, look, we know how to compute the distance that's been traveled by this car that's moving in a "jerky" fashion. That's what we taught ourselves before. So, how do we compute that? We say well, the car is going at 0 miles per minute for 1/2 minute; so, it hasn't gone anywhere during that first 1/2 minute. During the next 1/2 minute it goes at 1 mile per minute for 1/2 minute; that's 1/2 mile that it travels. During this next 1/2-minute interval, between time 1 and time 1.5, we're seeing that this car is going at 2 miles per minute for 1/2 minute. That's 1 mile. And so on throughout the intervals of time we see that it will go a total of 7.5 miles; it being the car that's going somewhat slower than the actual car whose velocity is always $2t$.

Well, that's fine, but we can do the same thing by overestimating, making an overestimate of how far the car goes by saying how about if we take a car that moves in the same kind of "jerky" way, but instead we consider it to be moving at a faster rate than the car actually moves during each sub-interval of time. So, in this case, we have a picture here, graphically, that we're considering a car that moves at the highest rate that the actual car attained during this first 1/2 minute, namely, 1 mile per minute during this first 1/2 minute, and then the highest rate that the actual car attained during the next 1/2 minute, namely, 2 miles per minute for the next 1/2 minute, and so on. That would be a car that went definitely further than the actual car during every single interval of time, and consequently, the totality would definitely be a larger number than the actual distance traveled by the car. If we do this computation of just the faster time × the interval + the faster time × the interval, and so on, we get 10.5 miles. That's how far this faster car would go during those three minutes.

Well, what can we conclude from this? We conclude that somehow the actual velocity of the car must lie between 7.5 miles, the distance traveled by the underestimate, and 10.5, the distance traveled by the overestimate. How can we get a better estimate?

Now, by the way, I hope that you now are sort of thinking back, and maybe with annoyance, that this same kind of repetitive strategy is going to happen that happened in the last lecture when we were talking about derivatives. How did we talk about derivatives? We said, well, if we take a smaller interval of time around the time of interest, we will get a different approximation that will be a better approximation of the actual instantaneous velocity. And now we're doing a similar thing. We're saying okay, we have a method for getting an overestimate of the distance traveled, and an underestimate of the distance traveled; we know the actual distance traveled is sandwiched between those two numbers.

How could we get a smaller sandwich? How could we get a smaller interval in which the actual distance has to appear? How? We simply take finer intervals of time. We divide our interval from time 0 to time 3 into tenths of a minute; tenths of a minute. Now, this is getting very laborious, but here's what we can do. We can take the interval of time between time 0 and time 3 and divide it into tenths of a minute. We have 30 intervals now. For each interval we can say we know what its initial time was and its final time was on our watch, so we know what the initial velocity of the car was and what the final velocity of the car was; it's going at $2 \times t$. So, if we want an underestimate of the distance traveled, we can assume that we have—we imagine we have a car that stays constantly at the lower speed for each of those 30 intervals of time. And, if we did that, we would have a long addition problem that involved 30 multiplications and additions, multiplying the velocity × the time + the velocity × the time + the velocity × the time, and this is capturing a picture of the car moving at little step values of velocity, but 30 of them this time.

Likewise, we can get an overestimate in the same way, dividing in these 30 intervals we can take a velocity that's slightly faster than the car actually moved during that 1/10 of a minute. And, if we do so, then during, for example, this first interval of time, the fastest—the first 1/10 of a minute, the fastest the car ever went during that 1/10 of a minute was .2 miles per minute; right at the end, that was its velocity. And so we assume, we say "What would have happened if a car had gone at that faster velocity during that first 1/10 of a minute? And the fastest velocity the actual car went during the second 1/10 of a minute?" And we added up those distances traveled, we would get a value of 9.3 miles. Therefore, we know that the actual distance the car must lie between our underestimate of 8.7 miles and our overestimate of 9.3 miles. So, you see, we're getting a narrower sandwich, a narrower window in which the actual distance must reside. Well, now you're getting the philosophy of the integral.

So, if we want to approximate the actual position of the car, the distance traveled by the car during that 3-minute interval, we could imagine dividing the interval of time into incredibly small intervals. And for each small interval, we could take the actual velocity of the car, the graph of its actual velocity, and imagine breaking the time interval between 0 and 3 minutes into very, very fine increments that we call Δt, and then for each one we could compute an overestimate by multiplying the velocity of an overestimate × the interval of time Δt—each width is Δt. So, we have the velocity overestimate × Δt, velocity overestimate × Δt, and so on. And, as we move along, you can see the highlight of where we are and the contribution to this summand of the distance traveled in each one as we add up this long, laborious addition process. We notice that the total distance, then, will actually come out to be sandwiched between an overestimate and an underestimate of this laborious addition process. [Total distance = $(v(t) \times \Delta t) + (v(2\text{t}) \times \Delta t) + (v(3\Delta\text{t}) \times \Delta t) + \ldots$ as Δt becomes small.]

As Δt becomes small, this addition process will become closer and closer to the actual value, you see, because at each moment of time if we have a step function that has just a little, tiny interval of time—like 1/100 of a minute—the actual car's velocity will not vary much at all between the step function velocity that's jerky but happening every 1/100 of a minute is not going to be much different from the actual velocity that changes slightly during those 1/100s of a minute intervals.

So, once again, what we're doing is we're devising a strategy for finding the totality of the distance traveled that is associated with a process of an addition problem, multiplication and addition, and then taking finer and finer increments of time gets us to the exact answer. Once, again, no one of our computations will give us the exact answer. Instead, the exact answer for how far we've traveled is a process of doing something infinitely many times. Namely, dividing the interval into small increments, multiplying how far you went during each of those small increments, if you stayed at a steady rate; and then taking the limit of those answers as we take finer and finer increments.

And, by the way, one of the concepts of calculus that is at the heart of both the derivative and the integral is this limiting process, and we'll see and talk about it a little bit in the next lecture, that this was one of the very big conceptual challenges for mathematicians for more than 100 years after calculus was invented, because this concept of coming down to one number that is approximated finer and finer by these approximations by taking smaller and smaller intervals, you know, what does it really mean? It's really quite a subtle kind of idea. But, we can see conceptually it makes a great deal of sense. If we wanted to figure out how far that person is one hour after he was kidnapped, the way we could do it is to divide our interval into tiny pieces and we'd see how far the van actually went.

Now, let's do some other kind of computation on this idea, and that is we were talking about the distance between time 0 and time 3, and we got the finest numbers actually computed told us that the actual distance traveled had to lie between 8.7 miles and 9.3 miles. If we had done that same laborious computation for 1/1000 of a minute, and then 1/10,000 of a minute, and so on, we would find that the sandwiching of those upper and lower estimates, over and under estimates, was getting closer and closer to 9 miles. And, so, in fact, the car went exactly 9 miles during that interval of time, between time 0 and time 3 minutes.

But, we might want to ask ourselves the question, can we think in a more—instead of always going to 3 minutes, suppose we looked at how far the van went between 0 minutes and, say, 1 minute? And how far did it go between 0 minutes and .5 minutes? And in each time we did this entire laborious process of dividing the interval up into little tiny pieces and adding them up, we would find that we have this same kind of a chart similar to the chart that we saw in the last lecture about derivatives; namely, there's a pattern here. If we look at the total distance traveled in the first 1/2 minute, we traveled .25 miles; that is, 1/4 of a mile. In the first minute, we traveled 1 mile; in the 1.5 minutes, we traveled 2.25 miles; in 2 minutes, 4 miles; in 3 minutes, 9 miles. If you notice the pattern, the pattern is that the distance traveled is just the square of the time elapsed between time 0 and the time that we're measuring. So, once again, once we see this pattern, we're in a position to instantly say how far that car went in any amount of car because we would just say between time 0 and time whatever it is, say, 4 minutes, you just square the 4 and that's the answer.

That is the strength of calculus; is that we're going to see that the formulas, for things like the velocity, give rise to a formula for the position of the car at each time. It's telling us something that could be computed by this very laborious process, and it's important to realize that the answer we get is the result of doing the laborious process whose result we can interpret as a meaningful thing; namely, how far the car traveled. That was a natural thing for us to do, and now we're seeing that it comes out to be a formula that makes it easy to compute that distance traveled.

So, the word of the day is the integral; and the integral is the, if you have a velocity function, then between any time a and a time b, we can compute the distance traveled by the car, the net distance traveled, by computing its velocity at time $a \times \Delta t$ + its velocity at just a slight moment later $\Delta t \times \Delta t + a + 2 \times \Delta t \times \Delta t$, and so on—this very long addition problem for any individual choice of Δt gives us an approximation of the distance traveled. [The integral of a function $v(t)$ between time a and time b is $v(a)\Delta t + (v(a+\Delta t) \times \Delta t) + (v(a+ 2\Delta t)\ \ \Delta t) + (v(a+3\Delta t) \times \Delta t) + \ldots (v(b) \times \Delta t)$ as Δt becomes very small.] If we take our Δt's increasingly small, that approximation gets finer and finer, and then, as we say, in the limit, we get this single answer. And that is the definition of the integral from a to b of this velocity function $v(t)$.

So, at this point then, we have introduced the two basic ideas of calculus: in the last lecture, the derivative, and in this lecture, the integral. Both of them associated with a car moving on a straight road. Because of the fact that we introduced them both in terms of a car moving on a straight road, in the next lecture we're showing the connection; we'll show the connection between the derivative and the integral, and see in what sense those are inverse processes of one another. So, in the next lecture I'll look forward to telling you about the Fundamental Theorem of Calculus. See you then.

The Fundamental Theorem of Calculus
Lecture 4

You can't take a derivative by just doing one thing and you can't take an integral by just doing one thing; they're both doing infinitely many things, dividing the time integral into finer and finer bits in order to get better approximations, and then taking a limit to get an actual one answer.

The Fundamental Theorem of Calculus makes the connection between the two processes discussed in the previous two lectures, the derivative and the integral. Again, this theorem can be deduced by examining the generative scenario of the moving car. The derivative and the integral involve somewhat complicated procedures that appear unrelated if viewed in the abstract; however, they accomplish opposite goals—one goes from position to velocity, the other goes from velocity to position. The duality between the derivative and integral is exactly what the Fundamental Theorem of Calculus captures. This convergence of ideas underscores the power of abstraction, one of the global themes of this series of lectures.

The two fundamental ideas of calculus, namely, the methods for (1) finding velocity from position (the derivative) and (2) finding distance traveled from velocity (the integral), involve common themes. Both involve infinite processes. Both processes involve examining a situation with increasingly finer time intervals. Both processes involve deducing a single answer from the whole infinite collection of increasingly accurate approximations.

We looked at the same situation—a car moving on a straight road—from two points of view. Knowing where we were at every moment, we deduced our velocity at every moment—the derivative. Knowing our velocity at every moment and where we started, we deduced where we were at every moment—the integral. The two processes are two sides of the same coin. Understanding implications of this relationship between these two processes is *the* fundamental insight of calculus. Indeed, it is known as the Fundamental Theorem of Calculus.

Specifically, suppose we are given a velocity function; that is, we are told how fast we are traveling at every instant between two times. We can find the distance traveled in a given time interval in two ways. First, we can compute the integral (by dividing the time interval into little pieces and adding up distance traveled over the little pieces). Second, if we know a position function whose derivative is the given velocity function, we can simply use the position function to tell us where we are at the end and where we were at the beginning, and then subtract the two locations to see how far we went. Method 2 is a lot quicker. It does not involve an infinite number of approximations as the integral does. These two ways of computing the distance traveled give the same answer. That's what makes the Fundamental Theorem of Calculus so insightful—it gives an alternative method for finding a value that would be hard or impossible or, at best, tedious to get, even with a computer.

The moving-car scenario presents a situation to analyze the Fundamental Theorem. Let's do so where the position function is $p(t) = t^2$, and the velocity function is $v(t) = 2t$. We can find the distance traveled from time 1 to time 2 via the integral. Remember all the sums involved: $(v(1) \times \Delta t) + (v(1+\Delta t) \times \Delta t) + (v(1+2\Delta t) \times \Delta t) + \ldots + (v(2 - \Delta t) \times \Delta t$, as Δt becomes increasingly smaller. Then we can find the distance traveled knowing that $p(t) = t^2$ by subtraction. Because $p(2) = 4$ and $p(1) = 1$, the car traveled 3 miles. We can consider other pairs of time values. The summing process of the integral will always yield the same result as just subtracting the positions, because all these processes are referring to the same scenario of a moving car.

Suppose someone told us that the velocity of a car moving on a straight road at each moment t was $v(t) = 2t$ but didn't tell us the position function and asked for the distance traveled in the first 3 minutes. We could either add up pieces via the integral process or we could find the position function and subtract. The fact that both processes yield the same answer is the importance of the Fundamental Theorem of Calculus. The process of the integral (summing up pieces) tells us that the answer is what we want to know. That process refers directly to the commonsensical way of finding the distance traveled given the velocity. If we can find a position function, it is much easier to just subtract.

The fundamental insight relates the derivative and the integral. The process of finding instantaneous velocity from position is the inverse of the process of finding position from velocity. We have two ways that give the same answer to the question of how far we have gone given the velocity that we have been traveling. One is by the infinite process of adding up (the integral). One is by finding a function whose derivative is the velocity function (thus, it is a position function) and subtracting. The insight is that both processes give the same result. From an arithmetic point of view, the Fundamental Theorem notes that the process of subtraction and division that is at the heart of the derivative is the opposite of the process of multiplication and addition, which is at the heart of the integral. This insight has many other applications, which we will see in future lectures.

Supporters of Newton and Leibniz had a lively and acrimonious controversy about who developed calculus first.

The development of calculus was an incremental process, as we saw when we spoke of mathematicians before Newton and Leibniz. Newton and Leibniz systematized taking derivatives and integrals and showed the connections between them. The development of calculus, however, was involved in considerable controversy. One type of controversy concerned who should get the credit for calculus, Newton or Leibniz. The second type of controversy concerned the validity of the ideas underlying calculus, particularly the tricky business involved in taking limits. Let's take a few minutes to talk a bit about each of these controversies.

Supporters of Newton and Leibniz had a lively and acrimonious controversy about who developed calculus first. Newton was quite averse to controversy, and this aversion made him reluctant to publish his work. Newton developed the ideas of calculus during the plague years of 1665–1666 when Cambridge was closed, but he did not publish those results for many years, in fact, not until 1704, 1711, and posthumously, in 1736. He did, however, circulate his ideas to friends and acquaintances in the 1660s. Leibniz was the first to publish his results on calculus. He conceived the ideas in 1674 and published them in 1684. In 1676, Newton learned that Leibniz had developed calculus-like ideas. Newton staked a claim on his priority in the invention of calculus by writing a letter to Leibniz. In this letter, Newton indicated his previous knowledge of calculus by writing an anagram.

The anagram consisted of taking all the letters from the words of a Latin sentence, counting them, and putting all the letters in alphabetical order, as follows: "6a cc d æ 13e ff 7i 3l 9n 4o 4q rr 4s 9t 12v x." The sentence was, "*Data æquatione quotcunque fluentes quantitates involvente fluxiones invenire, et vice versa*." This means, "Having any given equation involving never so many flowing quantities, to find the fluxions, and vice versa." Even the English version is not much of a hint of calculus. Some British supporters of Newton felt that Leibniz got the idea of calculus from Newton's manuscript during a visit that Leibniz made to England in 1674. Newton's supporters hinted at foul play in 1699. Modern historians believe that Newton and Leibniz independently developed their ideas. In any case, the controversy led to a downhill trend in relationships between the supporters of the two men. The controversy had a bad effect on British mathematics for a long time.

Newton, on limits

But the answer is easy, for by the ultimate velocity is meant that with which the body is moved neither before it arrives at its last place, when the motion ceases, nor after, but at the very instant when it arrives. And, in like manner, by the ultimate ratio of evanescent quantities, is to be understood the ratio of the quantities not before they vanish, nor after, but that with which they vanish.

The other controversy associated with calculus involved its validity. One thing that we have to understand is that, at the time of Leibniz and Newton, ideas that we consider absolutely fundamental to even starting to think about calculus today simply did not exist at all, for example, the idea of function. Most vague, though, was the idea of the limit. Neither Leibniz nor Newton had firm ideas of the limit. The concept of the limit was not resolved until the mid-1800s. ■

Suggested Reading

Any standard calculus textbook, section on the Fundamental Theorem of Calculus.

Boyer, Carl B. *The History of the Calculus and Its Conceptual Development*.

Questions to Consider

1. The derivative of the position function $p(t) = t^4$ yields a velocity function $v(t) = 4t^3$. Given that fact, use the Fundamental Theorem of Calculus to compute the distance a moving object will have traveled between time 0 and time 2 if its velocity at each time t is $4t^3$. If you like, you can check your answer by using the definition of the integral to compute the distance traveled.

2. We are here at one moment and there at another time. Thus, we know how far we traveled. Now let's look at this scenario dynamically, namely, how did we get there? Explain how the Fundamental Theorem of Calculus shows the connection between the dynamic and static views of the world.

The Fundamental Theorem of Calculus

Lecture 4—Transcript

Welcome back to *Change and Motion: Calculus Made Clear*. We've, in the last two lectures, been introduced to the two fundamental ideas of calculus, the derivative and the integral. Those two fundamental ideas are for the purpose of finding the velocity from the position function; if you know where you are at every moment, how do you find your instantaneous velocity? That was the derivative. And, two, finding the distance that you traveled if you know the velocity that you're moving at every moment of time; that was the integral. And these two fundamental ideas involved some common themes. First of all, they both involved infinite processes. You can't take a derivative by just doing one thing and you can't take an integral by just doing one thing; they're both doing infinitely many things, dividing the time integral into finer and finer bits in order to get better approximations, and then taking a limit to get an actual one answer.

So, both processes also involve this concept of taking a scenario of a car moving down a straight road. They were introduced that way. And in fact, we were looking at this scenario of a car moving down a straight road, but we looked at it from two points of view. From the one point of view, we said if we know where the car is at every moment of time, we can deduce how fast it's going; what its velocity is at each moment. And on the other hand, in the last lecture, we saw that if we had a picture of the speedometer, that is to say, we knew the velocity at every moment of time, and we know where we started, then we could compute where it is that that car would end up. That was the process of the integral.

Well, realizing that both of these processes involved the same car moving down the same road tells us that these two processes are two sides of the same coin; that they somehow are both reflecting one common underlying reality, that car moving down the road. Well, when we understand the implications of that relationship between the two processes, that insight is the fundamental insight of calculus. In fact, it's called the Fundamental Theorem of Calculus. It's putting those two ideas together.

So, let's think, specifically. Suppose that we're given a velocity function; that is to say, we're thinking about a car that's moving on a straight road, and at every time we're told what the velocity is of that moving car. Well, we can find the distance traveled in two different ways, okay? Here are the two different ways we can find the distance traveled. First of all, we can use the method of the integral. The method of the integral is to take the velocity at every moment and dividing it up into little moments of time and seeing how far we traveled during each little interval of time, and adding them all up to get the total distance traveled. But, there's another way that we could get the distance traveled. Suppose that we knew the position function of this car moving on a straight road. In other words, at every moment we knew where it was now, and where it was then. Then our strategy of computing how far we went between one time and another time would be much more straightforward. All we would do is we'd say well, where did we end up? Where did we begin? Subtract.

Well, that second method—the where are we at the end; where are we at the beginning; and subtract—is a lot quicker because it doesn't involve that infinite process of dividing the time into little bits and doing these infinite number, I mean, many multiplications and then adding them together and then taking a limit. Those are two different ways of getting the same answer; namely, the distance traveled.

Well, now what we need to do is to say what the Fundamental Theorem of Calculus gives us. It gives us an insight for an alternative way of computing a value that would be difficult, or maybe even impossible, to compute—at best it would be tedious, even with a computer—to compute the distance traveled by that integral method; by taking little intervals of time and adding them together; that would be a long, laborious kind of process.

But, this moving car scenario, therefore, gives us a situation in which we can present the Fundamental Theorem of Calculus so that it is very clear. So, let's go ahead and see if we can actually understand it in a particular situation. Namely, suppose that we're in the situation of a car moving on the road where at every moment of time the position function is given by t^2. This was the scenario that we analyzed in Lecture Number Two. The speed was given at every time t, that is the position, not the speed, not the velocity, the position at every time t was given by t^2. We looked at our watch, we squared it to see the position of the car at that moment; and we saw that a way to compute the instantaneous velocity was to look at this characteristic procedure of taking a difference in position divided by the difference in time, and computing that as this Δt, the difference in time, became smaller and smaller. And, we discovered in taking, for example, the velocity—instantaneous velocity—at time 1; we did these computations; all of which told us that the instantaneous velocity of time 1 was 2 miles per minute.

We went on to compute the instantaneous velocity at many different times, and we saw that there was a relationship if the position function at every time was t^2, then the velocity function at each time was $2 \times t$. And this chart sort of summarizes some examples of values. Sometimes when we looked at our watch, if our position were t^2, we'd see that the speedometer would be reading $2 \times t$. So, here's t; here's the position; here's the velocity at each time.

So, the derivative was a process by which we took a position function and deduced the velocity function. The integral, on the other hand, was the inverse process. It took the velocity function and gave us a position function. Or, actually, it gave us the net distance traveled during a particular time. We'd have to know where we started to see what the position was.

So, here is a little chart that summarizes the values for the function $p(t) = t^2$, and the velocity, $v(t) = 2t$.

So, let's go about computing the distance traveled between two specific moments of time. The time $t = 1$ and the time $t = 3$, just to pick arbitrary numbers. Well, looking back at our chart it's very easy to see what—this is at time $t = 1$, the position is 1 on the road. We look out the window it says we're at mile marker 1. At time $t = 3$, we look out the window and it says we're at mile marker 9. So, it's very easy to compute the distance traveled between time 1 and time 3. Namely, it's $9 - 1 = 8$. So, the distance traveled is 8.

On the other hand, the distance traveled could be computed in an alternative way. The distance traveled could be computed by doing this process of taking the infinite—infinitely dividing the interval of time between time 1 and time 3—and looking at this characteristic sum of the speed × the time interval traveled, adding it together to the next approximation, the next approximation, so that we can use this definition of the definite integral to get the total distance traveled. So, the Fundamental Theorem of Calculus tells us that if we have a position function, it entails the existence of a velocity function. That's what the derivative does. If we know the position function that generates the velocity function, then we know that the integral process, all that multiplying and adding process that gives us the integral that will definitely tell us the distance traveled between the time we start and the time we end up, we can see that that would also be given to us by taking the position function that generated that velocity function and just plugging in the final value minus the beginning value to find the net distance traveled.

Okay. Let me see if we can do another example here. Suppose that we are traveling with a velocity function given by $v(t) = 3t^2$. Now, what that means is that every time we look at our watch and we look at our speedometer and we see that the velocity given by that speedometer, or velocitometer, is given by three times the time told on our watch squared. Now, suppose that we set ourselves the challenge of saying how far did we travel between time $t = 1$ and $t = 4$? Well, there are two methods to do it. The two methods to do it would be 1) to use the definition of the integral, the integral process, because, remember, that was the natural way to compute the distance traveled. We broke the interval up between 1 and 4 into small intervals of time; we approximated how far we went during each of those small intervals of time, and added them up. And then we took finer and finer intervals of time and took a limit to get one single answer. That was our strategy for figuring out the distance traveled. So, that definitely gives us the distance traveled.

On the other hand, another way we could find that same answer would be to say can we find a position function so that the derivative of the position function—that is, the derivative, remember, is giving us the velocity at each time. Well, if we could find a position function whose derivative is the velocity function with which we were first trying to deal, then we could use this much simpler way of finding the distance traveled by just saying, where does the position function tell us we are at the end, at time $t = 4$, and where does it tell us we were at time $t = 1$, and just subtract.

So, that is the beauty of the Fundamental Theorem of Calculus. It lets us avoid this infinite addition kind of process—it's infinite because we have to use smaller and smaller intervals—and, instead, if we can find a function, in this case the function is $p(t) = t^3$, with the property that the instantaneous velocity at every moment for a car that's at position t^3, its instantaneous velocity at every moment is $3t^2$. Knowing that, we know that the total distance traveled is the position at time 4 minus the position at time 1, and that's the answer. The position at time 4 is 4^3— $4 \times 4 \times 4$, that's 64, $- p(1)$ is 1^3—that's 1—so the total distance traveled is 63. And, if we were to do this long, laborious addition process of the integral, we would get 63. So, that is the insight of the Fundamental Theorem of Calculus. It shows that the derivative and the integral are inverse processes to each other.

So, notice that what we've seen here is that the infinite process of adding things up, which is the integral, gives us this distance traveled; and the other was this derivative process, that is, finding the velocity function from the position function, is a question of subtracting and then dividing. So, the insight is that both these processes give the same result. That's what I've just been saying. But, think of it from an arithmetic point of view. So, the Fundamental Theorem notes that the process of subtraction then division that's the heart of the derivative gives us the opposite thing to the process of multiplication and addition, which is the integral. So, this insight has many, many applications because what it allows us to do is to take this integral process, which is laborious but it tells us what we want to know—in this case how far the car went—and we'll see that it tells us all sorts of other things that we want to know, but that we can actually compute the answer by finding a function whose derivative is the thing that we want to add up in that laborious integral kind of way.

Well, at this point, I thought it might be a good break for us, now that we've seen the three fundamental ideas of calculus. We've seen the derivative and the integral, the two fundamental ideas of calculus are those two; and, then, the connection between them is the Fundamental Theorem of Calculus. So, we can sort of take breather here and declare a victory over these really wonderful ideas. And, in all of the future lectures, we are going to see how rich these ideas really are because they apply to so many things that can be interpreted in so many different ways. But that's for the future lectures.

I thought right now would be a good break for us to talk a little bit about the history.

We've already shown that the history of the calculus, of the development of the calculus, was definitely an incremental process. There were ancient roots to it; Eudoxus and Archimedes both used processes that were very reminiscent of the integral. Then Fermat and Isaac Barrow and many others developed ideas that were close to the derivative. And it was Newton and Leibniz who actually systematized the taking of derivatives and integrals; and they were the ones who showed the connection and pointed out the connection that we just saw in the Fundamental Theorem of Calculus.

This idea of calculus and the Fundamental Theorem is really a wonderful accomplishment, and in fact, it's been celebrated in many different ways. One of them is a couplet from a poem by Alexander Pope, which is I think really a wonderful tribute to Newton. It says the following: "Nature and nature's laws lay hid in night; God said, 'Let Newton be,' and all was light." So, there's a tribute to Newton. I wanted to say what Leibniz had to say about Newton. Leibniz was the other co-inventor of calculus, and he had a comment to say about Newton. I want to tell you that I've been on many committees in the mathematics department to hire people or to promote people, and you get these letters of recommendation, of course, about people. Letters of recommendation often are very glowing. They say "Oh, this is the best—one of the three best people in the world." In fact, there are many people who are among the three best people in the world, it turns out, even though they're math… But, here is a letter of recommendation that you don't read every day, and if I were on a hiring committee and I read this one, it would certainly get their attention. So, this is what Leibniz had to say about Newton. He said, "Taking mathematics from the beginning of the world to the time when Newton lived, what he has done is much the better part." So, there's an amazingly comprehensive assessment of Newton's contributions.

But, we don't have to go back to the time of Newton and Leibniz to see their—how important these contributions are. There's a book called, *The Hundred Most Influential Persons in History*. It's sort of amusing. This author just decided to write down what he thought were the hundred people who contributed the most in history. And on that list, number two is Isaac

Newton, because of the calculus and, of course, his laws of physics, many of which were related.

But, the development of the calculus actually received considerable controversy. It was involved in a lot of controversy, and the controversies were really of two types. One type was the controversy about who should get the credit for calculus; should it be Newton or should it be Leibniz? And the second kind of controversy about calculus concerned the validity of the ideas that underlie calculus itself; and, in particular, the tricky business that's involved in this taking of limits. That was a very big conceptual obstacle for people. And so, what I'd like to do is just take a few minutes and talk about each of these two controversies.

We'll begin with the personal one about credit. It turns out that Newton was, apparently, sort of pathologically averse to controversy, ironically. And in fact, it's partly because of his aversion to controversy that he was embroiled in possibly the biggest controversy concerning priority—credit—for a discovery of any controversy in the history of mathematics and, maybe, science. And, in a sense, it was cause and effect because his aversion to controversy made him extremely reluctant to publish things. So, he would come up with ideas—and he would write them down, by the way—but he wouldn't publish them; he wouldn't make them public. So, then, the question of who actually came up with ideas first was sort of problematical, as we'll begin to see.

Newton actually developed the basic concepts of calculus during the middle of the 1660s. He was a student at Cambridge University, and it closed for two years, in 1665 and '66, owing to the plague. So, during that time Newton went to his aunt's farm and spent these two years thinking. During that time he developed these seminal ideas, his seminal ideas not only of calculus, but also of physics. Of course, as a professor it does make me think, "How well would our students do if we simply close the school and let them go back and think for themselves?" But, that's another story. But, in 1669, Newton actually wrote a paper on calculus, but it wasn't published; he didn't publish the paper, he just wrote it. In 1671, he wrote another paper on calculus; didn't publish it. He wrote another paper in 1676 and didn't publish it. In fact, all three of these papers were eventually published. The one he wrote in 1669 was published in 1711; that's 42 years after he wrote it. The one he wrote in 1671 was published in 1736, and that's nine years after he died. And the paper he wrote in 1676 was published in 1704; it was actually a part of the appendix to his *Opticks.*

But he did talk about the calculus somewhat in the *Principia*, which was written in 1687, but the arguments that he gave in the *Principia*, for his physical arguments, he basically discovered them using the methods of calculus, and then translated them into the much more laborious and old-fashioned methods of mathematics that were around before calculus, because calculus was not, at that time, a normally accepted method. So, none of his works on calculus were published when he developed those ideas. But, we know that he did develop those ideas because he wrote them and circulated those ideas to his friends and acquaintances. So, he definitely had developed the ideas of calculus in the 1660s; and, he used the techniques of calculus in all of his scientific work, and they appear in the *Principia.*

Well meanwhile, Gottfried Wilhelm von Leibniz independently invented calculus. And, his invention of calculus, he claims to have been sometime in the middle of the 1670s, so people think that it was probably somewhere in the neighborhood of 1674 that he got the ideas of calculus, and he published them in 1684; that's 10 years after he got the ideas. But, notice that 1684 is three years before Newton's first account of calculus appeared at all; his first account being in the *Principia.* So, I wanted to read you the title of Leibniz's paper on it and a little commentary on about how well it was received. Leibniz wrote his memoir on calculus; it was six pages long and it appeared in *Acta Euroditorum* of 1684. The title of his paper is this: "A New Method for Maxima and Minima, As Well As Tangents, Which Is Not Obstructed by Fractional or Irrational Quantities." One of the things that both Newton and Leibniz did were to generalize methods that generally had existed about derivatives and so on to involve more general classes of functions, and so, this was what he advertised in this paper. But, I thought it would be amusing for you to hear how it was received. The Bernoulli brothers were—I mentioned that there was a family of eight Bernoullis who did a lot in developing calculus, particularly on the Continent, and one of the Bernoulli brothers had this to say about Leibniz's paper of 1684. He said it was—it being the paper—was "an enigma rather than an explication."

Apparently, the paper was extremely difficult to make any sense of. For one thing, it contained many misprints; and, in the other, it used this strategy of exposition that is, unfortunately, still too common in mathematical circles of trying to be extremely terse and consequently not explaining how the ideas came about.

But, then, we come to the question of the controversy about the priority, about who came up with these ideas. When Newton began to realize that Leibniz had the basic ideas of calculus—and Newton began to be aware of this in the 1670s, in the middle of the 1670s—Newton's response, he wanted to make sure that he got credit for calculus. So, he wrote a letter—and he actually wrote it—it eventually went to Leibniz, but it went to Oldenburg first in 1676, and Newton wanted to indicate his previous knowledge of calculus. The way he did it was he took a Latin sentence, and then he scrambled the letters; in other words, he made an anagram of the letters in this one Latin sentence. So the anagram consisted of removing all the letters and just putting them in order. So, here is what Newton wrote. He said there were six As, two Cs, a D, 13 Es, two Fs, seven Is, and so on; and these were the letters of a sentence. Now, the sentence that he wrote was the following. It says "*Data æquatione quotcunque fluentes quantitates involvente fluxiones invenire, et vice versa*," which means, "Having any given equation involving never so many flowing quantities, to find the fluxions, and vice versa."

Now, even to a mathematician, this means very little. This sentence—it encapsulated Newton's thinking about derivatives, but it is a little bit obscure. And, when you read it, it doesn't mean that much. But, it does capture his concept of the idea of the inverse property of the derivative and the integral, fluxions was his word for the derivative. So, he tried to establish his priority in that fashion. But, then, later, some proponents of Newton made accusations that Leibniz had actually read some of the manuscripts of Newton's before he got his ideas.

Well, modern historians believe that the two inventors were independent—Leibniz and Newton independently thought of calculus, and you can see that it is incremental and it came about close to other work. But, Leibniz had published first, so people who sided with Leibniz said that Newton had stolen the ideas from Leibniz, and it became just a huge mess. It embroiled the British mathematics in opposition to mathematics in the Continent, and the two marked camps didn't talk to each other much and it really impeded the development of calculus in Britain compared to the Continent. So, it really was a sort of sad thing.

Well, there was another controversy associated with calculus, and that had to do with the validity of the reasoning. Now, you have to understand that the way we understand calculus at this time is, you know, fairly clear, and the things that we think of as absolutely fundamental to even starting to think about calculus didn't exist for them. They didn't have a clear idea of function. We're going to be talking about functions and graphs of functions; that wasn't something that they had a clear idea of until the 1690s, until after calculus was invented. But far more vague than that was the idea of limit. Remember, the idea of limit for both the integral and the derivative was taking these smaller and smaller things, getting approximating numbers, and then seeing that those numbers came to one value. The problem was that neither Leibniz nor Newton could possibly have—and did not have—a clear idea of this concept of limits. I wanted to read you a quote about Newton, talking about limits that might show you the clarity with which he mentioned them. So, here's Newton on limits. He says—he was asked about the meanings of the terms evanescent quantities and prime and ultimate ratios. Evanescent quantities—these are sort of disappearing quantities. And, he said the following:

But the answer is easy, for by the ultimate velocity is meant that with which the body is moved neither before it arrives at its last place, when the motion ceases, nor after, but at the very instant when it arrives. And, in like manner, by the ultimate ratio of evanescent quantities, is to be understood the ratio of the quantities not before they vanish, nor after, but that with which they vanish.

You see, this is the idea of the limit, and he didn't have this very clearly in mind.

Well, people derided Newton and Leibniz for talking about infinitesimals, which they did, little, tiny infinitesimally small amounts of time, and about—there was a phrase that was used to ride them, calling Newton's quantities—

evanescent quantities—of the ghosts of departing quantities. You see, these things were not at all clear, you see?

There was a particular opponent to calculus by the name of George Berkeley, who wrote an attack on calculus called *The Analyst,* and I just wanted to read you the title of this just because it's fun. Here's what it says. The title is "The Analyst," but the subtitle is the fun part. It says, "The Analyst; or A Discourse Addressed to an Infidel Mathematician."—and by the way, this referred to Newton's friend, Edmund Halley, not to Newton himself, who was a very devout religious person—"Wherein it is Examined Whether the Object Principles and Influences of the Modern Analysis Are More Distinctly Conceived or More Evidently Deduced Than Religious Mysteries and Points of Faith. 'First Cast the Beam Out of Thine Eye, and Then Shalt Thou See Clearly to Cast Out the Moat Out of Thy Brother's Eye.'" That was the subtitle; apparently longer titles than occur in this day and age.

In any case, the concept of the limit wasn't actually resolved until the middle of the 19th century, in the 1850s. So, it really was a great challenge. People didn't have a good idea of what the real numbers were, and they didn't have a good idea of the limit process until that much later period. So, there really was a substantial issue involved in understanding the basis of calculus.

Well, in the next series of lectures we'll return to the mathematics and talk about the derivative and see it in various applications, particularly graphically, and then algebraically. I look forward to seeing you then.

Visualizing the Derivative—Slopes

Lecture 5

Today, what we're going to do is begin a series of three lectures about the derivative as it relates to and reflects its presence in different manifestations. Today we'll talk about it's relation to graphs, graphical relationship.

Motion and change underlie our appreciation of the world, both physically and in many other realms. Change is so fundamental to our vision of the world that we view it as the driving force in our understanding of most anything. Frequently, the dependency of one variable on another is most easily described visually by a graph, for example, a graph that shows position as a function of time. The concept of the derivative provides a method for analyzing change. We explore the relationship between the graph of a function and its derivative. For example, we observe that an upward-sloping graph signals a positive derivative. Superimposing the graph of the derivative on the graph of a function reveals a visual relationship between a function and its rate of change.

In this lecture, we will look at the derivative and its relation to graphs. Graphs show a relationship between two dependent quantities. For example, when referring to our moving car, our interest in the derivative is to try to understand how the change in time affects the change in the position of the car.

Change through time is of fundamental interest in many settings. Physical motion is change in position over time. Understanding such change over time is important. Cars moving, the idea of velocity, is a basic example. In biology, we can consider the change in human population in the world from 1900–2000, for example. In economics, we consider changes over time to prices, employment, production, consumption, and many other varying quantities. With a graph, we can show the changes in the Dow Jones Industrial Average over the last century.

Understanding many important issues involves analyzing change in a characteristic over time. A graph can display the power output of the Chernobyl nuclear power plant on April 25–26, 1986. The steepness of the curve at 1:23 a.m. tells us how quickly the power output changed in a short amount of time, leading to the Chernobyl disaster. These changes can be visualized using graphs.

Let's again analyze the velocity of a car given its position on a straight road. Let's look at the position graph and see how the velocity is related to the graph. First, looking at any position, we can figure out how fast the car is going. From the graph, we can deduce features about the motion of the car. If the graph is going up, the car is moving forward. If the graph is going down, the car is moving backward. The top or bottom of a graph means the car is momentarily stopped. Its velocity is 0. A steep graph corresponds to high velocity. A straight line means constant velocity—horizontal ↔ stopped; 45° upward to the right ↔ velocity of 1 mile per minute forward; 45° down to the right ↔ velocity of 1 mile per minute backward.

Rate of Change of a Function

This ratio—this vertical change in position, divided by the elapsed time—was the ratio that we discussed in the definition of the derivative; and that ratio is also the slope of the straight line that we could draw between these two positions of the car.

$$\frac{p(t+\Delta t)-p(t)}{\Delta t}$$

Let's be more quantitative in our description. The steepness of the graph corresponds to the velocity. How exactly does it correspond? A straight-line position graph corresponds to a velocity equal to the **slope** of the line, where the slope is a quantitative measure of the steepness. Slope of a line is just the ratio of upward motion over sidewise motion, or the vertical change divided by the horizontal change. A straight line going upward to the right has a positive slope. A straight line going downward to the right has a negative slope. A horizontal line has slope 0.

Consider now motion with varying velocities. Recall the idea of the derivative and how that process gave the velocity. Let's see it with a curved graph of position for a car with varying velocities. We look at nearby values and draw a straight line. The slope of that line is the change in position divided by the change in time. As Δt becomes increasingly smaller, those straight lines converge to a **tangent line** whose steepness is telling us the instantaneous velocity. Let's magnify the graph. Magnified, the graph looks more like a straight line, and the tangent line and the graph appear to coincide. In general, the derivative at a point gives the slope of the line we would see if we magnified the graph. Derivative, then, gives two equal quantities: (1) the velocity of the car and (2) the slope of the tangent line. An important concept to remember is that smoothly curving lines, when viewed very close, look like straight lines. That is why the Earth looks flat to us even though we know it is curved.

Relationships between Functions and Their Derivatives:

- A function is increasing if and only if its derivative is positive.
- A function is decreasing if and only if its derivative is negative.
- A function is "flat" if and only if its derivative is zero.
- A function is concave up if its derivative is increasing, or equivalently, if its second derivative is positive.
- A function is concave down if its derivative is decreasing, or equivalently, if its second derivative is negative.

We can see **acceleration** in the graph of a moving car. Looking at examples of graphs, we can see where the velocity is increasing and decreasing. We see that an upward-cupping graph corresponds to an accelerating car and a downward-cupping graph corresponds to a decelerating car. Acceleration measures the change in the velocity over time. Acceleration is itself a derivative—the derivative of velocity, or equivalently, the second derivative of position. Given a graph of a derivative, we can sketch the graph of the function, and likewise, given a graph of a function, we can sketch the graph of the derivative. By sliding a tangent line along a graph and recording its slope at each point, we can generate the derivative graph.

Let's look at the whole trip. If you take a trip, you can easily compute what your average velocity was by taking the total distance between the starting and ending points and dividing by the time it took to cover that distance. On a graph, that process is figuring out the slope of the line between the beginning and ending points. Although your velocity may have varied during the trip, at some point, your instantaneous velocity will be exactly equal to your average velocity. This reasonable observation is known as the *mean value theorem* (where *mean* means "average" rather than "cruel").

Let's summarize the relationship between a graph and the derivative. The derivative of a graph at any point is equal to the slope of the tangent line. If we magnify a smoothly curving graph, it will look like a straight line—the tangent line. The perspective of a smooth curve looking like a straight line allows us to deduce various implications. One is known as *L'Hôpital's Rule*, after the man who wrote the first calculus textbook. L'Hôpital's Rule states that the limit of a ratio of two **smooth functions**, both of which approach 0, is equal to the limit of the ratio of their derivatives. A historical feature of L'Hôpital's Rule is that **L'Hôpital** did not discover it. He bought it from one of the Bernoulli brothers.

Understanding many important issues involves analyzing change in a characteristic over time.

When Newton and Leibniz defined the derivative in the 17th century, they used different words and different notation. Newton used a dot over a varying quantity to stand for the derivative. The problem with that notation is that an errant fly was capable of taking derivatives if it left its mark in the wrong place. Most people now use the notation for derivative that Leibniz introduced. Leibniz's notation includes the fundamental feature of the derivative as a quotient of changes, with Δ's becoming *d*'s. For example, if $p(t) = t^2$, then $dp/dt = (d/dt)(t^2) = 2t$.

Often, y is a function of x, and the notation for derivative is dy/dx. This notation reminds us that the derivative arises from looking at ratios. The letter *d* reminds us of "difference," suggesting that the top value is, for example, a difference in the position of the moving car at two times, while the bottom is the difference in the time. If the function is presented as $p(t)$, then another notation for the derivative is $p'(t)$. ■

Name to Know

L'Hôpital, Guillaume François Antoine (1661–1704). French marquis, amateur mathematician, and student of Jean Bernoulli. L'Hôpital provided one of the five submitted solutions to Bernoulli's **Brachistochrone** problem. He was the author of the first calculus textbook (1696), written in the vernacular and based primarily on the work of Jean Bernoulli. This text went through several editions and greatly aided the spread of Leibniz's calculus on the Continent.

Important Terms

acceleration: Rate of change of velocity; a measure of how fast the velocity is changing. The second derivative of position. Units are distance/time^2.

Brachistochrone: A curve traced between two points (not atop one another), along which a freely falling object will reach the bottom point in the least amount of time.

slope (of a straight line): The ratio of distance ascended to distance traversed, sometimes known as "rise over run."

smooth function: A function that is continuous and whose first derivative, second derivative, and so forth are all continuous.

tangent line: A straight line associated to each point on a curve. Just grazing the curve, the tangent line "parallels" the curve at a point.

Suggested Reading

Any standard calculus textbook, section introducing derivatives as slopes of tangent lines and section describing the connection between the graphs of functions and the graphs of their derivatives.

Questions to Consider

1. When we look at a circle, we see a curve. Why is it that when we magnify a circle a great deal, it no longer looks curved?

2. Understand slopes of lines. That is, how does the slope measure the steepness of a line? Why is the slope of a line the same at each point of a line? Is the angle from the horizontal of a line doubled if the slope is doubled?

3. Draw a graph of a function—any function—and call it $p(t)$. Where is it increasing, decreasing, and constant? Where is it concave up or concave down? Can you sketch its derivative, $p'(t)$? Suppose what you drew first is actually $p'(t)$, can you sketch $p(t)$?

Visualizing the Derivative—Slopes
Lecture 5—Transcript

Welcome back to *Change and Motion: Calculus Made Clear***.** In the previous lectures we've introduced the basic concepts of calculus—the derivative and the integral—and talked about how they're related to each other through the Fundamental Theorem of Calculus. But today, what we're going to do is begin a series of three lectures about the derivative as it relates to and reflects its presence in different manifestations. Today we'll talk about it's relation to graphs, graphical relationship; in the next lecture we'll talk about it's algebraic—the algebraic way of looking at derivatives; and, then, in the next lecture we'll talk about derivatives as they apply to things, other things in the world. The strength of the derivative and of the integral is that all of them can be viewed in these different scenarios, and in each case we see a richness and a relationship of this concept—of in this case, derivative—with, in today's lecture, the graphs.

So, let's first of all take a moment to think about how—what we're talking about when we talk about a graph. What is a graph? There are many instances in the world in which we are trying to relate two dependant quantities. In the case that we were talking about of a car moving down a straight road, we said at every time, the car is in a given position. So, that is an example of a function, because at every time you have a position.

Well the interest of the derivative is to try to understand what the change is; it's measuring how quickly a change in the time affects a change in the position, how quickly does the change in the position—what's the instantaneous velocity? Change in position with respect to change in time. And understanding change with respect to time is a very important concept that comes up in many other settings besides simply a car moving down a road; some settings that are not physical settings. For example, if we talk about the population of the world over time, we can draw a graph that captures this population change of the world over time, and we notice that in this graph of the population of the world in the century 1900 to 2000 that we actually have a steeply increasing curve. It's not a straight line; it's a curve that population slowly grows at the beginning, and then the rate at which it increases moves sharply upward. The population increases as a fast rate near the year 2000, and by looking at this graph, we can get a sense of how the population is changing with respect to time.

Here's another example of a graph. This is the Dow Jones Industrial Average over the last century; and it, too, has this property of having places at which it is increasing more slowly or more quickly.

Here's a very interesting graph. This is a graph that measures the power output at a nuclear power plant. At this nuclear power plant, this is at a particular day, August 25, 1986, and on this particular day there was the normal operation level of the power plant, how much output there was. It started fine; and then, at 1 a.m. it declined. So it was producing less power than ordinarily was the case. At 2 p.m. of that day it leveled off; and at 11:10 p.m. it made a sharp descent of less and less power being generated. Suddenly, at 1:23 a.m., the graph tells the story. Suddenly there were huge amounts of power being put out by the power plant; and this is, in fact, this particular day, and it's a particular place; this is the graph of the Chernobyl Power Plant on the day that it went critical. So, sometimes graphs tell us very compelling stories, particularly looking at how the steepness of the curve, which is telling us how quickly things are changing; in this case, how quickly the power output is changing, in just a short amount of time.

So, we're going to be discussing the concept of looking at a graph, and then trying to see how the derivative is associated with the graph. Let's take our example, again, a specific example here, of a car moving in a straight road—and a graph that captures the information of that car moving on a road. So, once again, let's imagine that our time axis is this horizontal axis, and at each moment of time we record where the car is on this straight road by making a mark and creating, thereby, a graph. So, if we wanted to record the motion of the car on this straight road, and if we thought of the straight road as the y-axis here—the vertical axis—we could record the motion of the car in the following way: We could move a dot along this curved graph, and the dot is moving so that its horizontal speed is constant, because we're thinking of the horizontal axis as telling us the way time is proceeding. And as it's proceeding, the car is moving on this straight road; it goes up to the 5-mile marker at time 1 minute, and then it descends back to the 1-mile marker at time 3 minutes, and then it descends up to the 5-mile marker at time 4 minutes … and you can see that the car is just moving on this vertical, straight road, but as we move the dot along the graph, it's displaying the fact of the car's motion on the straight road. So, that's what the graph is telling us.

Let's look at the graph and try to interpret aspects of the graph with respect to the motion of the car. Well, the first thing to notice is that if the graph is going upward; that is, as we move to the right, if the graph is increasing, that means that the car is moving forward on the road. It's moving vertically. If the car, if the graph, is moving downward, that corresponds to the car's motion backward on the road. That corresponds to a negative velocity; meaning that we're thinking of negative being downward, positive being forward. And, at a place like this at the top, that's a place where the car is stopped momentarily as it changes direction from going forward to going backward. Notice that if the car has moved quite a long distance in a short amount of time—like going from the 1-mile mark to the 5-mile mark in 1 minute—that corresponds to a rather quick speed; whereas, if the graph is more horizontal—where it's not moving very far in a period of time, like between here and here—those correspond to slower speeds. It's now our—the next goal of the lecture is going to be to pin this down. Can we be absolutely precise about how quickly the car is moving by looking at the graph?

Well, let's look at some examples of graphs of cars moving on a straight road, but in this case some simpler graphs. Suppose that we have a graph of a car that just looks like a diagonal line, a 45-degree line. Well, this is an example of a car that is moving at exactly 1 mile per minute because it proceeds 1 mile along the road for every 1 minute of elapsed time. And so, the graph that is a straight diagonal line of 45-degrees corresponds to a velocity of 1 mile per minute, and a constant speed of 1 mile per minute.

Let's look at this example of a car moving on a straight road. Here, the graph has a steeper slope to it. In fact, for every 1 unit of horizontal distance, this straight line changes by 2 units of vertical change. That corresponds to a car where every minute corresponds to a 2-mile change, which is a speed of 2 miles per minute. Now, notice, that it's the steepness of the line that is telling us the velocity of the car, and the steepness of the line is measured by taking the—of a straight line—is measured by taking how much vertical distance is accomplished divided by how much horizontal time it took to accomplish that. It's a ratio of the vertical change divided by the horizontal change. And that ratio has a name. It's called the slope of a line. Notice that a straight line, if it's going upward, has a positive slope and, in this graph, you see a line going diagonally down to the right, and that has a negative slope because for a positive 1-minute increment, there is a negative change; and therefore, the ratio of rise over run is a negative number divided by the elapsed time, which gives a negative value. So, slopes going upward to the right have positive slope; that is, lines going upward to the right have positive slope. Lines going down to the right have negative slope. And, a horizontal line has slope zero—that's a place where the car is standing still.

Now we're going to face a challenge of dealing with a part of the curve where the line is not straight; that is, where the graph is not a straight line. This is a case where the car is not proceeding at a steady speed or a steady velocity. The car is changing its velocity. But, how are we going to capture the instantaneous velocity of that car at such a moment?

So let's look at our graph here of the car that was moving at varying speeds and just focus on one little part of this graph, which we'll capture in a different picture, to see how we could specify how fast the car is going at a moment such as this moment. So, here we look at a blown-up picture of just that small part of the graph. We can see that the graph is a curved line, you can see this curved line here, and we're trying now to think about the instantaneous velocity of the car at this particular time. Well remember, our whole analysis of how we computed that instantaneous velocity in Lecture Two. Our strategy was we said where is the car at just a tiny amount of time after the time that we talked about?

In other words, we're looking at the time $p(t + \Delta t)$, and then we subtracted from that the position of the car that the time t we were focusing on. So, this difference of $(p(t + \Delta t) - p(t)$ represents the distance between one point on the graph, and the position of another point on the graph. In other words, at time t, we were at this location on the road; at time $t + \Delta t$, we were at this position on the road. So, the two coordinates of these two points are: For this point, its first coordinate is t, and it's second coordinate is $p(t)$. This point has two coordinates; its first coordinate is $t + \Delta t$, and it's second coordinate is $p(t + \Delta t)$; that is, the position of the car at time $t + \Delta t$. So, the vertical distance here is the difference between its position, the car's position, at the time of t plus Δt minus the car's position of time t, that's the vertical distance. The horizontal is the time Δt, that's the elapsed time. So, the ratio, this vertical change in position, divided by the elapsed time—that ratio was the ratio that we discussed in the definition of the derivative, which was telling us an approximation of the velocity, and that ratio is also the slope of the straight line that we could draw between these two positions of the car.

$$\frac{p(t+\Delta t)-p(t)}{\Delta t}$$

Now, notice that the curved line—the actual positions of the car—does not correspond to the straight line between those two points. And, this is where the concept of the limiting process that we were introduced to in the definition of the derivative comes into play because, what did we do next? After taking a particular time Δt, we said that's not the best possible approximation we can get. We need to look at a closer time, a smaller Δt, and see what approximation to the velocity is at that smaller Δt. But we proceed with exactly the same analysis. We see where the car was at this time, even closer past the time we're interested in, t; and we look at—we construct yet another small triangle; and we compute the rise over the run of that triangle—and notice that that rise over the run is the slope of a line that is getting even closer to the curved line. There's less curve; there's less distance for that curve to deviate from that straight line when we pick a smaller distance.

In fact, let's do the process of magnifying our picture. When we magnify our picture, notice that the curved line becomes less curved; and by magnifying the picture, let me be very specific on what I mean by that. I mean that we take the horizontal axis and expand it by, for example, multiplying everything by 2; making everything twice as long; and, then, taking the vertical axis and multiplying everything on it by twice as long; so, exactly the same expansion for both the horizontal component and the vertical component. When we make the same expansion, then a straight line will remain at the same slope. So, when we do this kind of magnification, and we go closer and closer and magnify our curve, which is the exact process of taking smaller and smaller Δt's, we find that the curved line begins to look much more like a straight line, and pretty soon becomes indistinguishable from a straight line. The limit—which is the derivative—is the limit as Δt approaches zero, is, therefore, approaching the value of the slope of the tangent line. That is, the line that just grazes this curve; the line that is of the slope that the curve would look like if we continued to magnify that curve, and it began to look more and more like a straight line at a certain slope. Well, that slope is the derivative of the slope at that point, and it's also the velocity of the car at that time.

So, I want to emphasize two insights that are absolutely fundamental to the derivative and its relationship to the graph of a curve. The first thing is that the graph of a curve is—at any point on the graph of a curve—the slope of the tangent line is equal to the derivative at that point. So, we can glance at a curve, and if it's a curve of the position of a car moving on a straight road, and we point to a particular point on that graph, we can estimate the speed of the car by estimating the slope of the tangent line at that point.

The other insight that comes from this is that smoothly curving lines, when looked at very close, look like straight lines. This is an insight. I want to point this out because it's interesting. I sometimes ask my students to take a circle and to draw on their paper a picture of a circle looked at very closely. So, this is an example of a circle. And, often when a student—I ask them this before I've explained this magnification concept, they will say a picture of a circle looked up close looks like this—it looks like a circle, like a curve. But, in fact, the correct answer is to take out your ruler and draw a straight line. That's what a circle looks like very close. But, all of us are familiar with this. This is not an unusual concept. We're all familiar with this idea because we live on the Earth. If we look around us—the Earth is round;—if we look around us on the Earth, it doesn't look as though the Earth is round. Locally, it looks like it's completely flat. That's because a large circle, when magnified, looks, locally, flat.

So, the derivative of a function gives us the slope of the tangent line at each point of the curve. Now, let's go back to our curve and look at various points on the curve and just see what these slopes are at these various points. In this curve, some places we can tell the slope easily. For example, at these peak points, the slope is 0 because the tangent line is horizontal. If we magnified this point very closely, we'd just see a horizontal line. Likewise here; this is a place where the tangent line is horizontal, and we would say the velocity is 0. At a place such as this one, we can draw the tangent line and we can see what the slope is of that tangent line.

So, where the curve is steep, we see that we have a faster velocity and that the velocity is equal to the slope of the tangent line. So, now, what I'm doing here is dragging along the point of the curve, and showing the tangent line at every point as we drag the point along the curve. And you can see that the slopes are varying from being positive and rather steep here; to 0 here; to being negative here; to being 0 here; to being positive, again, over here. So, what we're measuring and looking at the slopes at each point is the actual value of the derivative at each point.

Now, there's more that we can get from this graph of the position function. We can also talk about the acceleration of the car. When you stop at a stop sign and then you go forward, you increase your speed by accelerating. You go from a speed of 0 to a speed of 60 miles an hour by going through intermediate speeds, and you're increasing your speed at some rate. Well, acceleration is talking about how quickly you're changing your velocity. So, acceleration is a derivative of a derivative of the position function. It's the change in the velocity. Well, we can understand that second derivative, that acceleration, by looking at this curve also.

Look here, for example. Here at the point (0,0) we're going rather fast. The slope of the tangent line has a positive and steep slope. Here at 1 it's 0. So, it has declined. So, this means that the second derivative, the change in the derivative, is going from a positive number down to 0. So, the second derivative is negative; it's decreasing. And it continues to decrease as we go over the top of the curve. And, in fact, if the curve is a curve that is concave down, what does that mean? That means that the slope of the derivative, the slope of the tangent line, is decreasing as we move in the positive time direction this way. So, since the slope is decreasing, that corresponds to the second derivative being a negative number. The velocity is decreasing.

Likewise, when we have an upward cupped part of the curve, that corresponds to—and I'm looking at it so you can see the positive axis going in this direction—the slope of the tangent line increases as we move to the right; gets steeper as we move to the right. That corresponds to increasing velocity; positive second derivative. And, increasing velocity is the same as positive acceleration; acceleration meaning an increase in your velocity.

Well, we've actually then seen all sorts of relationships between the graph of a function and the derivative. Here they are; we've encapsulated them in this chart. When the function is increasing, the derivative is positive; when a function is decreasing, the derivative is negative; when the function is flat, the derivative is zero; when the function is concave up, the derivative is increasing; that is, the slope of the tangent line is increasing when it's concave up, which is the same as saying the second derivative is positive. And the physical interpretation of the second derivative, when we're talking about the position curve, is the acceleration. The acceleration is positive. When the function is concave down, that's saying that the derivative is decreasing. But, since the second derivative is the derivative of the derivative—that means that the second derivative is negative. We're really just referring back up here to these previous insights.

So, given a function, given the graph of a function, we can just take a piece of paper and sketch the graph of its derivative by, at each point on the curve, just measuring the slope of the tangent line and then making a point at the value of that slope. So, if we go back to the curve that we've been looking at several times, as we move a dot along this curve, we'll look at the value of the slope of the tangent line; that is, the value of the derivative; and plot it. That is, we just make a dot. So, for example, at this point we plot the point 0 right here. We're plotting the derivative, the slope of the tangent line, at each point. So, it's 0 here; it's 0 here; it's negative here; it's positive here; and, in fact, if we combine these together as we move the dot along the curve, we're drawing—we're capturing the slope of the tangent line and just marking a dot at that value. We'll see that we're drawing a parabola-looking curve, it's a cup-shaped curve and that is the derivative curve. So, the derivative of this curve of motion is a cup-shaped curve.

There's another insight that comes from derivatives that's rather interesting, and this has to do with the global picture of a trip that you might take in a car. This happens if you're driving on the Pennsylvania Turnpike, and you get on the Turnpike at one point and they give you a time-stamped ticket. Now, suppose you drive 100 miles on the Turnpike and you get off the Turnpike and they take your ticket and it's exactly 1 hour later. Okay? So, you've gone 100 miles, it's 1 hour later, and you drive slowly into the Turnpike and they say, "Well, sir, yes you owe the Turnpike thing, and you also owe a fine because you were speeding." And you say, "Well but no, I wasn't speeding, because you saw me just drive slowly into the turnpike thing." But then they say, "Well yes, but the fact that you went 100 miles in 1 hour means that at some point you had to be going at least 100 miles per hour." This is, actually, the insight of the mean value theorem; that if you are traveling along an interval of time, and your average velocity is a certain amount, then at some point during that time you actually were going that average velocity, and we can see this graphically.

So, this is the graph that expresses the mean value theorem. We start at one point in the graph, at a certain time, that is one point on a road at a certain time; we end up at a future time at a different point on the road; and, during the course of that, we have gone different positions on the road. We sometimes went backward; sometimes went forward. And the claim is that at some particular point, and maybe several different times, our instantaneous velocity was exactly equal to the average velocity during the whole period. And the proof is very simple. The proof is that we just draw the line from our beginning position to our final position on the graph; the slope of that line is equal to the average velocity during our trip; and, now, we just take a line that is parallel to that steepness and just let it float down until the first time it hits our curve. And where it hits our curve will be a point where it is the tangent line at the curve. So, at that point, the derivative is equal to the slope of that line which is parallel to the average velocity; and, therefore, its instantaneous velocity is equal to the average velocity.

I wanted to tell you a brief story about a rule about derivatives called L'Hôpital's Rule, and I'm not really going to explain the actual rule of it, but it's a theorem in mathematics that relates a ratio of derivatives to a ratio of the functions. And, with certain conditions, we can see that the limit of the ratio of derivatives is equal to the limit of the ratio of the functions. L'Hôpital was a mathematician who wrote the first calculus textbook. He wrote this book in 1696, and he wrote this book, and appeared this particular theorem—which all students who take calculus these days will see, L'Hôpital's Rule. So, he's become famous throughout time for this rule. Well, it turned out that L'Hôpital did not prove this rule, he bought it. He had a financial relationship with one of the Bernoulli mathematicians that he would take credit for the mathematical results that Bernoulli was able to produce, and one of them was L'Hôpital's Rule, which everybody knows as the rule associated with L'Hôpital. So, I think his investment definitely paid off, whatever he paid for that.

I just wanted to say two words about notation. Newton and Leibniz had different notations for the derivative and, incidentally, for the integral. Newton's notation was that he took a variable, like x, which he thought of, really, as the function of time, and he just put a dot over it. That was the derivative. Now, there was one problem with this notation for the derivative and that is that a fly, if a fly lands at a particular place on the page, can actually take a derivative. That was a bit of a drawback for this notation. But, Leibniz had a notation for derivative that captures the defining quotient property that we've talked about several time; that is, the change in position divided by the change in time, and one writes this whole fraction-looking thing as the derivative. Now, it has its own problems. Namely, it's not really two different quantities; it's the whole fraction that's one idea. So, that's sometimes confusing to people. But this is Leibniz's notation for the derivative, and a very common notation that we'll be using throughout the course is just to put a prime mark after a function to represent the derivative of that function. So, if we have the function $p(t) = t^2$, then we would note $p'(t) = 2t$, and this is probably the most common notation that is used for derivatives today.

In the next lecture, then, we will be talking about the algebraic manifestation of derivatives. I'll look forward to talking to you then.

Derivatives the Easy Way—Symbol Pushing

Lecture 6

Most of the functions that are used in physics, economics, geometry, or almost any area of study are expressions that involve basic arithmetic operations—addition, subtraction, multiplication, division, or taking exponents, or trigonometric functions, like sines and cosines.

Much of the practical power of calculus lies in dealing with specific functions that model physical and conceptual situations. We now have a good conceptual sense of what the derivative means both physically and graphically. In this lecture, we'll see how to compute derivatives algebraically. Most functions that are used in physics, economics, geometry, or almost any area of study are expressions that involve basic arithmetic operations—addition, subtraction, multiplication, division—or exponents or trigonometric functions. Here, we see how these functions give rise to neat expressions for their derivatives and how these expressions agree with the geometric properties of the graphs we observed in the previous lecture. If we have an algebraic expression that tells us the position of a moving car, then we can deduce the algebraic expression for the velocity of the car at each moment without having to carry out the infinite process involved in taking derivatives at each point.

In this lecture, we'll look at the derivative as it is manifested in algebra. Most functions used in almost any area of study are expressions that involve basic arithmetic operations. Here, we see how these functions give rise to expressions for their derivatives and how these expressions are obtained in a mechanical way. We will also see how these expressions agree with the geometric properties of the graphs we observed in Lecture 5.

Derivatives would be of no practical value if we had to do an infinite process at each point of time. Fortunately, we don't. The simplest function describing a moving car would occur when the car is moving at a steady velocity. In this case, the position function is $p(t) = ct$, where c is a constant. The velocity, which is the same as the derivative, is just c. Thus, if $p(t) = ct$, then $p'(t) = c$.

Let's look at a function we have already seen: $f(x) = x^2$ (which we have seen previously as $p(t) = t^2$).

Notice in the table that for every number x, the derivative value was $2x$. How can we see in general that this is the case?

Function	Derivative
x^2	$f'(x) = 2x$
0.7	1.4
1	2
1.4	2.8
2	4
3	6

We can check this derivative algebraically by considering the defining quotient for the derivative, namely,

$$\frac{f(x+\Delta x)-f(x)}{\Delta x}=\frac{(x+\Delta x)^2-x^2}{\Delta x}=\frac{x^2+2x\Delta x+(\Delta x)^2-x^2}{\Delta x}$$

$$=\frac{2x\Delta x+(\Delta x)^2}{\Delta x}=\frac{(2x+\Delta x)\Delta x}{\Delta x}=2x+\Delta x.$$

Now we take the limit as Δx gets smaller and smaller. Whether the function is $p(t) = t^2$ or $f(x) = x^2$, the answer is the same: $p'(t) = 2t$ or $f'(x) = 2x$.

Here are some related functions:

If $f(x)=5x^2$, then $f'(x)=5(2x)=10x$.

More generally, if $f(x) = cx^2$, then $f'(x) = 2cx$.

Let's consider other functions of the form x^n. For the function $f(x)=x^3$, $f'(x)=3x^2$. We can also understand this derivative algebraically by considering the defining quotient for the derivative, namely, $\dfrac{(x+\Delta x)^3-x^3}{\Delta x}$, and looking at the limit as Δx gets smaller and smaller. We can see the pattern for finding the algebraic formula for the derivative of powers. If $f(x)=x^n$, then $f'(x)=nx^{n-1}$.

We see that if $h(x) = f(x) + g(x)$, then $h'(x) = f'(x) + g'(x)$. Likewise, you can take one given function whose derivative you know and multiply it by a constant to get another function: For example, if $h(x) = cf(x)$, then $h'(x) = cf'(x)$. This allows us to take derivatives of such functions as $5x^3 + 2x$. We see this as a sum of two different functions, $5x^3$ and $2x$, and we see each of those as a product of a constant times a function whose derivative we know. Therefore, if $h(x) = 5x^3 + 2x$, then $h'(x) = 15x^2 + 2$. And if we add a constant to the same function, for example, $h(x) = 5x^3 + 2x + 3$, then we still have $h'(x) = 15x^2 + 2$. Why? Because the derivative of a constant is 0, so it does not change the derivative.

Derivatives would be of no practical value if we had to do an infinite process at each point of time. Fortunately, we don't.

Note that the derivative of a product is not simply the product of derivatives. The product rule and the quotient rule (if you have a function that is the quotient of two functions) are both rather more complicated algebraic equations. At this stage, we can take the derivative of any polynomial, that is, a function of the form $a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$. For example, let's take the function $f(x) = x^3 - 6x^2 + 9x + 1$. Note that the graph depicting this function starts down at the left, goes up, then down, and then goes up again. The derivative of this function is $f'(x) = 3x^2 - 12x + 9$. Note that in the graph of this derivative, we see that it is the form of a parabola, going down and up again.

The derivative reduces the number of "bumps" in the graph of a function. A typical third-degree polynomial has a graph that goes up-down-up. Its derivative has a graph that goes down-up. A typical fourth-degree polynomial has a graph that goes down-up-down-up. Its derivative has a graph that goes up-down-up. We can see this geometrically by tracing the moving tangent line and recording the changing slopes. In looking at these examples, we see that the derivative has one less bump.

Let's consider a function that is defined geometrically on a circle. If we take a right triangle whose angle is θ, the **sine** of the angle θ can be thought of as the ratio of the length of the side opposite θ divided by the length of the hypotenuse. The **cosine** of the angle θ is the ratio of the length of the side adjacent to θ divided by the length of the hypotenuse. On a circle of radius 1, the hypotenuse is 1; thus, sine of θ is just the vertical coordinate, and cosine of θ is the horizontal coordinate. That is, (cos θ, sin θ) are the coordinates of the point on the unit circle corresponding to the angle θ. As θ changes, so does the sin θ: When θ is small, so is sin θ; when θ approaches 90 degrees (or $\pi/2$ radians), sin θ approaches 1, and $\sin(\pi/2) = 1$. Similarly, when θ is small, cos θ is close to 1; when θ approaches 90 degrees (or $\pi/2$ radians), cos θ approaches 0, and $\cos(\pi/2) = 0$. Notice that the graphs of sine and cosine oscillate because their values repeat each time we move around the unit circle. What is the derivative of sine? Look at the rate at which the line opposite the hypotenuse is changing in relation to a change in the angle. The derivative of sine is cosine, and the derivative of cosine is negative sine. We can graph these functions and see geometrically why their derivatives are related as they are. Neat.

An interesting question is this: Does there exist a function that is its own derivative at every point? That is, is there a function $f(x)$, such that $f'(x) = f(x)$ at every x? Equivalently, we are looking for a function y that satisfies $dy/dx = y$. It turns out that the *exponential function* $f(x) = e^x$, where $e = 2.718281828\ldots$, satisfies these conditions.

This leads us to the following table of derivatives:

Function	Derivative
$f(x)$	$f'(x)$
1	0
x	1
x^2	$2x$
x^3	$3x^2$
x^n	nx^{n-1}
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
e^x	e^x

If we are trying to find the answer to a question that involves derivatives, we will be able to solve it in a practical way. These equations help to mechanize the process of finding the derivative. If a law of nature, for example, involves derivatives, then we will often be able to express that law with a simple, computable formula. We will use this ability in the next lectures that show some applications of the derivative. ■

Important Terms

cosine: A function of angle θ giving the ratio of the length of the adjacent side to the length of the hypotenuse of a right triangle, as well as the horizontal coordinate of a point on the circle of radius 1 corresponding to the angle θ.

sine: A function of angle θ giving the ratio of the length of the opposite side to the length of the hypotenuse of a right triangle, as well as the vertical coordinate of a point on the circle of radius 1 corresponding to the angle θ.

Suggested Reading

Any standard calculus textbook, sections on differentiation formulas.

Questions to Consider

1. Suppose $f'(2) = 3$ and $g'(2) = 4$. If $h(x) = f(x) + g(x)$ for every x, why is $h'(2) = 7$? This shows that the derivative of the sum is the sum of the derivatives. Warning: This pattern does not hold up for products.

2. Why do you believe that many aspects of nature and human creations are so well described by rather simple functions?

Derivatives the Easy Way—Symbol Pushing
Lecture 6—Transcript

Welcome back to *Change and Motion: Calculus Made Clear***.** In the previous lecture we saw how the derivative relates to the slope of a tangent line; the derivative is the slope of the tangent line to a graph of a function. But, much of the practical power of calculus lies in dealing with specific functions that model physical and conceptual situations. So, in the last lecture, we got a good graphical sense of what the derivative means, and we talked about the car moving on a road to get a physical sense of what it means, but in this lecture we'll see how to compute derivatives algebraically.

Most of the functions that are used in physics, economics, geometry, or almost any area of study are expressions that involve basic arithmetic operations—addition, subtraction, multiplication, division, or taking exponents, or trigonometric functions, like sines and cosines. So, here, we're going to see how these functions give rise to neat expressions for their derivatives so that we can see how those expressions are obtained in sort of a mechanical way. That's what this lecture is about.

We'll also see how these expressions agree with the geometric properties of the graphs that we observed in the previous lectures. So, the point is that if we have, for example, an algebraic expression that tells us the position of a car moving along a straight road, then we can deduce the velocity of the car by deducing an expression, an algebraic expression, that tells us the velocity of the car each moment without having to go through this laborious process of computing the limit at every single moment; or the process of looking at the graph of the curve and then just estimating the slope of the tangent line. So, this lecture, then, is about the derivative as it's manifested in algebra. And it's one of the areas that most students think of as taking derivatives. What they're really saying is, "If we give you an algebraic expression, can you find the algebraic expression for its derivative?"

So, let's begin, as always, with a simple example, and then move to more complicated ones. Let's begin with this example. Suppose that we have a function, which we could think of as a position function, or we could just think of as just a general abstract function, $p(t)$ = a constant $c \times t$. Notice that if the constant is 1, this will be a 45-degree line; if it's 2, it'll be a line with slope 2; and so on. We saw that the derivative at each point of such a function is just equal to the constant *c* because it's equal to the slope of the line at every point; and since it's a line, the slope is the same at every point on that line; and, therefore, the derivative function is a constant *c*, which has a graph that's just horizontal. So, that's our first and simplest example of a function that has a nice derivative expression.

Now, let's go to a slightly more difficult one, but one that we've dealt with before. Suppose we have a function $f(x) = x^2$. Now, we saw this expression before in dealing with a car moving on a road, and we saw that the way to compute the derivative was to evaluate $\frac{f(x+\Delta x)-f(x)}{\Delta x}$. And, then, look at that fraction as Δx became increasingly small and to see if those numbers approached one common limit; and then we called that limit the value of the derivative.

Now, notice, by the way, that in our previous familiarity with this function, we called it $p(t) = t^2$ for position; the position at time *t* was t^2. And now I've changed it to $f(x) = x^2$. That makes no difference from the abstract concept of a function, it doesn't matter if we're calling the name of the function *p* or *f*, or if we call the variable *t* or *x*; it doesn't matter. And so, this is just a tiny step in the direction of abstracting the concepts of the car moving on the straight road to now realizing that any two dependant quantities—any function $f(x)$—that may relate any two kinds of dependent quantities, can have the same analysis of taking derivatives.

We saw before—that we made a chart where we had actually gone through the laborious process of computing this characteristic difference and quotient that's associated with the derivative. We saw that for a function $f(x) = x^2$, that the instantaneous—that the derivative at every point were the values in this chart. Namely, at .7 it was 1.4; at 1 it was 2; and in each case, the derivative was exactly $2x$.

If we wanted to graph the relationship between the graph of the function and the graph of the derivative, here is the graph of the function, $f(x) = x^2$, and here is the graph of its derivative, $2x$. So, we see that if we use the perspective of the moving tangent line, we can see that it accords with the graph of the straight line with slope $2x$. Because, when we look at this left-hand part of the parabola graph, we can see that the slope is a negative number; and, sure enough, the value of the derivative is a negative value. When it gets to $x = 0$, we see that the tangent line is horizontal; that is to say that the derivative has value 0, and here it is, value 0; and, then, when we moved to the positive values of x, we see that the slope of the tangent line is a positive number and increasingly steep, and that accords with this derivative function having slope 2 and continuing to rise as we get larger values of x.

So, once again, we see an algebraic expression for the derivative. But, why is that algebraic expression true? Let's go ahead and do the actual mathematical derivation of the derivative. Remember that the derivative, $f'(x)$, is equal to—and I'm now going to actually use the notation associated with taking the limit. That is it's expressed by saying LIM for limit, as Δx approaches 0, with that arrow; Δx with an arrow towards 0 means that we look at this expression that follows that limit symbol, namely, $\frac{f(x+\Delta x)-f(x)}{\Delta x}$, and evaluate that difference quotient for smaller and smaller values of Δx, as we choose Δx to be increasing small and getting closer and closer to 0. We evaluate that quotient and see that those numbers get closer and closer to one value.

$$p'(x) = \frac{p(x+\Delta x)-p(x)}{\Delta x}$$

$$= \frac{(x+\Delta x)^2 - x^2}{\Delta x}$$

$$= \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x}$$

$$= \frac{2x\Delta x + (\Delta x)^2}{\Delta x}$$

$$= 2x + \Delta x$$

$$= 2x \text{, as } \Delta x \text{ becomes small}$$

Well, let's see what value they get closer to. So, the limit as Δx approaches 0 of this expression, well, what is $f(x + \Delta x)$? We're looking at the function $f(x) = x^2$. So, $f(x + \Delta x) = x + \Delta x^2$. And, the function of x is x^2. So, we have—this is the numerator, and then Δx remains as the denominator of that fraction; and, then, just doing the expansion of this expression. Just in other words, multiplying out, using the distributive law, we can see that $(x + \Delta x) \times$ itself is equal to $x^2 + 2x\Delta x + (\Delta x)^2 - x^2$, all over Δx. The x^2's conveniently cancel; leaving $2x\Delta x + \Delta x^2$. The Δx's can be factored out, so that we have in this expression $2x\Delta x + \Delta x$, divided by Δx. The Δx's cancel, leaving the limit as Δx approaches 0, which, of course, has preceded in each of these lines, of $2x + \Delta x$. All that means is that if we take any value of x, and we plug in any value Δx that's not 0, but maybe close to 0—it would be true even if it weren't close to 0, but we're interested in those values that are close to 0—if we plug it in, all of this algebra is just telling us that the value of this quotient is exactly just $2x + \Delta x$. That's what it says.

Now, what happens as Δx becomes increasing small, increasingly close to 0? Well, this expression, then gets increasingly close to $2x$, and then in the limit it's equal to $2x$. That is the derivation of the formula for the derivative of the function of $f(x) = x^2$. Now, this is a good example of the kind of derivation involved in actually finding algebraic expressions that equal the derivative of a given algebraically expressed function. So, I hope that you enjoyed that.

Now, let's look at some related functions and just see how we get their algebraic derivatives. We just saw this one. Suppose that we took our expression, instead of x^2 we multiplied it by some number, like 5. What is the result? Well, it's easy to see by just looking at the derivation and just thinking about what it means to multiply by the constant 5; that means that the values in the vertical direction are all expanded by a factor of 5, which means that every single quotient will be five times as tall on that vertical part of the triangle as it was in the previous derivation. And consequently, the value of the derivative is going to be five times as great as the value of the derivative before. And in general, there is the general statement that if you have $f(x)$ = a constant $\times$ x^2, the value of the derivative—and you can think of it as the slope of the tangent line—is always going to be just 2 $\times$ that constant $\times$ x. In other words, the constant just flows through the taking of the derivative.

Let's do one more example of taking a derivative to get a spirit of how these go for functions of the form x to a power. So, this one is the function $f(x) = x^3$; and we're going to just follow the exact form that we did before; namely, we realized that the definition of the derivative of that function, $f'(x)$, as always, is the limit as Δx approaches 0 of the function evaluated at $x + \Delta x$

– the function evaluated at $\frac{x}{\Delta x}$. When you think of derivative, you should always, immediately, come to this expression; this characteristic fraction.

Well, what is the function of $x + \Delta x$? Well, the function we're talking about just takes any value and cubes it to give us the value there, so we just cube $x + \Delta x$. This is cubed – $\frac{x^3}{\Delta x}$, multiplying it out. Once again, the x^3's cancel; we're left with some expressions, each of which has a Δx in it. After we factor out the Δx's and cancel the denominator, we're left with only one expression that fails to have a Δx. The two other expressions, $3x\Delta x$ and $-\Delta x^2$, are expressions each of which contains a Δx in them. Now, notice, as Δx goes to 0, that's what the limiting process is. It says what happens when you take Δx to be .0000001, these expressions here are going to become negligible while this expression stays the same because it doesn't have a Δx in it. So, if x is equal to 5, for example, this would stay 3×5^2—it would be 75—plus some value that if I chose little tiny Δx's, those would become .0000001; they become negligible, and in the limit, disappear. Consequently, the derivative of $x^3 = 3x^2$.

$$p'(x) = \frac{p(x+\Delta x) - p(x)}{\Delta x}$$

$$= \frac{(x+\Delta x)^3 - x^3}{\Delta x}$$

$$= \frac{3x^3 + 3x^2\Delta x + 3x(\Delta x)^2 - (\Delta x)^3 - x^3}{\Delta x}$$

$$= 3x^2 + 3x(\Delta x) - (\Delta x)^3$$

$= 3x$, as Δx becomes small

Well, maybe we're beginning to see a pattern here; and the pattern is so far, the derivative of x—remember, that has slope 1, so its derivative is 1; the derivative of x^2 is $2x$; the derivative of x^3 is $3x^2$; and the pattern continues. In general, for any function $f(x) = x^n$, its derivative is equal just to $n \times x^{n-1}$ power. And this algebraic way of taking derivatives becomes very habitual for students. In fact, many of them believe the derivative is taking down the exponent, multiplying and subtracting 1 from the exponent, without—failing to remember that the derivative is telling us something of meaning; mainly, the slope of the tangent line, or the velocity of a car moving on a road, if the function is telling us the position of the car on this straight road on every moment. So, we have our chart of derivatives for exponential—for functions that are just x to a power.

Much of the way that we develop our strength of taking derivatives of more complicated functions is to see how, if you can get a function as being constructed from other functions by various operations, such as adding two other functions together to get a new function or multiplying two functions together to get a new function, that we want to see how the derivatives of the newly created function relate to the derivatives of the individual component part functions. So, we'll start that process by talking about the sum of two functions. In other words, suppose that we have a function $h(x)$ that equals the sum of two other functions, $f(x)$ and $g(x)$. In other words, to get the value for a function h, you first compute the value of f and you get a number; you compute the value of g at that same point, you get a number; you add them together to get the value of h. Then, the derivative, as you would expect, is just equal to the derivative of f plus the derivative of g, at that point; at every point.

Similarly, let's take another way that you could construct a function. You could take one given function, whose derivative maybe you know, and multiply it by a constant to get another function. We already saw that in the case of a constant times x^2, and it more generally applies to any function; that is to say that the derivative of a constant times a function, that is, if we create a new function by taking an old function, $f(x)$ and multiplying everything by a constant, then the derivative of that new function is just the constant times the derivative of the old function, for the same reason that we talked about before, just the vertical components being expanded by the factor c.

This allows us to take derivatives of functions such as this one: $5x^3 + 2x$; because, you see, we see this as a sum of two different functions, $5x^3 + 2x$, and each of those is the product of a constant times a function whose derivative we know; like, the derivative of x is equal to just 1, so 1×2 is 2; and the derivative of x^3 is $3x^2$; and, consequently, $5x^3$ will have a derivative that is five times as big as the derivative of x^3; so, it's just $3 \times 5 = 15x^2 + 2$. So, this is the way of understanding the derivative of a function that is created by the sum or a constant times a given function whose derivative we already know.

$$h(x) = f(x)g(x) \text{ then } h'(x) = f'(x)g(x) + f(x)g'(x)$$

Notice this slight variation on this. Let's look at a function that's the same function except that we add a constant. What's the derivative of a constant? Well, the derivative of a constant, a constant has a graph that is just horizontal, that derivative is 0 at every single point; and, consequently, it does not change the derivative. The derivative is exactly the same if we add a constant. By the way, the look of a function, the graph of this function compared to the graph of this function, all we've done is shifted the graph up by three units. That's what it means to add a constant. And, since we've just shifted the graph up by three units, the tangency at the same point is exactly the same; and, since the slope of the tangent is what the derivative is measuring, we get the same answer when we add a constant.

$$h(x) = \frac{f(x)}{g(x)} \text{ then } h'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{\left[g(x)2\right]}$$

When it comes to products of functions, we have a surprise. You might think that the derivative of the product of two functions is the product of the derivatives, but this is false; this is simply false. One good way to realize it must be false is the derivative of x^2—x^2 is $x \times x$—but it's not equal to the derivative of $x \times$ the derivative of x. The derivative of x is 1; so, it's not 1×1; in fact, it's $2x$. So, that is not—that shows the simple product rule, the naive one that you might think of, is not correct. The actual product rule is the following—it's a little more complicated—that if you construct a function, $h(x)$, by taking the product of two previously known functions, $f(x)$ and $g(x)$, then the derivative of the product is the first times the derivative of the second plus the second times the derivative of the first.

Likewise, there's a quotient rule. If you have a function that's the quotient of two functions, then its derivative is a rather complicated expression. It's the denominator function times the derivative of the numerator function minus the derivative of the denominator function—the one downstairs—times the numerator function divided by the square of the denominator function.

Now, each of those can be deduced by just writing down the definition of derivative and then doing some algebra.

At this position, now, at this stage, we can take the derivative of any polynomial. A polynomial is a function of the form something times x to a power plus some other constant times x to another power plus something times x to another power. Let's look at a specific example. Suppose we look at the example $f(x) = x^3 - 6x^2 + 9x + 1$. The graph of this function, as you see, it starts down to the left, it goes up, it comes back down, and back up again. Notice, actually, this is exactly the same function that we used in the previous lecture. Suppose that we take its derivative; that is, we actually compute the derivative using now the formula rules that we've deduced. The derivative of $f(x) = 3x^2 - 12x + 9$; and, if we graph that derivative on the same graph as we see the graph of this function, we can see that it corresponds the way we saw visually that it must correspond in the previous lecture. Namely, at this peak point of the original function, that is the point on top where the derivative is—where the tangent line is horizontal, that is to say the derivative is 0. Well, we notice that the derivative function is equal to 0 at that point. Likewise, at this minimum value, the point where it's at the bottom of this trough; that is, the graph of the function has this trough. Once again, the tangency is horizontal at that point; that is, the derivative has value 0; and, sure enough, the value of the derivative function at that point is 0 as expected, as it must be. And also notice that the graph of the derivative function has just—it's called a parabola; it just goes down and up again. It just has one change in direction; whereas, the function that we started with, was a cubic function. It went up, down, and up again. It's characteristic of polynomials that higher-degree polynomials have more oscillations, and that when we take the derivative, the number of their oscillations declines by one. I always find this an interesting thing, that you can actually graphically see that the number of oscillations will decrease by one, and it is expressed algebraically as well. It's manifested by the algebra and by the geometry together.

Here's an example of a fourth-degree equation, and we take it's derivative and draw the graph of the derivative on the same graph, and we see once again that the number of this fourth-degree equation goes, in this case, down, up, down, and up again; and, then, the derivative function, which is a third-degree polynomial, starts low because you see here, the derivative has—the tangent lines are negative and steeply negative. So, the value is down here, and then it goes up, down, and up again. So, once again, we have a manifestation that the derivative, the formula for the derivative, is giving us something that has geometric significance as well.

Now we're going to turn our attention to two additional classes of functions that are of interest. The first one is the trigonometric functions. Now, the trigonometric functions, particularly the sine and the cosine, are two of the most fundamental of the trigonometric functions; and let me, first of all, remind you how the trigonometric functions are defined. The sine of an angle Θ can be thought of as the quotient of the—if we take a right triangle whose angle is Θ, it's the ratio of the opposite side of Θ divided by the length of the hypotenuse. That ratio is the sine of Θ, and the cosine of Θ is the adjacent side divided by the hypotenuse.

Another way to talk about such functions is the following: Let's just take a unit circle, so that the hypotenuse has value 1; then all of the points on the unit circle have two coordinates. The first coordinate is the cosine of the angle that's measured by starting at the horizontal axis and measuring counter-clockwise around to get to the angle that we're talking about. The x-axis is equal to the cosine, actually, of Θ, because it's the adjacent side over 1. So, this distance here, which is the first coordinate, is the cosine of Θ. This second coordinate is the sine of Θ. So, the cosine Θ, sine Θ are the two coordinates of a point on this unit circle. By the way, for convenience, we measure this angle Θ by the arc length around the circle, of the curve. So in other words, as we go along the curve, remember a unit circle has circumference $2\pi \times 1$, so just 2π; and so this measure Θ goes from 0, for example, up to this distance, it's $\pi / 2$; that's a quarter of the way around; halfway around is π; then $3\pi / 2$; and, then, 2π, after you've gone once around—and notice that if we go twice around, we get the same value as not going anywhere. Every time we go 2π or more around, we get the same value.

Now, the sine and the cosine, you see, oscillate because every time we go 2π around, they oscillate between what they started with, and then they come back to that same value. So, if we graph the sine and the cosine, we'll see that the sine varies between 0, at $\Theta = 0$; it gets up to 1 at $\Theta = \pi / 2$; it goes back to 0 at π; it goes to −1 here; and then it gets back to 0 here; and then it oscillates. So, the graph of the sine function goes up and down and up and down, oscillating forever.

It turns out that the derivative of the sine function is the cosine function. It oscillates also, but it starts at 1 and it goes down to—the value of the cosine at 0 is 1 because, you see, the horizontal coordinate is 1, and 0, and then it oscillates going back and forth, and you can see just graphically that as we slide the tangent line along the sine curve, the slope of that tangent line is 1, and that's the value of the cosine at 0. Likewise, when the sine reaches its maximum at 1, here at $\pi / 2$, the maximum value of 1 at $\Theta = \pi / 2$; the cosine is down at 0. And, so, those two things oscillate together in making the derivative, the cosine is a derivative of the sine. So, here we can summarize this that the derivative of the sine is the cosine, and the derivative of the cosine is negative of the sine function.

I wanted to talk about one other class of functions, and these are interesting functions that are exponential functions. Could we have a function whose derivative is exactly equal to itself? Well, that's an interesting question. Suppose we started at the point 1 and we said could we draw a function whose derivative is exactly equal to the value of the function? Well, if we started our function, just arbitrarily, at value 0, having slope equal to 1; in other words, it starts up at a 45-degree angle, then we could start drawing the function by just saying well, it goes up by about 45 degrees, but then if we want its derivative to be equal to itself, then since the value is increasing of the function, you see, because it's going up, then the derivative has to be steeper. So, it has to become steeper, and it has to become ever steeper as it goes up, and that creates an exponential function. And, in fact, we can find such an exponential function just by drawing it. It has the name e^x; and

e^x is a function with the interesting property that its derivative is equal to itself. And, in fact, it's equal to the function where there's an actual number associated with it if you take the number 2.718281828459…. you will find that when you raise it to a power and you actually compute the derivative, you will find that the derivative at every value is equal, exactly equal, to itself. Of course, the actual number goes on forever, it's not a number that stops at any given point.

Exponential functions in general have the property that the derivative of taking any constant number raised to a power, like 2^x power, by looking at this algebra, the definition of the derivative, we see that since we can factor out the a^x itself, we see that, in fact, the derivative of any exponential function, a constant to the x power, is that same value times a constant. This leads us to a table of derivatives. So, so far we have this following table of derivatives; we know how to take the derivatives of x^n power; the derivative of the sine is the cosine; the derivative of the cosine is minus the sine; and the derivative of e^x is e^x.

So, the purpose of this lecture was then to give us a mechanical way for taking an algebraic expression that may express some relationship in nature or in science or in economics, and if we can express it in some relation using the functions that we have, then we can automatically understand what the derivative of those functions are in this mechanical way.

In the next lecture, we'll see the derivative in many different application areas. I'll see you then.

Abstracting the Derivative—Circles and Belts

Lecture 7

The true power of the derivative lies not only in its ability to help us understand change with respect to time, like the car moving down the straight road, but also to deal with dependencies of all different types—whenever we have one quantity dependent on another, that's a place where the derivative can come into play.

In the previous lecture, we saw how the derivative behaves algebraically, so we can easily compute algebraically the derivative of the function that tells us the area of a circle of a given radius. Here, we show how to interpret the answer geometrically. That is, we see what the derivative means visually when we apply the concept of derivative to formulas for area and volume. Our geometric insight leads to a surprising realization about the roominess of belts around the Earth.

In this lecture, we continue our study of the derivative. In Lecture 5, we saw how the derivative is expressed graphically. In the last lecture, we saw how to express a derivative algebraically. In this lecture, we will see how to interpret a derivative geometrically.

The concept of the derivative can be generalized to apply to any two dependent quantities. The derivative can be applied to all cases of two interrelated quantities. The derivative describes how a change in one quantity entails a change in the other. For example, the area of a circle is dependent on its radius. We can discuss quantitatively the rate at which an increase in the radius of a circle would change the area of the circle. The derivative measures the rate at which a change in a variable causes change in a dependent quantity.

We will now see that the algebraic formulas for derivatives developed in the previous lecture make perfect intuitive and geometrical sense in situations other than motion, as well. We start with the square. The area of a square is $A(x) = x^2$, where x is its side length. Algebraically, we know that the derivative is $A'(x) = 2x$. That is, for every unit increase in the side length x, the area increases by roughly $2x$. But where does this come from?

Let's see visually what the derivative means. That is, we want to compare the area of one square with the area of another square that is slightly bigger. Traditionally, we use the term Δx to measure the change in the length of the side. At a given side length, we can compute how much the area would increase if we increased the length of the side. The influence of a change in the length of the side varies depending on how big the square is to start with. Note the fraction we get when we look at the ratio of the change in area over the change in the length of the side: $\frac{(x+\Delta x)^2 - x^2}{\Delta x}$.

We find the exact rate of change by taking the limit as Δx goes to zero and arriving at the derivative. For the larger square, we take the augmented rectangular area of one side, which is $x\Delta x$, and add it to the augmented area of the other side, $x\Delta x$, and finally, we add the area of the smaller square where the added lengths meet $(\Delta x)^2$.

We then divide by Δx, so we have:

$$\frac{(x\Delta x + x\Delta x + (\Delta x)^2)}{\Delta x} = 2x + \Delta x.$$

As Δx gets smaller and approaches 0, it becomes negligible, and essentially, we arrive at $2x$, exactly as predicted by the algebra.

Let's now consider the volume of a cube. The volume is $V(x) = x^3$, where x is its side length. Algebraically, we know that the derivative is $V'(x) = 3x^2$. That is, for every unit increase in the side length x, the volume increases by roughly $3x^2$. But where does this come from? As with the square, at a given side length, we can compute how much the volume would increase if we increased the length of the side. Let's look at how much volume we add if we increase the length by Δx on all three sides of the cube. The influence of a change in the length of the side varies depending on how big the cube is to start with. In any case, it adds a layer or "slab" of extra volume on three faces of the cube. The added volume for one side is $x^2\,\Delta x$, where Δx indicates the added thickness. We do the same for all three sides and arrive at $3x^2\,\Delta x$. To find the rate of growth, we divide by Δx and arrive at $3x^2$. We have four other tiny additional pieces of volume right at three edges of the cube and at one corner, but as Δx grows smaller and approaches 0, these tiny parts of the volume become increasingly negligible and tend to disappear, and we see the rate of growth is $3x^2$, exactly as predicted by the algebra.

Analyzing dependencies and change is one of the most fundamental things we do to try to understand our world.

Let's consider how the derivative would be defined in the case of the area of a circle viewed as dependent on its radius, $A(r) = \pi r^2$. Again, from the previous lecture, we know $A'(r) = 2\pi r$; that is, for every unit increase in the radius, the area increases by roughly 2π, or a little over 6 units. Why does this make sense? At a given radius, we could compute how much the area would increase if the radius were increased a certain amount Δr. The ratio of that change in area divided by the change in radius Δr would yield a rate of change of the area with respect to a change in the radius. The additional area (called an *annulus*) is in the shape of a thin washer or ring or belt whose width is Δr and whose length is the circumference of the circle, $2\pi r$. Thus, the additional area is approximately equal to the thickness of the ring times the length around the circle; that is, the additional area is about $2\pi r\Delta r$. When we divide that change in the area by the thickness of the ring, Δr, we get a number that is approximately equal to the circumference of the circle, $2\pi r$. This ratio of change of area divided by change in radius again approximates the influence of radius on area. Thus, we obtain an intuitive understanding for $A'(r) = 2\pi r$.

Suppose a belt is put around the equator of the Earth (about 25,000 miles). Suppose we increase the length of the belt by 7 feet so that it hovers above the Earth. How far away from the Earth will it be? The circumference of a circle of radius r is $2\pi r$. From the previous lecture, we know that the derivative is 2π. This means that every unit increase in r results in roughly 2π (or just over 6) units increase in the circumference. Most people guess wrong about the belted Earth, assuming that the longer belt will hover just slightly off the ground. In fact, the derivative tells us that adding 7 feet to the circumference of the belt will result in the belt being more than a foot away from the Earth! In other words, we have added more than a foot to the radius of the belt. We can understand this result better if we look at the derivative, 2π. A 1-foot change in the radius of the belt brings about roughly an addition of 6.28 feet to the length of the circumference of the belt. Therefore, adding 7 feet to the length of the belt would indeed result in the belt being more than a foot away from the Earth.

Now let's consider a stalactite growing in a cave. How quick does a 20,000-year-old stalactite acquire volume compared to its 10,000-year-old cousin? Depending on the environment, stalactites usually grow in cone-like shapes at the rate of 1–4 inches every 1,000 years. We'll assume our stalactites grow at the rate of 3 inches every 1,000 years. We'll also assume that the ratio of the stalactite's base radius to its height is 1:6. If a stalactite is x thousand years old, its height is $3x$ inches and its radius is $x/2$ inches. Therefore, at 10,000 years, the stalactite is 30 inches high and has a radius of 5 inches; at 20,000 years, the stalactite is 60 inches high and has a radius of 10 inches. The volume or weight of the stalactite when it is x thousand years old is $W(x)$ = 1/3 base times height, or $W(x)=\frac{1}{3}\pi\left(\frac{x}{2}\right)^2 3x=\frac{\pi x^3}{4}$.

What is the growth rate of the weight of the stalactite? It is precisely the derivative of the weight. Using our table of derivatives, we see that $W'(x)=\frac{3\pi x^2}{4}$. Therefore, the derivative of the weight at 10,000 years is 75π, while at the 20,000-year mark, it is 300π. Thus, even though the 20,000-year-old and the 10,000-year-old stalactites increase in *height* at the same rate, the older one increases in weight at 4 times the rate of the younger one!

Let's consider how we would define the derivative for dependencies involving supply and demand. We see that the supply curve shows how many items producers would produce if the price they could sell the item at were p. The demand curve shows how many items consumers would buy if the price were p. In this graphical representation of the curves, notice that if the price increases, the supply increases. Also notice that if the price increases, the demand decreases. The place where the demand curve and the supply curve cross is the point where we expect the price to be. What does the derivative measure, when applied to the supply and demand curves? If the curves are steep near the crossing point, slight increases in price would cause major increases in supply and major decreases in demand. These are elastic goods. Goods that are insensitive to price, for example, gasoline, are called inelastic.

We have now seen the importance of the derivative. The derivative analyzes change. The derivative gives a quantitative view of how one quantity will change when another quantity on which it is dependent changes. Analyzing dependencies and change is one of the most fundamental things we do to try to understand our world. The derivative in general arose from first analyzing the simple case of a moving car. That example illustrated one example of dependency—position was dependent on the time. We then saw that the same analysis would serve in many cases where one quantity is dependent on another. ■

Suggested Reading

Any standard calculus textbook, sections on applications and interpretation of derivatives. (These explanations may be difficult to find in standard texts.)

Questions to Consider

1. Consider an animal whose weight varies according to height according to the function $w(h) = h^3$. Suppose the strength of the animal's legs at a given height is given by $s(h) = h^2$. Could similar animals exist that are 10 times as tall? Does this help explain the limit to the size of land animals?

2. Suppose the population increases by 1% each year. Does the derivative of the population function (that is, the function whose value is the population at each time) increase, decrease, or stay the same over time?

Abstracting the Derivative—Circles and Belts

Lecture 7—Transcript

Welcome back. As you recall, we're in the midst of a sequence of three lectures about the derivative and seeing the manifestation of the derivative in many different settings. In the last two lectures, we saw the derivative first geometrically, associated—that is, graphically associated with a graph and the fact that it records the slope of the tangent line at each point of a graph. And, then, in the last lecture we looked algebraically how if we're given an expression, an algebraic expression of a function, we can compute an algebraic expression for its derivative.

Well, the true power of the derivative lies not only in its ability to help us understand change with respect to time, like the car moving down the straight road, but also to deal with dependencies of all different types—whenever we have one quantity dependent on another, that's a place where the derivative can come into play. For example, the area of a circle is dependent on its radius; and we can write a description of an area in terms of its radius, a familiar formula for the area of a circle. And then we can discuss the rate at which the area of the circle changes when we change the radius. That's an opportunity for the derivative to give us some insight.

In the previous lecture, remember, we saw how we can compute things algebraically, and so we can easily compute algebraically a formula for how the radius of the circle is going to influence the area of the circle; that is, a change in the radius influences the change in the area. So, what we're going to do today is to look at some applications of the derivative and think about the geometric manifestations of them, first, and then in some more abstract settings at the end of the lecture.

So, the first one we're going to do is start with the square. Now, the first thing to think about with a square is we're going to remind ourselves of the derivative formulas that we deduced in the last lecture, and, in particular, this one here, that the derivative of $x^2 = 2x$; meaning that if we draw the graph of x^2 for example and we focus our attention at any point x and we go up to the graph, we'll find that the slope of the tangent line at that point is exactly $2x$.

Well, let's think about the area of an actual square. So, here we have an actual square; and the area of the square is the length of one side times the length of the other—they're the same, so it's just $x \times x$. So, when we think of x^2 as meaning something, geometrically, it means that if we have a length x and we construct an actual physical square whose sides are each x, then the total area of that square will be x^2.

Well, let's see what the derivative means in terms of this geometry. I find this sort of interesting, by the way, that if you take this square and now let's go ahead and do the derivative procedure. What is the derivative procedure? The derivative procedure is to say that we take—we want to compare the area of square with the area of a square that's slightly bigger; and, traditionally, we use the term Δx. In other words, we're saying what is the area of a slightly bigger square? Now, this is a square each of whose sides is $x + \Delta x$. So, here we have this bigger square. Now, let's look at the definition of the derivative. The definition of the derivative says that we look at the value of the function at the bigger moment, $x + \Delta x$, minus the value of the function at point x, and divide by Δx. And, then, take the limit as that fraction goes—as Δx goes to zero—we look at that ratio of the difference between the bigger, from the original one, and then divide by Δx and see how that ratio behaves.

Let's interpret that ratio in a completely physical way. Here we have this larger square. So, on our graphic we see that the larger square has sides $x + \Delta x$, and then we're going to subtract the area of this bigger square, from that we're going to subtract the area of the original square we started with. And, we can see geometrically that the area, the augmented area, the additional area, is shaded here; it is the area that sits on these two sides of the square.

Well, now let's see if we can understand the idea of taking this additional area and dividing by Δx. Well, let's see, the area of this rectangle that goes up to the top here and down, it has width Δx, and we're going just up to the top of the original square. So, this height is x; the width is Δx; so, the area of this square right here, this rectangle right here, the area of this rectangle right here, is x $\times \Delta x$. Likewise, the area of this rectangle up here is $x \times \Delta x$. Those two rectangles encompass most of the additional area. Then we have this little square up here in the corner that's also added. It's $\Delta x \times \Delta x$.

Notice that when we divide by Δx, we divide this additional area by Δx, well, if we take this rectangle and divide by Δx; we take the area of that little thin rectangle and divide by Δx, what do we get? We get x. When we take this rectangle, which is—the area of this rectangle, it's $\Delta x \times x$; when we divide by Δx we simply get the length, x. So, what we're getting when we take this additional area and divide by Δx, we're going to get three contributions to that difference quotient. One of them is going to be x, contributed by this rectangle; another one is going to be x, contributed by this rectangle; and then we have this $\Delta x \times \Delta x \div \Delta x$, which is just Δx, which is getting tinier and tinier and becomes negligible as Δx becomes zero. But whereas, these two—the area of this one divided by Δx—for any Δx—will continue to give us this value of x.

The point is that the derivative of the area function can be viewed geometrically. It's saying that if you take a square and you imagine that square growing in size. So, we're imagining this square getting bigger and bigger. You see how we see this square growing bigger and bigger and bigger, that the rate at which the area is increasing is going to be equal to, geometrically, this length plus this length. That's the rate, meaning that when we take those lengths times Δx, that's going to be a good approximation to the additional area that we're creating. So, it's the rate of growth of this square.

So, I find this interesting because it shows that the derivative of the area formula, by viewing this dynamically, you see, it's this dynamic vision of the way we think of this extremely static-looking square, but if we think of it as a growing object, starting at this one corner and growing up bigger and bigger squares, the derivative is telling us the rate at which the area is growing. And, we see it geometrically by the length of those two lines.

So, this is the first example of a geometric interpretation of the meaning of the derivative. Let's do another one just to get in the swing of it. Going back to our formulas for our table of derivatives, we saw that the derivative of x^3 is $3x^2$. We saw algebraically why that's true. We actually went through the derivation and saw that that is the derivative. Let's interpret it once again geometrically. Well, the formula x^3 is referring to the volume of a cube whose sides are x. So, here is an example of a cube. If its side is x, all three of the lengths are x in every direction, then the volume of this object is $x \times$

$x \times x$. That's the volume of a cube. Now, can we envision in our minds what the derivative of x^3 means with respect to this geometrical manifestation of the function x^3?

Well, what does it mean? It means that if we think of this cube as growing; that is, this is a small cube, suppose we had a bigger cube, a bigger box. We could imagine this cube growing from one corner to become a bigger cube. Now, as it's growing, the sides, each of the sides is growing. It becomes longer. And, we're going to ask ourselves, what's the rate at which the volume is increasing? Well, let's look at it geometrically and see the rate at which the volume is increasing.

Well, if you take a cube, and you think about anchoring the cube at one corner, and having the lengths of the three sides increasing, then we see that when we add—we start at a given length x and add Δx to that length; we add Δx to the height; we add Δx to this length coming out the back, we can see that the additional volume, the change in the volume, is a layer of material on three sides of the cube. It adds three layers on the side of the cube: a slab on this face, a slab on this face, and a slab on the back face. Now, we're adding that amount of material, but then to see the rate at which the volume is increasing, we divide by Δx the amount of increase. Well, what happens when we divide by Δx? Let's look at the slab on this face of the cube. That face has area $x \times x$—x^2—and then it adds the thickness Δx; when we divide by Δx, we simply get the area of the face of that cube, x^2. Likewise, on the top face, doing the same procedure, the volume of this additional material just on the top is $x \times x$ — x^2 — $\times$ the thickness, Δx. When we divide by Δx, we just get x^2 again; and the same for the back.

Now, notice there are three other—or four other pieces of volume that we have not accounted for in this geometric visualization of the derivative process. The three additional pieces of volume occur right along this edge, right along this edge, right along this edge, and then right in the far corner, the diagonally opposite corner. Notice the part on this edge is a little—it's actually a square times x; and on this side, a square times height x. It's a little tiny parallel pipe; a very thin box. We have three of those, and then we have a little tiny cube here in the corner, which is $\Delta x \times \Delta x \times \Delta x$. Notice that as Δx goes to zero, this proportion of the volume becomes increasing negligible;

and, when we divide by Δx, those things disappear. The main contribution to the volume is occurring on these three faces. The fact that there are three faces, each of which have sides x^2, is telling us that the rate at which the cube is growing in volume is $3x^2$. So, notice that we have once again seen a correspondence between the derivative formula that we previously had deduced from the last lecture; that the derivative of $x^3 = 3x^2$, and now we've seen physically what that means. What does it mean, $3x^2$? It's the three faces that are growing when the cube is growing.

Okay, so this is the case of a cube. Let's go ahead, since we're doing these geometric objects, let's go ahead and do the circle. The area of a circle, as you know, is πr^2. Now, what that means is that if you take a circle, such as this one, and you measure the radius of the circle, and then you just take that radial measurement, whatever it is, you square it, and multiply it by the number π, which is 3.1415926… and so on, which is the ratio between the diameter of a circle and its circumference. πr^2 will automatically give you the exact area inside this circle. That's what it means to have a dependency of the area on the radius, πr^2. So, you don't need to take each individual circle and measure how many square inches it has; you can just find out how many inches the radius is, and then do this computation, π times that radius times itself, πr^2.

Let's see geometrically what—well, let's see, first of all, algebraically; let's remember algebraically what the derivative of that area formula is. The area formula is πr^2, and we saw in the last lecture that the derivative is obtained by bringing down this exponent 2, the constant π —π is just a constant number, and so it's preserved in the derivative. The derivative of the formula that gives us the area of the circle as a function of the radius, that derivative is $2\pi \times r$. So, that we saw algebraically; let's look at it geometrically and see what it's telling us.

Well, if we have a circle of radius r, and then we increase that circle—we take the circle of radius r, so it has a certain area. Then, we increase the radius by Δr, so we increase the radius by Δr; and we're trying to see the rate at which the area of the circle is increasing as a function of the rate at which the radius is increasing. In other words, for a unit increase in the radius, how much increase in area do we expect to have happen? Well, let's look at it geometrically. Geometrically, what we're saying is that if we look at a circle that's a slightly bigger circle than our circle of radius r. It's slightly bigger by having a radius of $r + \Delta r$. Then that circle just circles around, it's slightly bigger than the original one, and notice that the difference in area, the additional area, is what's called an annulus; it's like a belt, whose width is Δr, and then whose length goes around the circumference of the circle. So, an approximation of the area of that additional area, an approximation would happen if we cut it and thought about straightening it out. Now, of course, you couldn't actually straighten it out because the inner length would be not quite the same as the outer length. The inner length, here, you see, is smaller, shorter, than the outer length. So it wouldn't quite be straighten-outable, but if you did that you would get an approximation to that area in that annular region.

Now, what's the process of the derivative? It's saying the rate at which that area is growing. So, we take that additional area and divide by Δr—that's the change in the radius—to get the rate; that is, the amount of area per added radius. Well, if we think about the approximation of this area as the circumference of the circle, $2\pi r$, that's the inner distance around this circle, times Δr, that's the—so, this is an approximation to the additional area; when we divide that by Δr to get the rate at which the additional area is getting constructed, we simply get $2\pi r$. So, once again, we see that there's a relationship between the formula for that boundary of the circle and the derivative of the area function. So, there's a correspondence between the geometric interpretation of the derivative and the algebraic interpretation of derivative.

We're going to do one more example of changes in a geometric form, and this example is an example that has to do with the Earth. So, suppose you took the entire Earth, and imagine the Earth to be literally a perfect sphere, of course it's not quite, but imagine it's a perfectly round sphere. Now, suppose we take the entire equator of the Earth, which is about 25,000 miles around, and we put a belt around the Earth—and you see this very attractive belt here in this graphic, just snuggly around the Earth. It's exactly along the Earth; and the Earth is perfectly round so it just touches it at every point. It just literally fits right there on that belt. It's exactly the length of the equator, about 25,000 miles. Now we say to ourselves, "Well, you know, I think that

that belt is a little tight. The poor Earth needs a little bit of breathing room." And so, we decide to take the belt and loosen it up a little bit; take the belt and loosen it up. The way we'll loosen it up is we'll just, as you do with a belt, we'll just add a little bit more length to the belt. And just to be specific, let's add another 7 feet to the belt. So we make it—it's about 25,000 miles + 7 feet. You know, this much extra. Then we center it around the Earth. So, it's slightly off from the equator of the Earth, you follow me? So, it's just a little bit off.

Now, I'm going to ask you a question to see how your intuition is. How far off from the Earth do you think that belt would be? So, you've got this 25,000 miles around, and then you've added 7 feet, and how far off the Earth would it be. In other words, if this were part of the Earth, would the belt be just a very tiny bit off, or would it be this far off, or would it be this far off—all the way around? You've centered it so that it's literally centered around the Earth, and it's not touching. I'm going to give you a few seconds to think about this and make a guess, and let me ask you specifically, do you think you could crawl under it, for example?

Well, most people, their intuition says that if you've got something that's 25,000 miles around, and you add 7 feet, most people thing that that belt is going to be just barely off the ground. True answer? The belt is going to be a little more than a foot off the ground; and you can actually crawl under it, everywhere on the entire equator, it's more than a foot off the ground. How can we see that? Well, we an see it by thinking about what that increase of the circumference of the circle, which is adding distance to the belt, what does that correspond to? We're asking the question of the relationship between a change in the circumference and a change in the radius. Because, you see, the change in the radius is corresponding to how far off that belt will be from the Earth. That's what the change in the radius is.

So, let's look at this formula. Let's look at the formula that makes a relationship between the radius of a circle and its circumference. The formula is very simple. It is that if you have a circle of radius r, then its circumference is $2\pi r$. So, in the case of the Earth, you can think that the radius of the Earth is something like 4,000 miles, and so the circumference of the Earth is 2π times that radius; π is about 3, so this is a number that is a little more than 6, more than 6.25, but less than 7. That is, 2π is less than 7.

We saw that the derivative of this function is very simply computed since the graph of this function is a straight line with slope 2π; the derivative is simply 2π. Now, what does that mean? That means that for every 1 foot change in radius, it corresponds to a 2π, or about 6.28 and so on feet change in circumference. So, those two things correspond. In other words, if you increase the radius by a foot, you increase the circumference by $2 \times \pi$, which is about 6 feet. So, we can see that, in fact, increasing the circumference by more than 7 feet actually makes the radius have to increase by a rate—by more than 1 foot. So, this is something that often strikes the intuition a little aslant. So, that makes it fun.

Let's now turn our attention to another physical phenomenon, and this is the growth of a stalactite in a cave. Suppose we imagine that we're in a cave and we see stalactites growing from the ceiling. Now, a stalactite is—we're going to be viewing just conical stalactites; that is, stalactites that look exactly like cones. So, this is an example of the shape of the stalactite that we're considering. And as the stalactite grows, it adds material to it and it becomes a larger and larger cone. Now, that is to say, that's the model that we're going to be discussing in this next session. So, we're imagining that the cone grows so that it maintains the same shape; that is to say, the same ratio of its radius to its height. And, specifically, let's write down a formula for that. Let's say that for a stalactite of age x 1,000 years; let's suppose that it's $3x$ inches in height. So, stalactites don't grow very fast, by the way, so in 1,000 years we're imagining that it grows 3 inches. Okay? And then we're imagining that it maintains it shape; the shape being that the radius is always 1/6 of its height. So, it's sort of a pointy cone. That is, its radius is $x/2$—if you measure the years in thousands of years, then the radius is going to be half of that many inches in radius.

So, for example, at the 10,000-year mark of the stalactite, its height will be 30 inches and its radius will be 5 inches. At the 20,000-year mark its height will be 60 inches and its radius will be 10 inches. Now, the weight of a stalactite is proportional to how much volume there is in that stalactite. So, we can write down a formula for the volume of the stalactite, which we can interpret as its weight, if we use the correct value of weight per volume, it's proportional to the volume, so we'll just call this the weight of the stalactite, and we can compute it because we know that—in fact, we'll prove in a later lecture that the volume of a cone is equal to 1/3 the area of the base times the height. Don't just take my word for it for now, but later you won't have to take my word for it; we'll actually demonstrate that that's the case. But, in this case, it's $\frac{1}{3}\pi\left(\frac{x}{2}\right)^2(3x)$, which is $\frac{\pi x^3}{4}$. That's the formula for the weight of the stalactite at every moment x where x is the number of thousands of years old it is. So, here we have it, and you can see that at 10,000 years old it's 250π pounds, and at 20,000 years it's $2{,}000\pi$ pounds in weight.

Now, here's a question for you. Do you imagine that the derivative of the weight function at 10,000 years is greater or smaller than the derivative of the weight function at 20,000 years? So, this is a question in interpreting the meaning of the derivative. It's talking about the rate of change of the weight of the stalactite with respect to time. Well, let's think about it. We know that the stalactite is growing at a constant height rate. But, then, material is being added to the outside as well. So, the volume increase for a unit change in the height, when the cone is smaller you're going to have a smaller layer of material, and therefore lighter, when you add 1 inch of thickness to that small stalactite. When you have a bigger stalactite, and you add 1 inch of thickness to it by adding 1 inch of length, then the material is greater. Therefore, what we're saying, the interpretation of it with respect to derivatives, is that the derivative of the weight at the longer time is going to be bigger than the derivative of the weight at the lesser age. So, this is an example, and we can actually compute it here using the formula for the derivative, we see that the derivative at the 10,000-year mark is 75π; that's the amount of additional weight per change in thousands of years; and at 20,000 years it's going to be 300π change in weight per change in year.

Okay. Let's do just one more quick example, and this is that we can also apply the analysis of looking at graphs and derivatives having to do with an economic situation. Suppose we have two graphs here, a supply and demand curve, where, depending on the price of an object, the suppliers are going to be inclined to produce more of them if the price goes higher. On the other hand, the demand curve is going to go down when the price goes higher. And, at some point, if these things cross, that is the place where you would expect the price to be, because if the price were higher than that, then more would be supplied than would be demanded, and so you wouldn't sell them all; and, if the price is less than that crossing point, then the supply is less than the demand, meaning that if the demand is bigger than the supply, then, if you increase the price, you still have more demand than supply, and so you could sell more at a higher price, and that would be an incentive for the people to supply more of these goods.

Well now let's look at how to interpret two different supply-demand graphs. If the supply and demand graphs cross sharply with sharp derivative, then this is an example of an elastic good, meaning that the price would tend to be very stable because any deviation from that price would lead to a greater imbalance. Whereas, if the supply and demand curves were almost tangent at the point of their crossing, then this is called an inelastic good, meaning that the amount of demand remains about the same for a large range of prices; things like how much gasoline you buy is of this character, because you have to have a certain amount of gasoline, regardless of the price of it, more or less; and so, there tends to be a big range in the variation of price of gasoline. So this is just an example of a case where by analyzing the slopes of functions and their relationship, we can deduce something about the world; in this case, in economics.

In the next lecture, we're going to turn our attention to some ancient developments that were precursors to the interval. I look forward to seeing you then.

Circles, Pyramids, Cones, and Spheres

Lecture 8

In this lecture, we're going to ask the question how do we find the formulas for areas of objects, such as circles, and for things like the volumes of solids that are more complicated than a cube—whose volume we can easily see—but, how about the volumes for things such as cones, or pyramids, or spheres?

We can deduce each of these formulas by dividing the object into small pieces and seeing how the small pieces can be assembled to produce the whole. The area of a circle, πr^2, is a wonderful example of a formula that we may just remember with no real sense of why it's true. But we can view the circle in a way that shows clearly whence the formula arises. The process involves a neat method of breaking the circle into pieces and reassembling those pieces. This example and others illustrate techniques of computing areas and volumes that were ancient precursors to the modern idea of the integral.

Greek mathematicians had a keen sense of integral-like processes. We look first at an ancient process for discovering the formula for the area of a circle. Remember that the number π is the ratio of the circumference of a circle to its diameter. A circle of radius r can be broken into small wedges. The wedges can be assembled by alternately putting one up and one down to create a shape almost like a rectangle. As the wedges are made ever tinier, the assembled shape more and more closely approximates a rectangle. The top plus the bottom of the rectangle is precisely the circumference of the circle, therefore, of total length $2\pi r$. Thus, the top is πr long. The height of the rectangle is ever closer to r. The rectangle is approaching a rectangle of height r and width πr; hence, it has area $r\pi r$ or πr^2, the familiar formula for the area of a circle.

The derivative gives a dynamic view of the relationship between the area of a circle and its radius. We find that the derivative of the area (the change in area divided by change in radius) must equal the circumference of the circle. When we add thin bands to a circle to increase its size, then divide by the increment that we made to the radius, that division gives us the length of the circumference. Therefore, the derivative of πr^2 equals $2\pi r$. The derivative measures how fast the area of the circle is changing relative to a change in the radius.

The area of a triangle is dependent only on the height and base, not on whether or how much it is leaning. The area of a right triangle is easy to calculate because it is half of a rectangle. We can see that the area of any triangle depends only on its base and height by sliding thin pieces to the side to create a right triangle. Therefore, we find that the area of any triangle is equal to 1/2 the base times the height. In fact, we can use this formula to calculate the area of a circle by imagining the circle divided into tiny triangles. All the triangles have the same height (r), and the sum of the bases of all triangles equals the circumference of the circle ($2\pi r$). Therefore, the sum of the area of all the triangles equals the total of the bases ($2\pi r$) times the height (r) divided by 2, or πr^2, again, the familiar formula for the area of a circle.

The area of a triangle is dependent only on the height and base, not on whether or how much it is leaning.

We can determine the volume of a tetrahedron (a pyramid over a triangular base) by thinking of sliding its parts as we did with the triangle. The volume of a tetrahedron is determined only by the area of the base and its height, regardless of where the top point of the tetrahedron lies. It is difficult to compute the volume of a tetrahedron, so we start with the volume of a prism. The area of a prism is the area of its base times the height. Three tetrahedra of the same volume fill up a prism; thus, the volume of a tetrahedron is 1/3 the area of its base times its height. Once we know the volume of a tetrahedron, we can determine the volume of a pyramid. A pyramid can be seen as two tetrahedra if its square base is divided into two triangles. The volume of the pyramid, therefore, is equal to1/3 the area of half of its base times its height plus 1/3 the area of half of its base times its height. In other words, the volume of a pyramid equals 1/3 the area of the base times the height.

The volume of a cone is easy to compute once we know the volume of a tetrahedron. A cone can be approximately filled up by tetrahedra each with the same point as the cone point and with the bases of the tetrahedra on the base of the cone. Given that the volume of each tetrahedron is 1/3 the area of its base times its height and the height is the same as the height of the cone, then the volume of the cone is also 1/3 the area of the base of the cone times its height.

The surface area of a sphere can be computed by breaking that surface into small pieces. The area between latitude lines on a sphere is the same as the area of a band around the surrounding cylinder if the band is contained between parallel planes that intersect the sphere on the two latitude lines. The smaller radii for latitudes near the North Pole are accompanied by the "slantiness" at those higher latitudes in such a way that the area on the sphere between parallel planes near the North Pole exactly equals the area on the sphere between parallel planes of the same fixed distance apart near the equator. We can see that this area equality holds by looking at a picture of a circle and finding two similar right triangles that tell the story. Thus, the surface area of the sphere is exactly the same as the area of the surrounding cylinder, or $4\pi r^2$.

These examples show ancient ideas that resemble the modern idea of the integral. The ancients did not have a well-defined idea of what happens at the limit, but their arguments are persuasive and can now be made mathematically rigorous. ■

Suggested Reading

Boyer, Carl B. *The History of the Calculus and Its Conceptual Development*.

Dunham, William. *Journey through Genius: The Great Theorems of Mathematics*.

Questions to Consider

1. In thinking about the surface area of the sphere, why couldn't we approximate the area of the northern hemisphere by just thinking of triangular wedges going from the North Pole down to the equator and viewing each wedge as a triangle?

2. Is there a philosophical reason that the formulas for areas and volumes are so relatively simple? In other words, could you imagine a geometric world in which the area of circle or the surface area of a sphere was not so simple? There is a whole different world of ideas associated with non-Euclidean geometries in which such formulas are not as simple.

Circles, Pyramids, Cones, and Spheres

Lecture 8—Transcript

Welcome back to *Change in Motion: Calculus Made Clear***.** In the last lecture we were talking about the derivative and seeing how the derivative is expressed in its relationship to the areas of objects and the volumes of objects and their formulas. But one thing that we didn't do last time is to say why the formula, for example, for the area of a circle is correct. So, in this lecture, we're going to ask the question how do we find the formulas for areas of objects, such as circles, and for things like the volumes of solids that are more complicated than a cube—whose volume we can easily see—but, how about the volumes for things such as cones, or pyramids, or spheres?

Well, it turns out that we can deduce each of those formulas by taking the object in question and dividing it up into many, many small pieces, and adding up those small pieces to get the formula for the whole assembled object. That concept, the concept of dividing things up into little pieces and adding them together, is the concept of the integral. So, what we're really going to be doing today is talking about the integral, an introduction to the integral, except that these are ideas that many of which occurred thousands of years before the integral was actually defined as we know it today. So, what we're talking about here is the concept that underlines the concept of the integral and how it appeared in deducing the formulas for areas and volumes. Let's get started.

Let's begin with the formula for the area of a circle. So, here we have a circle, and, of course, we all know the answer. The area for a circle is r^2. Now, first if all, let's remember what π is. π is the number just above three—3.14159 and so on, just above 3; it's the ratio of the circumference of a circle to the diameter. In other words, the circumference is π times the diameter. Now, our question is: Why is it that a circle has area, π , that ratio, times the radius squared? Well, here's a very clever way to see that that formula is correct. Let's just take a circle and break it up into wedges. This, by the way, is important to do because that reminds us that π, if we think of the circle as a pie, then this is a good way to remember, breaking it up into pieces. So, we break it up into pieces, and here I've taken this circle and broken it up into eight pieces, and I'll just put two of them back here so you can see that, in fact, these are sectors of the circle, there are eight of them, and all I'm going to do is take them and rearrange them, put them cleverly one up, one down, one up, one down, and arrange them in this kind of a pattern. Well then we see that the area of these eight—one, two, three, four, five, six, seven, eight—these eight pieces, the area is, of course, the same as the area of a circle. And it's beginning to look a little bit like a rectangle.

Well, what happens if instead of dividing the circle into eight pieces, suppose that we divided it up into 16 pieces? Now we'll go to graphics so we can see this more effectively. Here's an example where we've divided the circle into 16 pie-shaped pieces, and put them alternately up and down, up and down, up and down. Now, notice something about this process. The circumference of the circle, which, remember is $2\pi r$ because 2 times r is the diameter and π times the diameter is the circumference of the circle, so we know that the total circumference of the circle is $2\pi r$, and yet that entire circumference, since half of the wedges are facing upward and half are facing down, we see that the total, sort of scalloped, top edge of this figure is exactly π times r in length; likewise, the bottom is also π times r in length; the totality being the circumference of the entire circle. And this slightly slanted length is r, the radius of the circle.

But now let's just think. What happens when we take this circle and divide it instead of into 16 pieces, into a million pieces, or a billion pieces, or a trillion pieces? When we do that, and then alternately put those pieces up-down, up-down, up-down, this becomes indistinguishable from just a straight line whose side is π times r in length; this will become a vertical line of length r; and the product or r times π of r, which is the area of a rectangle, is exactly equal to the area of a circle. So, this is a derivation of the formula for the area of a circle.

By the way, let me remind you, just as we did in the last lecture, that the derivative of the area does equal the circumference of the circle, and the reason for that is, as we discussed the last time, that when we add a small bit to the area of a circle, and then divide by the increment that we made to the radius, then that division is going to give us something that is, basically, just the length of the circumference. And, as you see, the derivative of πr^2 is equal to $2\pi r$. So, everything is fitting together.

Well, let's just move on and consider some other very familiar shapes and talk about how we're going to deduce the areas of these familiar shapes. Let's start with a very simple one, namely, a triangle. Suppose that we take a triangle—here, I'm going to move this picture of a triangle here. Now, if I take a triangle, and it's a right triangle, the area of a right triangle is very easy to compute, because if we put two right triangles together, we simply get a rectangle. And the area of a rectangle is very simple because it's just the base times the height. So, in that way, we see that for a right triangle, the area is equal to the base times the height, divided by 2. So, it's very simple to find the area of that kind of a triangle.

But, suppose that we have a triangle that is not a right triangle. How are we going to find its area? Well, here's what we're going to do. We're going to take our triangle and divide it up into little tiny pieces, and here we've given an indication of these horizontal slices that are making up our triangle. Now, notice something about a triangle and how it relates to other triangles in the following way. Suppose I take this triangle, and I simply slide the pieces over into a different configuration—see? When I slide the pieces over to this different configuration, notice that every single one of these horizontal strips simply slid over and is exactly the same length as it was before. And the reason for this is simple: that, if we have the same base to a triangle, and the same height, if we consider a band right in the middle, the area of that band is going to be exactly half—that is, the length of the band that's halfway up is going to be exactly half of the length of the base, because they're similar triangles. So, that means that when we do this sliding, we're going to get the same area as an area of a right triangle that had the same height and the same base.

Let's look at some graphics that illustrate this. We can slide the top of the peak of this triangle, slide it over to the side, and get sort of a slaunch-wise triangle, and notice that each of these horizontal bands just slides over neatly and fills up that slaunch-wise triangle. So, the area of a triangle is determined entirely by the length of its base and its height. It doesn't matter where that top point is in relation to the triangle. And, yet, we know that a right triangle has area 1/2 its base times its height, so we see that the area of every triangle is just the base times the height, divided by 2. By the way, it also gives us another way to see what the area of a circle is, because if we imagine a circle being made of tiny little triangles put together, we can see that the area of each triangle is its base times its height divided by 2. If we add up all of these triangles, we just get a bunch of triangles who have the same height, namely the radius r, and the sum of those bases is going to equal the circumference of the circle, $2\pi r$. So, the sum of all of those little tiny triangles would just equal the total of the bases, which would just be $2\pi r$, times the height, r, divided by 2, because each triangle has base times height over 2 as its area.

Now, let's take our philosophy of sliding things over and look at another kind of an object whose volume is a little bit trickier to figure out. Let's consider this object. This is a tetrahedron. It's like a pyramid over a triangular base. This is a tetrahedron. We'd like to find what the volume of this tetrahedron is, and the way we're going to analyze this is by realizing that a tetrahedron can be—although we don't know what it's volume is, we don't know a formula for its volume, that's what we intend to deduce in a minute. But one thing that we do know is that the volume of a tetrahedron that with a certain fixed base, as we have here, and a certain height, will be the same as a tetrahedron that has the same height and the same base, regardless of where the top point of the tetrahedron lies. Isn't this amazing? Because, you see, as I slide the top of the tetrahedron to different positions at the same height above the base—I'm trying to keep it at exactly the same height—you notice that all of these horizontal bands just slide and neatly fill up that tetrahedron. So, what that tells us is that the volume of a tetrahedron is determined by the area of its base and its height, and it's not determined by, for example, where that top point is. First of all, let me just say, I think this is a really neat kind of an illustration that shows when you move that top along all of those, like that middle one just slides and fills up exactly the correct position. It just finds its way.

So, our goal is to find the volume of a tetrahedron. Well this is a little bit tricky. The way we're going to do it is we're going to take an object that is a prism—now, this is an example of a prism. Its like—you've seen prisms that break light into colors. It's easy to compute the volume of a prism, because it's just the area of the base times the height. What's difficult is to determine the equation for the volume of a tetrahedron. But, what we're going to do is be very clever and see how a prism is made up of tetrahedra. So, here we have a picture of a prism, and we're going to take that prism, realize that its volume is the area of its base times the height—that's the volume of an entire of a prism, the area of the base times the height. If we take that prism, we're going to divide into three tetrahedra. Now, this is a little tricky to see, so you're going to have to focus for a second here. What I've done here in this illustration is to take our prism and illustrate that it's the union of three tetrahedra. One of the tetrahedra is this red one. The red one is obtained by taking the triangular top of the prism, and then taking the cone point down at this bottom vertex. So, you see that red thing is a tetrahedron. This is its base on top, and then the cone point is down here. Likewise, the white tetrahedron is constructed by considering the base of the prism, that's a triangle, and then taking this vertex at the top and coning up—remember a tetrahedron is like a pyramid over a triangular base. Well, the triangular base is the white base of the prism.

The tricky one to see is this blue one. This blue thing is, in fact, a tetrahedron, and the way to see it is this: the front part of the tetrahedron is a triangle—do you see this blue triangle? It's half of the rectangular side of this prism. So, this half is a triangle—that's a blue triangle—and then we cone down to this opposite point, and that creates a blue tetrahedron, and that blue tetrahedron creates the third tetrahedron that makes up the prism. The question is: Why do we think those three tetrahedra have the same volume? But, they do.

It's easy to convince ourselves that the volume of the white tetrahedron is equal to the volume of the red tetrahedron. The reason is that they both have identical sized bases, namely the top and the bottom of the prism, and then the height of the tetrahedron, the red one, is just the height of the prism and the height of the white tetrahedron is the height of the prism again. So, the white and the red tetrahedra are clearly the same volume. But now look at the white tetrahedron compared to the blue tetrahedron. The white tetrahedron can be viewed, instead of having its base at the bottom of the prism and then its cone point up at the top; instead, view it as the base of the tetrahedron being this side, this triangular half of this rectangular face of the prism, this white triangle, and then coning to this point. You see, a tetrahedron—you think of it as this being the cone point over a triangular base, but there's no reason we can't think of this as the base and then coning to this vertex. Any one of the sides of a tetrahedron can be viewed as the base and coning to the opposite vertex. In this way, we can see that this white tetrahedron can be viewed as having a base over here and this as the cone point, but then it's clear that its volume is the same as the blue tetrahedron, because, you see, the blue tetrahedron has an identical area base as this white one does, because they're just the two halves of a rectangle, and they also have exactly the same cone point. So, their height is also identical, therefore the volume of the blue is the same as the volume of the white, which is the same as the volume of the red tetrahedron. So, those three tetrahedra have the same volume as each other; and, consequently, the total volume of one tetrahedron is just one-third of the volume of the entire prism, which is the area of the base times the height. So, this has allowed us to deduce this formula for the volume of a tetrahedron. It's 1/3 the area of the base times the height.

Well, once we know the volume of a tetrahedron, we can also figure out what the volume of a pyramid is because, look, a pyramid can be written as just two tetrahedra. Now by the way, remember that a pyramid has a square base, and then it has a cone point over it, like the Great Pyramid at Giza. So, here is this great pyramid—you can divide it in half by simply taking the base and drawing a straight line down the middle, and then you can see that you're coning down to one of those triangles, making a tetrahedron half of it, and the other one is the other tetrahedron, another half. So that means that the volume of this pyramid is equal to 1/3 the height of the pyramid times the area of half of the base, plus 1/3 the height of the pyramid times the area of the other half of the base. By the distributive law, since both of those involve 1/3 times the height and we're just adding the areas of the base together, we see that the volume of a pyramid is just the area of the base times the height, divided by 3.

Well, that concept allows us to also compute the volume of a cone, because just think of a cone as being made up of many, many little triangles, and, of course, we might have to have infinitely many of them to fill up the base of the cone, but a cone, you see, has an area of the base—which could be thought of as adding a bunch of triangles together—and then it has a height, so the volume of the cone is equal to 1/3 the area of the base time the height, because it's just the sum of this formula for each of the tetrahedra that we can imagine to be making up that cone.

So, this process has allowed us to compute the volumes: formulas for the volume of a tetrahedron, for the volume of a pyramid, and the volume of a cone. Well, that's very nice. In fact, we can take these ideas and now we're going to try something that's a little bit trickier, perhaps, and that is to compute the surface area for the surface of a sphere. So, here we have a sphere and we imagine this sphere to have radius *r*, and our question now is going to be, what is the surface area of this sphere? Now, we can think of it as, for example, the Earth, and we're trying to find what is the surface area of the Earth. This is actually quite a tricky kind of concept because, of course, it's curved—it doesn't have a flat bottom with a cone or anything like that—we've got to figure out how to compute the surface area of the sphere.

Well, our strategy—as our strategy has been for all of these methods—is to break something up into small pieces and add those small pieces together. And that, by the way, is the philosophy of the integral, and that's why we're doing this lecture, because it's introducing the concept of the integral, which is to break things up into little pieces and add them together. So, let's go ahead and see if we can find a way to compute the area of the surface of a sphere.

The way we're going to think about it is this, and I'm going to pose a question to you: suppose that somebody were to make you an offer to buy some land on the Earth, and here's the way they made the offer, they said, "I'm going to take a slab of the Earth," and what I mean by that, I take a straight line from the North Pole to the South Pole, and I'm just going to cut it perpendicular to that line with a plane. So, that is a plane through a fixed latitude line—do you follow me? We're just going horizontally through the Earth in a plane parallel to the latitude. Now, we're going to take another latitude that's, say, a mile down. So, these are two parallel slices, go through two latitudes that are 1 mile vertical distance down, and they cut through the Earth, see, in a slab, and the boundary of the Earth, then, has this thin strip that goes all the way around. Now, here's a question for you: suppose somebody said to you, "I'm gong to give you an area of the Earth, and you can take it either right at the equator, where you have vertical distance of 1 mile, and it goes all the way around the equator, and all the area in that—you can take that, that is one of your options; or, alternatively, right about halfway in between here, the top and the bottom, we'll take this vertical distance of 1 mile, again, slice through and you could have that area, if you'd prefer. Or, even, right at the very top, you can take the area that goes from tangent to the North Pole down 1 mile, vertical distance 1 mile, and take the area of the Earth between those two bands. Which one would you choose to get the most area?"

Now, let me point out why this is sort of a challenge. At the equator, when you take this band that has this vertical distance 1 mile, the Earth is basically vertical at the equator. So, you're getting a strip that's sort of like a vertical belt. But, when you're halfway up in between the equator and the North Pole, the Earth is tilted somewhat, so although the circle at that latitude has a smaller radius, nevertheless, since it's tilted, the tilt gives more area. You see, because it's tilted, you have a longer strip along the Earth, and yet you have a smaller radius. So, these things have to be balanced. At the very top, you'd have just a little—like a northern polar cap—cap; it would be like a little cap—which one do you think you'd prefer, in the sense of just surface area? Of course, we're not talking about what's cold and what's warm, and you like the tropics—we're not talking about that; we just want to know which has the most area.

Well, the surprise is that, in fact, all of those options have exactly the same area. That is, if you take a sphere and you cut it by two parallel planes, the area, the surface area on that sphere between any two is going to be precisely the same, if the distance between the parallel planes is the same. Let's see why this is true.

Here's a picture of where we're imagining cutting our surface in this fashion, here between these two parallel planes, and I'm going to put a cylinder around the sphere, and the cylinder is just snuggly fits on the outside of the sphere—as you see—and the claim is going to be this: That, no matter where we cut the sphere, the area between the cuts on the sphere is precisely the same as the area of this band on the cylinder that's on the outside of the sphere.

I want to make absolutely sure you understand this, so let me draw this picture of what's going to happen when I cut right here between the sphere and, you see, I'm going to have a band here, and then I'm going to have this tilted blue part here; it's tilted because at that, you know in the northern hemisphere, the Earth, or a sphere, is going to be somewhat tilted. Now, we're going to compare those areas and we'll see that they're exactly the same. Now, here's how we're going to do it: imagine cutting the sphere by a vertical circle, and if we cut it by a vertical circle—so, let me make sure you're following this, we have this cylinder going around it like this, and now I'm going to cut with a vertical circle, right through this way, and it's going to cut through this outside cylinder in two vertical lines—a vertical line here and a vertical line here—and then it's going to cut the sphere in a circle, a great circle. Here we have it on the graphic—here are the vertical lines where it's cutting the cylinder and here is that vertical circle that cuts through the sphere. Now we're going to look at this diagonal piece of the sphere that's being cut, and then we have this vertical piece between these two vertical planes, here and here; we're going to see that this diagonal piece has a certain relationship to the other parts of the figure.

At this height, the radius of the sphere is s, that is if I go from this point right in the middle out to the outside; since it's near the top, its radius is smaller than r, and, in fact, we're going to call that s. This is the radius, s. The radius of the whole sphere, and therefore this circle, is r. The diagonal line, we're imagining we're taking such a small part of the sphere that the curve of the sphere is indistinguishable from a straight line, and we'll call its length Δs, and Δh is the vertical height. Now, this figure may look familiar to you because we actually saw it in a previous lecture, and notice that this triangle here is similar to this triangle up here, and I'll tell you why. They are both right triangles, and this angle here is equal to the opposite interior angle of a line cutting two parallel lines; so this angle right here is equal to this angle here; Δs is on a line that is tangent to the circle and therefore perpendicular to r; and, consequently, this angle is 90° minus this angle; and therefore this angle is exactly the same as the angle at the origin here.

In other words, if you didn't follow all those details, the fact is that this little triangle is similar to this bigger triangle; and, since they're similar triangles, we see that the hypotenuse of the small triangle, Δs, divided by Δh, the leg that is next to our angle that is the same as this angle, that that ratio is equal to the same thing as the hypotenuse of this bigger triangle divided by s, the radius at this height of the circle. Therefore, we have this little formula.

Well, once we have this formula, just a little tiny bit of algebra, taking this formula and cross-multiplying, multiplying both sides by 2π tells us that $2\pi r$

Δh, which is equal to the area of this belt around the cylinder, is equal to $2\pi\ s\Delta s$, which is this diagonal length times the circumference of a circle of radius s. Consequently, as promised, the area on the sphere of that slanted slice is exactly equal to the vertical area on the boundary of this enclosing cylinder. Since that's true for every single slice, we can add them up and see that the area of the sphere is exactly equal to the area of this enclosing cylinder, but the enclosing cylinder has circumference $2\pi r$, it has height $2r$, so its area is $4\pi r^2$. So, the sphere, the surface area of the sphere, is also $4\pi r^2$—four times the area of the circle of radius r.

Well, I think that all of the strategies for finding these equations for areas and volumes that we saw involved breaking up pieces of what it was we were trying to get the formula for, and breaking them up into little bits and adding them together. Those are ancient ideas that very closely resemble the modern idea of the integral. And the ancients had this wonderful concept that they used all the time, but they didn't define it in the clear way that we see as the limit of a process. But their arguments were very persuasive and lead to mathematically correct ideas that we can now make completely mathematically rigorous.

I look forward next time to showing you an amazing example of this kind of logic that Archimedes used to deduce the equation of the volume of a sphere. I'll see you then.

Archimedes and the Tractrix

Lecture 9

Today we're going to look at a couple of examples of this strategy that preceded the definition of the integral, and then some examples that actually are very modern.

In the 17th century, **Bonaventura Cavalieri** analyzed shapes using his *method of indivisibles*. If one thinks of the surface of a sphere as comprised of many tiny triangles, then the volume of the sphere can be viewed as made up of many tiny tetrahedra with those triangles as bases and the center of the sphere as the top of each tetrahedron. Because we know that the volume of each tetrahedron is 1/3 the area of the base times the height, then the volume of the sphere will be 1/3 the surface area of the sphere times the radius (the height of each tetrahedron). Because the surface area of the sphere is $4\pi r^2$, as we saw in the last lecture, the volume of the sphere is 1/3 of the product of that area times the height of each tetrahedron, which is r, thereby giving the formula for the volume of the sphere, namely, $(4/3)\pi r^3$.

Archimedes had an amazing way to discover the formula for the volume of a sphere of radius r. His method involved a lever. He balanced a cone (with a base of radius $2r$ and a height of $2r$) and a sphere (of radius r) on one side of the lever with a cylinder (with a radius of $2r$ and a height of $2r$) on the other. Archimedes's method for showing that the objects balance involved dividing the sphere, the cone, and the cylinder into thin slices and hanging those slices on the lever. Originally, one could picture the cylinder being in its horizontal position with the cone and sphere neatly inside it. Archimedes's insight was that if we take a thin slice through the cylinder (thereby cutting through the sphere and cone also), that thin slice of the cylinder (by itself, left where it is, at point x distance from the fulcrum) would be exactly counterbalanced if the slices of the cone and sphere were both moved to the other side of the lever at distance $2r$ from the fulcrum. Because that insight is true for each slice, the totality of all the slices all the way along the cylinder shows that the cylinder, cone, and sphere will balance on the lever as described above.

In our demonstration, we can see that placing the sphere and cone exactly $2r$ from the fulcrum balances the cylinder. The cylinder lies horizontally along the lever with one end at the fulcrum and the other at $2r$. Both the cone and the sphere are hung from the same point on the other side of the lever, namely, at the point that is distance $2r$ from the fulcrum. If we know that the cone, sphere, and cylinder balance, and we know the volumes of all the objects except the sphere, then we can deduce the volume of the sphere. Thus, Archimedes found that a sphere of radius r has a volume of $(4/3)\pi r^3$. Today, we would formalize this procedure of slicing up a sphere by using integrals.

In the 21st century, Mamikon Mnatsakanian devised an ingenious method for computing areas by breaking up regions into pieces that are sectors of a circle.

Let's examine the relationship between the surface area of a sphere and its volume. Knowing the meaning of the derivative, we know that the derivative of the volume is telling us the rate at which the volume is changing relative to a change in the radius. Geometrically, that rate of change is the volume of a thin layer over the surface of the sphere divided by the thickness. As the thickness gets tiny, that fraction will simply equal the surface area. Using our knowledge of derivatives, we know algebraically that the derivative of $(4/3)\pi r^3$ is $4\pi r^2$, which is the formula for the surface area of a sphere. The derivative of the volume of the sphere must give a formula for the surface area of the sphere—and we see it does.

In the 21st century, Mamikon Mnatsakanian devised an ingenious method for computing areas by breaking up regions into pieces that are sectors of a circle. The area between two concentric circles can be computed in two ways. Just subtracting the area of the smaller circle from the area of the larger circle is one method to arrive at the area of the ring (annulus). The Mnatsakanian method is to view the annulus as a polygon with many, many sides. We can sweep and shift the small triangular segments that make up the annulus and see that the sum of those segments will be the area of a circle. We find that the area of the annulus is πa^2, where a is the distance from a tangent point on the small circle to the outer circle. This method also provides a proof of the Pythagorean Theorem.

This method can be used when computing the area under a tractrix. A *tractrix* is a curve created by pulling one end of a string along the x-axis while the other end, attached to a pen, starts on the y-axis and is dragged along to create the curve. One way to compute the area under the tractrix is to view that area as a sweeping of tangent lines and to approximate the area as segments of a circle. The method shows that the area is simply equal to a quarter of a circle $(1/4)\pi a^2$. The hard way to compute the area under the curve involves finding the formula for a tractrix.

Both Archimedes's and Mnatsakanian's methods involve breaking up an object into small pieces and adding up their contributions. This strategy is the fundamental strategy of the integral. ■

Names to Know

Archimedes (c. 287–212 B.C.). Ancient Greek mathematician, physicist, astronomer, inventor, and prolific author of scientific treatises. He studied hydrostatics and mechanics and discovered the general principle of the lever, how to compute tangents to spirals, the volume and surface area of spheres, the volume of solids of revolution, many applications of the method of exhaustion, and an approximation of the value of π, among other work. Archimedes was killed by a Roman soldier when Syracuse was conquered during the Second Punic War.

Cavalieri, Bonaventura (1598–1647). Italian mathematician; professor at Bologna; student of Galileo. He developed the method of indivisibles that provided a transition between the Greek method of exhaustion and the modern methods of integration of Newton and Leibniz. He applied his method to solve a majority of the problems posed by Kepler.

Suggested Reading

Boyer, Carl B. *The History of the Calculus and Its Conceptual Development*.

Dunham, William. *Journey through Genius: The Great Theorems of Mathematics*.

Mnatsakanian, Mamikon. *Visual Calculus by Mamikon*. www.its.caltech.edu/~mamikon.calculus.html.

Questions to Consider

1. Experiment with a level to see that the formula for balancing on a lever is correct. See how putting two weights on one side is equivalent to putting one weight at a different location. What location?

2. Deduce the formula for the area of a circle using Cavalieri's method of indivisibles.

Archimedes and the Tractrix

Lecture 9—Transcript

Welcome back. Remember that in the last lecture we began a process of seeing how we could find the formulas of the areas and volumes of familiar figures using techniques that involved chopping up those figures into little piece and adding them together. Today we're going to look at a couple of examples of this strategy that preceded the definition of the integral, and then some examples that actually are very modern.

We'll begin in the 17th century, when a mathematician by the name of Bonaventura Cavalieri analyzed shapes and found the formulas for the areas and volumes of various geometrical figures using a method that was called his "method of indivisibles." So, I wanted to just tell you how he deduced the equation for the volume of a sphere using the kind of knowledge that we just figured out from the last lecture.

Remember, in the last lecture, we deduced what the equation was for the volume of a tetrahedron. Remember that the volume of a tetrahedron turned out to be the area of the base times the height divided by 3; 1/3 the product of the area of the base times the height. Now, what Cavalieri did was to say let's imagine a sphere as composed of a collection of tetrahedra, basically, where the base of the tetrahedron is on the surface of the sphere, and they just cone down to the center of the sphere. So, here in this graphic, we can see an example of one of these little imagined tetrahedra that cone down to the center of the sphere. Well, if we imagine covering the entire surface of the sphere with little tiny tetrahedra, so that the fact that the tetrahedras's bases are flat doesn't make any significant difference, the total of the volume of each tetrahedron is the area of its base times the height divided by 3, but since we're talking about summing over the sum of the bases of the tetrahedron being equal to the entire surface of the sphere, then we can see that the volume of the sphere will just be the surface area of the sphere times the radius, which is the height of each of those tetrahedra that we're thinking of putting together to compose this total solid sphere, divided by 3; that is, the surface area of the sphere, which last time we computed to be $4\pi r^2$, so that's the total surface area of the sphere—and think of it as the sum of the areas of the bases of all of these tetrahedra that are filling up the sphere—

times the height of each tetrahedron, which is just r, each one is r, then times 1/3. So, the volume that we get for the equation that Cavalieri's method gives us for the volume of the sphere is then just $1/3r$ times the area of the sphere, $4\pi r^2$.Combining them gives the correct formula that the volume of a sphere is $\frac{4}{3}\pi r^3$.

So, that's one method for finding the formula for the volume of the sphere, but the one that I think is really amazing is a method that Archimedes devised. Archimedes was a mathematician from Syracuse, who lived in the 3rd century B.C., and he was just amazing and did just all sorts of incredible things, but this one may take the prize, because in this one he used his method of levers to find the formula for the volume of a sphere. You may remember the way levers work—this is the way a lever works: If you have a lever, for example a teeter-totter, and you have a heavy person on the teeter-totter and a little child on the teeter-totter, in order to get the teeter-totter to balance, the heavy person stays closer to the fulcrum and the lighter person goes further out, because a teeter-totter will balance if the product of the distance times the weight on one side is equal to the distance times the weight on the other. That's the principle that Archimedes used in computing the volume of a sphere, and let me show you how he did this.

The story starts with three objects. One is the sphere, whose volume we wish to find a formula for. So, this is a sphere. This is a cone. Now, this cone has exactly the same height as the diameter of the sphere; that is $2r$. So, this height right here, from the top of the cone down to the base is $2r$, and notice that this is exactly the same height. Then, the base of this cone, by the way, is exactly $2r$ in radius; in other words, the diameter of this is equal to the radius of the base of this cone. Likewise, this is a third object. This object is a solid cylinder. It has radius, once again, $2r$; the radius is $2 \times r$, that is the diameter of the sphere, and you can see that the height of this cylinder is exactly the same as the height of the sphere, and it's exactly the same as the height of this cone.

Okay, why do we have these three things? It's not at all clear. So, let me explain how Archimedes thought about this. He imagined putting all three of these things in the same location. Now, of course, you can't actually do this, but we're imagining this cone as being inside of this 3-dimensional solid, and this sphere as being right in there as well, and arranging them in this following way, which we can visualize, even though, of course, we can't literally put them—because they're solid we can't put them in the same place, but we can visualize them as being in the same place—and this is a picture of this big cylinder here, with radius 2*r* enclosing the sphere, whose diameter is 2*r*, and enclosing this cone, here having base the same as the base of the cylinder, and coning to the fulcrum point.

Now, of course, if we put all three of these objects on one side of this lever, it's obviously not going to balance. So, this is where Archimedes starts to analyze how to balance pieces. The way he thinks is the following: Let's just chop this entire picture by a plane, a very thin slice that we take that is some distance away from the fulcrum. If we do that, it slices through all three of our objects; it slices through the big cylinder, and it creates a disk whose radius is 2*r*. It slices through part of the cone; and it slices through a part of the sphere. Here's a picture of those three things; they're just concentric disks inside one another. This is what Archimedes observed. He observed that if you take the disks that come from the sphere and from the cone—the disks that were exactly at this place and move them over to exactly 2*r* distance on the other side of the fulcrum—so he hangs a disk over exactly point 2*r*, and he hangs another disk, the disk from the cone, from the cone and the sphere, they are hanging at distance 2r from the fulcrum, and then he just leaves the big disk from the cylinder exactly where it is, at this point x distance away from the fulcrum. The fact is that those objects balance one another. They will stay in exact balance if they're hung from this fulcrum.

The reason that they hang from the exact balance, we can actually compute this using this diagram. This is a sideways slice of seeing our sphere here, the cone coming out, and then the big cylinder is not pictured. Using this diagram, we can compute what the radius is of the part of the sphere that is distance x away from the fulcrum. And the answer is its radius is $\sqrt{2\boldsymbol{x}} - x^2$ and the radius of the cone—the radius of the disk that comes from the cone—well, since this distance is x away from the fulcrum, and the cone has the property that when it goes 2*r* distance this way, it goes 2*r* distance up, then that means that if we're x distance away from the fulcrum, then the radius of the cone at that point is also x.

Using that formula and the Pythagorean Theorem, we can actually deduce that, in fact, those disks put at distance 2*r* from the fulcrum will exactly balance a disk of radius 2*r* that's left alone at that place x. Let's not worry about the details, but realize the implications. The implication is this: That if we do this balancing of the part of the cylinder at every point from the fulcrum out to 2*r*, and we accumulate all of the disks that we get from the sphere and all of the disks that we get from the cylinder, we'll have the entire sphere here, the cone, all hanging from distance 2*r* away from the fulcrum, and then the big cylinder is just going to be fixed where it began, and the claim is that those two things will balance.

Now, what I'm going to do now is demonstrate that they balance by actually, physically doing it. Here we have the cylinder, the sphere, and the cone, and I'm going to attempt to balance them. So, I'll take them over here. So, on one side of this lever—here we have this lever; it's a balanced object here—on one side we have the entire cylinder sitting exactly where it began. Now, balancing the sphere and the cone exactly distance 2*r*, that is, the diameter distance away from the fulcrum, exactly balances this cylinder. Well, first of all, it's, I think, rather impressive that this actually works. So, that is something to celebrate. I'm impressed that this is balancing, but what it does is it allows us to take our knowledge of what the volume equation is for a cone, and what the equation is for the volume of a cylinder, and the fact that they balance will allow us to deduce what it is that is the formula for the volume of a sphere. Let's go ahead and do the computation. Here it is:

We know that these two objects, the sphere and the cone, are both hanging at a distance 2*r* away from the fulcrum. Now remember, in order for something to balance, the distance times the weight on one side has to equal the distance times the weight on the other. Well, the distance away is 2*r*, and then the weight is the weight of the sphere plus the weight of the cone. On this side of the equation, we have this cylinder balanced. Now, of course, it's spread out between zero distance all the way up to 2*r* distance, but that's the same as having all of its weight exactly distance r away from the fulcrum. So, its distance from the fulcrum is r times the weight of the cylinder and these two things balance because they balance on this lever. So, we know the formulas for the cone—it's 1/3 the area of the base times the height; we know the formula for the cylinder—it's just the area of the base, π x $2r^2$ x its height, and it's distance r away, and this side is $2r$ away. So, these things balance. Now we just do a little bit of algebra, and going through there we solve for a sphere, which is the volume of a sphere, and we see that the volume of a sphere is, indeed, $\frac{4}{3}\pi r^3$. Now, this is an amazing way to deduce the formula for the volume of a sphere. So, Archimedes was an amazing, amazing mathematician, and this is just one of many wonderful things that he accomplished.

By they way, let's just notice something about the equation for the volume of the sphere, that if we imagine taking the derivative of the volume of the sphere—remember, the formula for the derivative is just taking the volume of a radius plus a Δr minus the volume of the sphere over Δr, we see that, on the one hand, we know that when we divide this incremental increase in volume by this incremental change in the radius Δr, we just get something that's a thickening up of the surface of that sphere divided by Δr; and, on the other hand, we get the fact that we can take the derivative using the formulaic method that we saw in the previous lecture, we saw that its derivative, the derivative of $\frac{4}{3}\pi r^3$, bring down the 3, reduce the exponent, we get $4\pi r^2$, which, remember, was the formula for the surface area of the sphere. So, once again, we see that the geometry fits together with the beautiful mathematics.

Now, what we're going to do is take a step into modern times, and this is actually rather amazing to me, that I'm now going to be talking about some work that was published in papers that were written in the 21st century—that is just a few years ago—by a mathematician by the name of Mamikon Mnatsakanian and Tom Apostle. The two of them wrote up this work, and these ideas are owing to Mamikon Mnatsakanian, and these are beautiful concepts. Of course, these did not precede the definition of an integral, but they very well may have been ideas that Archimedes himself had thought of. But, of course, much of Archimedes's work was lost, and Mnatsakanian himself has said that maybe Archimedes actually did think of these ideas. But, I think that you'll find them very intriguing. Once again, they illustrate the concept of breaking up regions into small pieces and fitting them together as a method for computing areas.

So, here's what Mnatsakanian observed. He said suppose that you have an object, for example an octagon. Now, here I've drawn an octagon; you can probably see this eight-sided figure here, and Mnatsakanian said the following: suppose we do the following procedure. That is, we simply take a sector of a circle and extend one of the legs of the octagon and just sweep down until we are touching that leg of the octagon. Then just shift if forward so that we can just sweep some more, sweep some more, shift it forward so we can sweep some more, and so on, all the way around the octagon. Now, if we do this, that is, we always are taking the same radial distance and sweeping it down, each of these pieces is a sector of a circle. The totality of all of these sectors we get, since the sum total of everything that we're doing is to sweep all the way around a circle; that is, we sweep, and then shift over; sweep, shift over; sweep, shift over; sweep, shift over; and so on, that all of these pieces put together will simply be a circle. That's pretty clear. And, by the way, it doesn't matter that it's an octagon; it could be a square, it could be a triangle, it could be a 100-sided figure, anything that we do as long as we take a straight line and then just sweep it and go to the next corner and sweep it and go to the next corner and sweep it—once we go all the way around we will, of course, have swept a totality equal to a circle whose radius is the distance of the sector that we swept.

This is interesting and a rather simple observation about things. Let's go ahead and on the graphics see that we can do it with a 16-sided figure. Something to notice, by the way, is that if we take a larger figure, for example a larger octagon, but kept the sectors of the circle exactly the same radius, then the sum of those sectors around them will still be the same total area, the same area of the circle. Now, so far these are rather simple observations. We can do the same thing with a small 16-sided figure and a larger 16-sided figure. So far those observations may not be all that interesting, but they have some very interesting consequences. So, let's look at the following:

Suppose we ask the following question: What is the area between two concentric circles? Well, of course, the natural thing to do is just to take the area of the outer concentric circle and subtract the area of the inner concentric circle, and then the difference would be the area of that ring. By the way, that ring-shaped thing is called an annulus. So, that's certainly one way to find it—just take the area of the outer circle minus the area of the inner circle—but let's think about it in a different way, and let's think about it using the Mnatsakanian strategy. You see, his strategy was we'll take this circle, and instead of thinking of it as a round circle, the inner inside circle, let's approximate it by a polygon with many, many different sides. So, for example, we could start with our 16-sided polygon, and then take our radius and put together those segments just as we did before, and we would see that the sum of those segments shifting and sweeping, shifting and sweeping, shifting and sweeping, we know that the sum of those segments, just as we saw in this original picture, will just be the same as a circle; the total area will be the area of a circle whose radius is the distance that we swept. Now, if we approximated this inner circle by a polygon that had, for example, a million sides—a million-sided polygon; in other words, it would be indistinguishable from an actual circle, look what would happen. We would have a little tiny sliver, a little sector, and then another little sector, another little sector, and sweeping those sectors around the totality of those sectors would be indistinguishable from this ring. Yet, the insight of Mnatsakanian is that the sum of those sectors could just be recombined to be just a regular circle. So, the area of this ring is just π x a^2, where a is the distance from a tangent point on the small circle out until it hits the outer circle.

Now, this is rather interesting because look, suppose that we think about having a bigger inner circle, and we ask the question: How big should a larger circle be so that the ring between those two concentric circles would have exactly the same area as the ring over here? Well, we know the answer, because if we use his exact same method, and we start with a larger circle that is hit at distance a away from the tangent line coming out from the inner circle, we know that we could sweep that line around, or, if you prefer, we could approximate this inner circle by a polygon of a million sides and think about putting together all these little sectors, and those would combine to be a circle whose radius is a. This is rather interesting because what it tells us is that if we want two rings to have the same area—it's a very simple thing to do—all we do is we make the distance from the top point tangentially off until it hits the outer ring to be distance a, and then those rings will have exactly the same area.

Well, that's rather interesting. It's interesting in its own right, but also it's surprising that it actually provides a proof of the Pythagorean Theorem.

Now, let me demonstrate this. You see, because what we showed was that if we have a big circle and a small circle, and we wished to know what the area of that ring is, we saw that the area of the ring was just equal to the area of a circle whose radius is this distance a; but, but on the other hand, we also know the original method for finding the area between the area of a ring, the original method that we thought about was to take the area of the outer circle minus the area of the inner circle. Well, let's just do that. The area of the outer circle—let's suppose the radius of the outer circle is R, then the area of the outer circle is πR^2; and if we subtract the area of the inner circle, the inner circle is πr^2; we will get πa^2. That is, we already know from Mnatsakanian's analysis what the area of that ring is—it's πa^2. So, we have this formula, using the analysis, just factoring out—canceling out—the π's, we see that $R^2 = a^2 + r^2$, which is the Pythagorean Theorem. So, we actually can deduce the Pythagorean Theorem from Mnatsakanian's analysis of the area of that ring.

Well, this is really, I think, a very clever concept, and I wanted to show you another example of how this kind of insight can be used to find an area that is very complicated to find using other methods. So, let's clear this off for a minute. What we're going to be talking about now is trying to find the area underneath a curve that's a little bit complicated to draw. So, let's begin by drawing this curve. Now, this curve that I'm about to draw has a name, it's called a tractrix; and I'll first describe what I'm going to do, and then do it. A tractrix is a curve that's obtained by the following thing: We take a point that's vertically above—thinking about it on the vertical axis, the y-axis; it's some fixed distance away from the origin, and what I'm going to do is just drag my hand to the right along the x-axis, and it will cause this pen to be pulled along. The word "tractrix," think of it as tractor, and so it's a way of pulling, and it creates a curve. Here we go, when we do this, I'm just pulling along the x-axis, and you can see it's drawing a curve just by being pulled along. This curve becomes closer and closer to horizontal and it continues to get closer and closer, and eventually gets very, very close to the axis.

Now, it was a very difficult problem for people to write down an equation for this curve, and then to find the area under the curve was a very complicated thing to do; that involved calculus and it can be done with calculus. But I want to point out something about it that allows Mnatsakanian's method to allow us to find the area under this infinite curve in a very clever way. What I'm going to observe is at any point of this, as it's being pulled, notice that this line segment that is pulling it is always tangent to the curve that it's drawing; and the reason that it's tangent to the curve that it's drawing is this: that when I'm pulling to the right on the x-axis, I could resolve the forces of how the forces are acting on this point where the pen is, by looking at one force that goes directly in the direction of the segment, and then a perpendicular force of sufficient size, so that the two of them together equal the force that I'm actually adding along the x-axis. When we have that kind of resolution of forces, we notice that the vertical force, the perpendicular one, has no effect on this curve because it's like a circle, it doesn't have any pulling effect. So that means that the pulling effect is exactly straight along this, the direction of the line between the point and the x-axis, and consequently the line is tangent to the curve that it's drawing.

Well, here's what Mnatsakanian's strategy is: His strategy is to say instead of thinking of a smooth curve, suppose I think of it as a bumpy curve. That is, I go straight for a while, straight for a while, straight for a while, straight for a while; and then I draw a sector of a circle where the radius of the circle is the distance between the point and the point that I'm pulling; that is, the pen and the point I'm pulling, that distance *a*; that's the original height here, this distance is *a*, and then I can approximate the distance underneath the tractrix as a sector, another sector of the circle, another sector of the circle, another sector of the circle; and, if I divide my tractrix into more and more smaller and smaller pieces, the sum of those segments can all be combined back together again to just form a quarter of a circle, you see? We swish a certain amount, we make a sector of a certain angle, and then we start there and we swish another certain amount, and we start there and swish a certain amount, and so on. As they get smaller and smaller, those become indistinguishable from just starting—they can be moved back to be starting at the same radial point, and just construct a quarter of a circle, because it becomes almost flat at the end. So, the total area under the curve is just a quarter of a circle, $\frac{1}{4}\pi a^2$.

Well, you'd be very much impressed with this analysis if you had first worked on trying to write down the formula for a tractrix, which involves hyperbolic trigonometric functions. In any case, I thought you would be amused to see all of these methods that are really examples of the philosophy of how the integral computes things—breaking things into small pieces and adding up their contributions to get their total.

I look forward to seeing you next time.

The Integral and the Fundamental Theorem

Lecture 10

In the 21st century, Mamikon Mnatsakanian devised an ingenious method for computing areas by breaking up regions into pieces that are sectors of a circle.

If a car goes at a constant velocity of 30 miles per hour, it is a simple matter to compute how far the car has traveled during an interval of time. We saw that to deal with varying velocity, we just break the total time into small intervals and add up approximations of how far the car traveled in each small interval of time. In this lecture, we will see the geometric implications of this integral process as we view it in graphical form. In particular, we see that the same process that computes the distance traveled by the car also computes the area between the graph of the velocity curve and the axis. We use Leibniz's notation for the integral because the long S shape reminds us that the definition of the integral involves sums.

After our introduction to the precursors to the modern concept of the integral, we will start a series of lectures that correspond to previous lectures about the derivative. We will see, first, the integral in its graphical interpretation. We will then study the integral in its algebraic interpretation.

Recall how the integral was defined in the case of the car moving down a straight road. We are given the velocity function $v(t)$ and want to compute the total distance traveled. For example, if we know the car was traveling a constant 2 miles per minute for 3 minutes, we know the car traveled a total distance of 6 miles.

Let's look graphically at the scenario of a forward-moving car. Notice that the process of finding the distance traveled involves finding products that are equal to the areas of rectangles. That is, the distance is equal to the product of the height of the rectangle (the line representing a constant 2 miles per minute) times the width of the rectangle (the 3-minute mark on the horizontal axis) of the graph.

Let's look at another velocity function in which velocity is two times the time, or $2t$, so we know the car will be traveling along at an ever-increasing velocity. Our graph in this case shows an upwardly sloping line. To compute the distance traveled, we break the interval of time into small bits and do some adding. We then approximate the velocity traveled within those small intervals of time (half-minute intervals) by assuming our car remained at a constant velocity during that time, then "jumped" to the next constant velocity at the next interval and so forth. We then add the distance for the first interval of time plus the distance for the second interval of time and so forth. Notice again that the process of finding the distance traveled involves finding products that are equal to the areas of rectangles. As the little intervals get smaller, the total of the area of the thin rectangles is getting ever closer to the area between the curve and the axis; that is, the approximations improve. This infinite process of taking ever-smaller intervals of time provides us with a single exact answer.

The definition of the integral just adds up the products, not of the height of the rectangles, but the signed height—positive if the function is positive and negative if the function is negative.

Let us add here that the notation for the integral is Leibniz's, namely, $\int_a^b v(t)dt$. The long S shape reminds us that the meaning of the integral involves *sums*. The *a* and *b* denote the starting and ending times, respectively. In the limit, the answer is exactly equal to the area under the curve. Thus, in integral notation form, $\int_a^b v(t)dt$, which we know is the distance traveled by the moving car, is also equal to the area under the graph of $v(t)$.

Let's look at the specific example where the velocity at each moment is $2t$. Then, $\int_0^3 2tdt$ is equal to the distance the car traveled between time 0 and time 3. But $\int_0^3 2t\,dt$ is also equal to the area under the graph of 2*t*. We can check that area geometrically because the area under 2*t* between *t* = *0* and t = 3 is just a triangle with base 3 and height 6. Thus, the car traveled (1/2)(6×3) = 9 miles. And the area under the graph of 2*t* from 0 to 3 is 9 square units.

We can think about the motion of the car to see some features of the integral. The integral from *a* to *b* plus the integral from *b* to *c* equals the integral from *a* to *c*. This is obvious because it simply says that we see how far we went during the time *a* to *b* and how far we went from time *b* to *c*; the total is how far we went from time *a* to time *c*. Suppose the velocity is negative. When the velocity is negative, we are traveling backward. Then, the integral is telling how far backward we traveled. More exactly, the integral of the velocity is telling us not how far we drove, but how far we end up from where we started. Examples of when we are going forward part of the time and backward part of the time illustrate this concept.

Let's look at the graphical interpretation of integrals again. If the function is below the axis, then the integral is negative. If the function is part above the axis and part below, the integral combines the two. The definition of the integral just adds up the products, not of the height of the rectangles, but the signed height—positive if the function is positive and negative if the function is negative.It's easy. When the graph goes below the axis, the integral is negative; when above, positive. The summation fact, *a* to *b* plus *b* to *c* equals *a* to *c*, works regardless of whether the graph goes above or below the axis.

The Fundamental Theorem of Calculus

$$\int_a^b F'(x)dx = F(b) = F(a)$$

Integrals behave "opposite" of derivatives graphically. For a function *f* (*x*), we can define a function *F* (*x*) as the integral of *f* from starting time *a* to ending time *x*. Think of *f* as the velocity and *F* as the mileage marker. Recall that if a function *f*(*x*) is increasing, then its derivative *f* ′(*x*) is positive, and if a function *f*(*x*) is decreasing, then its derivative *f* ′(*x*) is negative. For the integral, it's the opposite: If the function *f*(*x*) is positive, then its integral *F*(*x*) is increasing; if the function *f*(*x*) is negative, then its integral *F*(*x*) is decreasing. If the function *f*(*x*) is zero, then its integral *F*(*x*) is constant—it's just $F(a)$ because we are not adding any area. In summary, notice that we have $F'(x) = f(x)$.

Now let's turn to an algebraic representation of this same idea. Given the equation for the velocity of a body, we can deduce the equation for its position using the integral. The integral of the velocity is the position. Let's look at an example where we know the answer. Suppose $v(t) = 2t$. We do some calculating and see that the distance traveled is the height times the width divided by 2, or t^2. If we stop at any time, the integral will give an answer. The answer is always t^2.

Recall that the Fundamental Theorem of Calculus relates the integral and the derivative. If $v(t)$ is the velocity at every moment of a car moving down a straight road, then the integral of $v(t)$ between one time and another equals the net distance traveled. Thus, $\int_a^b v(t)dt$ equals the net distance traveled between time a and time b. We saw before that if we can find a position function $p(t)$ whose derivative is $v(t)$, then the integral is easy to compute. It is simply the position at time b minus the position at time a. If $p'(t) = v(t)$, then $\int_a^b v(t)dt = p(b) - p(a)$. Whether the variable is t or any other letter, the relationship is the same. Thus, the Fundamental Theorem of Calculus states:

$$\int_a^b F'(x)dx = F(b) = F(a).$$

Suppose we want to do the integral process (which involves doing infinitely many approximations, each of which is a sum) for some function $f(x)$. If we can find an **antiderivative** $G(x)$ for $f(x)$ (that is, a function such that $G'(x) = f(x)$), then we can get the answer to the integral process just by doing one subtraction, $G(b) - G(a)$. An antiderivative is a function whose derivative is the function whose integral we are trying to take. When we are faced with an integral, our first thought is, "Can we find an antiderivative for the function that is under the integral sign?"

Let's look at some examples. Every derivative formula leads to an antiderivative formula, because we can just go backward. For example, we know that the derivative of x^2 is $2x$. Thus, an antiderivative of $2x$ is x^2. The derivative of a constant times x, $f(x) = cx$, is just c. That is, $f'(x) = c$, or $(d/dx)(cx) = c$. Thus, an antiderivative of the constant function c is just cx. The derivative of $(x^{n+1})/(n+1)$ is x^n. Thus, an antiderivative of x^n is $(x^{n+1})/(n + 1)$.

Why do we say *an* antiderivative rather than *the* antiderivative? Any two functions that differ by a constant value will automatically have the same derivative at each point. We can see this fact graphically. If two functions differ by a constant, then their graphs are merely shifted up and down. The slope of the tangent line above each point will be precisely the same. The derivative is just measuring the slope of the tangent line. For each function, instead of just one antiderivative, we really find one antiderivative, then add any constant to indicate that any shifting of the antiderivative is also an antiderivative of the same function. Just as we saw the table of derivatives in Lecture 6, now we have the table of antiderivatives.

Function	**Antiderivative**
$f(x)$	$F(x)$
x	$x^2/2 + C$
x^2	$x^3/3 + C$
x^3	$x^4/4 + C$
x^n	$(x^{n+1})/(n + 1) + C$
$\sin x$	$-\cos x + C$
$\cos x$	$\sin x + C$
e^x	$e^x + C$

The most important thing to remember about integrals is what they mean. The integral is a number that is the result of doing an infinite process. The process involves approximating the answer. Each approximation is obtained by taking a sum. As we divide the interval up into smaller subintervals, the approximation gets better. For positive functions, the value of the integral is equal to the area under the curve and above the axis. ■

Important Term

antiderivative (of a function *f*): A function with a derivative equal to *f*.

Suggested Reading

Any standard calculus textbook, section defining the definite integral and exploring its properties.

Questions to Consider

1. The integral is defined in terms of sums of products. One number in the product is the value of the function in a small interval. Why does it not matter which point in the interval you choose in defining the integral?

2. Explain why an integral that gives the volume of a solid is somehow adding up pieces of area that have no volume and combining them to create volume. That almost paradoxical perspective was an obstacle to understanding the integral for many years.

3. Use the Fundamental Theorem of Calculus to evaluate $\int_1^2 x^3\,dx$.

The Integral and the Fundamental Theorem
Lecture 10—Transcript

Welcome back. During the last two lectures, we saw some examples of ancient work that were precursors to the modern concept of the integral. In this lecture we're going to start a series of lectures that correspond to the previous lectures that we had about the derivative. As you recall, we explored the derivative's manifestations in various places. We saw its relationship to graphs; and then we saw an algebraic manifestation of the derivative; and then applications of the derivative, interpretations of the derivative in many application settings.

In this lecture we're going to do the first two of those manifestations of the integral; namely, the integral and its graphical interpretation and its algebraic interpretation. So, to begin, let's begin by recalling the definition of the integral, and, in particular, remember that it was generated—the concept of the integral—was generated by our viewing the car that was moving down a straight road where we were looking at the velocity of the car at every moment, and then computing from that information the total distance traveled, the net distance traveled, by the car. And, so, here we go. If, for example, the velocity was at a constant 2 miles per minute, then we would discover that after 3 minutes if we wanted to know how far the car had traveled in 3 minutes, we would simply multiply the 2 miles per minute × the 3 minutes to get a total of 6 miles. And, by the way, I'm going to use this notation and talk a little about it in just a minute; this notation of the integral of $v(t)$.

Notice that that computation, the computation that gave us the total distance traveled, has another interpretation in this graph; namely, the interpretation is that it's the area between the graph of this function, just this very simple constant function, 2 miles per minute, the horizontal line, it's the area underneath that graph and above the horizontal axis, the time axis. Because the height is 2, that's the velocity; and the width is 3, that's the number of minutes that passed. And so 2 × 3 is the area under the curve. So, we see a correspondence between the computation that computed the distance traveled, velocity × time, and a computation that tells us an area, the area underneath the graph that we're taking the integral of. We'll see that this same procedure occurs for the integral in general. Suppose we have another velocity function; velocity at every moment is 2 × the time; so, our car is constantly speeding up as time goes from 0 forward. If we draw a graph of that velocity function, it's just this diagonal line; it's a straight line with slope 2. And, if we wish to compute the distance traveled, recall we had a strategy for that. Our strategy was that we broke the interval between 0 and 3, if we're trying to figure out how far we went during those first 3 minutes, we broke that interval into small increments of time, in this example we've broken it into 1/2-minute intervals, and then we approximated the speed that the car traveled in each of those sub-intervals of time. We approximated it by assuming that the car just stayed at a constant speed during each of those sub-intervals, and then jumped to the next constant speed, and the next constant speed, and so on throughout the interval.

So, we have this characteristic sum, where we have a product of the velocity at a steady velocity for the first interval of time plus a steady velocity for the second interval of time. We multiplied the two to get a distance traveled. If the car had been going at a steady speed for this interval of time, the distance traveled during that 1/2 minute of time would be the rate of speed, that is, the velocity at that time times how long we assume that the car went at that time.—in this case, 2 miles per minute × this 1/2 minute. That's the distance traveled if the car were traveling at that constant speed.

Now, I want to point out two things about this. First of all, notice that the procedure by which we are deducing the distance traveled by this approximating vehicle, that the distance traveled is also computing something else—it's computing the area of each of these rectangles, because each of these products is the product of the height of—that is, the value of the velocity is equal to the height of this rectangle, and then the width is the width of the rectangle. We multiply those together to get the area of each of those thin rectangles, and then we add them together. That is the fundamental defining property for the integral, and when we think of an integral, we should think of it as a process of summing products and sums.

So, here we go. I wanted to make a small comment about this notation. The notation has an elongated letter *s* [$\int$]; it's a long—it's really quite attractive, actually, this long letter *s*—and this was one of the innovations of Leibniz, again. Leibniz had a flair for good notation, which makes a big difference in our ability to understand it. In particular, in the case of the integral, the big *s* reminds us of several things. First, it's the first letter of the word sum, and that's not an accident, it was the first letter of the German word *summe*, which meant sum. Therefore, when we see this long letter *s*, we should think we're going to add up things. Then the question is what are we going to add up? Well, we're going to add up products, and you see that inside the integral side we have velocity, $v(t)$ and then dt. Well, that reminds us that what we do is we look at the velocity at each increment of time, that's the height of one of these rectangles, and then we multiply. This dt was the Δt represents the width of each of those rectangles. So, the product of the width of the rectangle times the velocity—that is the height of the rectangle—gives us the area of that rectangle. And then we're adding them up—that's the *s*—as the *t* varies between the lower limit of integration, the 0, and the upper limit of integration, the 3. So, this notation really does capture and remind us of the defining features of the integral.

Now, we're trying to make a graphical correspondence to this method for figuring out the distance traveled, and the graphical correspondence is that this same procedure is telling us the area under the velocity curve. The area under the curve is going to be exactly the same because as we got finer and finer intervals, we see that, in fact, that long sum of products is just approximating the area under the curve. Well, knowing that, we can see that there are alternative ways for figuring out the distance traveled for our car that's moving at $2t$ miles per minute, because we know that the distance traveled is going to be equal to the area under the curve. Well if a curve is really just a straight line with slope 2; it actually just represents a triangle. It's a triangle whose base is 3, whose height is 6, and we know how to compute the area of a triangle. The area of a triangle is just the base times the height divided by 2. Consequently, we know that since the area of this triangle is 9 units, we also know that the distance that the car traveled, if it traveled always at $2t$ miles per minute, would be exactly 9 miles during that 3-minute time interval. And, in fact, instead of thinking of the distance traveled between time 0 and time 3 if we just thought of a variable time, we just chose any time at all, we would be able to compute the area of the triangle very simply because we would know what the height of the triangle was. If, for example, if we had some time *t*, its height would be $2t$; and, consequently, the width would be t, the height would be $2t$, and so we would see that the area of the triangle would just be t^2.

In general, then, because the characteristic picture of the integral is obtained by just drawing these approximating rectangles, and as the rectangles get smaller and smaller, the product of their height times their width plus the height times width plus height times width plus height times width is just giving us the sum of the areas of the rectangles. We can see that the integral from the point A to the point B of some function $v(t)dt$ will just give us the area under the curve between A and B. So, that is the graphical representation manifestation of the integral; the interpretation of the integral.

Now, this interpretation leads to many simple observations about the integral. One thing is if we have any function at all, and we take the integral between one point, A, and another point, B, and we add to that the integral from the point B to some further point C, we will obtain the integral from the point A to the point C. Why? Because the integral from A to B of our function is just telling us the area under the curve and above the horizontal line axis of that function; and, then, the integral from B to C is telling us the corresponding area in the next part; and, consequently, adding them together will give us the area under the curve between A and C.

We need to make some comment about times when the function goes below the axis; that is, when the value of the function which we're integrating is negative. In the model of the car moving on the straight road, that means when the car is going backward down the road, it has a velocity in the reverse direction. Then we call that the negative velocity, the velocity would be a negative value, but the method of computing the integral is exactly the same as before; namely, we take the velocity times the small intervals Δt and add them to get the distance—it's the signed distance traveled, sign meaning plus or minus distance traveled. Since you're moving in the negative direction—if the velocity is negative, you're moving backwards on the road—then, you're computing for values where the velocity is negative, the value of the integral will be the negative of the area; it will be the negative of the area between the graph and the horizontal line, if the graph goes below the axis. But, the point is, that if you take a graph where some of the time you're traveling in the positive direction, if this is a velocity curve, every point where the velocity is positive means that you're moving forward along the road. When the velocity is negative, you're moving in the reverse direction. If we compute the net distance traveled from some time A to some time C, we can compute this in the following way: During the time when you're velocity is positive, you'll be moving forward along the road; and, during that time, whatever the area is under that part of the curve will be the distance between where you started and the forward distance that you end up. So, in this case, it's 5 miles if we assume the area under these curve is 5.

Now, we're returning backwards. At this point, the velocity is 0; and when it descends below the axis, that means we are now driving in the reverse direction. And the area under the curve between B and C will accumulate the value of the distance traveled in the reverse direction. So, the total net effect is that if we travel from time A to time C, we will have traveled a total that is a net distance of 3 miles. We actually traveled a total distance of 7 miles, 5 forward and 2 back. So, an integral is not a good way to compute your gas usage if you're going sometimes forward and sometimes back. But, it is a good way to tell the difference between where you started and where you end up. Notice that it maintains the relationship that the integral from A to C is equal to the integral from A to B plus the integral from B to C; this relationship is true no matter where the value or the function is, positive or negative.

So, at this point, we have seen the graphical interpretation of the integral as the area under the curve. And, regardless whether it's a velocity curve or if it's just a general curve, $f(x)$, the integral is defined the same because of the idea of mathematical abstraction. There's no reason that we can't substitute the variables x for t or f for p; the integral is defined in the same way.

Now, I'd like to turn our attention to trying to understand the integral from a different point of view; and, to accomplish that, let's first of all think about the idea of the integral as an accumulation of distance traveled over varying distances, varying times, which we will call x. So, suppose that this is our graph of some velocity function that we are considering. So, this is a velocity function; here we're going forward; here we're standing still for a while, the velocity is 0; down here we're turning in the backwards direction, and so on. The integral from A, this fixed number A, this site, this starting time

A. If we measure how far the accumulated, the net distance traveled is for any distance in time, for any time in the future; we're really computing the integral of this velocity function, starting at A and moving forward. So, we have the area of this or the area of this. As our x moves along this axis, the integral of that is going to tell us the net distance traveled. And, if we graph that, we will actually be graphing our position function on the road. In other words, we start here at the beginning. Since this is telling us that the velocity is positive and forward, the integral up to a given point will be some positive value; and, then, here when the velocity is 0, we don't change our position, we're not moving forward or backwards. Here the velocity is negative; consequently, we're moving backward along our road. Then, after we get to this point where the velocity becomes positive, then we begin to move forward on the road again, and that gives us a function. Now, notice that this function is related to the given function that we—the velocity function that we started with; and, this function will actually have its derivative equal to the velocity function. Because, you see, this is a position function determined by the velocity function. So, the velocity function is the derivative of this position function.

Well, we can interpret the graphical meaning of this integral in the following way: If we think of this as a velocity function, then the integral from this fixed point A up to x will have various properties as we think about how the car is moving. So, this is basically summarizing what I just said. That if our velocity function is positive, when it's positive, the accumulated distance will be increasing; the integral is an increasing value. During the times when the function is decreasing, that is, the velocity is going in the backwards direction, then we would expect the accumulated distance, that is to say, the integral, will be decreasing; the integral from a fixed point up to the point x of what you think of this is the velocity function, that it will be decreasing. Likewise, when the velocity function is 0, then the integral stays constant during that time. So, this is showing us that at different points x, whether we would expect, given these velocity properties, what we expect from our accumulated distance, which is our integral.

Now, let's turn to an algebraic representation of this same idea. Suppose that we think about having a velocity function $2t$, and then we say, well, at each point x, what is the distance traveled? Well, the distance traveled is the area under this curve, which is the height, which is $2x$; times the width, which is x; and the area of a triangle is the height times the width divided by 2, which is, if the speed is $2t$, so this is a curve of slope 2, then we find that the integral from 0 to x is the area of that triangle, which is simply x^2.

Now, let me remind you of the Fundamental Theorem of Calculus and how it relates to this issue of the velocity function. Recall that we showed that by taking this sum process associated with the velocity, now we know we get two things when we do that. We get the area under the curve, with the plusses and minuses, depending on whether the curve is above or below the axis, of course; but, it's the area under the curve. But, in addition to it, it is, if this is the velocity function, if we consider this to be the velocity of the car at each moment of time, that integral is going to tell us the net distance traveled between the time A and the time B. Well, we saw the insight of the Fundamental Theorem of Calculus was that if we can find a position function, that is a function that tells us where we are at each moment, with the property that its derivative is equal to the velocity function we're given, then there's this other, simpler way of finding the net distance traveled; namely, just saying well, where were we at time B, and where were we at time A; let's just subtract the two and that will give us the net distance traveled. So, in other words, the area under the curve of a certain function is equal to taking another function whose derivative is what's inside, what you're trying to find the integral of, and subtracting those two values, the value at the top limit of integration minus its value at the bottom limit of integration. That was the Fundamental Theorem of Calculus. It associated the derivative with the integral.

So, the point of the Fundamental Theorem of Calculus is that if our challenge is to find an integral of a function, it suffices to find some other function, in this case capital F, such that the derivative of capital F is the function whose integral we're trying to take. So, our challenge in trying to take an integral, we could say well, what we need to do when we're trying to take an integral is divide up the area between A and B into tiny, tiny parts, and then add up all those little areas, and then take a limit as we take finer and finer divisions of the interval. That is what the integral is telling us. But, that value is equal to the value we get if we can find a function whose derivative is what it is we're taking the integral of.

Now, the point is that this tells us that the burden of taking an integral is solved if we can find what's called an "antiderivative." An antiderivative means a function whose derivative is the function whose integral we're trying to take. We're trying to do the opposite of the derivative process. This, then, is where the algebraic manifestation of integrals is going to appear. If we're trying to take the integral of $2x$ from one place to another, that's what we talked about before, we recognize that x^2 is a function whose derivative is $2x$. And, therefore, if we're trying to take the integral of $2x$ between two points, let's say 1 and 3, knowing that an antiderivative is x^2, we can plug in the upper limit of integration, 3, to get 9; plug in the lower integration, 1, to get 1; and we see that $9 - 1 = 8$ is the integral of $2x$ between 1 and 3.

Let's consider another one. Suppose that we want to take the integral of just x. Well, we find a function whose derivative is x. Here's an example, $\frac{x^2}{2}$ is a function whose derivative is x, and we saw that from the algebraic methods of finding derivatives. And, consequently, we are in a position to quickly and easily take the integral of x between any two values.

What's the antiderivative of a constant function? That is, a function whose graph is simply horizontal. Well, an antiderivative is that constant times x, because the derivative of a constant times x is just the constant.

In general, if we are given a function x^n, and our goal is to try to find a function whose derivative is x^n, we have recourse to looking back at our derivative formulas and realizing that if we take the function $\left(\frac{x^{n+1}}{n+1}\right)$, and we take it's derivative—remember how we took the derivative; we took the exponent and we bring it down, and then we take x to one lower power, and that $\frac{1}{n+1}$ is just a constant, so then the $n+1$'s cancel and we're left with x^n. So, here we are. We see that the derivative of this expression $\frac{x^{n+1}}{n+1}$ is exactly x^n; so we found an antiderivative of x^n.

Now, you may note that in each of these cases, I have talked about an antiderivative. Why don't I say the antiderivative? Why don't I say the antiderivative? When I say, "What is an antiderivative of $2x$?" I said, "x^2 is an antiderivative of $2x$, x^2 is a function whose derivative is $2x$." Well, why didn't I say, "The antiderivative of $2x$ is x^2?" Well, the answer is: that if you have any function, if you add a constant to that function, you'll get another function whose derivative is exactly the same at every single value. So, in the case of x^2, for example, if I take $x^2 + 5$, $x^2 + 5$ has exactly the same derivative as x^2 does. It also has derivative $2x$. So, any constant that I add to a function gives me exactly the same derivative because it's just shifted up the slope of the tangent line at every point is exactly correspondent and is exactly the same slope. Therefore, it has exactly the same derivative. That's the geometric way of seeing that it has the same derivative. An algebraic way of seeing that if you have a function and you add a constant, you get the same derivative. It's just realizing that the derivative of the sum of two functions is the sum of the derivatives, and that the derivative of a constant is 0, because a constant is just a horizontal line.

We're now in a position to take our table of derivatives and associate it with every derivative and antiderivative formula. You see, you may recall that our table of derivatives, we saw that when we had the function x had derivative 1; the function x^2 had derivative $2x$; the function x^3 had the derivative $3x^2$; and, in general, x^n had derivative nx^{n-1} power; and we can then make a corresponding antiderivative chart by just looking at those things backwards. How can we manipulate them to say, "If I'm given a function, what is another function whose derivative is what I started with?" And the idea is just to look at this column and take the derivatives and see that what we get is in this column.

Likewise, notice that the derivative of the sine is the cosine; we mentioned that. Therefore, the antiderivative of the cosine is the sine. Since the derivative of the cosine is –sine, the antiderivative of the sine is –cosine; because the derivative of –cosine is the sine. And, in all cases, we can add a constant and continue to get a function whose derivative is as we want it.

The whole purpose of what we've said here is that we are now in a position to do things, such as take an integral of a value, such as what is the area under the graph of the curve x^4 between 1 and 4? That's a curved line. It would be very difficult to get the exact area in any mechanical kind of a way. But, we can take the antiderivative of x^4, which is $\frac{x^5}{5}$; plug in the upper limit of integration, 4; plug in the lower limit of integration, 1; compute this antiderivative at the value 4; subtract the antiderivative at the value 1; and, get the answer. So, this strategy of taking an antiderivative and plugging in the upper limit of integration and the lower limit of integration is what students come to learn as what it means to do integrals.

So, we've seen now a graphical manifestation of the integral and the algebraic manifestation of the integral; namely, antiderivatives. See you next time.

Abstracting the Integral—Pyramids and Dams

Lecture 11

In this lecture, we're going to be talking about how the integral can be used to apply to questions that are beyond just the idea of a car moving on a straight road.

We saw the power of the derivative in its applications beyond motion: the dynamic view of areas and volumes, the growth of stalactites in caves, and supply and demand curves in economics. Similarly, the integral, when viewed abstractly, is an important tool for understanding diverse dynamical situations such as (again) areas and volumes, as well as engineering. In this lecture, we work out the volume of a pyramid, the volume of a cone, and a solution to an engineering problem: the hydrostatic pressure on a dam.

The integral is important because the process of summing that the integral is performing is precisely what we need to do to solve various problems in various settings.

Areas and volumes are natural applications of the integral. Consider a square with side length x and area $A(x) = x^2$. When we studied the derivative, we saw that $A'(x) = 2x$ essentially represents the change in the area when we increase the side length by 1. This means that for each Δx increase in the side length, the area increases by approximately $2x\Delta x$. Thus, the total area of a square of side length 5 is the sum of all pieces for each Δx between 0 and 5, and it can be realized as an integral and evaluated by the Fundamental Theorem of Calculus: $A(5) = \int_0^5 2x dx = 5^2 - 0^2 = 25$.

Now consider a cube with side length x and volume $V(x) = x^3$. When we studied the derivative, we saw that $V'(x) = 3x^2$ essentially represents the change in the volume when we increase the side length by 1, remembering that adding to the length of the side adds a layer of extra volume on three faces of the cube. Thus, again, for each Δx increase in the side length, the volume increases by approximately $3x^2\Delta x$. The total volume of a cube of side length 5 is simply the integral of this function: $V(5) = \int_0^5 3x^2\,dx$. We can compute the integral using the Fundamental Theorem of Calculus. Because an antiderivative of $3x^2$ is x^3, we simply evaluate x^3 at the upper limit of integration and subtract x^3 evaluated at the lower limit of integration: $V(5) = \int_0^5 3x^2\,dx = 5^3 - 0^3 = 125$

Now, let us look at more complicated examples, such as computing the volume of a pyramid with a square base of side length 200 ft and a height of 200 ft. We can approximate the pyramid by a stack of slightly thickened squares placed on top of one another, in which the squares get smaller as we get near the top of the pyramid. The volume of each thickened square is easy to write down, namely, it is the area of the square times the thickness. The area of a square h units from the top is h^2, so the volume of each slice is approximately $h^2\Delta h$. Adding up the volumes of those thickened squares gives an approximation to the volume of the pyramid. Thus, the total volume of the pyramid is $\int_0^{200} h^2\,dh$. To evaluate the integral, we use the Fundamental Theorem of Calculus. An antiderivative of h^2 is $h^3/3$, so $\int_0^{200} h^2\,dh = \frac{200^3}{3} - \frac{0^3}{3} = \frac{200^3}{3} = 2,666,666\text{ft}^3$.

Let's compute the volume of a cone of base radius 3 and height 4. We can view the sideways cone as skewered on the x-axis, and we can think of slicing it up into thin slices as we would do to a loaf of bread Each slice is approximately the same volume as a slightly thickened disk, and the total volume of the cone is approximately equal to the sum of the volumes of those small slices. Using similar triangles, we can see that at point x, the radius of the disk is $3x^3/4$, so the area of the disk is $\pi\left(\frac{3x}{4}\right)^2$, or $\frac{9\pi}{16}x^2$; thus, the volume of a slice is $\pi\left(\frac{3x}{4}\right)^2\Delta x = \frac{9\pi}{16}x^2\Delta x$. The total volume of the cone is $\int_0^4 \frac{9\pi}{16}x^2\,dx = \frac{9\pi}{16}\left(\frac{4^3}{3}-\frac{0^3}{3}\right) = \frac{36\pi}{3} = 12\pi$, or 1/3 times the area of the base ($\pi 3^2$) times height (4). Again, note the use of the antiderivative of x^2: $x^3/3$.

The integral is important because the process of summing that the integral is performing is precisely what we need to do to solve various problems in various settings. Suppose we are building a dam and want to know the total hydrostatic force on the face of the dam. The pressure (that is, force per square foot) varies with depth. The pressure is greater near the bottom of the dam than at the top. The total force is obtained by adding up how much force there is at each depth. Suppose the dam is 100 feet wide and 40 feet deep. Then, we could think about dividing the face of the dam into narrow strips of Δh height for the width of the dam. The amount of force on a strip at depth h is the product of the area of the strip, $100\Delta h$, and the pounds per square inch of water pressure at that depth. The water pressure at a depth h is about $62.5h$ pounds per square foot.Thus, the force on a strip of height Δh that lies at depth h is about $(62.5h)(100\Delta h)$ pounds, so that the total force is $\int_0^{40}(62.5h)100\,dh$. We can evaluate this integral via the Fundamental Theorem of Calculus. Given that an antiderivative of h is $h^2/2$, we can calculate that the total hydrostatic force on the face of the dam is 5,000,000 pounds. ■

Suggested Reading

Any standard calculus textbook, section on applications of the integral.

Questions to Consider

1. How does the integral help us view the area of a circle in a dynamic way?

2. Hooke's Law states that, within certain limits, the force exerted by a spring that is stretched to x units beyond its resting length is a constant times x. Suppose that for a given spring, the spring constant is 5. *Work* in physics is just force times distance. Write an integral whose value equals the work done to stretch this spring from its resting position to one that is 3 units longer. (Note: If a constant force F were applied over 3 units, the work done would simply be $3F$. With the spring, the force varies; thus, $F(x)$ is the variable force where $F(x) = 5x$, and where x is the amount the spring is stretched. The integral is perfectly designed to add up force times distance products as the distance varies from 0 to 3.)

Abstracting the Integral—Pyramids and Dams

Lecture 11—Transcript

Welcome back. During the last lecture, we saw the integral interpreted in terms of its graphical interpretation as the area under a curve, and then we also saw how taking antiderivatives gave an algebraic significance to the integral. In this lecture, we're going to be talking about how the integral can be used to apply to questions that are beyond just the idea of a car moving on a straight road. As you may recall, when we were talking about derivatives, we talked about how you could interpret the meaning of the derivative in terms of things like the change in the area of a square or the volume of a cube, and we're going to do the similar thing here in having us look at the interpretation of the integral as a process of summing. And, I'm very anxious that when you think of the integral, you think about its applicability to any kind of a situation that involves dividing things up into little tiny pieces and adding them up. That's really the power of the integral, and it particularly has made manifest in issues about areas and volumes. So, that's where we'll start today.

Let's begin by looking at a square, which we've done before. But, we'll just take this square right here and ask ourselves the question, suppose we were faced with the question of finding the area of the square. Now, I understand that we already know what the area of the square is, so this is a—I'm doing this for the purpose of illustrating the strategy of using the integral rather than getting an answer we don't previously know. So, we know that the area of the square is x^2, but let's see if we can think of it in a dynamic sort of a way.

Suppose we think of this square as evolving, as growing from increasingly large squares that evolve up and grow into bigger and bigger squares. What are we doing in the sense of changing the area of the square? In other words, we're accumulating material to the square; and what amount of material are we accumulating? Well, we're accumulating little pieces like this and like this. We're adding on, so to speak, strips along these two sides, and then other strips along these two sides, and if we thought about adding sort of infinitely thin strips; and, of course, we don't really want to think about infinitely thin, we think about very thin ones, adding up the strip, strip, strip—adding them up as it starts from 0 and it goes to any size that we want. We would accumulate the area of the entire square. So, the idea in our minds is to say, "Ah, the way we get the total value of a square would be to add up this strip plus this strip, and then this strip plus this strip, this strip plus this strip." We can see that adding up that number of strips gives us the square. That gives us a hint that we should be talking about integrals, because integrals are a process of summing, and when we are trying to sum up things in order to get the total value, that's a place where we say, "Ah, use an integral."

So, here is the integral associated with the area of a square of size x. In fact, let's be specific. This is an integral associated with a square of size 5. What is it? Well, all we do is we take the integral from 0 to 5 of $2(x)dx$; $2(x)dx$. Now, why is that? Because we can interpret it visually. We're saying that if we think about this whole square, we can think about it as a strip for every value x between 0 and 5. What are we doing? We're saying take the value $2x$ and thicken it up by a little, tiny width thing, Δx, which turns into dx in the limit. Imagine just a little thin strip of two strips like this, and we're thinking of adding them up to accumulate the area of the square, so that visual meaning of adding up what we see inside the integral sine and thinking of the integral as a process of summing tells us that this integral will, in fact, equal the value of the area of the square. Now, notice that we can actually take this integral. By taking the integral I mean doing what we did in the last lecture; finding an antiderivative—an antiderivative of $2x$ is x^2; plugging in the top value, that's 5; plugging it into x^2 gives 25; subtracting the bottom value, that's 0; giving a total of 25. It tells us an answer that was obvious from the beginning and we knew it already, but we have now an idea of how we can interpret a problem in terms of integrals by dividing up whatever we're after into little bits and adding them up. And that tells us what integral we should take to get the answer we seek.

Okay, let's do another example with a cube. Here's a cube, and once again we're going to do exactly the same thing. We say to ourselves suppose we want to find the volume of a cube of a certain size, like 5; $5 \times 5 \times 5$. Now, of course, we know how to do this; we would just multiply $5 \times 5 \times 5$, but let's think of it instead as an accumulation of layers that make us this cube. In other words, we think of starting at a point, and then adding up a little collection of three boundaries of a certain size—this area plus this area plus this area, and then another layer on top of that—always 3, always 3, always

3; coming out to create a bigger and bigger cube. We can see this in our picture; we can take our cube and divide it up for every value x between 0 and 5, what we're going to do is look at the layer which consists of three slabs that are associated with that number x. The three slabs are this thin layer here on this face, the thin layer on this face, and the thin layer in the back. Each of these has a certain amount of incremental volume.

What is the incremental volume associated with that one particular collection of three slabs, associated with the distance x away from the base point? Well, what's its area? It's $x \times x$, so it's x^2, and then thickened up by Δx. So, we basically have $x \times x$, $\times \Delta x$, $\times$ 1, 2, and, in the back, 3. So, the incremental volume is $3x^2\Delta x$, and we're adding those up for as x varies between 0 and 5. So, the integral from 0 to 5 of $3x^2dx$ is going to give us the volume of a cube because we know what the integral means; it's a process of summing. It's summing the $3x^2$'s thickened up by Δx as the x varies between 0 and 5. And, we can see this laminated picture of a cube that fills up this cube of size 5 by increasingly bigger layers of cubes, each one with three sides growing. Then, of course, once again, this is an example of an integral we can actually take. We realize an antiderivative for $3x^2$ is x^3; and, consequently, we know that this integral, which is really obtained by an infinite summing process, that is, you need to do infinitely many because you do a finite number and then you do finer, you do another number, and then you take a limit. But, that laborious process can be computed in a more easy way by finding an antiderivative and then plugging in the upper limit of integration, which is 5; so the antiderivative is x^3, $5 \times 5 \times 5$ is 125; and then we subtract the lower limit of integration, which is just 0, 0^3 is still 0; and we end up with a value of 125 for this integral, which is telling us the volume of the cube.

Okay, so far so good. Let's see if we can do one that may be a little bit—slightly more challenging. Let's consider computing the volume of a pyramid. So, here is a beautiful pyramid, and we can imagine it's like the pyramid at Giza, and we have this beautiful pyramid and our goal is to find what the volume is of this pyramid. Well, the strategy of the integral, the perspective of the integral, is to take what it is you want to know and divide it up into thin laminates, and then, if we can write down a sum that would approximate this volume of the total pyramid by a bunch of layers, then often we can write an integral whose value is exactly equal to the volume of the pyramid that we're trying for. So, let's go ahead and see if we can accomplish this. Let's be specific and have a particular pyramid in mind. Suppose we have a pyramid whose height from the apex down to the base is exactly 200 feet; and, so, the base is 200 feet by 200 feet. So, it's a square base, 200 feet by 200 feet, and the height is exactly 200 feet. What is the volume of that pyramid? How can we conceive of it as an integral problem?

Well, the answer is that we view this pyramid as being a collection of layers of squares thickened up. So, in other words, we go from this smooth-sided pyramid, which is—we, instead, change our view to those step pyramids that come up and straight and then up and straight because those are squares thickened up, square thickened up, square thickened up that become smaller toward the top, bigger toward the bottom, that add up to a volume that approximates the volume of this actual pyramid. Now, just look at it. You see, we can imagine the layers at the top and lower down and lower down and lower down, each having a thickness of Δh. At a given distance down from the top, how big is the square that's contributing to the volume of the whole pyramid but that's at distance h down from the top? Well, let's just think about it. If the total pyramid from the top to the bottom is 200, and the base is a square base of, side length, 200; then, if we go down h feet from the top, we'll be at a level of the pyramid that is $h \times h$. So, that means that this slab, a slab at height h, is going to have an area of $h \times h$—that's how big the pyramid is when we go down h units down from the top and we cut it across. We have a square that is $h \times h$, and it's h high. Well, it's h from the top, but the thickness of this slab is just Δh. We're just thinking of a thin slab. Now, look, we're thinking to ourselves okay, if each slice—as h varies between 0 and the total height of the pyramid, 200, if each slice contributed is $h^2 \times \Delta h$, then we want to add up those incremental pieces of volume so that they accumulate to be the volume of the entire pyramid. So, here we go. That's an integral.

So the total volume is the integral as h varies between 0 and 200 of h^2dh. Whenever you see an integral like this, you must interpret it in your mind as its defining value. That is, that's a sum of products. What products are they that you're adding up? Well, you're saying okay, here's the variable h, dh, and h is going to vary between 0 and 200, the limits of integration, and what am I going to add up? I'm going to add up the value of h, h^2 for every number. I'm going to take the h^2 and then Δh + another $h^2\Delta h$, as h is varying between 0 all the way up to 200. So, we have this long summation whose value is approximating the integral. And, you can see that that summation corresponds to the volumes of each of these slabs, which are $h^2 \times \Delta h$ in volume.

Okay, now once again, the point is that you can interpret this integral to give you the value you're seeking. It's giving you the value of the volume of this pyramid, but now we can actually compute that integral by using the insights we have from the previous lecture and the Fundamental Theorem of Calculus. Namely, if we want to compute an integral, what we seek is an antiderivative for what's inside. What is an antiderivative of h^2? Well, an antiderivative—remember, we're seeking a function whose derivative is h^2. Well, that function is $\frac{h^3}{3}$. And we test ourselves with it. We say, okay, $\frac{h^3}{3}$, when we take the derivative of $\frac{h^3}{3}$, we bring down the 3; the 3 cancels with the denominator 3; and we're left with h^2.

So, we see that computing the integral between 0 and 200 just amounts to plugging in the 200 into the value of the antiderivative, which is that cubed over 3; and then plugging in the bottom limit of integration, which, in this case, is 0; and we get that this value is the total volume of a pyramid whose height is 200 and whose base is 200 by 200.

So, we're always going to take what it is we're seeking and think about dividing it up into little pieces and adding them together. That is the burden of our song for this lecture. Let's try yet another one.

Here's one, a cone. This is a cone. A cone has a circular base and a point, it comes to a cone; this is a cone; and, we're trying to compute the volume of this cone. Now, I know we've already previously computed the volume of this cone; we know what the answer is. We did that in a previous lecture, but we're going to do it again here to see how it could be done using the concept of the integral. So, the way that we want to think about it is to take this cone and slice it into little thin disks that all are layered on top of each other and accumulate to be the volume of the total cone. Whenever we see something that we're trying to compute the volume of, if we can slice it into little pieces that we add together to accumulate to that volume, we should be thinking to ourselves: integral. That's probably an integral. And, indeed, it is. So, here we go, let's go ahead and think of this.

Here we have a cone, and to be specific, we'll make our cone have specific dimensions so that we can focus our attention on this. We'll imagine our cone to have a height of 4 units—from here to here is 4 units—and then the base of the cone has a radius 3; and then the cone point comes down here and is at the origin of our picture. So, our cone is sitting, from your perspective, sideways here with the x-axis going right through the cone like this. Where this radius is 3, and this height here is 4. Now, how can we imagine this as computing the volume by adding up slices? Well, we say look, for every value of x between 0 and 4, we can imagine cutting the cone at that number x, so here at this point x between 0 and 4 we imagine slicing the cone and approximating the volume of a thin slice by a cylinder, a thin cylinder, whose area is a circle that we get by actually cutting the cone, and then we just thicken it up so that it no longer has a side that's going sideways. Instead, it's once again a step kind of function. It goes straight across for a little bit of distance. So, that little disk, that little disk is a contribution to the volume of the entire cone.

Well, what is that slice volume? Well, at every point x, and we can just draw this point x, so we say okay, suppose I have a value x somewhere between 0 and 4; what is the value of the area of the disk that I obtain by cutting the cylinder exactly at that distance away from the cone point? At exactly x distance? Well, in order to do that, we need to realize that the dimensions of that are—well, the area of it is determined by its radius; it's going to be a circle; and the dimensions of it are determined by its radius; and the radius is going to be told to us because we know the ratio of the entire height of the entire cone to the radius of the base of the cone. Its height is 4 and its radius is 3. So, that means an intermediate place, like height x, what will the radius be at height x? It will be 3/4 of x, just like at distance 4 here, when the height is 4, the radius is 3/4 of 4; 3. So, at an arbitrary point x, the radius is 3/4 of x. That means that the area of the circle that is obtained by cutting the cone right at that point is $\pi \times 3/4\, x^2$. That is the formula for the area of a circle of radius3/4x.

And, then, what do we do? We take the area of that circle and thicken it up by Δx to get an incremental volume, a contribution to the volume, that is the exact volume of a cylinder whose circular base has radius $3/4x$, and whose height or thickness is Δx. When we have such a concept of the slice volume at each value x, we imagine having such slices for every x value between 0 and 4. That tells us how to make an integral. An integral is, then, the integral as x varies between 0 and 4 of $\frac{9\pi}{16}$ that just means $3x$ of 4^2 times πdx. So, the total volume is the integral as x varies between 0 and 4 of our slice volume added up. So, in other words, inside we have $\pi \times 3/4x^3$; and our Δx turns into a dx, meaning that we take arbitrarily fine divisions, and we take a limit as we take smaller and smaller Δx's. That integral, then, is the total volume of our cylinder. And we can, once again, take an antiderivative by realizing that the antiderivative of x^2 is $\frac{x^3}{3}$, and computing—plugging in the top value 4 and the bottom value 0 to get the total value of the volume of that cylinder—of that cone. So, the point is that if you can take something that you want and divide it up into little pieces that add up to an approximation of it, then taking the integral will often give you the exact answer.

Let's move on now to a different kind of a scenario. This is one where instead of talking about physical objects, areas, or volumes, let's talk about a different kind of a situation that shows the strength of the integral is that it can apply to things even that aren't physical and geometric ideas, like area and volume, or even like speed and distance traveled. Instead, let's think about the hydrostatic force of the water against the side of a dam. That is to say, what is the total force of that water pushing against the dam?

Well, let's think about what that means and how it is that we might go about actually trying to figure out such a thing. The problem with trying to figure out the force of water on a dam is that at the shallow levels, the force of the water against the dam is not as great as it is at the bottom levels. We all know this. When you go diving in a swimming pool, the pressure near the bottom of where you dive is stronger than near the top. You can feel it in your ears and you know that the pressure is stronger. So, the question is, if we're trying to figure out the total force against the wall of the dam, we've got to somehow deal with the idea that the force of the water near the bottom is different from the force of the water near the top. In other words, the pressure is not the same near the top as near the bottom. How are we going to cope with that?

Well, okay, let's just see. What do we think about when we do such a thing? To ground our discussion in a specific case, let's assume that the dam is 40 feet deep and 100 feet wide. So, we have this water over here, and it's pressing against the dam, and our challenge is to find the total hydrostatic force, that's the force by the water, against the side of the dam. Now, one thing we know about water is that at any given depth the force, that is the pressure, the pounds per square inch of that water, is the same at the same depth because it's really just talking about the weight of the water that's above it that's pushing down. That's where the force comes from; it's the weight of all that water that's pushing against it. And, then, something that we need to know from physics is if you are at a certain depth, the water pushes to the side in the same amount that it pushes downward. So, the hydrostatic force against the wall of the dam is just determined by the water above it; the depth of the water.

Well, then, what that means is that if all the water were at the same depth, we would be in a position to figure out what the total force against that part of the dam is, because if it were at the same depth, we would say well, what is the pressure at that depth, and then just multiply by the area. See, the pressure is the pounds per square inch. So, if we had a strip along of the dam at a certain depth, let's say h feet down from the top, we would be able to say well, the force of the water at that depth is so many pounds per square inch, or per square feet, and if we had a certain number of square feet at that depth, then we'd know what the total force is at that depth. So, that's exactly what we do. We say to ourselves, "Ah, wait a minute. Then I know how to do this problem." The way I would do this problem is I would look at a thin strip at a certain depth h. I would imagine myself going down h feet down into the water; I'd take a thin strip of Δh, with the concept in mind that during that thin strip, the difference in the force from the top to the bottom of the strip is not too much because it's more or less at the same depth; not exactly, but more or less.

What would the area of that strip be at depth h? Well, let's see. It would be the area of the strip, which is Δh, times the width, which we assume to be

100 feet, times the pressure, that is, the pounds per square foot of the water at depth h. Well, it turns out water weighs 62.5 pounds approximately per cubic foot, and so at depth h the pressure is $62.5h$ pounds per square foot. So, therefore, that means that $62.5h$ is the pounds per square foot at depth h, and the strip that we're talking about has $100\Delta h$ square feet in it. That's this thin strip here, has $100\Delta h$ square feet in it. So, this product is the force that lies on that thin rectangle. Well, now look. If we add up the force on this rectangle and the force on a rectangle right below it, and right below it, and right below it, all the way down; rectangle, rectangle, rectangle, rectangle, horizontal rectangle all the way to the top, we would get the total force on that dam. Well, we've got to say to ourselves integral; this is an integral. Because, what we're doing is we're saying for every h varying between 0 and 40, what we're doing is taking the width of this strip, which is 100, times Δh, which in the limit becomes $\mathrm{d}h$, because our strips become thinner and thinner, so that variation in pressure, even from the top of the strip to the bottom of the strip, although very small, is not completely negligible, but in the limit we get the exact answer. So, the $100\ \Delta h$ would refer to the area of a strip at depth h. We multiply it by the pounds per square foot at depth h feet, $62.5h$; and that integral from 0 to 40 of $62.5h \times 100\mathrm{d}h$ is going to give us the total force on the dam.

Now, we can actually do this integral. We can actually do it because of the fact that we know that 62.5 is a constant, 100 is a constant; we're really just taking the integral of h times a constant. The integral of h is just $\frac{h^2}{2}$; we can find an antiderivative. So, we can find that antiderivative. So, we have $62.5 \times 100 \times \frac{h^2}{2}$; of course, we can simplify this. And then we plug in 40 to the h; that is, we square it and then subtract what we get when we plug in 0; of course, we get 0. And, then, that computation gives us the total force on the dam.

So, this lecture has been about interpreting the integral as a process of summing. And when we have a problem that can be visualized, and this very often happens, by taking some question and dividing it up into pieces that add up to what it is that we're after, then that's calling out for the integral. In our next lecture, we'll see an application of the integral to probability. I'll see you then.

Buffon's Needle or π from Breadsticks

Lecture 12

It should come as no surprise that calculus also is useful in many branches of mathematics. So in this lecture, we're going to explore an example in which calculus is used to compute a surprising result in probability.

Here, we explore an example in which calculus is used to compute a surprising result in probability. What's especially surprising is that we can compute a definite number, namely, the number π, using a random process. Random processes can lead to unrandom conclusions. In this lecture, we will explore an experiment called *Buffon's Needle*, which involves dropping needles randomly on a sheet of paper. In order to analyze this scenario, we will need to use the sine and the cosine functions, so we begin today's lecture with a review of what the sine and cosine functions are and what their derivatives and integrals are.

In this lecture, we will compute a definite number, namely, the number π, by exploring an experiment called *Buffon's Needle*, which involves dropping needles randomly on a sheet of paper. To solve the Buffon's Needle problem, we will use the Fundamental Theorem of Calculus to compute the integral of the sine function over a certain interval. First, however, we need to get a better understanding of sine. Recall that sine is a function that is defined geometrically on the circle and that associates a number with every angle. We use radian measurement of the angle to tell us the distance along the unit circle from the point (1,0) counterclockwise up to the point in question. Sine of the angle θ is the height of a right triangle with angle θ and hypotenuse 1; the cosine is the width of that triangle. As θ changes, so does the $\sin \theta$: When θ is 0, sine is 0; when θ is small, so is $\sin \theta$; when θ approaches 90 degrees (or $\pi/2$ in radian measurement), $\sin \theta$ approaches 1. More generally, the $(\cos \theta, \sin \theta)$ are the coordinates of the point on the circle of radius 1 corresponding to the angle θ radians. As we rotate our angle around the circle, the sine value varies from 0 up to 1, back to 0, down to –1, back to 0, and so forth. We can graph the sine function and see that it is an oscillating curve. The cosine function is the analogous computation for the horizontal distance to every point on the circle. The cosine measure of angle 0 is 1, and it then oscillates back and forth.

Now let's try to understand the derivative of sine. Look at the rate at which the line opposite the hypotenuse is changing in relation to a change in the angle. We ask ourselves how quickly the sine of angle θ will change as we change the angle a small amount. Notice that the graph of the cosine captures the slopes of the tangent lines on the sine graph. That is a visual indication that the derivative of the sine is the cosine. We use the fact that the tangent line of a circle is perpendicular to the radius and find similar right triangles in a figure of the unit circle that illustrates the sine function. We see that the derivative of the sine is the cosine, and the derivative of the cosine is negative the sine. From the graphs of these functions, we see geometrically why their derivatives are related as they are. Neat.

Now that we have derivatives of sine and cosine, by the Fundamental Theorem of Calculus, we also have their antiderivatives: An antiderivative of $\cos x$ is $\sin x$. An antiderivative of $\sin x$ is $-\cos x$.

We now look at something entirely different by taking a brief excursion to probability. We can quantitatively describe the chance of an uncertain event. For example, the chance of rolling a 3 when rolling a die is 1/6. In general, the probability of an event measures what percentage of the time that event will happen. One way to measure probability is to do the experiment many times and just count the fraction. For example, in the die case, we could roll a die many times and see what fraction of the time we get a 3. I did this experiment with the help of my children. They rolled the die 1,000 times and counted 164 3s. The fraction of 3s rolled was actually 164/1000, which in decimal form is 0.164. This result is close to the probability that we reasoned it must be, 1/6, which in decimal form is 0.16666… In general, the more times we perform an experiment, the closer the experimental fraction will be to the actual probability. This concept is called the **Law of Large Numbers**.

Now we use probability and calculus to understand Buffon's Needle. The 18th-century French scientist **Georges Louis Leclerc, Comte de Buffon**, asked a question about a random experiment. The experiment involves dropping needles (or, in our case, breadsticks) on a lined paper. Drop a needle randomly on a lined page where the distance between lines is equal to the needle length. What is the chance that the needle will hit a line? Repeat the process of dropping the needle a number of times and count the times it hits a line. The number of times the needle crosses a line divided by the number of times we dropped the needle is a measure of the frequency with which we hit a line. If we drop the needle more and more, that measure of the frequency should get close to the actual probability. What we will see is that by doing this experiment, we can estimate the value of π.

In general, the more times we perform an experiment, the closer the experimental fraction will be to the actual probability. This concept is called the *Law of Large Numbers*.

We can use calculus to deduce what the exact probability should be. Let's describe where the needle could land. There are two parameters we will consider associated with each needle's landing. One is the angle at which it lands (the angular measurement). If it lands exactly parallel to the parallel lines, its angle is 0. As it rotates, we will measure its position in radial angle from 0 to π. We will also measure where the center of the needle lands relative to the lines. The center could be on the line or somewhere between the lines. For convenience, we will say that the distance between the lines and the length of the needle are both 2 units. In this way, the position and the angle tell the story. Our challenge is to see how many of those positions hit the line. If the angle is close to 0, then the center must be very close to the line to cause a hit. If the angle is about $\pi/2$, then the center can be far away and still cause a hit. Can we make that specific? For any angle θ, if the center's distance is less than $\sin \theta$, the needle will hit the line. Every angle has a particular distance where a needle at that angle first starts to encounter the line.

We have, then, a rectangle describing possible positions of the needle. Within that rectangle, those positions under the $\sin \theta$ curve are positions that hit, and those above the curve are positions that don't hit. The total rectangle has area π. The question is: How much area is under the curve? Calculus comes to the rescue. The integral of $\sin \theta$ from 0 to π is 2. Thus, the probability that the needle will hit a line is $2/\pi$.

This experiment shows a method for estimating the value of π. We now know that the probability in the abstract of the needle hitting a line is $2/\pi$. If after many experiments, we find that a needle hits the line x times in y droppings, then we would expect that x/y is about equal to $2/\pi$. That is, π is about equal to $2y/x$.

Let's see what happens with the data we collected before. When we dropped the needle 100,000 times, we hit a line 63,639 times. Using a calculator, we see that $\frac{2 \times 100,000}{63,639} = 3.1427269...$, quite close to π, which is 3.1415926….

Buffon was able to give estimates for π by, we kid you not, throwing breadsticks over his shoulder on a tiled floor and seeing how often they hit the grouting. Hundreds of years after Buffon tossed his breadsticks, atomic scientists discovered that a similar needle-dropping model seems to accurately predict the chances that a neutron produced by the fission of an atomic nucleus would either be stopped or deflected by another nucleus near it—even nature appears to drop needles. Buffon's Needle used calculus and gives one way of estimating π. Another way to use calculus to estimate π is by doing an infinite addition problem, as we will see later. Perhaps this story could be called: Randomly dropping a needle from the sky gives us the ability to approximate π. ■

Name to Know

Buffon, Georges Louis Leclerc, Comte de (1707–1788). French naturalist and author of *Histoire naturelle*. He translated Newton's *Method of Fluxions* into French. He formulated the Buffon's Needle problem, linking the study of probability to geometric techniques.

Important Term

Law of Large Numbers: The theorem that the ratio of successes to trials in a random process will converge to the probability of success as increasingly many trials are undertaken.

Suggested Reading

Any standard calculus textbook, sections on applications of the definite integral.

Burger, Edward B., and Michael Starbird. *The Heart of Mathematics: An invitation to effective thinking.*

Questions to Consider

1. In what sense do repeated trials of an experiment lead us to conclude the probability of an event happening? Why do more trials result in increasing accuracy?

2. Find a website about Buffon's Needle and try the virtual experiment.

Buffon's Needle or π from Breadsticks
Lecture 12—Transcript

Welcome back to *Change in Motion: Calculus Made Clear***.** As you know, calculus finds applications in many corners of the world; many applications in science, in economics, and all sorts of things, many of which we'll see in the future lectures and have seen in the past ones. So, it should come as no surprise that calculus also is useful in many branches of mathematics. So in this lecture, we're going to explore an example in which calculus is used to compute a surprising result in probability.

What I find especially surprising is that probability—that is, a random process—can be used to compute a definite number; namely the number π. In other words, a random process can lead to a completely unrandom conclusion. So, in this lecture, we're going to explore an experiment that's called "Buffon's Needle" experiment. This involves dropping needles randomly on a sheet of paper in order to analyze how often those needles are going to hit some lines. But, before we get to that, it turns out that in analyzing that experiment, we need to understand the sine and the cosine functions a little bit. I know we've mentioned them briefly before, but I thought it would be a good idea to begin today's lecture with a review of the sine and the cosine functions, what they are, and what their derivatives and their integrals are. So, we'll begin with the sine and the cosine.

Recall that the sine and the cosine functions are functions that are associated with every angle. The way that the sine function is defined is that we imagine a unit circle. This is a circle centered at the origin of the plane that has radius 1. Every point on that circle, then, is associated with an angle, and that angle is associated with the distance around the circle that we go in order to get to that point. So, we're going to use what's called "radian measurement of angles." So, the radian measurement just tells what the distance is along this unit circle up to the point in question. So, that is to say, an angle of θ means that there is distance θ along this unit circle to get up to this point.

Now, of course, every point in the plane has two coordinates. It has a coordinate on the horizontal direction and a coordinate in the vertical direction. The vertical coordinate is that height of the point at angle θ is called the sine of θ. Another way to think of the sine of θ is the ratio of the height of a right triangle opposite an angle, the height of the leg opposite the angle divided by the height of the hypotenuse. Since in this case the hypotenuse is length 1, the sine in this case for the unit circle is just the vertical height; that is to say, the *y*-coordinate of the point on that unit circle.

Now, let's just understand how the sine varies as the angle θ varies. First of all, when the sine is 0, when θ is 0—that is to say, when we're talking about the angle that just is 0 length; it just is this point right here on the horizontal axis, well, its *y*-coordinate is 0. So, the sine of 0 is 0. The entire circumference of a circle is—remember, the circumference of a circle is 2 × *π* × the radius of the circle. Since the radius is 1, the entire circumference is 2 × *π*. That means that if we go a quarter of the way around the circle—that is, up to the top—we would have gone $\frac{\pi}{2}$ distance along that curved circle. The sine at this point—that is to say, if we have an angle that is of what we think of as 90°, but in radial measurement it's $\frac{\pi}{2}$—then it's sine is equal to 1 because the vertical height there is equal to 1. When we continue around the circle, the sine at *π* is equal to 0, and when we go all the way down to the bottom here, that's the sine of $\frac{3\pi}{2}$, then its value is –1 because the vertical distance is in the negative direction—so, it's -1—and when we return to 2*π*, the sine is once again 0.

Let's look at some other numbers in between. So, for example, an angle whose sine we can actually compute would be accomplished by looking at this figure. Suppose we consider a hexagon inscribed in the circle as you see pictured here. A hexagon has 6 equilateral triangles. They fit neatly around a circle because each angle, remember, of an equilateral triangle is 60 degrees, or in radian measurements, $\frac{\pi}{3}$. Well, if we configure the hexagon as you see pictured here, then you can see that half of the length of this—this is an equilateral triangle, so the radius is 1, and this length is also 1; consequently, this distance from the horizontal axis up to this point is equal to 1/2. And the angle is just half of 60°, if we're thinking in degree measurements; or, that is to say, half of that is $\frac{\pi}{6}$ radian measurement. So, the sine of $\frac{\pi}{6} = \frac{1}{2}$. So, that is just an example of a particular angle whose sine we can compute.

The point is, as we rotate our angle around the circle, the value of the sine varies from 0, up to 1, back to 0, down to –1, back to 0, and then continues.

If we graph that, we will see that the graph of the sine function is this oscillating curve that is rather attractive, and it oscillates up and down after π radians, and then comes back up after 2π radians, and then just repeats the pattern again.

Likewise, the cosine function is the analogous computation for the horizontal distance to every point on the unit circle. That is to say, it's the first coordinate of every point on the unit circle. So, the cosine measure of angle 0 is 1, and then it proceeds downward and oscillates back and forth, as we see.

Well, what we're interested in doing right now is to compute the derivative of this function. Now, recall, the derivative is measuring the rate of change of the function with the change of the variable. Or, in the graphical terms, it's measuring the slope of the tangent line at each point.

Let's recall the definition of the derivative of any function, the definition of the derivative, recall, was that you took the value of the function—you were trying to compute the derivative at a particular value—in this case, we're trying to compute the derivative of the sine function at a particular angle θ. In other words, we imagine we've gone up some angle θ here, and we're asking ourselves how quickly will the sine of θ change as we change the angle, the radial measurement of the angle, a small amount? What's the rate at which the sine changes given a change in the radial distance? Well, the characteristic difference quotient associated with the definition of the derivative is that we look at the value of the sine of $\theta + \Delta\theta$, thinking of $\Delta\theta$ as a small increment in the variable θ; we subtract sine θ to see the total difference in the sine function for a difference of $\Delta\theta$; and then divide by $\Delta\theta$ to see the rate at which the sine is changing given a change in $\Delta\theta$.

Well, let's look at the picture of what the sine of θ is. It's this vertical distance for an angle θ on a unit circle. Then look at the picture of what the value of the sine of $\theta + \Delta\theta$ is; this is $\Delta\theta$, a tiny additional angle, and the value of sine of $\theta + \Delta\theta$ would be the vertical height at this point on the unit circle; that's $\theta + \Delta\theta$ around the unit circle. So, the difference between sine of $\theta + \Delta\theta$ and sine θ is the vertical height of the change in the vertical height from here to here. In other words, it's the length of this vertical leg of this small triangle.

Now, notice that this small triangle and—recall, when we thought about a circle and looking at a circle up close, we saw that it looked like a straight line. Therefore, we can imagine this hypotenuse to just be the hypotenuse of a straight line of length Δθ. Notice that any curve along a circle is perpendicular to the radius. So that is to say that this angle here is a right angle. Well, there's something interesting about this little, tiny triangle. Namely, its hypotenuse is Δθ; its vertical side is sine θ + Δθ – sine θ. But, let's look at the angle.

This angle is θ. The angle between the tangential line and the radius is a right angle. Consequently, this angle here is also θ; it's the angle between two parallel lines. We have a parallel line here and a horizontal line here; we draw this radius, which is cutting those two lines; so, the opposite interior lines are equal. So, this angle is also θ. So this little angle inside here is $\frac{\pi}{2}$ – θ; which makes this angle up here θ again. In other words, this little triangle, this little angle way up here, is the same angle as this one, which makes this little right triangle similar to this big right triangle. Well since those angles are similar, we can say that the adjacent leg next to the angle θ divided by the hypotenuse of the small triangle will be the same thing as the adjacent leg divided by the hypotenuse of this big triangle. But, in the small triangle, that ratio is exactly the ratio associated with the derivative. It's the sine of θ + Δθ – sine θ ÷ Δθ. And, yet, it is equal to the similar triangle, the similar ratio, the leg next to the angle θ, that's cosine θ, divided by its radius, which is 1. So, we have this ratio right here which says that this number, which, as Δθ approaches 0 is approaching the derivative of the sine, is simply the cosine. So, what we have proved is that the derivative of the sine is equal to the cosine by this very neat geometric interpretation; and likewise, we can do exactly the same analysis to see that the derivative of the cosine is –sine.

Now, let's look at our graphical representation of the sine and cosine and see that these answers make sense. That is to say, notice that the sine at 0 is going up to the right, and the slope of that line is 1. Well, the value of the cosine at that point is 1. Likewise, as we proceed up on the sine's curve, the slope of the tangent line becomes less until at the very top here at $\frac{\pi}{2}$, the value of the slope of the tangent line is 0. So, the derivative of the sine at $\frac{\pi}{2}$ = 0; and, lo and behold, the value of the cosine at $\frac{\pi}{2}$ is 0, as predicted. Consequently, we have established that the derivative of the sine is the cosine; similarly, the derivative of the cosine is –sine; and, if we look at it backwards—that is to say, finding antiderivatives—we can conclude that the antiderivative of the cosine is the sine. In other words, the derivative of the sine is the cosine, and an antiderivative of the sine is –cosine because the derivative of the cosine is negative to the sine, so putting a –sine there, the negatives cancel out, and the derivative of –cosine x = sine x.

Okay, this is exactly what we need to analyze our experiment that we're going to now undertake associated with Buffon's Needle experiment. So let me just take one short interlude, now, and tell you a bit about probability and what probability is. Probability is making a measurement of the likelihood of some event occurring, when that event is a random event. The most basic example of this occurs in games of chance. For example, when we take a die, such as this die, it has six sides to it; if we roll the die, then whatever side comes up is as likely as any of the other five sides coming up—if it's a fair die, of course. And, so, we say the probability is 1 out of 6. That's the probability that any particular side will arise.

One aspect of the measurement of probability is that if we repeat an experiment many, many times—for example, we roll a die many, many times—and we see how many times—we record how many times it comes up a 1, a 2, a 3, or a 4. If we roll it many times, we expect that any particular face of that die will arise approximately 1/6 of the time. And, indeed, we expect that if we roll the die many, many times, we would expect the ratio of times it comes up, say, a 3 to approach closer and closer to the probability of its coming up a 3; namely, 1 out of 6 times. Now, I actually did this experiment and—well, it's not quite true that I did this experiment, I had my children do this experiment—of taking a whole bunch of dice and rolling them a lot of times. We rolled them 1000 times. And I had them keep careful track of how many times they got a 1, a 2, and a 3; in particular, I was asking about a 3. I said, "How many times will you get a 3 in rolling these die over and over again?" Actually, we used lots of dice and we'd roll them over and over again. And the answer was, after rolling the dice 1000 times, that 164 times a 3 came up on the face of the die. Well, let's see if that seems about right. You see, our expectation is that roughly 1/6 of the time we should see a 3. Well, 1/6 of 1000 =166 2/3. So, the fact that it came up 164 times was pretty close to the actual expected number of times that it should come up a 3. Now, I do have to confess that if I had done this experiment and they had found that only 50 times it came up a 3, I would not have reported that experiment. So, you'll just have to take my word for it that this is a legitimate experiment.

But, the basic feature of probability I want to emphasize right now is that there's a concept called the Law of Large Numbers, which asserts that if one performs an experiment many, many times, then the successes—in this case, a success is defined as getting a 3—divided by the number of trials, the number of successes divided by the number of trials, as we do the experiment more and more times, that ratio should get closer and closer to the actual probability. So, that's the Law of Large Numbers, and that will come up when we talk about the Buffon Needle experiment.

So, let's turn now to an 18th-century French scientist by the name of Georges Louis Leclerc Comte de Buffon. He asked a question about a random experiment, and here was the experiment he considered:

Suppose that we take a striped—in this case we have a table, and we have two lines on the table, and we take what has come to be known in the literature as a needle, but in this case we have a French breadstick; we have a needle whose length is exactly equal to the distance between these parallel lines.; and you can imagine the parallel lines going on forever on all sides of forever, but the experiment consists of the following thing: We take this needle and randomly throw it on the table, and ask ourselves the following question: Does the needle hit the line? Sometimes it does, one of the lines. Or, sometimes we throw it down and it lands so it doesn't hit any of the lines. He wanted to investigate the question of what is the probability that the breadstick will hit a line. Okay?

Now, it sounds like it's—I don't know how interesting that sounds in itself, but it actually turns out to be quite interesting because what we'll see is that by doing this experiment, we can actually estimate the value of the number π, the ratio of the diameter of a circle to its circumference. We can actually get an approximation of that number π by doing this random experiment. Let's see why.

What we're going to do is to analyze the probability of this randomly thrown needle hitting a line. And, the way that we're going to analyze it is to record all possible ways that this needle could land. Well, first of all, it could land anywhere. So, there are two parameters that we're going to identify associated with the possible positions of its landing. One parameter is the angle at which it lands, and we'll measure the angle in the following way: We'll say that if it lands exactly parallel to our parallel lines, we'll say its angle is 0. Then, as it rotates this way, we'll measure our angle, in radians of course, as we rotate round. So, if it's this angle, we'll say that the angle is $\frac{\pi}{2}$, because the radial measurement from this direction on the horizontal line up to this point is $\frac{\pi}{2}$. And, likewise, we'll continue around—we're rotating from the center of the needle—we rotate around, so that it can go all the way to radial measurement π. So, its radial angle will always be some number between 0 and π. Because we're going at the center, you see, once we get to π, this end is getting closer to 0 again. It's some number between 0 and π. That's the angular measurement.

Then we're also going to ask the question about where the center can land relative to these lines. Well notice that the center is going to be somewhere—it could be right on the line, or it could be the most distant place, which would be halfway between two of the lines. For convenience, we're going to say that the distance between two consecutive lines is exactly 2 units. In other words, we're thinking of our needle as being exactly 2 units, which is also the distance between these parallel lines. We're picking the number 2—I'll tell you in advance—we're picking the number 2 because 1/2 of 2 is 1, and then we're going to be talking about sines, and so it would be convenient for half of this needle to be 1 unit in length.

So, let's consider our needle and see how we're going to analyze the question of under what circumstances the randomly thrown needle hits the line or does not hit the line. Can we record that information? Well, first of all, notice that if the angle is very shallow—for example, suppose the angle is 0—then the only possibility for its hitting the line is if it actually lands on the line because if it's even a little bit off the line, if the center is off the line and its angle is 0—that is to say, horizontal parallel to the line—it won't hit the line.

Whereas, if we have some intermediate angle, such as this one; this looks like about $\frac{\pi}{4}$. Remember, $\frac{\pi}{2}$ is 90°, so $\frac{\pi}{4}$ is 45°. If we have this angle at 45°, then if the center point is 0 distance from the horizontal line, of course, it hits. Likewise, it continues to hit, as we think about that center point landing at a more distant value from the horizontal line. At some point, it just barely hits; and then, after that, it quits hitting. So, let's see if we can record that exact moment where it hits, obviously, when it's 0 distance away, and it continues to hit; as we slide this down, keeping the angle fixed, it'll continue to hit until it just barely hits. Now, can we analyze what that just barely hitting is? When will it just barely hit? Well, let's think.

The center of this breadstick—remember, the center is 1/2 distance from either side because the breadstick is 2 units long, so this is—unit measure, this is distance 1; and, think about the vertical distance here. Well, if this angle here is our measure θ, then this angle is also θ, so this vertical distance here is the sine of θ. So, the time when you will just hit, just barely hit, the horizontal line is when the distance of the center point of the needle away from the line is exactly sine of the angle θ. So here we have this animation that illustrates this very clearly. It's right here at this point where the center of the needle is sine θ distance away from the horizontal line that the needle will cease to hit if we're thinking of moving away from the line.

Now let's see if we can record all of this information in a graphical form. Here is a rectangle that's recording all of the possible landings of the needle in the following way: The angle can be anything between 0 and π; and the distance away—that is, of the center—away from the horizontal line, the nearest horizontal line, is some distance between 0 and 1 unit. So, every point in this rectangle corresponds to one possible landing position of the needle. Let's think about any particular angle θ that the needle could land in, and ask ourselves the question: For what distances away from the nearest line would that center point hit the horizontal line and for what distances would it miss? And the answer is, as we already discussed, that for the distance up to sine of θ that corresponds to a position of the needle in which the needle will hit; and, then, if the distance is bigger then the sine of θ, then the needle will miss. So, in other words, those points in this rectangle of possible landing positions of the needle, those that correspond to the needle hitting one of the horizontal lines is in the white area here, the light area here. Whereas, the area outside are those places where the needle has failed to hit the line.

Now all we need to do, then, to compute the probability of hitting the line is to say, "Well, what's the ratio between the area under the curve—that is, the hits—compared to the total possible area, which is the possible landing sites of the needle?" The total area of the whole rectangle is π wide, 1 high; so, it's π units in area. What is the value of the area that corresponds to a hit, a needle hit? Well, it's the area underneath the sine curve between 0 and π. Now, we're in the domain of calculus. We want to figure out what is the area under the sine curve.

The area under the sine curve is the integral from 0 to π of sine *xdx*. We saw that in previous lectures. But, now, we saw earlier, that the derivative of the sine is –cosine; and we saw from the previous lecture about the Fundamental Theorem of Calculus that the strategy for computing an integral is to simply plug in the value of the antiderivative at the upper limit of integration, so it's –cosine of π, and then subtract what we get by plugging in the antiderivative at the lower limit of integration. So, negative –cosine of π minus –cosine of 0—well, the cosine of π is –1; minus that is +1; negative negative –1—two minuses in a row—gives another +1; 1 + 1 = 2. So, the area underneath the sine curve between 0 and π is exactly equal to 2. That's rather remarkable.

So, that means that the probability of our needle hitting a line is $\frac{2}{\pi}$ because the area of the hits is 2 and the total area of all the possible landings is π. So, the answer is the probability of this randomly thrown needle is $\frac{2}{\pi}$.

Now we return to the idea that when you do an experiment many, many times, the actually number of hits divided by the total number of experiments you do is going to approximate the probability. But, in this case, we've actually computed the probability on theoretical grounds. So, we know the probability is $\frac{2}{\pi}$. So if we actually do the experiment many, many times, we can expect that the number of hits divided by the number of throws will approximate to $\frac{2}{\pi}$. Just doing the cross multiplication, that means that π will be approximately equal to 2 times the number of throws divided by the number of hits.

Well one can look on the computer, on the Web, and actually perform this experiment—simulate the performance of this experiment—many, many times. So, I went to one of these websites and performed it 100,000 times. In that 100,000 times, 63,639 times the needle hit the line. We know that that is an approximation to $\frac{2}{\pi}$. Consequently, we can guess that π is approximately equal to $2 \times 100{,}000 \div 63{,}639$l, which is 3.1427269…. Well, π is actually 3.1415926…. You can see that we actually got π to two decimal places by simulating this random experiment.

Now I was told that Buffon himself actually performed this experiment by taking breadsticks and throwing the breadsticks over his shoulder on the floor and, counting the number of times, actually estimated the value of π. Now, I have trouble believing this because I think the breadsticks would break and you'd have to do so many experiments to get close that I don't think that he could actually do that. But, in fact, hundreds of years now after Buffon tossed his breadsticks, or didn't, atomic scientists have recently discovered that a similar needle-dropping kind of model seems to accurately predict the chances that a neutron produced by the fission of an atomic nucleus would either be stopped or deflected by another nucleus near it. So, it appears that even nature appears to drop needles.

Anyway, we've seen from this whole discussion today that Buffon's Needle experiment is an interesting way to apply calculus to something that seems to be completely unrelated—in this case, estimating the value of π. I think maybe this whole story could be summarized by saying *randomly dropping a needle from the sky gives us the ability to approximate π.*

In the next lecture, we'll begin to explore some of the foundational concepts that underlie the mathematics of calculus; namely, the limit, continuity, and differentiability. I'll see you then.

Achilles, Tortoises, Limits, and Continuity

Lecture 13

The limit plays a technical role in making the derivative and the integral possible. But the intuitive idea of the limit has existed since ancient times; and, in fact, much of the development of calculus occurred without understanding what it is that the limits really are in the sense that we understand them today, the rigorous sense, the mathematical definition of the limits.

In a race between Achilles and a tortoise, the tortoise gets a head start, say to position 1. Achilles zooms along to position 1, but the tortoise has moved a bit forward, to position 2. Achilles proceeds to position 2, but the tortoise, though slow, does move a bit to position 3. Achilles must then go on to position 3, but the tortoise has moved to position 4, and so on, forever. Zeno's paradox is that Achilles will never pass the tortoise, because Achilles must always catch up, while the tortoise has moved forward. Because there are infinitely many times that Achilles must catch up, the tortoise is confident he will win the race—that is, until Achilles passes him by. How can the infinite number of times that Achilles remains behind be reconciled with the experience that Achilles wins the race? This paradox illustrates the idea of *limit*, which makes the infinite processes in the definitions of the derivative and the integral meaningful and precise. The notion of the limit is also essential to understanding which kinds of functions are susceptible to the methods of calculus and which functions are not.

The intuitive idea of limit is that if a quantity gets closer and closer to a fixed value, then the fixed value is the limit. Thus, we can state that .999999999… (repeat 9s forever) is the same as 1.

The *limit* plays an essential role in calculus. We could not really define derivatives or integrals without the concept of limit. Historically speaking, however, limits were hard to formalize. The intuitive idea of limit existed since ancient times. Much of the development of calculus occurred without an understanding of limits that today we would view as rigorous. Newton and Leibniz did not know the definition of limit. More than 150 years elapsed between the invention of calculus and an adequate definition of limit.

In this lecture, we will discuss not only the limit but also continuity and differentiability. One of Zeno's paradoxes leads to an idea that is central to the mathematical underpinning of calculus—the idea of limit. The paradox involves a race between Achilles and a tortoise. The tortoise gets a head start, but Achilles zooms along and gets to the place where the tortoise started. However, the tortoise keeps moving along at the same time. Zeno's paradox is that Achilles will never pass the tortoise! The relationship between the infinite number of subjourneys and the finality of arrival captures the idea of limit. The intuitive idea of limit is that if a quantity gets closer and closer to a fixed value, then the fixed value is the limit. Thus, we can state that 0.999999999… (repeat 9s forever) is the same as 1.

The derivative and the integral entail taking limits. Let's recall what the derivative means for the function $f(x) = x^2$ at the point $x = 1$. We choose small increments of x, by tradition denoted Δx.

Our notation is:

$$\lim_{\Delta x \to 0} \left(\frac{f(1+\Delta x) - f(1)}{\Delta x} \right).$$

As we choose values of Δx that get closer and closer to 0, we find that the values of $\frac{f(1+\Delta x) - f(1)}{\Delta x}$ are getting closer and closer to 2; hence, the derivative of $f(x) = x^2$ at $x = 1$ is equal to 2. Likewise, the integral is a limit of the approximating sums.

The formal definition of limit is challenging but illustrates an interesting approach to understanding. The intuitive idea of limit is that if a quantity gets closer and closer to a fixed value, then the fixed value is the limit.

For example, consider the value of the expression $\frac{2(x^2-4)}{x-2}$ as x is chosen close to 2. Plugging in such values as $x = 2.1$, $x = 2.01$, and other values getting increasingly closer to 2, we see that $\frac{2(x^2-4)}{x-2}$ has values that get increasingly close to 8. Essentially, given any tiny neighborhood of 8, no matter how small, choosing numbers for x near enough to 2 will make $\frac{2(x^2-4)}{x-2}$ have a value in that tiny neighborhood of 8. Thus, we say that the limit as x approaches 2 of the expression $\frac{2(x^2-4)}{x-2}$ is 8.

The Formal Definition of Limit

$\lim_{x \to c} f(x) = L$ means for every $\varepsilon > 0$, there exists a $\delta > 0$, such that if x differs from c by less than δ, then $f(x)$ differs from L by less than ε.

The formal definition of limit, finally formalized in the middle of the 19th century, makes that intuitive notion precise. This definition captures the idea that for every challenge $\varepsilon > 0$, there is a response $\delta > 0$ that satisfies a condition. Every challenge has a response. We can apply this formula in the example we saw before: $\frac{2(x^2-4)}{x-2}$. Because we know that limits are essential to the methods of calculus, we can ask what functions are susceptible to the methods of calculus, that is, what functions have derivatives.

One idea associated with the limit is the concept of *continuity***.** To hope to have a derivative, a function must be continuous: There are no gaps and there are no sudden, discrete changes. In the case of motion, position is dependent on time. From moment to moment, the position changes little by little; there are no jumps. Graphs show our position on a line at various times. Jumps on the graph showing a change in position are not physically possible. Likewise, predicted values cannot be different from the actual values of a position function. *Continuous* means that at every point along the graph, the value is predictable in the sense of being the limit of the neighboring values. Graphically, a function is continuous if the graph is connected and can be drawn without lifting the pencil. Consider temperature changes at a given location. Temperature may change quickly but not instantaneously. We can record these temperature data with a graph.

Now we come to the concept of *differentiability*. When does a function have a derivative? Not all **continuous functions** have a derivative. A **differentiable function** is a function where you have a smooth curve, and if you magnify the curve, it begins to look like a straight line. Such a function has a derivative. A function that takes an angled turn and does not smoothly change direction, such as the path of light being reflected in a mirror, is not a differentiable curve. It will not have a derivative. Functions that have sharp points are not differentiable, for example, the stock market. Graphically, a function is differentiable if it is continuous and has no "sharp points," that is, kinks or cusps. It is easy to identify the graphs of functions that are not differentiable functions because they have sharp points. Many functions are differentiable, including polynomials, trigonometric functions, exponential functions, logarithmic functions, and combinations of these. In a precise mathematical sense, most functions are so infinitely jagged that they have no place where they are smooth enough to talk about a derivative there. Yet smooth functions have been the ones on which calculus relies and from which all the wonderful developments and understanding we have seen arose. ■

Important Terms

continuous function: A function that has no breaks or gaps in its graph; the graph of a continuous function can be drawn without lifting the pen.

differentiable function: A function whose derivative exists at every point where the function is defined; a continuous function without kinks or cusps.

Suggested Reading

Any standard calculus textbook, explanation of the limit.

Cajori, Florian. "History of Zeno's Arguments on Motion," *The American Mathematical Monthly*, Vol. 22, Nos. 1–9 (1915).

Questions to Consider

1. How does the limit concept avoid the problem of division by 0 in the definition of the derivative?

2. Could a limiting process be used in nonmathematical settings, such as labor-management negotiations?

3. Draw a smooth function. Now modify the graph so that your function is continuous but not smooth. Finally, modify the graph so that the function is discontinuous. Which of these three do you think occurs most often in nature?

Achilles, Tortoises, Limits, and Continuity

Lecture 13—Transcript

Welcome back. In this lecture we will come to grips with this fundamental idea of calculus; namely, the limit. As you recall, when we defined both the derivative and the integral, we used the idea of the limit; and when I described that, I have to admit, I described it rather vaguely. In fact, if you sort of remember the definition of the derivative, for example, we defined it and we said things like this: as we choose increasingly smaller values for Δt, this fraction will approach some value, and that limiting value is the derivative. In a similar way, we did the same thing with the integral. We said things like this: as we divide the interval of time into increasingly smaller pieces, and for each one we took the value of that small piece times the height of the rectangle, and we added them all up. Then that sum approached a single value, and that is the integral.

Well the problem is that both the definition of the derivative and the integral are a little bit vague. In fact, according to the standards of mathematical discourse, they are really extremely vague. Historically speaking, the limits were very difficult to formalize. So, that's what our goal for the day is. In fact, I wanted to tell you that when I was first beginning to develop these sequence of lectures way back before the first edition of this course, I was going to introduce the course in the following way: instead of saying, as I have during this course, "There are two ideas of calculus, the derivative and the integral," I was going to say, "There are three ideas of calculus, the derivative, the integral, and the limit." But the limit plays a different kind of a role. The limit plays a technical role in making the derivative and the integral possible. But the intuitive idea of the limit has existed since ancient times; and, in fact, much of the development of calculus occurred without understanding what it is that the limits really are in the sense that we understand them today, the rigorous sense, the mathematical definition of the limits.

Newton and Leibniz did not know the definition of limit. In fact, it required between the time of the introduction of calculus in 1665 to the middle of the 19th century, the 1850s, before a rigid definition of limit was actually created. In this lecture, what we're going to do is we're going to actually define what the limit is and discuss some of these foundational ideas that are the basis of calculus; not only the limit, but we'll also talk about continuity and differentiability.

But before we begin, I wanted to read to you a little bit more about Newton's view of the limit to underscore the vagueness associated with the limit at the time of Newton and Leibniz. This is from his *Principia,* where he talks about this concept of limiting process, and I've read part of this to you before in a previous lecture, but then there is some additional material. Here's what Newton said:

> By the ultimate ratio of evanescent quantities, i.e. ones that are approaching zero, it is to be understood the ratio of the quantities not before they vanish, nor afterwards, but with which they vanish. Those ultimate ratios with which quantities vanish are not truly the ratios of ultimate quantities, but limits toward which the ratios of quantities decreasing without limit do always converge, and to which they approach nearer than by any given difference, but never go beyond, nor, in effect, attain to, until the quantities are diminished in infinitum.

Well, it's a little bit vague. In fact, he went on to try to define the limit concept in these words: "Quantities and the ratios of quantities, which in any finite time converge continually to equality, and before the end of that time approach nearer to each other than by any given difference, become ultimately equal."

You can see that Newton, by the way,—and Leibniz—had wonderfully clear and precise concepts of the limit. They did not make mistakes—maybe they made a few mistakes, I don't know, but basically they had very clearly in mind what it is that this limiting concept means. But, it didn't become mathematically rigorous until, as I said, nearly 200 years after that.

Well let's begin our discussion of trying to pin down these ideas of limit and some of its consequent ideas by beginning with one of, another of, Zeno's paradoxes. Zeno's paradoxes are rich sources for insight, and so here is the paradox of Achilles and the tortoise. Here was Zeno's paradox of Achilles and the tortoise; here we have Achilles dressed up as a Jedi knight; and here we have a tortoise; and they are about to start a race. Now, of course, Achilles, as you know, is much faster than the tortoise and, consequently, it's only fair that the tortoise be given a head start. So, let's suppose that this tortoise is given this head start and we begin the race. Now here is what Zeno pointed out; he said in this race Achilles zooms along and finally soon gets to the place where the tortoise began. But the tortoise, though slow, does make some forward progress. So the tortoise remains ahead. Achilles proceeds to catch up to the position the tortoise now is, and then the tortoise moves forward slightly more. Now of course, these are going on continuously; it's not that Achilles moves first and then the tortoise. They both are moving, and the tortoise, you see, stays ahead. Zeno points out that then Achilles has to catch up to that point, and the tortoise moves slightly forward, and so on, and this has to happen infinitely many times. Consequently, said Zeno, the paradox is that Achilles will never pass the tortoise.

Well let's imagine, to be specific, in this scenario of Achilles and the tortoise, the tortoise is going exactly 1/10 of the speed of Achilles. And, let's imagine that they start with Achilles starting at the point 0 and the tortoise starting at the point 1, and just see how this race proceeds. So here is Achilles; here is the tortoise; and it proceeds in the following way: At the time when Achilles has gotten to point 1, the tortoise has moved on to point 1.1—going 1/10 as fast. Then when Achilles goes that next 1/10 of a unit, the tortoise goes on 1/100 of a unit, and, consequently, the tortoise finds himself at 1.11—I should say himself or herself. Then Achilles goes on for another unit of time, and the tortoise finds himself at—that is, 1/100 of a unit of time—and the tortoise finds himself at point 1.111, and so on.

Now here is where the first concept of limit comes in. The concept of limit is that there is a point at which Achilles and the tortoise are at exactly the same point. And, namely, we can describe that point, in modern notation about the real numbers, that point would be described as the point 1.11111 forever; which, incidentally, is exactly equal to 1 and 1/9; 1.11111 forever. So, the concept of the real numbers was something that was not well developed at the time of Newton and Leibniz, and certainly, of course, not in the time of Zeno. So, one of the concepts of limit is that that number is actually attained; that Achilles and the tortoise actually attained that number in a finite amount of time; which, of course, we know in reality occurs—Achilles does, in fact, really catch up to the tortoise.

Let me give you another example of a real number limiting process that comes up all the time. The real numbers are the decimal numbers. That is, if you put a decimal point and then any string of digits forever, that defines one real number. But that concept is a little bit difficult to grapple with, and one of the consequences is that the real number .99999 forever is precisely the same number as 1. It's not a number that's just barely less than 1; it is 1. In fact, we can prove it is 1 in a very simple way. We'll set x =.99999—that's the value; we're wondering what that value is. And then we multiply x by 10, and we see that $10x = 9.99999$ forever. Subtracting the two, we see that $9x = 9$. And, so, $x = 1$. So, this little demonstration proves that, indeed, .99999 is exactly equal to 1. But this is a concept of the limiting value having an actual numerical value.

Okay. Let's proceed to see how the limiting process occurs in our applications to derivatives that we've seen before. Let's consider the function $f(x) = x^2$ and investigate the derivative of that function at the point 1. Well we've seen before how to compute that value. The notation for this limiting value is to say limit with Δx approaching 0 with an arrow to 0 of the characteristic difference quotient associated with the definition of derivative; namely, we evaluate the function at $1 + \Delta x$, we subtract $f(1)$, and divide by Δx. Plugging in the what the function—we're considering the function $f(x) = x^2$—so this $f(1) + \Delta x$ is squaring $1 + \Delta x$ and multiplying it out, the Δx's cancel, and we end up with $2 + \Delta x$. The limit as Δx approaches 0 of that value is just equal to 2.

So, that's an example that we saw previously when we were talking about the definition of the derivative; how the limiting process comes up. But, once again, we have just, at this point, just talked about it intuitively. Let's look at another example, intuitively, of a quantity that has a limiting value. Consider the following fraction: $\frac{2x^2-4}{x-2}$. This is a quantity—at different values of x we get different numbers. Notice that this fraction is not defined for the number $x = 2$. It's not defined because there at $x = 2$, if you plug that in, you would be dividing by 0, which is not allowed to do. But, if you choose values of x very close to 2, we will see that that fraction has different values. We can see that, for example, at the point 2.1, the value of that expression is 8.2; and at the value 2.01 the expression will have value 8.02; at 2.001 the expression is 8.002. So, we can see that the values, as we choose numbers *x*, that get increasingly close to the number 2, intuitively we sense that that expression $\frac{2x^2-4}{x-2}$ approaches the number 8, and that is what we are going to try to make precise by actually defining the meaning of that limiting concept.

So, let's go ahead and now formally define what the limit means. The context of the definition of the limit is that we have a function. A function means that for every number *x*, the value of $f(x)$ is some number. We'll say that that number approaches *L*, a limit, as *x* approaches *c*—first of all, intuitively, what that means is as we choose numbers *x* that are close to *c*, and then we evaluate the function *f* at those numbers that are close to *c*, the value of the function is getting close to the number *L*. In the previous example, for example, when we chose numbers *x* close to 2, the function $\frac{2x^2-4}{x-2}$ got closer and closer to 8. So, the limit in that case is 8.

Okay, so how are we going to formally define this? Well, we're going to define it like this. We say that the limit of this function $f(x)$ as *x* approaches $c = L$ means the following: it means that for any challenge distance around the limiting value *L* you choose to give me, I can give you an interval around *c* that's so small that all of the function values within that small interval of *c* are that close, the challenge distance close, to the limiting value *L*.

So, let me do this graphically. The way it's phrased is—by tradition, έ is used as the challenge value—for every έ bigger than 0, there exits a Δ bigger than 0 such that if the distance between *x* and *c* is bigger than 0 and less than Δ, then the function value at $x - L$ is less than έ. Now, this is a very abstract kind of definition because it is associated—it doesn't come out and say something, it says that every challenge has a response. If this were the graph of our function $f(x)$, what we're talking about is that at the number *c*, as you choose values of *x* close to *c*, the function values are close to *L*. In other words, if I covered up this vertical line and I didn't show you what the value of *f* is at *c*—in fact, what the value of *f* is at *c* is irrelevant—if I said the values as I get close to *c* are getting increasingly close to *L*, then that would satisfy the definition. In other words, if I take an έ neighborhood of *L*, I take an interval around *L* that's έ on each side, then I can find bands around *c* that are so tight that every value of x inside those bands lands between these two bands around L. That is the concept of limit.

Now by the way, this definition of limit was created in about the 1850s. As I said, it took almost 200 years to pin down this idea that was really clearly understood in some sense way back in the time of Newton and Leibniz, and even in ancient times to some extent. This definition appears in every calculus book, and it's an early chapter because, you see, it's really needed, logically necessary, before you can define the derivative. So, it occurs in the early chapters of every calculus book. The only problem is since the time it has been formalized, in 1850 until now, millions of students, of course, have taken calculus. This definition has been presented to millions of students. Since 1850, there have been, I believe it's about three students who have understood this definition of the limit. It requires a lot of thinking to really pin this down, and most mathematicians certainly did not understand the definition of limit when they were first presented with it. It's a very tricky kind of a concept. But, I wanted to give you a flavor of what the definition actually is; this is a formal definition; and let me actually apply it in this particular case that we had before to show you how the definition is used to prove a limit.

So, this is an actual proof that the limit of that example we saw before was, indeed, 8. In other words, the function $\frac{2x^2-4}{x-2}$. We can see that in this expression it is not defined for $x = 2$, but it is defined for other values. As we choose values of x close to 2, we do, in fact, get values close to 8. How would we formally prove that the limit is 8? Well, if somebody gave you a challenge value έ bigger than 0, and think in your mind—that έ, is for example, is .01; it's a small positive number—it is my job to give you a Δ response so that any number that's that close, Δ close, to the number 2, the function value is έ close to the number 8. So in this case, if I choose Δ to be έ/2, in the case where έ is .01, then Δ would be equal to .005. If I choose a number x that is closer than Δ—more than 0 but closer than Δ away from the number 2—that is, if I choose it less than .005, in this specific case, away from the number 2, and I plug in this value and I just compute it, I will find that its value is less than έ; in this specific case, .01. The fact that I can accomplish a choice of Δ for every choice of έ tells me that I have verified that the limit of this function as x approaches 2 of $\frac{2x^2-4}{x-2}$ is, indeed, equal to

8. So, this is a formal proof of a limit, and we take advantage of this formal proof when we prove such things as this: that the derivative of x^2 is the $2x$ that we know it to be.

How do we do it? We say as Δx approaches 0, what is the value of this characteristic's difference quotient associated with the derivative? Working our way through it, we see that the limit is actually equal to $2x$ as Δx approaches 0; and, consequently, that is the formal way of proving that the derivative of x^2 is equal to $2x$. It requires using the definition of the limit.

Now, I want to tell you some ideas that are associated with the limit. One of the ideas associated with the limit is the concept of continuity. Continuity is the following: that if you can predict the value of a function by its neighboring values, then you say it's continuous at that value. Let me tell you what I mean by this. If you are driving a car on a road, you're driving a car on a straight road—suppose I told you where you were on that road at every moment except for precisely 1:00. Well, you would know where you were at 1:00 because the car wouldn't be very close to this position at points getting up to 12:59.99999 minutes; it wouldn't be getting extremely close, and then, suddenly, you're somewhere else. You would have to be at the place that is predicted by the values nearby. We can formalize that by saying that at a given place, the value that you are at that moment of time, the position that you are at that time, is the limit of the positions that you are at neighboring times. So, we're returning now to our definition of limit; that is, the definition of limit said that a value is a limit if the values nearby are approaching the limiting value. In the case of an actual life, where you're talking about the position function of a car, then at every moment it is predicted by the values nearby.

Suppose somebody said, "This is where I was yesterday on my position function. At these times I was up here, and then, suddenly, I jumped to be at some completely different position." You'd say, "That didn't happen." You could not—for example, let's suppose this is 1 mile apart—you cannot go instantly from one position at say, 1:00 to immediately after 1:00 being a mile down the road. You had to take the intermediate steps.

So, the actual definition of the limit of continuity is the following: that we say that function is continuous at a value c if its actual value at c is the predicted value—meaning, is the limit of the values as x approaches c. So you'd formally write that as the limit as x approaches c of $f(x)$ is equal to $f(c)$. In other words, you don't have a situation like this that jumps, nor do you have a situation where the predicted values are something, but the actual value of the function is something different. You couldn't be just off at that one point. So, continuous means that at every single point along your graph, the value is predictable in the sense of being the limit of the neighboring values.

Another way to think of a continuous function is a function that you can draw without lifting the pencil, whose graph you can draw without lifting the pencil, because if you draw the graph of a function without lifting your pencil, what does it mean? It means that the stream of graphite that you're laying down and drawing the graph, the next point is the next point you get to as you drag the pencil along. So, that's a good way to think about what a continuous function is. There's no jumping involved.

There are many examples in life besides just the position of a moving car that naturally demonstrate the concept of continuity; such as, if you measure the temperature during a day. That will vary continuously during a day; it won't just suddenly jump to some different position. So, many things in real life exhibit continuity.

Now we're going to turn to the question of differentiability; when does a function have a derivative? This is something that I have talked about—we've discussed the concept of having a derivative at each point, but I was always a little bit vague about it. In fact, you may have detected, or, probably not because I didn't make a point of it. I said, if you have a function—and, I usually said, if you have sort of a smooth function, I used words like "sort of smooth"—that was a way for me to assuage my mathematical consciousness about saying if you have a differentiable function, then something is true. Remember, I talked about the idea that if you have sort of a smooth curve that had a derivative and you looked at it very, very close, you magnified it, it became like a straight line. Well, that's true. For functions that are differentiable, that's exactly the way they look. But some functions don't look that way. Some functions are not differentiable. So, let's look at some examples.

If you have a curve that has this sort of smooth characteristic to it, then as you magnify the curve, it will, indeed, look like a straight line and it will have a derivative. But, suppose that you draw a function like this: It goes straight up for a while and then it takes an angled turn. It doesn't smoothly change its direction but, instead, it just has an angled turn. Let's look at what would happen if you blew up the picture at that direct angled turn. In other words, it's two straight lines that meet at a point. Suppose you blew it up; it would still look like that angle. Suppose you blew it up a million times; it would still look like that angle. It would not begin to look like a straight line like a smooth curve does. So, a differentiable function is a function where if you do magnify it, the graph does begin to look like a straight line, whereas if you have a kink in it, such as this, no matter how much you blow it up, it will not ever look like a straight line, and the limit that is associated with taking the derivative does not exist. It simply does not exist. Because, for example, with a kink, on this side, if you take points nearby on this side, it would be telling you that the slope is going up; whereas if you take points nearby on this side, it would tell you that the slope is 0. So, there's no one value to which that characteristic difference quotient is converging, and therefore, it's not differentiable.

There are many things in real life that we, conceptually at least, think of as taking paths that are not differentiable paths. For example, if we imagined a light beam coming down hitting a mirror and going back up. Conceptually, we see that as having a straight line down and a straight line reflected back up. Well that is not a differentiable point at that place of reflection because there's no one direction that it's going at that point. Once, again, it changes direction too quickly. So, it's not differentiable at that point. So, some functions are differentiable and some are not.

Anything with a kink is not differentiable; anything that has sort of a cusp to it, like this, a sharp point, not differentiable because if you magnify them it will not be straight. If you have any differentiable function, however, a differentiable function must be continuous. In other words, differentiability is a stronger condition than mere continuity. It's continuity plus some more things.

But many functions in the world are differentiable functions. The functions that we actually use to describe the physics of our world and other features of our world often are automatically differentiable. For example, all polynomials are differentiable; and, in fact, not only do they have derivatives, but their derivatives have derivatives, and their second derivatives, and their third derivatives—you can differentiate them forever. Trigonometric functions, like the sine and the cosine—those functions are differentiable. Exponential functions are differentiable. Logarithmic functions are differentiable.

And, you can combine all these in any ways you want. You can take the sums of any two, the differences of any two, the products of any two, the quotients of any two differentiable functions, and, at any points at which such functions are defined, they, too, will be differentiable. So, many of the functions in practice that we use are differentiable.

You might think that functions that are not differentiable are rather anomalous because you could have a kink at one point, but notice that a function that just has a kink at one point is differentiable everywhere except that point. Well, it turns out that it's possible to have a function that at no point is it differentiable; but, yet, it is still continuous. So, you could draw it, if your pencil had infinitely many jiggle-jaggles in it, you could actually draw this function without lifting your pencil to make a function that is continuous but nowhere differentiable. We can see such functions, graphed like this; they would be extremely jagged, so that no magnification at any point is differentiable. But, as I say, the functions that we actually use in practice to describe the world are often differentiable in most places.

So, in this lecture, we have seen some mathematical foundational ideas associated with calculus—the idea of limits; the idea of continuity; and the idea of differentiability. In the next lecture, we're going to see how calculus is used in having our calculators and computers do some of the amazing things they do. I look forward to seeing you then.

Calculators and Approximations
Lecture 14

In this lecture we're going to see how our calculators are able to work some of their magic. In fact, it turns out that the way calculators work and some of the processes that they do actually involve a resolution of another of Zeno's paradoxes, namely his arrow paradox.

Zeno's arrow paradox shows us that an infinite addition problem, 1/2 + 1/4 + 1/8..., results in a single number, 1. In Zeno's case, we know the answer in advance. However, for π and the square root of 5 and others, we may not know the whole decimal expansion of the number, yet we may be able to show that it is equal to a specific infinite sum, and we can approximate this sum by merely adding up, say, the first few hundred terms of the infinite sum. This is exactly what our calculators do to compute the answer when we press the sin key or square root key. Where do the infinite sums come from? Some come from calculus, and we illustrate them both graphically and numerically. We also show what happens when we ask a computer to *solve* an equation numerically. Here, an infinite process dating back to Newton's time is automated and iterated sufficiently many times to produce an answer that is as close to the true answer as needed for the application.

In the case of Zeno's arrow, we know in advance what the sum is, namely, 1; however, the real significance of infinite series is that they are used to approximate quantities that we don't otherwise know.

Zeno's arrow paradox considers an arrow flying through the air traveling toward your heart. The arrow goes 1/2 the distance from bow to heart, then 1/4 that distance, then 1/8 that distance, and so on, forever. Because the arrow does, in fact, arrive at its destination, the totality of those fractions is 1. Every point between the bow and the target is passed during one of the fractional distances traveled. Written as an equation, we have noted that $1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \ldots$ We call this an *infinite series*. Notice that if we stop the summing process after hundreds or thousands or millions of terms, the sum of the distances traveled will be as close to 1 as desired. In other words, the sums of only finitely many of the fractions are as close to 1 as desired but will never equal 1. The limit of the finite sums, however, is 1. In the case of Zeno's arrow, we know in advance what the sum is, namely, 1; however, the real significance of infinite series is that they are used to approximate quantities that we don't otherwise know.

Examples of infinite series whose values involve π:

$$\frac{\pi}{4} = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \ldots$$

$$\frac{\pi^2}{6} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots$$

In each case, we can see how adding up the first terms of each series allows us to deduce closer and closer approximations of π. Each of the above equations is obtained using calculus, but for now, we leave their genesis as a mystery.

Polynomials are easy to calculate because they involve only addition, subtraction, multiplication, and division. We often want to approximate a nonpolynomial by a polynomial. The sine curve can be approximated by polynomials. That is, we can write down an infinite polynomial that gives us the precise value of sin x for every angle x given in radians.

Specifically,

$$\sin x = x - \frac{x^3}{3\times2\times1} + \frac{x^5}{5\times4\times3\times2\times1} - \frac{x^7}{7\times6\times5\times4\times3\times2\times1} + \ldots$$

Using the more compact factorial notation, this is

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

This means that we can take any number x and approximate its sine by plugging in x to the right-hand side of the equation and doing the multiplications and additions. The more terms we use, the closer we will be to the exact answer. The value of the sine at any angle is equal to an infinite sum. Let's consider some examples, such as the angle 30°, which is $\pi/6$ radians. Notice that sin $\pi/6$ (which equals 1/2) is increasingly better approximated by adding more numbers in the series.

Let's look at various curves that are becoming increasingly like the sine curve. That is, we can compare the graph of the sine function with the graphs of the polynomials obtained by using more and more terms of the right-hand side of the equation. We can see better how this process works using an example of a car moving on a straight road. Suppose the car never moves faster than 1 mile per minute per minute. If the car travels at its maximum speed, after 1 minute, the car would have traveled 1 mile. If the car starts traveling from a full stop and accelerates at 1 mile per minute per minute, then after 1 minute, the car would have gone 1/2 mile. If you have a function whose derivative starts at 0, and its derivative is 0, and the derivative of that is 0, and so on with infinitely many derivatives all equal to 0, the car would not travel at all. Using series is exactly how calculators compute values of trigonometric functions, such as sine.

Where do these approximations come from? We can compute the first few derivatives of sin x, evaluate them at $x = 0$, and write an infinite polynomial that has the same values for its derivatives evaluated at 0 as sin x. This gives the following:

$$\sin x = 0 + x - 0 - \frac{x^3}{3!} + 0 + \frac{x^5}{5!} - 0 - \frac{x^7}{7!} + \ldots = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

In addition to using derivatives to compute such values as π or sin 1, calculators also use derivatives in the **Newton-Raphson Method** to solve equations. The Newton-Raphson Method for approximating a solution to an equation involves making an initial guess, then using derivatives to refine the guess.

Let's actually work an example of finding a root of an equation using this method. Let's try to find the square root of 5. The square root of 5 is a solution to the equation $x^2 - 5$. Let's try a first guess of 2. The method gives a second guess of 2.25, whose square is already pretty good, namely 5.0625. Using 2.25 as our next guess produces a third guess of 2.236111111…. That guess is really good, with a square of 5.000192852…. Using 2.236111111… as our next guess produces a fourth guess of 2.236067978…. This is as close as my calculator gets. Now you know what really happens inside your trusty Casio, TI, IBM, or Macintosh once you press the sin or SOLVE button! ■

Important Term

Newton-Raphson Method: An iterative technique for finding solutions of an equation using graphs and derivatives.

Suggested Reading

Any standard calculus textbook, sections on infinite series and the Newton-Raphson method.

Questions to Consider

1. Consider this infinite sum: 1 – 1 + 1 – 1 + 1 – 1 + …. Is there a single number to which it tends? Not all infinite sums converge to a number.

2. Suppose you add up an infinite number of terms and they converge to a single answer. Why must the terms that you are adding have a limit of 0? Why is that condition not a guarantee that the infinite series converges?

3. Use the Newton-Raphson Method to approximate the cube root of 2. How close can you get with your calculator?

Calculators and Approximations

Lecture 14—Transcript

Welcome back. In this lecture we're going to see how our calculators are able to work some of their magic. In fact, it turns out that the way calculators work and some of the processes that they do actually involves a resolution of another of Zeno's paradoxes, namely his arrow paradox. So, let me begin this lecture about how calculators work, and computers in some arenas, with the description of Zeno's arrow paradox.

Suppose that somebody is standing over there with a bow and arrow and shoots the arrow right at my heart. Now ordinarily, you would think I would be apprehensive at this moment, but I am very calm if I am Zeno because Zeno reasons in the following way: Zeno reasons that if the distance between the bow and my heart is viewed as a unit distance, that the arrow must first go 1/2 of a unit, getting halfway to my heart; and then it must go 1/4 of a unit, half the remaining distance; and then it must go 1/8 of a unit, half of the remaining distance; and so on, forever—1/16, 1/32, 1/64, and so on. But I'm quite calm because of the fact that it must accomplish infinitely many of these steps before it gets to my heart.

Well let's look at this arrow paradox in terms of an infinite addition problem. So here we have a graphic that shows the arrow heading from its originating place toward the number 1; and the resolution of Zeno's paradox is that, in fact, it does, in fact, reach my heart. It does make it the entire distance. But it doesn't make it until we have allowed ourselves to add infinitely many terms. In other words, it's the limit of a sequence of how far the arrow got at all of the different intermediate stages, and it actually accomplishes that limit. That is one of the resolutions of this paradox of Zeno; that, in fact, the limit is actually reached.

So, let's just look at this limit again in this addition form. That is, the arrow goes the following distances: first it goes 1/2 of unit, halfway; then it goes 1/4 of the way; then it goes 1/8 of the way; plus 1/16 of the way; plus 1/32 of the way; and so on forever. The fact of its reaching its target is the fact that this summation is equal to 1. Now, this is an infinite summation, but it doesn't make any sense to add infinitely many numbers, and in fact, that was the problem with Zeno's paradox. He was not in a position to say what does adding infinitely many numbers mean? It doesn't really mean anything. The only thing that we really can do is add finitely many numbers. So, when we write an infinite addition problem, such as this—which is, incidentally called an infinite series—when we have an infinite series, that is, an infinite addition problem, this is really shorthand for doing infinitely many finite addition problems, and then taking a limit that we talked about last time.

Here we go. If we look at the value of the sum of the first two terms of this infinite summation, that's 3/4. Well, that's a number, 3/4. Now we look at the sum of the first three terms—1/2 + 1/4 + 1/8 = 7/8. If we add the first four terms, we get 15/16. If we add the first five terms we get 31/32, and so on. The pattern continues. Well that gives us an infinite sequence of numbers, and those numbers are converging toward one fixed number; namely the number 1. So, the concept of an infinite series is that this infinite addition problem can be reinterpreted as doing finitely many increasingly larger finite numbers of the additions of these infinite number of terms, and that those values are getting closer and closer to the final value. That is, they converge to the final value. In this case we know what the final value is; it's 1.

Well this is the basic concept of an infinite series. Let's see how it applies to cases where we have infinite summations that are heading towards some number whose value we don't know in advance. Here's an example: One can compute using methods of calculus, some of which we'll talk about in a little while; one can sometimes write down that a certain value, in this case $\frac{\pi}{4}$; π is the number that's the ratio, as always, the diameter of the circle to the circumference, or the circumference to the diameter, the circumference is exactly π times the diameter of any circle. One can theoretically deduce that $\frac{\pi}{4}$, 1/4 of that ratio, is exactly equal to the infinite series 1/1 – 1/3 + 1/5 – 1/7 + 1/9 – 1/11 and so on forever. What that means—and, by the way, I should say the sequence is attributed to Leibniz, it's known as Leibniz Series for $\frac{\pi}{4}$, and using a calculator or computer, we can see what happens when we add up a certain number of these terms.

For example, if we add up the first 10 terms, and then we multiply by 4, we would get this approximation to π: 3.0418 and so on. If we add up 100 terms—do you follow me what we're doing by adding up 100 terms? I mean, we're just following this pattern, and since we just have a finite number of fractions, we can get those numbers exactly. That is, a computer can get them exactly; I would certainly make plenty of mistakes. So, if 100 terms, we get this value, getting closer to π; with 1000 terms, we are at 3.14 and so on; at 1 million terms we get 3.1415916535. The actual value of π is 3.1415926 and so on. So, you can see that we differ only at this term; at this stage. And if we took more terms beyond 1 million, we would get increasingly close to π.

So, this is an example where we can compute a value such as π, whose digits we may not know beyond a certain place. There's a process by which we can get as close as we like to that value π by simply adding up more numbers. Now, I haven't told you anything yet about how it is that these numbers come about, but that will come up a little bit later in the lecture. Let me give you another example.

Another example of an infinite series that gives a value associated with π is this one: $\frac{\pi^2}{6}$ is equal to the sum of the reciprocals of the squares. That is, if we add up a certain number of sums of reciprocals of the squares, we're getting answers, we're getting numbers, and those numbers are getting closer and closer to the exact number $\frac{\pi^2}{6}$; $\frac{\pi^2}{6}$ is an exact number. And as we add up more and more of these numbers, for example, if we add up 10 values—by the way, the reason I have the chart here is we are familiar with the number π; we are not familiar with the number $\frac{\pi^2}{6}$—so, what I'm saying is we add up 10 terms in this case, and then we multiply by 6 and take the square root to get the approximation of π that we see in this column. With 10 terms, we get 3.04; 100 terms, 3.13; 1000 terms; 10,000 terms; and 1 million terms, we see, once again, that we get π to 3.14159261, again, still slightly off at this point in the decimal representation.

So, these are some of the techniques by which a calculator, for example, is able to compute a value such as π. When you have your calculator in front of you, it may actually have a key on it, which says π; and if you push that π key a number comes up. Well, it didn't store those digits of the number π; instead, it knows that it can use this kind of a technique to get an approximation of π. So, this is one of the tricks of the trade.

One thing about calculators they can do that's really rather remarkable is that they can take functions, such as, for example, the sine function, which we met in previous lectures—remember, the sine function is telling us the value of the vertical coordinate of a point moving along the unit circle. So, the sine is this oscillating function. Well, how does the calculator know—when you push the sine key, and then you push an arbitrary number, like 37—how does the calculator possibly know what the sine of—or, in fact, you can make it much more complicated—sine of 37.127856? You can plug that in and you push enter and it tells you the answer. How does it know? How does it know? Did you ever think about that? Nope, you probably never thought about that. How does it know? It does not store those values, how does it know what you're going to type in? It doesn't know how to do that. But what it does know how to do, what calculators do and their par excellence at doing is they're good at doing arithmetic. They're good at evaluating polynomials, because they know how to add, they know how to subtract, they know how to multiply, they know how to divide. That's what calculators are extremely good at. So, the question is how can this calculator know the answer to the sine of an arbitrary value that you put in when all it can do is do multiplication and summations and so on?

The answer is this: that the value of the sine of any quantity x, for any radial measurement x—so, let me remind you, that means that we move along the unit circle x distance, and that's measured as x radians of radial measurement, and then we look at the height of that point, and that is the value of the sine of x; it's the y-coordinate, the vertical coordinate, of that point. It turns out that for any value x, that value, the sine of x is equal to an infinite series, and this is the infinite series: $x - \frac{x^3}{3}$ factorial + $\frac{x^5}{5}$ factorial – $\frac{x^7}{7}$ factorial + $\frac{x^9}{9}$ factorial, and so on, forever. Now, when we say forever, remember what we mean about an infinite series. What we mean by forever is that we add some finite number of those terms. For example, we may add 100 of those terms, and we get a number, and that number is close to the value of the actual value of the sine. Then if we add 1 million of those terms we get even closer to the value. By the way, notice that this expression for the sine of x, $x - \frac{x^3}{3}$ factorial (3 factorial just means 3×2) + $\frac{x^5}{5}$ factorial (5 factorial is just shorthand for multiplying all the numbers 5 and lower, $5 \times 4 \times 3 \times 2 \times 1$) – $\frac{x^7}{7}$ factorial (7 factorial, again, is just $7 \times 6 \times 5 \times 4 \times 3 \times 2$). So, that's just a shorthand notation for that multiplication process.

So, what we assert is that for any number that you put in, a radial value *x*, if you plug it into this equation, this infinite series, and you add up a certain number of the terms, you'll get a value that's very close to the sine of that number. For example, let's take the first example, the simplest example we can think of, how about the sine of 0. Well, the sine of 0, if you plug in 0 to all of these *x*'s, all of these terms are just 0. So, the answer is 0, and, indeed, that is correct. Let's see if we can look at an example whose answer we know so that we can verify it. Here's an example.

Remember we met this picture a couple of lectures ago, and we noted that the sine of 30°, which is $\frac{\pi}{6}$ radial measurement, the sine of $\frac{\pi}{6}$ = 1/2. We saw that before, so let's take that for granted. And let's see, now, why it is that our calculator can compute the sine of $\frac{\pi}{6}$ by plugging its values into this infinite series approximation for the sine of *x*. Namely, what we do is we take $\frac{\pi}{6}$ and say that the sine of $\frac{\pi}{6}$ is exactly equal to what we would get if we took exactly $\frac{\pi}{6}$ – exactly $\frac{\left(\frac{\pi}{6}\right)^3}{3!}$ + exactly $\frac{\left(\frac{\pi}{6}\right)^5}{5!}$, and alternating sum forever. Now, of course, the value of π also has to be computed, but calculators can compute the value of π; they can do that as well. And when we do this and we look at the different terms, we'll find that it will compute the value of the sine of $\frac{\pi}{6}$, whose actual value is exactly .5000 forever, that if we add just the first few terms, like the first term here, $\frac{\pi}{6}$. Well, it's some number that's close to half, but too big; if you add the first three terms you get this value; if you add the first five terms—already we have the value of the sine of $\frac{\pi}{6}$ to a large number of decimal digits, probably as many as need to be displayed on our calculator's display. So, by doing just a small number of addition processes, the calculator has been able to figure out the value of the sine function.

Now, let me just demonstrate graphically why it is that these polynomials are approaching the sine function. The graph of the sine function is this wavy function, as you know. The graph of these first two terms of this approximating series is this up and down cubic graph, and you can see that in the neighborhood of 0 it's very close to the sine function. When we add another term, we get a graph of this 5th degree polynomial that is a better approximation in the middle. When we go on to the 7th degree polynomial, it's still approximating the sine even greater distance beyond the origin. This is the 21st power. When we go all the way to the 21st power, look at the graph of this 21st degree polynomial. It comes like this and it's indistinguishable from the sine graph all the way out to here, and then it goes off. So, you can see that we're getting a very good approximation to the sine for values of x close to 0.

Now, what I'd like to do is give you some idea of why it is that this process works, and this is rather subtle, but it's, I think, quite interesting. Let's go back and remember a car moving on a straight road. Suppose that we have the property that our car never moves at a velocity greater than 1 mile per minute. Let's ask the question how far could that car have gone, if it starts at the 0 point on our road and it never goes more than 1 mile than minute, how far could it have gone in 1 minute? Answer: 1 mile. If it goes at its maximum speed of 1 mile per minute for 1 minute, it will have gone 1 mile.

Now suppose that we say to ourselves we have a car; it's stopped at a stop sign. Its velocity is 0; and its acceleration is never more than 1 mile per minute per minute. In other words, the change in the velocity never exceeds 1 unit. So, the graph of the acceleration of the car is going to be just the horizontal line, and the graph of the velocity curve is going to be this diagonal line. In other words, if it's accelerating at 1mile per minute per minute, then the velocity starting at 0—we're assuming he starts at a stop—you increase the velocity according to that diagonal line and it is following the graph $v(t) = t$, and it attains the velocity of 1 mile per minute after 1 minute, but you weren't traveling that fast until 1 minute had elapsed.

Now, recall that the position function—if you know the velocity function—the very first introduction of integral was that if you knew the velocity function, the integral of that gave you the position. So, if you start at 0 and your velocity is according to $v(t) = t,$ then the position function is $\frac{t^2}{2}$. So, after 1 minute you will have gone 1/2 mile. What we're doing is following a pattern of assuming that we are stopped at the origin and then we're increasing our speed according to certain restrictions. Suppose that, for example, both the velocity is 0 and the acceleration is 0 at when we start, at time 0, but the change in acceleration never exceeds 1 mile per minute; in fact, it's just equal to 1 mile per minute for the entire thing. The acceleration itself, that's the derivative of the acceleration; it's changing at 1 unit per minute. Well then we can see how far—the maximum distance you would have traveled would be—the acceleration would be this diagonal line, $a(t) =$

t; the velocity would be the integral of that because we're doing the opposite of derivatives each time, the velocity would be $\frac{t^2}{2}$; and your position function would be $\frac{t^3}{2\times 3}$, $\frac{t^3}{6}$—that's the integral of $\frac{t^2}{2}$.

Suppose we took yet another derivative and we did the same thing; that is, the rate at which things are changing. We're taking the position function, saying suppose the position is 0 at 0; the velocity is 0 at 0; its acceleration is 0 at 0; the derivative of that is 0 at 0; the derivative of that is 0 at 0; and we continue on, and doing this process we find that the position is unable to change very much. In other words, if we start with all of the derivatives starting at 0, and then we just allow one change of the derivatives four derivatives down, then the maximum, if we do this process one more time, we see that the position function is $\frac{t^4}{2\times 3\times 4}$; $\frac{1}{24}$, if it were 1 minute down the road. The point is that if you have a function whose derivatives—you start the function at 0, and its derivative is 0, think of it as velocity if you think of it as a car moving, if you think of the position of a car—if its derivative is 0 and the derivative of that equals 0 and the derivative of derivative equals 0, and so on, and if you did that for many, many times, then you would find that the car could not travel at all after 1 minute, if you had infinitely many derivatives all equal to 0, then the car could not travel at all. In other words, the value of the function would still be 0 one minute later. You couldn't move.

This is actually quite interesting because what it does is it allows us compute to tell us why that equation for the sine function is correct; and I'll tell you why. If we look at this infinite series approximation to the sine function, this side over here is a polynomial. Sine of x is the thing that we're trying to approximate. If we subtract those two functions from each other, we get a function that has the properties that its value at 0 is 0. Well, that's easy to see because the sine is 0 and this number is 0, they're both 0, $0 - 0 = 0$. But it's also true that the derivative of the sine minus the derivative of this at 0 is 0. That is second derivative of the difference of the sine minus this infinite polynomial; its derivative is 0, its second derivative is 0, its third derivative at 0 is 0, and what we've just seen before is that we have written a polynomial; this polynomial here has the same derivatives at 0 as the sine does. Therefore, the difference of those two things at 0 have exactly 0 as their derivatives—the first derivative, the second derivative, third derivative at 0 is all 0, and we've just shown before that that means that that function is just flat. It never changes. So the effect is that since the derivatives of this polynomial at 0 are the same as the derivatives of the sine at 0, then, in fact, the sine of x for every value is equal to the value of this infinite polynomial at 0. That's where the kind of reasoning associated with the sine of x is equal to what it is.

Well, I wanted to do one other thing in this lecture about how it is that calculators and computers work their magic. The next thing I want to talk about—so, this is a new breath of life. So, on the off chance that you're ready for a break, this is now a new handle; we're going to start again with a new topic. So, here is the topic:

One of the real issues about equations is that you often want to solve them. That is, you want to see when the value of a particular function is equal to 0. How does a calculator do that? In other words, in some calculators or computers you can type in the function and say solve, meaning find a number x so that when you plug that number x into the function you get 0. That's what it means to solve that equation. Well, there's a very tricky way of getting approximations to the solutions of an arbitrary equation that was actually thought of by Newton and the title of it is the Newton-Raphson Method. Two mathematicians, Newton and Raphson, back in the very early days of calculus, used this method to approximate the solution to a function.

Suppose you have a function $f(x)$ here. You can see that it is crossing the x-axis at this point c, and we're trying to find what the value of c is. Here is the strategy. You just make a guess as to what that solution is. In this case, we guessed $x1$. You can see that it's too small; it's less than c. But, we don't know c yet. That's what we're trying to find out, where it actually crosses; where the value of the function is equal to 0. Here's what we can do. We can look at this point, $x1$, $f(x1)$. Those are the two coordinates of the point on the graph at our first guess, $x1$. Now what we're going to do is just draw the tangent line at that point and see where the tangent line crosses the x-axis. The concept being that if we are close to the place we're at 0, and we just head out in a tangent direction, we'll get closer to the right answer; we'll get closer to the place where the function actually crosses the x-axis where it is equal to 0.

Now, it's really quite simple to find out where the tangent line to the point $x1$, $f(x1)$ crosses the x-axis because we know that the derivative is the slope of the tangent line. So, the derivative is equal to the rise over the run; the rise being $f(x1) - 0$, because we're looking for the point whose y-coordinate is $0 \times x1 - x2$. We're trying to find $x2$; that's the thing we don't know at this point. We've guessed $x1$; we're trying to find $x2$. So, we know that $x2$ satisfies this condition, and doing this just little bit of algebra, we see that $x2 = x1 - f(x1)$ over the derivative of $x1$. That gives us our point $x2$.

Now what we do is we iterate the process; that is, we do it again. We find the value at $x2$, and then we say that's not quite 0, so we'll follow the tangent line down and see where it hits, and that gives us $x3$, using exactly the same way to get the next approximating value. Then we go up to the line, find the derivative, follow it down, and see where it hits.

Now let's ground this in an example to see that those numbers can converge very quickly to the actual place where the graph of the function crosses the axis. So, here we go. Let's consider the specific function $f(x) = x^2 - 5$. We know algebraically that the value is—$\sqrt{5}$ is the number where the value of the function is equal to 0, because when you square $\sqrt{5}$, you get 5; $5 - 5 = 0$. So, we know the answer is $\sqrt{5}$. Let's see how the Newton-Raphson method would give us a numerical approximation of that value.

The first step was to just make a guess as to a number for which the function would be equal to 0. Let's make our first guess equal to 2; that's a pretty good guess. We make it equal to 2. What does the Newton-Raphson method say that we do? We draw the tangent line and we compute where that tangent line crosses the axis again; in this case, just plugging in the value $2 - f(2)$ over the derivative of the function at 2. By the way, the derivative of the function is just $2x$, so it's easy to compute the derivative at any point. So, this derivative is 4 here. We plug it in and we see that $x2 = 2.25$. After we get 2.25, we do the same process again. That is, to get the third approximation, we take this value, 2.25, $f(2.25)$. We plug that in to the equation. We divide by the derivative of 2.25. And we get the number 2.23611111 as our next approximation. So, graphically, we're going up to here, then we're taking the tangent line down to here, and that gives us the point $x3$. Then we continue. We find $x4$ to be this value, and so on.

Now after we've done that, we discover that we can compute the square of our approximating values to see how close we're getting to the actual $\sqrt{5}$. We find that our first guess gave us 4; $2 \times 2 = 4$. The second approximation, 2.25, gave us 5.0625. And, then, you can see that the third approximation, the fourth approximation, and the fifth approximation are getting very, very close to having numbers whose square is equal to 5. This is a method by which calculators or computers can actually solve equations.

In this lecture, then, we've seen how calculators work in methods of getting answers to various things by taking infinite series and by taking iterative methods that approach one single value. In the next lecture, we'll see how calculus is going to be used to optimize possibilities. That is, we'll see how calculus can help us find the best of all possible worlds. I'll see you then.

The Best of All Possible Worlds—Optimization

Lecture 15

Today we're going to explore one of the most practical applications of calculus, and derivatives in particular, namely the idea of trying to optimize things.

One of the most practical applications of calculus and of derivatives in particular is *optimization*. Suppose we want to build a box that holds 8 cubic feet with material that costs $2 a square foot for the bottom, $1 per square foot for the sides, and another cost for the top. What shape will minimize the cost of the box? This type of problem brings students to tears, but it illustrates a process of enormous importance in the real world, namely, selecting, from a range of possible designs, the design that optimizes some feature—in this case, cost savings. In this lecture, we'll look at several kinds of optimization problems, including optimizing the area that can be enclosed given a certain amount of fencing and optimizing the shape of a soda can.

Optimization enables us to select, from a range of possible designs, the design that optimizes some feature, perhaps cost savings, time, materials, or use of space. The strategy for solving such problems involves an intriguing application of derivatives. This kind of problem is the bane of the lives of calculus students everywhere.

Suppose you have 600 feet of fencing with which you want to enclose a herd of camels in a rectangular field, one edge of which is bounded by a straight river. What dimensions should you make the fence to enclose the largest possible area? The interesting strategy involved in analyzing this question is that, in a sense, we look at all possible answers at once. The question is one of trying to maximize a quantity—the area of the field. Let's think about some possibilities for the shape of the field—long and narrow, tall and thin, and shapes in between. We can experiment with different choices. Thinking about listing all the choices gives us the realization that any allowable choice of length and width determines an area. The area depends on length and width. The width can be any value from 0 (laying the fence right along the riverbank) to 300 (having the fence just go up and back right next to itself). Of course, the answer will lie somewhere in between.

Calculus is, of course, the hero or heroine. The clever strategy is to ask a question about *change* in area (as a function of *change* in width), rather than the more natural question about values of the area. If we make a guess for an optimal width, what would be the effect on the area if we changed our guess a little? A change in width results in a change in the area, because the area is determined once we know the width. This situation is exactly what calculus is for. Calculus (in particular, the derivative) deals with cases where the change in one quantity produces a change in another. For example, suppose we make the reasonable guess that the optimal shape for the field would be a square. The area would be 40,000 square feet. If we change the dimensions to 201(length) × 198(width), we have an area of 39,798 sq ft, so we know the square has more square feet. However, if we change again to 199(length) × 202(width), the square footage is 40,198. Thus, we know the square is not the optimal fencing shape. All possible configurations of the area enclosed within the fence can be expressed in the following formula, where w = width: $A(w) = (600 - 2w) \times w = 600w - 2w^2$.

The clever strategy is to ask a question about *change* in area (as a function of *change* in width), rather than the more natural question about values of the area.

Let's organize the possible widths and areas in a convenient form. For every choice of width, we can record the area by graphing the function of area as it depends on width. Looking at the graph of possible areas, it is easy to pick out by eye the **maximum** area; it's the one on top. Notice that at that peak point, the blown-up picture of the graph would look like a horizontal line. When the graph looks like a horizontal line, the derivative is 0. Thus, we know that the maximum area will occur at a point where the derivative is 0.

How can we find a point where the derivative is 0? The answer is to find an expression for the derivative and set it equal to 0. In this case, the derivative of $600w - 2w^2$ is $600 - 4w$. We set the derivative equal to 0 and solve for w, as follows:

$$0 = 600 - 4w$$
$$4w = 600$$
$$w = 150$$

Here is an overview of our method: The strategy for finding the maximum area was not to compute the area for different values of possible widths. Instead, we considered the whole collection of possible areas created by all possible choices of widths. Among those, we looked for a value of the width at which the rate of change of the area with respect to width was 0. In other words, we looked for features of the dependency of the area on the width. In particular, we looked at the rate of change of the area when the width was changed and found a place where the rate of change was 0.

This strategy of finding *maxima* and *minima* is extremely useful. The strategy is to realize that *maxima* and *minima* will occur at places where the derivative is 0. Generally, in a graph there aren't very many places where that happens, so we don't have too many values to look at. This max-min strategy is valuable in many settings. Here is an optimization problem from human physiology: When we cough, the trachea (windpipe) contracts to increase the velocity of the air going out. How much does it contract to achieve the maximum velocity for the escaping air? Biologists have observed that a good model for the velocity of the air escaping from the trachea as a function of its radius is given by the equation $v(r) = c(r_0r^2 - r^3)$, where c is a constant and r_0 is the radius of the trachea at rest. The strategy is then to find where $v'(r)$ is zero: $v'(r) = c(2r_0r - 3r^2) = 0$ exactly when $r = (2/3)r_0$, that is, when the trachea is about 2/3 contracted. Amazingly x-ray photographs confirm this observation!

Our final example comes from product design. Suppose we are designing a soda can that is to hold 12 fluid ounces. What dimensions minimize the amount of aluminum used? We will assume that our soda can is cylindrical. If it has base radius r and height h, its volume is $V = \pi r^2h$ and its surface area is $A = 2\pi rh + 2\pi r^2$. One fluid ounce is approximately 1.8 cubic inches, so our volume V is 21.6 cubic inches. Solving for h, $h = 21.6/\pi r^2$, and substituting into the equation for A, we find $A = 43.2/r + 2\pi r^2$. Finding the derivative, setting it to zero, and solving for r, we find $r = 1.51$ inches, which in turn, yields $h = 3.02$ inches. Does this agree with the actual dimensions of soda cans? In reality, many other factors must be taken into consideration when designing soda cans, changing the function to be optimized and, thus, resulting in different measurements. ■

Important Term

maximum: The largest value of the outputs of a function. The *y*-value of the highest point on the graph of a function. It does not always exist.

Suggested Reading

Any standard calculus textbook, section on max-min problems.

Questions to Consider

1. Find two numbers such that their sum is 20 and their product is maximum.

2. Issues of balance and proportion are often problems of maximizing or minimizing something. What is the balance of daily activities, such as work, social activity, and exercise, that maximizes happiness? Do you think that the philosophy of the mathematical strategy for finding maxima and minima is useful when considering such completely nonmathematical issues as human happiness?

The Best of All Possible Worlds—Optimization

Lecture 15—Transcript

Welcome back to *Change and Motion: Calculus Made Clear***.** Today we're going to explore one of the most practical applications of calculus, and derivatives in particular, namely the idea of trying to optimize things. Let me give you an example of a kind of a problem that we're going to deal with. Suppose we want to build a box, and we want the box to hold 8 cubic feet, and we're going to build it with material that costs $2.00 per square foot for the bottom and $1.00 per square foot for the sides, and maybe another cost for the top—$3.00 per square foot for the top. What shape should we make the box so that the cost of the material is minimal, but it still contains 8 cubic feet?

Now, this is a great question and this is the kind of question that brings students to tears. They just hate these problems because they don't know what to do; they don't know how to set it up. It's a word problem, and word problems see are the bane of existence of students. But, this kind of problem illustrates a process that's of enormous importance and it actually is used in the real world in many different arenas. The idea is to take from a range of possible designs for something to find the design that optimizes some feature of it. In the case of the box that we were talking about, we wanted to minimize the cost; so we're optimizing the cost efficiency of the box. We're going to be looking in this lecture at several different examples of problems of optimization, including trying to optimize the area that can be enclosed by a given length of fencing, and optimize the size of a soda can so that it will contain our 12 ounces of delicious soda with the least use of aluminum. All of these problems illustrate a particular strategy for solving problems that is really an intriguing application of derivatives.

So, let's begin with a specific problem involving fences, camels, and a river. Here is the problem: suppose that we have 600 feet of fence, and let's just suppose this cord here is 600 feet of fence, represents 600 feet of fence. We wish to create a rectangular enclosure next to a river. That is, the river will form one side of the enclosure, but we want to use just 600 feet of fence and we want the enclosure to be rectangular. So in other words, the fence goes out from the side of the river, and then it goes across, and then it goes back to the river, as you see pictured here. Well, we have 600 feet of fence and we wish to enclose the maximum area because we have a herd of camels that we want to enclose in that area and we want them to have as much food as they can and room to roam. Now, we choose camels; you may wonder why do we choose camels to illustrate this problem with, and the reason is that camels, I am told, can't swim. So, this means that the river will be an effective enclosure part for the camels.

Well, here's the strategy for working on this problem. The goal of the problem is to find the optimal dimensions to put this 600 feet of fencing in to enclose the camels. The strategy that we're going to use is to imagine all possible shapes for this enclosure. From a shape that's very, very thin, like this; this is an extremely thin kind of shape that uses 600 feet of fencing where, basically, we just put the 600 feet of fencing straight along the river. Of course, that would not be the optimal amount of area enclosed because of the fact that if it's right next to the river, no camel could fit there; there's almost no area there. Likewise, we could adjust it—these are intermediate kinds of shapes—and we could adjust it until we got a very thin and tall enclosure. Well, if it's extremely tall, like this, then, once again, it has almost no area. So, somewhere in between, the amount of area enclosed by the fence is the maximum amount of area, and it is our challenge to find what the dimensions are that maximize that area.

Well, of course, the hero of this story is calculus. The strategy that you might think of doing is to examine different possible configurations of the fence and find out what the area is contained inside that amount for that dimension. What is the area contained here? What is the area if the shape of the fence is here? And we can do that kind of analysis to make a guess as to what the maximum area is. But, it turns out that the strategy that's most effective to find that maximum area is not to look at what area is enclosed in the various possible dimensions and possible designs, but, instead, to imagine all of the different designs simultaneously, and for each design we're going to see how the area changes for a change in the choice of the design. So, that we're looking—instead of the absolute quantity of area contained for each of the different design choices, instead we're going to see how that area changes for different choices of the design, and then we're going to take a derivative—we're going to take a derivative that tells us how the area is changing for a change in the design; and using derivatives, we're going to find the optimum. And that's what we're going to get to in a minute.

So, let's begin, though, by starting with an example of a likely shape for the fencing that might work. Let's try a square. It seems as though a square would be a good shape for the fence, and we might think that's the optimal size. Well, if we have a square of the fence—let's just look at what the dimensions are. It would be 200 feet by 200 feet by 200 feet. So, here we go. Here's our example where we have the river is on the bottom, and then we go out 200 feet, over 200 feet, and down 200 feet. That's one possible shape of the enclosure using 600 feet of fencing. Well notice that its area that it encloses is 40,000 square feet—200 × 200—40,000 square feet of area.

We can think about this square as our possible shape in the following way: let's think about what happens if we alter the shape just a little bit. In other words, suppose that instead of a square, we make a rectangle that's a little bit higher, that is farther away from the river, and less wide. Now here on this graphic you can see that we have altered it so that now it's only 201 feet going out from the river, and then 198 feet across, and then 201 feet back down to the river. Well notice that what happens is if we multiply 198 × 201, that will give us the area of that potential shape of the enclosure, and we see that that area is only 39,798 square feet. So, this is making us feel good that the square is a good shape because it encloses more area than when we made the enclosure go out further from the river.

Let's do another exploration of instead of making the square taller—that is, further away from the river—let's make it a little bit wider. In other words, instead of having it just 200 feet wide on the river, let's make it where we have 199 feet going out from the river, which would leave 202 feet across, and then 199 feet down back to the river. So, in other words, we took our square and we widened it a bit, just by a tiny bit. Well look what happens. If we multiply out to see what the area of that new alternative enclosure is, we multiply 199 × 202—the length times the width of the rectangle, and we see we get 40,198 square feet. We get an added number of square feet. That tells us that the square is not, in fact, the optimal configuration for the fencing to enclose the maximum area because we demonstrated that by making it wider, we increased it; we increased the area.

We can think about all possible configurations of our fencing by introducing a variable w that refers to the distance between the river and going out from the river. For any choice of w that we make, that determines the shape of the rectangle that we can construct with 600 feet of fencing. Namely, if it's w feet out from the river, then, of course, we have to have w feet of fencing used on the opposite side, which leaves a total of $600 - 2w$ fencing for the sideways direction of that rectangle. So, this introduction of the variable w is allowing us to consider every possible value of the configuration of the fence; from a very small w, which we would have if we used a small w and leaving a very wide area to a tall w leaving a less wide area.

For any choice of w, the area enclosed by that choice of dimension is—the area for the choice w is the width, $600 - 2w \times w$, which is $600w - 2w^2$. So, for every choice, that is the area; for every choice w, this is the area of that particular choice of design. We can graph that relationship; that is, for every choice of the width—and by width, of course, the term width—I'm sort of struggling with the term width because since sometimes the w is a smaller number than the sideways dimension and sometimes the w is a larger dimension than the sideways dimension, the usual word width refers to the narrower one for a rectangle; but I'm talking about width, meaning the distance from the river out and the other dimension is the other dimension of the rectangle.

So, the distance away from the river is the variable w, and for every choice of such a width between 0 feet, which just lays the fence right on the river, to 300 feet, which just puts the fence right next to each other, 300 feet up and 300 feet back, for each such choice we have computed the area of that particular choice of design according to the formula, $a(w) = 600w - 2w^2$. Just looking at the graph, of course, we can answer the question what choice of w gives us the maximum area? Well we can see in the graph it's the top point; that's where the graph has the biggest area, because this is a graph of the area. It occurs at the top point. But, notice something about that top point. Notice something about the derivative; the derivative of this area function with respect to w. At the top point, of course, the derivative is 0 because the tangent line is horizontal. Tangent line horizontal corresponds to derivative being 0. Well this, then, is the key to the strategy of max-min problems, as they're called, maximum-minimum problems, max-min problems in calculus. We introduce a function that is recording all possible values, all possible designs; in this case, all choices of the width *w*. We see the value of the thing that we're trying to maximize or minimize; in this case, the area enclosed. Then we notice that the optimal value is going to take place—in this case, the highest value—will take place at a place where the derivative is 0. So, our challenge now becomes to find a place where the derivative is 0.

Well, of course, how do we do that? If this is our function for the area for every choice *w*, $600w - 2w^2$, which it is, we can just take the derivative. We know how to take the derivative of this kind of polynomial. The derivative of $600w$ is just 600; the derivative of $-2w^2$ is $-4w$. So, we now have an expression that tells us the slope of the tangent line at each choice of *w*. We're looking for a place where that slope of the tangent line is 0, because it's at the top the slope will be 0. So, we set the derivative equal to 0; so, $0 = 600 - 4w$, and simply solve for *w*. In other words, we find what *w*—in this case, $w = 150$ feet—we find the place *w* where the derivative is 0. Now notice that in this computation here, we never looked at the actual area of the enclosure. Instead, our analysis had to do with how our different design options were changing. So, this is what the derivative is telling us, how the design options are changing with a choice of *w*. At a maximum point, a change is always going to not increase or decrease the value of what it is that we're trying to optimize. So, $w = 150$ feet.

Now, let's go back and just confirm, by the way, that our solution is indeed the maximum. We'll confirm it by actually computing the area and just seeing what happens. If *w* is 150, then the width is exactly 300, and *w*, again, 150 here. So, the area of the enclosure that is 150 by 300 is 45,000 square feet. If we make *w* a little bit taller; that is, we make it 151 feet out from the river and, therefore, make it only 298 feet across and 151 feet down to the river, we can see that the area of that enclosure, 151×298, is 44,998 square feet. So, it has, in fact, declined; decreased in area. Likewise, if we make it wider; that is, we go less far out from the river, 149 feet instead of 150 feet, then the width along the river would be 302 feet, and, so, the area of that enclosure would be 149×302, which is 44,998 square feet again.

We have successfully solved the problem of what dimensions to make the fencing to enclose the biggest area. But, much more important than that, we've introduced a strategy; a strategy of finding the optimal values for a varying function. Namely, if we have any function at all which you can view as something which you are trying to optimize or minimize—optimize, in this case, the area of the enclosure, or minimize, like minimize the cost of that box we talked about in the introduction. If you can write an expression that tells us for some design parameter it tells you the value of that variable that you're trying to optimize, then instead of having to look at all the different places along those potential design options, you can just find those places where the derivative is equal to 0 and just check those, because you know that the optimal values are going to occur at the places where the derivative is equal to 0.

Let's do another example. This is an example from biology. Suppose it turns out that by doing experiments people have found that when you expel air from your mouth, that the velocity at which that air comes out is a function of how constricted the trachea is. They have discovered that the velocity of the escaping air is approximated by this function—that the velocity is determined by the radius of the trachea—and it's according to the following function: It's proportional to—that is to say, there's a constant, times the radius of the trachea that it started with at rest, $r0$ is the radius of the trachea when it is at rest, times $r^2 - r^3$, where the r is the radius to which it has been constricted. So, this tells us the velocity of escaping air for different choices of the radius of the trachea.

Suppose we want to say, how much should one constrict your trachea in order to expel air at the fastest possible speed? For example, if you're coughing and you're trying to get stuff away from you, how do you expel it at the fastest possible way? Well, you contract your trachea to the level at which the velocity of escaping gas will be maximum. So, it's a max-min problem. So, how do we do this? We have a function here that describes the velocity, so what do we do? We know that the maximum is going to occur at a place where the derivative is equal to 0. For all possible radii, we're going to find that place where the derivative is equal to 0. Here is the function that is telling us the escaped velocity. We take the derivative—well, it's easy to take the derivative of this function—$r0$ is a constant, it's the rest radius; c is a constant. So, to take the derivative, we just bring down the 2 and we bring down the 3, and r^2; we get that the derivative is equal to $c \times 2 \times$ the resting radius × $r - 3r^2$. We set that derivative equal to 0 because we want to know at what radius will the derivative be 0 because that is going to be at a place where there is a maximum. Setting it equal to 0 gives us this expression; we can factor out the r, as you see. Solving for it we have $r = 0$ or $r = 2/3r_0$.

This means that the $r = 0$ is not a reasonable answer, and sometimes there, by the way, you can have a function where it has a place where the derivative is equal to 0 that is not at the maximum. It may be some other place. It can be not the correct answer. What you need to do is realize that the correct answer will be at a place where the derivative is 0, but not all places where the derivative is 0 are the correct answers. So, you have to also think about what you are doing.

In this case, the correct answer is that $r = 2/3r_0$. In other words, if you contract your trachea to 2/3 of the rest diameter, rest radius of your trachea, then the expulsion will be maximum. Amazingly, people have actually taken x-ray photographs of people coughing that confirm that nature, herself, has discovered that this is the optimum and that one's trachea does, in fact, constrict to 2/3 its resting radius at the time that you cough in order to expel the things you're trying to get rid of at the maximum velocity. So, that shows that nature knows calculus. Let's do one more example of optimization. Here is an example of optimization of a type that is used in business. Namely, suppose that we wish to design a soda can that contains a particular volume, a certain amount of liquid; and let's say 12 ounces. Our goal is to design that soda can so that it uses the least amount of material. Now, this is a very practical kind of problem. If you're going to make billions of soda cans, you want to use the least amount of material, so you want to ask the question: What shape would we use the least aluminum? How would one design a soda can to contain a certain volume and have the least surface area of the can? That's the question.

We'll restrict ourselves to a soda can that is cylindrical. Suppose we want a cylindrical soda can. The volume of a cylinder is $\pi r^2 h$; πr^2 is the area of the base of a cylinder, times its height; that gives the volume of a cylinder, $\pi r^2 h$. What's the surface area of a cylinder? Because the surface area is the part that's going to be the material. Well the surface area has two parts; it has the wall part, that is the up and down part. Here, I have this soda can. So, it has the up and down part, the wall part. Well the wall part is going to be the circumference of the circle times the height. The circumference of the circle, if it is radius r, is $2\pi r$—$2\pi r$ is the circumference of a circle of radius r—$\times$ h. So, the wall part, the up and down part, is $2\pi rh$, and the top is just the area of a circle radius r, so it's πr^2. The bottom is also area πr^2. So, the total area—top, bottom, and side walls—is $2\pi rh + 2\pi r^2$.

We now have an expression for every possible choice of r, for the radius, we have an expression that tells us what the area is. Let's now solve the problem of what would be the optimal shape of a soda can to contain 12 ounces. Well, it turns out that 1 fluid ounce of liquid is about 1.8 cubic inches, occupies that many cubic inches. The volume that we're seeking is then 12 times that, which is 21.6 cubic inches of volume. So, in our example we have, then, that the volume is $\pi r^2 h$, which is 12 ounces, 21.6 cubic inches. So, that is the volume that we're trying to enclose in this optimal-sized can.

We have, then, an expression, $\pi r^2 h$ equals some number of cubic inches, so we can solve for h. Solving for h; that's simply dividing by πr^2 on both sides of this equation. We have that $h = 21.6 \times \pi r^2$. Now why do we do this? We do this because originally the area of the sidewalls of our soda can involved two variables, a choice of r and a choice of h. But we'd like to imagine that we're considering all possible shapes of this soda can with just one variable, and the variable that we're going to use is r. In other words, as we choose we could have a very squat kind of a can that had a big r; or we could have a very tall, thin kind of a can that enclosed this 21.6 cubic inches of liquid, and that would have a small r. So, we're imagining ourselves choosing the different r's, and you can see if you choose a given r, then the h is determined if you want it to contain exactly 21.6 cubic inches. This expression for h is telling us that a choice of r determines h for a given volume.

We know that the area of a can is $2\pi rh + 2\pi r^2$. That's the sidewalls plus the top and bottom. We know that $h = 21.6/\pi r^2$. So, we can just put that in here. Now, we cancel the πr from both sides here, and we end up with $43.2/r + 2\pi r^2$. So, now we have an expression for the area for any choice of r that we make. This is good, because what it tells us is this is the expression that we wish to optimize—in this case, minimize—because what we want to do is have the choice of r that gives us the least area. Now contrary to one's first inclination would be to just put in different r's and see what the areas are. Instead, what we're going to do is find the point on the graph where the derivative is equal to 0, because that's pointing to the place at which the area is minimum. We can, indeed, do that. We take the derivative of our expression; the derivative of the expression $a(r) = 43.2/r + 2\pi r^2$. The derivative of that is equal to $-43.2/r^2 + 2\pi r$.

Having taken the derivative, we're seeking what design option makes that derivative 0. So, we set this expression equal to 0 and simply solve for r. That is, we multiply through by r^2, we solve this expression, and we get that $r = 1.51$ inches. Once we know what r is—and you can see that in this graph that shows us for every possible design choice, for every choice of r, we can see that it's right about here where the minimum occurs; that is, where the area of the soda can is minimum. The actual answer here is 1.51 inches. The height is—just plugging it back into our expression that tells us what the height is in terms of the radius—is $21.6/\pi r^2$—we see that that value is, just calculating it out, is equal to 3.02 inches. Now, what that is is a can, then, that looks like this. So, here is a can that uses the least amount of surface area to enclose the same amount of material as in this actual soda can.

Now I recommended to The Teaching Company that they start producing soda pop, you see, in cans that minimize the amount of surface area. You could have this beautiful soda can from The Teaching Company and send it off with The Teaching Company courses. But, they didn't think that was a good idea; so, I don't think that's going to happen.

Why is it that soda cans are not this shape and are, instead, this shape? Well, one reason is that there are other considerations. For example, if you're going to make a soda can, the top of the soda can has to be cut out of, maybe, a rectangular sheet of aluminum. So, you have certain wastage. So, that would mean that the tops of the soda cans shouldn't be valued the same as the sidewalls of the soda cans. So, considerations such as that and manufacturing issues can actually change the actual answer. But, in solving all of the different ones, you just have a different equation of what it is that you're trying to minimize.

This was an introduction to a strategy for finding maximum and minimum problems; looking at all possible designs and finding a place where the derivative is 0 to focus your attention on how the change is operating rather than the actual value of what you're trying to optimize. This gives us an actual dynamic perspective of how it is that you look at the world from the point of view of calculus. In the next lecture we're going to see some applications of calculus to economics and to architecture. I'll see you then.

Economics and Architecture

Lecture 16

In the last lecture we saw how to optimize various quantities by looking at how those quantities were changing and looking, in particular, at the derivative; setting the derivative equal to 0. In this lecture we're going to illustrate these same concepts of optimization, but apply them in the context of other things: economics and architecture.

Many aspects of business and economic conditions are described by functions that relate one feature of the situation to another. Heads of business wish to optimize and maximize the profits for their companies. Remember: Maximization is a process that calls for calculus! The cost function $C(x)$ tells us how much it costs to produce x items. The revenue function $R(x)$ tells us the total revenue expected if we sell x items. Profit $P(x)$ is simply the revenue minus the cost: $P(x) = R(x) - C(x)$. Our job is to maximize that difference. If we change the number of items produced, the change in the cost can be measured as a derivative of the cost curve. The change in cost is called the marginal cost, *MC*, and it measures the additional cost of producing one more item when x are already produced. *MC* is the derivative of the cost function. Similarly, the marginal revenue, *MR*, is the rate at which the revenue grows, and it measures the additional revenue we get by selling one more item when x are already sold. *MR* is the derivative of the revenue function.

The maximum profit will occur when the marginal cost and the marginal revenue are equal. Why? If, at a particular production level, the marginal revenue is greater than the marginal cost, then producing one more item will yield greater profit—the increase in revenue will exceed the increase in cost. On the other hand, if the marginal cost is greater than the marginal revenue, then decreasing production will lead to more profit, because the cost will be reduced more than the revenue is reduced. Mathematically, we know that the maximum of the function $P(x)$ occurs where $P'(x) = 0$. However, $P'(x) = MR(x) - MC(x) = 0$ exactly when $MR = MC$. Thus, mathematical reasoning agrees with our intuitive understanding.

Here is an example of a business problem with the goal of choosing production and price levels that will maximize profit. Suppose our company wishes to manufacture and sell a product. The fixed cost for designing and modifying manufacturing equipment is $100,000. After those fixed expenses, the cost for producing and marketing the item is $30 per unit. Experience has shown that if the items are sold at price p, then approximately 20,000 – 50p will be sold. Notice that this means that if the price is $400, no one will buy any, and if the price is extremely low, about 20,000 people will buy them. As we have learned, *MC* is 30. *MR*, a somewhat more complicated figure, is calculated to be 400 – x/25. What price should be set, and how many items should be manufactured? The answer is 9,250 at a price of $215 each.

Even an architectural issue, such as where to stand to get the best view of a painting in a museum, is a problem of maxima and minima. First, we need to draw a picture to see what we are trying to optimize. In this case, we are looking to optimize the angle at which we can best view the painting; therefore, our goal is to find the distance from the wall at which the angle is biggest. Once again, we use the derivative—in this case, the derivative of the arc tangent—to arrive at the answer. A similar calculation can be effective in determining where to sit at a movie theater to optimize the angle to the screen.

A strange property of the catenary is that if you made a road of inverted catenaries, you could ride a bicycle with square wheels on that road and be perfectly level.

Challenge problems were a feature of European mathematics in the 17th century. We'll look briefly at two such questions. One such problem was to find an equation to describe the shape of a hanging chain. **Galileo** had mistakenly thought that such a chain hung in the shape of a parabola, but Jungius, another mathematician, had disproved that proposition in 1669. In 1690, **Jakob Bernoulli** posed the challenge to find the correct equation for the shape, and his challenge was solved by Huygens, **Johann Bernoulli**, and Leibniz in 1691; of course, the solution used calculus.

The shape is called a *catenary*, and it has several interesting properties. The horizontal forces on each link are in equilibrium. An inverted catenary is a particularly good shape for an arch because the forces are tangential along the arch. A beautiful example of an inverted catenary arch is the St. Louis Arch. A strange property of the catenary is that if you made a road of inverted catenaries, you could ride a bicycle with square wheels on that road and be perfectly level.

Galileo was almost right about the parabola shape. A catenary and a parabola are similar in shape, and a catenary actually turns into a parabola if you construct a bridge using hanging cables, then suspend weights evenly along it. For example, at the Golden Gate Bridge in San Francisco, the tension in the cable must support a certain load, so the arch that was a catenary became a parabola after the road was hung from it.

The other challenge problem is called the *Brachistochrone problem*. In 1696, Johann Bernoulli posed the problem; namely, down what shaped curve between two points would a ball reach the lower point fastest? Newton was tired from a hard day at the mint when he received the problem. Because he was intrigued and annoyed by the problem, he stayed up until 4:00 a.m. until he had solved it. Five people solved the Brachistochrone problem correctly, and all are famous in calculus: Johann Bernoulli, Jakob Bernoulli, the Marquis de L'Hôpital, Leibniz, and Newton. The solution is a cycloid, which is a curve created by a rolling wheel. A straight line is not the answer. We can perform the experiment to demonstrate this fact. Intuitively, it is reasonable to think that if the curve starts steeply down, the ball will gain velocity faster. This curve is also called an *isochrone* because, no matter where on the curve you release the ball, it arrives at the end at the same time.

Virgil's *Aeneid* refers to an optimization problem now known as *Dido's problem*. Queen Dido was granted the amount of land that could be enclosed by an ox's hide. She cut the hide into a long, thin strip and made a large circle. How do we know that the circle encloses the biggest area given the length of the string? The solution to the optimization question is to find the pattern of a curve and not just a number. The solutions to this problem and the Brachistochrone problem involve a method of calculus known as *calculus of variations*. ■

Names to Know

Bernoulli, Jacques (often called Jakob or James) (1654–1705). Professor of mathematics at Basel and a student of Leibniz. He studied infinite series and was the first to publish on the use of polar coordinates (the *lemniscate of Bernoulli* is named for him). He formulated the Law of Large Numbers in probability theory and wrote an influential treatise on the subject. Together, Jacques and brother Jean were primarily responsible for disseminating Leibniz's calculus throughout Europe.

Bernoulli, Jean (often called Johannes or John) (1667–1748). Swiss mathematician. He was professor of mathematics at Groningen (Holland) and Basel (after the death of his brother Jacques). He was a student of Leibniz and applied techniques of calculus to many problems in geometry and mechanics. He proposed the famous Brachistochrone problem as a challenge to other mathematicians. Jean Bernoulli was the teacher of **Euler** and L'Hôpital (who provided Jean a regular salary in return for mathematical discoveries, including the well-known *L'Hôpital's Rule*).

Euler, Leonhard (1707–1783). Swiss mathematician and scientist. Euler was the student of Jean Bernoulli. He was professor of medicine and physiology and later became a professor of mathematics at St. Petersburg. Euler is the most prolific mathematical author of all time, writing on mathematics, acoustics, engineering, mechanics, and astronomy. He introduced standardized notations, many now in modern use, and contributed unique ideas to all areas of analysis, especially in the study of infinite series. He lost nearly all his sight by 1771 and was the father of 13 children.

Galilei, Galileo (1564–1642). Italian mathematician and philosopher; professor of mathematics at Pisa and at Padua. He invented the telescope (after hearing of such a device) and made many astronomical discoveries, including the existence of the rings of Saturn. He established the first law of motion, laws of falling bodies, and the fact that projectiles move in parabolic curves. Galileo made great contributions to the study of dynamics, leading to consideration of infinitesimals (eventually formalized in the theory of calculus). He advocated the Copernican heliocentric model of the solar system and was subsequently placed under house arrest by the Inquisition.

Suggested Reading

Any calculus textbooks on applications to economics and calculus of variations.

Questions to Consider

1. Recall your last trip to the supermarket. What function are you optimizing as you shop for groceries? What is your constraint for the optimization? Do you think it's possible to write down equations for both functions?

2. What other shapes, coming from nature or architecture, do you think solve some kind of a physical optimization problem? (Think of arches, for example.)

Economics and Architecture

Lecture 16—Transcript

Welcome back. In the last lecture we saw how to optimize various quantities by looking at how those quantities were changing and looking, in particular, at the derivative; setting the derivative equal to 0. In this lecture we're going to illustrate these same concepts of optimization, but apply them in the context of other things, economics and architecture.

We'll start right in with an example in the realm of economics. If you're the head of a company, and you wish to do well, you're trying to optimize the profit of your business. Perhaps you're selling a certain amount of goods and you produce a certain amount of goods; it costs you a certain amount to produce the goods; and then you sell them for a certain amount of revenue, depending on how many goods you produce and sell. So you have a revenue function and you have a cost function, and the difference is the profit. If you make more revenue than it costs you to make it, that's profit; and you're trying to optimize that profit. So, that is the basic concept, the basic goal, of a firm making goods.

Let's go ahead and try to analyze this situation to see how it is that you might use calculus to make a decision about how many goods you should create in order to maximize your profit. Let's look, here, at an example of three functions that are relevant to this question.

First is the cost. For any number x of items that you can contemplate producing, you know that it's going to cost a certain amount of money. Generally speaking, that amount of money is associated with fixed costs, to start the production cycle, and then a certain amount to produce incrementally more items. That produces a function called the cost function. However it's involved, it is capturing the idea of how much does it cost to produce x items.

On the revenue side, we can produce another function that depends somewhat on how many goods you can sell and what price you can sell them at that are involved in making a function that tells you how much revenue you could get by selling x items of a particular good.

Then the profit is simply the revenue minus the cost. That's the profit for the choice of selling exactly x items, producing and selling x items. So, our goal, here, then is to maximize this profit; that is the difference between the revenue and the cost. Once again, this whole strategy that we met in the last lecture, and that we're going to see again today, is that we're going to imagine hypothetically all possible choices of how many goods to produce, and then we're going to look at what effect those different choices have on the profit, and try to maximize the profit. Let's go ahead and proceed.

As before, it's the change in these quantities that's going to be significant; in particular, the derivative of the cost function means—what does the derivative capture? It's capturing the rate at which some function is changing. That's what derivative does. So, if the function is the cost function, then the derivative of the cost function is telling us how much additional cost is involved in producing one extra item.

On the other hand, the revenue function is $r(x)$, so its derivative is telling us the amount of additional revenue we would expect by producing and selling one additional item. That is how much income we get, not subtracting the cost, but just what the income we would get from one additional item. So, that is the marginal revenue. In economic terms, the marginal cost is the term that refers to the derivative of the cost function, and marginal revenue is the term that refers to the derivative of the revenue function.

Here's an example of the kinds of a graph that captures these two different quantities. We have the revenue function here; the revenue function goes up like this. Notice, that here is where 0 items are sold. If 0 items are sold, then, of course, you bring in 0 revenue. As you sell more items, of course, your revenue goes up. At a certain point, you reach a maximum amount of revenue that you can get from your items. Maybe people quit buying the items beyond a certain point, and so your revenue does not continue to rise forever. Likewise, the cost function is a rising function. We have a fixed cost here to start the production process, and then typically the cost will rise with a certain amount of cost per additional item produced.

Our goal is to try to maximize profit, which is revenue minus cost. Let's think about different places that we might consider. For example, how about if we consider right here producing this number of items. Would that maximize our profit? Well, we can analyze it using our analysis of the derivative of the cost function and the derivative of the revenue function. At this point, the slope of the tangent line for the revenue graph is steeper than the slope of the tangent line of the cost graph. What that means is that for an incremental item, the cost increase is less than the revenue increase for producing that additional item. Consequently, this would not be the place for optimal profit, because you might as well move forward because your increase in your revenue is greater than the increase in the cost.

When we get to a point like here, here the revenue increase is very minimal. The slope of the tangent line of the revenue curve is low; consequently, since the slope of the cost function is greater than the slope of the revenue function, then we'd spend more to produce one additional item than we would gain in revenue. So that would be a poor number of items to produce and sell. There is one point here at which the slope of the revenue curve, that is, the derivative of the revenue curve, is exactly equal to the derivative of the cost curve; and that is the place at which the profit will be maximized because that is telling us that at that point it doesn't do us any good to produce more items and it doesn't do any good to produce less items. There's no benefit in producing fewer or greater number of items. Consequently, that is the place where the profit is maximum. It's also, by the way, just the difference between the revenue and the cost. If we actually took the revenue function minus the cost function that difference is a function in itself and, of course, the maximum will occur where the derivative is 0.

Maximizing the profit is a matter of finding a place where the derivative of the revenue is equal to the derivative of the cost, because we're trying to find a place where the profit has derivative equal to 0. That's our strategy of maximizing any function. Let's do an example.

Suppose that we have a fixed cost of \$100,000 to start, and a variable cost of \$30 per unit in our production. So, our cost function is \$100,000 + 30x. The derivative of this function, it's a very simple function to take a derivative of, is 30; that is, for any value x it costs \$30 additional to produce 1 additional item.

The revenue function is typically a little bit more complicated because the revenue is dependent on the price. At a certain number of items, the price varies. In other words, if you're trying to sell a lot of items, you produce many, many items, generally the price goes down per item. So, at a given price you can sell a certain number of items. So, if we have a function 20,000 – 50*p*, where *p* is the price per item that would tell us how many items we could sell at that price *p*. Taking this equation and turning it and solving for *p*, we can see that the price is going to be equal to 400 – *x*/50, where 50 is the number of items that you are selling. So, if you want to sell *x* items, you sell it for price *p*.

Well, then, the revenue is just the price times the number of items. In this case, the price is 400 – *x*/50, so multiplying that by *x,* we have $400x - x^2$ /50 is the revenue that you would obtain if you were selling *x* items. So, the marginal revenue, that is, the derivative, is just 400 – *x*/25. The maximum profit occurs, as we saw, when the marginal revenue is equal to the marginal cost, which is just another way of saying that the difference between the revenue and the cost, the derivative of the difference, is going to have 0 derivative. But these two derivatives, setting these derivatives equal to each other, we simply solve these equations and see that *x* = 9,250. That is the number of items you should produce to make maximum profit, and the price at which you sell each item is $215. So, that's just an example of how you use calculus to maximize your profit for a company and make a decision about how many items to produce.

Well, now let's shift gears and turn to another issue. Namely, suppose that we're now going to talk about an architectural kind of issue. Suppose that we're in a museum, and in this museum there is a work of art, and the work of art is high on the wall. It's above our line of sight. Our question is: Where should we stand so that we get the best view of that picture? If we stand very, very close to the wall, the angle from our eye to the top and the bottom of the picture will be very, very small because, remember, I'm assuming the bottom of the picture is higher than our line of sight. So, if we're very close to the wall, the angle from our eye to the top and the bottom of the picture is small. Likewise, if we move way, way back to the far end of the museum, the angle from our eye to the top and the bottom of the picture is, again, a small angle. If we want to have the picture appear biggest to us, we want to find that distance from the picture in which the angle from our eye to the top and the bottom of the picture is maximized. So, we're faced with a max-min problem. How do we do this problem?

The first strategy for doing this is that we—and, always, in doing these kinds of problems, we draw a picture because a picture is going to tell us what it is we're trying to optimize. In this case, we're trying to optimize this angle, and we can imagine ourselves at all possible distances from the wall. At each distance x from the wall, we can ask ourselves what is the angle θ of x that would go from the top of the painting to the bottom of the painting? Then our goal is to find that distance x away from the wall at which θ of x is maximum; the biggest angle.

So, how do we do this? We just start writing down what we see. If we assume that the height of the painting is h, and the distance above our line of sight is d, then we have this picture of this triangle with another triangle inside it that we're trying to analyze. And, we just write down what we see. Namely, that the tangent of this angle α, this small angle α; the tangent is the opposite side over the adjacent side of this right triangle, so it's just d/x at distance x; that's α's angle. And then the angle $\theta + \alpha$ has a tangent $h + d/x$. The tangent of $\theta + \alpha = d + h/x$. Writing those down, remember our goal is to optimize the angle θ.

So, what do we do? Now we do some algebra. How do we solve for θ, so to speak, in these equations. Well, we take what's called the arc tangent of both sides and then we find a function of θ with respect to x that is this complicated looking thing that involves arc tangents. And, then, what we know our strategy; our strategy—and I purposely picked a question that involved complicated functions because our strategy is the same regardless of what those functions are—our strategy is that we take the derivative of this complicated expression. This is for every distance x away from the wall, this is telling us what our angle θ is in terms of x. Then, our strategy is to say let's imagine every possible distance x that we can think of, and then see at what value of x the angle is biggest by taking the derivative. So, we need to take the derivative of this expression and then set it equal to 0.

Here is the way the graph of that function looks. Our goal is to find the place where the derivative is equal to 0. The problem is that this function whose derivative we need to take is a complicated function. It's the arc tangent function. So far in these series of lectures we've talked about the derivative of the sine and the cosine, but we never talked about the derivative of something as complicated as the arc tangent. This is where this book comes in. This is a typical calculus textbook, and what actually happens in a calculus course such as one I will teach in college is that much of the time is spent allowing us to take the derivatives of many different kinds of functions, including the inverse tangent function. Each one, each category of function, requires a separate analysis and that's what's done in order to allow us to be able to solve a broader range of problems.

In this case, we need to find the derivative of the arc tangent function, as it's called, and that gives us this complicated looking derivative here. It's not really that difficult to find the derivatives of all these different functions. There are ways to do it; they're not that complicated. But each one needs to be done in order to be able to take the derivative and then solve the problem that we're interested in.

So, once we have the particular problem, if we think of the—here's a specific example: If we have the height of the picture is maybe 2 meters, a very big picture; and the height of the bottom of the picture above our eye is 1 meter; then we can actually solve this problem by writing down the equation, taking the derivative, setting the derivative equal to 0, finding what value of x makes that derivative equal to 0, and that will be the distance away from the wall that we should stand to maximize our angle.

I wanted to show you that as just an example of the kind of problem you can solve. Another kind of problem is suppose you're sitting in a movie theater. Where should you sit on this kind of bank movie theater to optimize the angle to the screen? Once again, what do you do? You draw a picture; you write down equations that describe the particular angle given the distance away that you want to consider sitting. You can draw a graph, but you take the derivative, you set the derivative equal to 0, and then you solve for that—find out what distance away gives you a maximal angle of view.

I want to shift gears a little bit right now and turn to challenge problems. In the 17th century when Leibniz and Newton were working, challenge problems were one of the features of European mathematics. So, I want to look briefly at two of these questions. The first of them that I want to look at is the question to describe the shape of a hanging chain, such as this one. Here's a chain. Suppose that we have this chain here, and the challenge was to ask what is the description, the mathematical formula, that describes what this shape is? Well, Galileo had mistakenly thought that if you hang a chain like this, it will hang in the shape of a parabola. But, mathematician Jungius disproved that proposition in the year 1669. You can see—we'll show you a graphic here that shows the difference between a parabola and the shape of a hanging chain. Well in any case, in 1690 or so, Jakob Bernoulli posed the challenge to find the correct equation, and his challenge was actually solved by three people, Huygens, Johann Bernoulli, and Leibniz. They solved this question in 1691, and of course, the solution is an example of a calculus problem.

I don't want to actually solve that question, but what I do want to do is say a few things about this intriguing shape. First of all, it's called a catenary. The shape of a hanging chain is called a catenary, and it really has several interesting properties. One is the horizontal forces on each link have to be in equilibrium; otherwise, it would change shape. So, that is, the right hand and left hand force have to be in equilibrium. As a result of that, if you take a catenary shape and you build something and then put it upright, so an inverted catenary, it is a particularly good shape for an arch because the forces are tangential along the arch. A beautiful example of that, of an inverted catenary arch, is the famous St. Louis Arch. In fact, the catenary equation for that arch is actually printed and displayed on the St. Louis Arch.

A historical note about this, by the way, is that the word itself, catenary,—which is as opposed to the older word was *catenaria*—but that word, centenary, was apparently first used by Thomas Jefferson when he was discussing alternative possible shapes for the arch of a bridge in a letter to Thomas Paine in 1788. So, that's interesting. I just love the fact that these people were interested in all sorts of matters.

A strange problem of the catenary, by the way, is that if you made a road of inverted catenaries, so your road, instead of being flat was made of inverted catenaries, you could take a bicycle that had square wheels, and ride on that road and your ride would be perfectly smooth. In other words, you would sit perfectly level, and we have an animation that will show you this person riding on a bicycle on this road. It really is a strange looking phenomena.

But, in fact, Galileo almost had it right about the parabola shape of the catenary, and that is because the catenary—first of all, they certainly looked very similar, but if you hang a weight from a catenary curve, for example, if you're building a bridge—such as the Golden Gate Bridge—and you start with a hanging thing from which you're going to hang a road, if you put weights at equal increments along the catenary, the tension of the equal weight, such as hanging a road, will actually turn a catenary into a parabola. So, if you look at the Golden Gate Bridge, it actually is a parabola and not a catenary because of the weight of the bridge.

So, we've now discussed a catenary, but I wanted to now go on and to discuss the second of the challenge problems from the 17th century. This second problem was called the Brachistochrone problem. In 1696, Johann Bernoulli posed this other challenge problem, and this was his problem. He said, "Suppose that you have two points in space,"—two points like this—"what shape would the curve be so that if you slid an object down the curve, as it goes down because of gravity, what shape would allow the ball or whatever to reach the bottom point in the quickest possible time?" You see? One possibility would be to just have a straight line; or, another possibility is to have a steeply curved thing like this; or, you could have it go up and down; there are infinitely many possibilities about how to do this.

This challenge problem, as I said, was posed in 1697, and Newton received this in a message. He was tired from a hard day at the mint where he was working at that time. He received this problem, but he found it intriguing—and annoying, by the way; he found these challenge problems annoying, apparently—but, anyway, he stayed up until 4 am until he had solved this problem. I wanted to give you a description of his solution from Eric Temple Bell's book called *Men of Mathematics*, which said the following. He said:

> In 1696, Johann Bernoulli and Leibniz between them concocted two devilish challenges to the mathematicians of Europe. The first is still of importance, the second not in the same class. The first

> problem is the problem of the Brachistochrone. After the problem had baffled the mathematicians of Europe for six months, it was proposed again, and Newton heard of it for the first time on January 29, 1696, when a friend communicated it to him.—[Or maybe 1697, by the way, it's not 100 percent clear.]— Newton had just come home, tired out from a long day at the mint, and after dinner he solved the problem, and the second one as well, and the following day communicated his solutions to the Royal Society, but he communicated them anonymously. But for all his caution, he could not conceal his identity. On seeing the solution, Bernoulli at once exclaimed, "Ah, I recognize the lion by his paw."

So, this is a famous quote.

There were five people who solved the Brachistochrone problem correctly, and all of them are famous in the history of calculus: Johann Bernoulli, Jakob Bernoulli, the Marquis de L'Hôpital, and then both Leibniz and Newton, both also solved the problem. What I wanted to demonstrate about this problem is that the answer turns out to be a curve that's called the cycloid. Now, a cycloid is a curve that is made in the following way: We take a circle here, and we're going to roll the circle around following the path of a particular point on the rim of the circle as we roll the circle against a straight edge. So, here we go. We roll the circle, and as we roll it I'm drawing the path, and the path that I'm creating is called a cycloid.

Now this path of the curve turns out to be the solution to the Brachistochrone problem. That is to say, that if you turn this over, and if you roll a ball down that curve, it will arrive at the bottom faster than any other path. Over here we have a demonstration to show this. It won't show that it's optimal, but it'll show that it's faster than the straight line path. So, let's suppose that we're trying to go from this point at this level down to this point at this level, and we're trying to see which path is quicker; rolling the ball down the straight-line path—which, of course, is less distance—or rolling it down this curved path, and notice that this curved path is a Brachistochrone and you see we've drawn it behind and we previously took a circle and drew it to create Brachistochrone.

So, now we're going to actually perform this experiment by putting a ball on this track and a ball on this track, and I will let them go at exactly the same time, and see what happens. Ready, set—notice that the Brachistochrone ball arrived at the base quicker than the straight-line ball. So, that is an example. In fact, of all of the different shapes, that shape is the one over which the ball will arrive at the base in the quickest possible time.

By the way, it has one other property I wanted to point out, and that is that if you start the ball on any point on this Brachistochrone, it will arrive at the base in the same amount of time. In other words, if you start at the top, it gathers speed and goes very quickly and gets to the bottom in a certain number of seconds; if you start right here, it doesn't accelerate as fast, it arrives at that base point at exactly the same amount of time. So, it's also called an isochrone curve. It's the same as the Brachistochrone.

I wanted to talk about another problem that involves finding the shape of a curve. In this case, we're going to look at a problem that came up in Virgil's *Aeneid*, in fact. Virgil referred to an optimization problem that has come to be known as Dido's problem.

Queen Dido was granted the amount of land that could be enclosed by the hide of an ox. So, what she did was to cut the hide into very thin strips and attach them together, and then had a long curve, and the question was in what shape should that curve be put to enclose the maximal amount of area? It could be put in the shape of a square; it could be put in the shape of an oval; or it could be put in any shape whatsoever. The question is, how in the world are you going to choose what shape is maximum? Well, you might very well correctly intuit that the answer is, in fact, that you put the curve in the shape of a circle. But if you think about how would you actually prove such a thing? It's not so clear. It involves matters such as saying if the curve had a bump in it, then you could increase the area by making the bump go outward instead of inward. If it had a sharp crease in it, you could increase it by making it a smoother curve; and there are lots of technical issues involved like the Brachistochrone problem, the answer to the optimization question, in this case, is to find the pattern of a curve rather than just the number. So, it's a little bit more complicated of a calculus optimization problem. It's a whole area of mathematics; it's called the calculus of variations.

In the next lecture, we're going to continue to look at the many applications and variations of the basic ideas of derivative and integral. In fact, in the next lecture we're going to talk about flying objects, such as baseballs. I look forward to seeing you then.

Galileo, Newton, and Baseball
Lecture 17

One of the most direct applications of calculus concerns the description of moving objects. In fact, we introduced the derivative and the integral by analyzing a car that was moving down a straight road. But, when objects are flying through the air, we have to deal with some complications.

It is vitally important to be able to predict where a projectile will land and what path it will take. Certainly in using any military device, such as a cannon, we need to know the path of the cannonball so we can figure out where to aim it. Perhaps more important, an outfielder needs to know where to stand to catch a fly ball. Galileo performed a famous experiment at the Tower of Pisa to demonstrate a feature of the velocity at which falling bodies fall; **Kepler** studied planetary motion; and Newton devised a general theory of gravitation. All these examples illustrate how methods of calculus can describe the path and velocity of projectiles from cannonballs to baseballs to planets.

One of the most direct applications of calculus concerns the prediction of where a projectile, whether a cannonball or baseball, will land and what path it will take. The work of Galileo, Kepler, and Newton all illustrate how methods of calculus can describe the path and velocity of projectiles. Aristotle wrote that heavier objects fall faster than lighter ones. Galileo refuted Aristotle's assertions by direct experiment. He may or may not have dropped balls from the Tower of Pisa to show that heavier bodies do not fall faster than lighter ones. In a vacuum, a feather falls exactly as fast as an iron ball. Galileo devised a formula for how far a body will fall in a given amount of time. Galileo also formulated the idea that a body in motion will tend to remain in motion at the same velocity and in the same direction until some force acts on it.

In 1665, the plague closed Cambridge University and Newton spent a couple of years on his family farm thinking about mathematics and the universe. (Perhaps years with less instruction would improve our creativity, as well.)

Newton devised the law of universal gravitation, namely, that the force of gravity between any two objects is proportional to the product of their masses and inversely proportional to the square of their distance apart. A falling apple actually may have suggested to Newton the idea that the force pulling the apple was the same as the force holding the Moon or the planets in orbit. Newton also formulated other laws of physics, the first one following Galileo. A body in motion will stay in motion unless a force is applied to it. The second law is written $F = ma$. This law connects force (such as gravity) to acceleration. A uniform force applied to a body creates a uniform acceleration. The third law is stated: For every action, there is an equal and opposite reaction. From the inverse square law of gravitation and his other laws of motion, Newton was able to deduce Kepler's laws of planetary motion, including the fact that the planets revolve around the Sun in elliptical orbits.

One of the most direct applications of calculus concerns the prediction of where a projectile, whether a cannonball or baseball, will land and what path it will take.

Using Newton's laws, we can analyze the motion of a falling body. The force of gravity at the surface of the Earth will cause a body to accelerate at about –32 feet per second per second, where the minus sign indicates that the acceleration is in the downward direction. From this insight, we can conclude that a falling body will accelerate at a constant rate. That is, if we drop a ball, after 1 second, it will be traveling –32 ft/sec; after 2 seconds, –64 ft/sec; after 3 seconds, –96 ft/sec; and so on. Notice that it is the velocity that is increasing with each second. The distance traveled is increasing far more each second. It is an easy matter, then, to write down the velocity of the object at each second. Namely, $-32t$ ft/sec, where t is the time in seconds since the ball was dropped. Now we can use our insights about the relation between distance traveled and velocity to compute the distance the falling ball will have fallen after any given amount of time. Notice that $-16t^2$ is a position function that would give $-32t$ as the velocity at each time, because the derivative of $-16t^2$ is $-32t$. The distance that the ball falls after t seconds is $-16t^2$. After 3 seconds, the ball will have fallen $-16(3^2) = -144$ feet.

Throwing balls lets us analyze the paths that projectiles take. Let's first consider throwing a ball straight up, then catching it. Let's say the initial velocity is 48 ft/sec. If we throw the ball up at v_0 ft/sec, then the velocity will decrease at 32 ft/sec each second, where a negative velocity is the velocity at which the ball will fall after it reaches its peak height. Thus, the velocity at time t will be $v_0 - 32t$ ft/sec. For example, if we throw the ball up at 48 ft/sec, it will rise for 1.5 seconds before it reaches its highest point. At 1.5 seconds, its velocity will be 0 right before it descends. The height of the ball at time t will be $v_0t - 16t^2$; that is a position function whose derivative would yield the velocity we know we have at each time, namely, $v_0 - 32t$.

Let's analyze the path of a fly ball with the help of a graph. Putting these ideas together, a ball thrown or hit at a certain velocity upward (in this case, 48 ft/sec) and a certain velocity forward (100 ft/sec) will rise and fall along a curved path. The height at each second t is given by $v_0t - 16t^2$, where v_0 is the initial vertical velocity. We know the ball will land when the height of the ball is 0 again. Using the calculation $48t - 16t^2 = 0$, we find that the ball will land after 3 seconds. The horizontal velocity will be a steady velocity equal to the initial horizontal velocity. Therefore, after 3 seconds, the ball will land 300 feet away from home plate (100 ft/sec × 3 sec).

Consider a fly ball. How does the outfielder know where to stand to be in the place where the ball will land? The answer is neat. At each moment, the outfielder can measure the slope of the line from the outfielder's eye to the ball. If the rate at which the slope is changing is a constant, that is, the derivative of the slope of the line to the moving ball is a constant, then the ball will land right in the outfielder's eye. Perhaps Willie Mays was doing calculus when he made an incredible catch in the 1954 World Series, winning the series for the New York Giants.

Another way to catch the ball would be available to a baseball player who was playing on a more theoretical field. Suppose the batter hit the ball, but the Earth itself was not there. Instead, suppose that all the mass of the Earth were concentrated in a single point at the center of the Earth. The baseball would then take on an elliptical orbit around that point. It would zoom down, whip around the concentrated center of the Earth, and return. The outfielder could leisurely stroll up to where the batter hit the ball and simply wait. In about half an hour, the ball would return, traveling at precisely the velocity at which it was hit but coming from the opposite direction. Conclusion: Every outfielder knows calculus. ■

Name to Know

Kepler, Johannes (1571–1630). German astronomer and mathematician; mathematician and astrologer to Emperor Rudolph II (in Prague). Kepler assisted Tycho Brahe (the Danish astronomer) in compiling the best collection of astronomical observations in the pre-telescope era. He developed three laws of planetary motion and made the first attempt to justify them mathematically. They were later shown to be a consequence of the universal law of gravitation by Newton, applying the new techniques of calculus.

Suggested Reading

Any standard calculus textbook, section on falling bodies.

Questions to Consider

1. If you are skeptical that every outfielder knows calculus, then how does the outfielder know where to run?

2. Is it surprising to you that such an easily contradicted assertion, such as Aristotle's assertion about heavy bodies falling faster, could nevertheless be believed for centuries?

Galileo, Newton, and Baseball

Lecture 17—Transcript

Welcome back to *Change and Motion: Calculus Made Clear*. One of the most direct applications of calculus concerns the description of moving objects. In fact, we introduced the derivative and the integral by analyzing a car that was moving down a straight road. But, when objects are flying through the air, we have to deal with some complications. Certainly, if we're going to use some military devices, such as, for example, a cannon, we'd want to know what the path of the cannonball is so we can figure out where to aim it. Perhaps much more important than that is that, if we're a baseball outfielder, we need to know where to stand in order to catch a fly ball.

Galileo was reported to have performed some famous experiments at the Leaning Tower of Pisa that were intended to demonstrate a feature about the way that falling bodies fall, the velocity with which the falling bodies fall. Kepler studied planetary motion and Newton devised a general theory of gravitation. All of these things are examples that illustrate how methods of calculus can be used to describe the path and the velocity of projectiles, whether it's cannonballs or baseballs or planets. We're going to, in this lecture, describe some of these ideas of the laws of motions of falling bodies, and we'll begin with Galileo.

Aristotle wrote that heavier bodies fall faster than lighter ones. Well, Galileo refuted Aristotle's assertions by doing a direct experiment. I don't believe it's known whether or not he actually did drop balls off of the Leaning Tower of Pisa to show that heavier bodies do not fall faster than lighter ones, but he may well have done so, and we can do this experiment very simply right now. So, here we have a heavier body and a lighter body—in this case, a baseball and a chicken—and we're going to drop these two bodies to see which one hits the ground first. Now, if we drop them, we notice that they, in fact, hit at exactly the same time.

Now, to be perfectly honest, I have a great deal of difficulty believing that people would accept Aristotle's view that heavier objects fall faster than light objects when the experiment that proves that is simply not true can be done completely trivially by anybody. But, that's what we read in the history books, so I suppose we have to accept it. But, the point is that the objects accelerate, owing to the gravitational forces, at a certain prescribed manner. In fact, if we were in a vacuum, for example, we could take a feather, and we would drop it, and a feather and an iron ball would drop at exactly the same rate.

Well, Galileo was investigating these kinds of issues, and he devised a formula that would tell how far a body would fall in a given amount of time. So, this was what he was investigating. Galileo was thinking about the motion of bodies, and he formulated many interesting ideas. One of them was that if you have a body that's in motion, that it would tend to remain in motion at that same velocity, and in the same direction, until some force acts on it. This was an idea that later Newton described as one of his basic laws. In fact , in 1665, Newton had been a student in Cambridge, and Cambridge University was closed because of the Plague. So, Newton had to spend two years at his family estate, family farm, thinking about mathematics and physics. He thought about the universe and he spent these years thinking of ideas. During those two years, he thought of the absolutely most incredible ideas that any human being probably has ever thought of in that amount of time. He not only devised the basic ideas of calculus, which we're devoting this entire course to, but he also invented the laws of planetary motion and universal gravitation and the laws of motion of bodies. It does make me wonder whether we would get a lot more creativity and education done if we simply closed our schools for a few years every once in a while and let people think for themselves. But in any case…

So Newton during this time, what did he devise? He devised the law of universal gravitation, and that law states that the force of gravity between any two objects is proportional to the product of their masses and inversely proportional to the square of the distance that they are apart from each other. In other words, if you have two objects that are further apart, they pull on each other less than if they are close together by the prescribed amount that is specified in Newton's law of gravitation.

Well, Newton was thinking about these things while he was sitting on this family property, and there's the old story that he saw an apple drop and the apple dropped on his head and he conceived the idea of universal gravitation, and, in fact, it may literally have been a falling apple that suggested the idea that the force that's acting on that apple is the same as the force that holds celestial bodies in place, such as the moon in its orbit. That conceit, the idea that forces happening locally that we see everyday are the same as the ones that are being at work in constructing the universe was a grand thought, and that's one of the features of Newton's physics.

Newton's laws of physics then, there were several of them, and maybe we should state them. One is that—following Galileo's lead—one of his principles of motion was that a body that is in motion will remain in motion at the same velocity and same direction unless a force is applied to it. So, this is the same as Galileo's law.

The second law of motion is that $F = ma$. So, this is the second law of physics of Newton. And, this connects force, such as gravitation, to acceleration. The a of $F = ma$ refers to the tendency for the object to accelerate. So, in other words, if you have a uniform force applied to a body, it will create uniform acceleration.

And then his third law of motion is that for every action there is an equal and opposite reaction.

Newton became aware of the law of the planetary motion of Kepler. In 1600 or so, Kepler was working in the astronomical observatory of Tycho Brahe, and Kepler devised his own laws of planetary motion of which one was that planets proceed around the sun in elliptical orbits. Now, when Newton learned this law, Newton said, "Well, I can actually deduce the law of elliptical orbits from my law of universal gravitation." From the point of view of Kepler, Kepler was simply describing what he saw in a statistical way. He just took the observations that had been made for years at this wonderfully precise observing facility of Tycho Brahe and then he formulated his law of elliptical orbits of the planets to accord with the data.

But Newton, on the other hand, said well, no, the elliptical orbits come not just by observation, not just by luck, but because if there is an inverse square law of universal gravitation, then, in fact, elliptical orbits are a consequence of that law. Bodies will have to fall around each other, that is, orbit one another, in the shape of an ellipse. In fact, there was a quote from Newton about Kepler's work. Kepler's writing was a little bit dense and obscure, and Newton was reported to have said about reading Kepler's work, that it is easier to deduce the law of elliptical orbits and the other laws of planetary orbits from scratch than to find them in the writings of Kepler. So, Kepler's writing was somewhat obscure.

In any case, what we're going to do today is to talk a little bit about how we describe motions of falling bodies. So, let's begin with an example of how a body that is flying around, how we might describe where it is at different times. So, let's start with a baseball that we're going to use, and just think about throwing the ball up in the air. The first example that we'll use is just a ball that just goes straight up in the air and then straight back again. So, this will be our first example.

Now, when we throw the ball up in the air, it has a certain initial velocity. Then the velocity is slowed down because of gravity. So, let's begin our investigation by the simplest possible description of a falling body, which would be to start with a body that's at a certain height and drop it. So, let's begin with the following: Suppose that we take a body near the Earth and we drop it, then the force of gravitation will tend to accelerate that body at a fixed rate, namely about 32 feet per second per second. And, if there's a minus sign in front of that, –32 feet per second per second, because the acceleration is downward. So, if we think of our measure as going up and down, downward is the negative direction.

Let's understand what acceleration means. What it means is that if we drop the ball so that its initial velocity is 0, so we're just holding the ball here and we drop it, what it says is that as time progresses in seconds, so we have time 1 second afterwards, then the velocity would have altered by being 32 feet per second smaller. So, in other words, it was originally stopped; when we let go after 1 second, the speed of the ball going downward is now –32 feet per second. After 2 seconds, the velocity of the ball is –64 feet per second. So, in other words, the ball is falling at that rate. And so on. At any time t, the ball will be traveling at the speed $-32t$ feet per second. This accords with our understanding about the way things actually happen when we drop a ball. It gets increasingly fast; as it's falling it gets increasingly fast.

Now, of course, this is all idealized in the sense that we're not accounting for things like wind resistance and things of that sort that make the actual formulaes a little bit more complicated. Also, notice that we're talking about a constant acceleration due to gravity, when I just told you a few minutes ago that, in fact, the force of gravity increases as the bodies get closer to each other. The reason that we can make that approximation that the force of gravity is constant is that we're just talking about things right close to the surface of the Earth. In fact, as we drop it, there's a slight change in how powerful the gravity is, it gets more powerful as it gets closer to the Earth, but it's negligible because the Earth is about 4,000 miles in radius. Consequently, there's not much difference in just a short drop. A good approximation is that the acceleration is constant.

If we drop the ball, we know what the velocity will be at every time t in seconds, –32 feet per second, if we start at a stopped position, it has no velocity, and we let go. Well we know that the derivative of the position function is the velocity function of a thing moving on a straight line, in this case a vertical straight line, so that we can say that the position function will be a function whose derivative is $-32t$, and that is $16t^2$. This is the position function that will tell us where the ball will be after any number of seconds that we like. In other words, if we hold the ball in its initial position, $p0$ here is just 0—so we call 0 the original position of the ball. We let go and the initial velocity is also 0. Then the only remaining term is $-16t^2$ feet, which is the distance that the ball will have fallen after t seconds. So, that is the first analysis that we've made here of a falling body. So, here we have, in seconds, we can see what the velocity will be of this falling body, and we'll see what the distance will be as the body drops at every second of time. So, after 1 second it will have dropped 16 feet; after 2 seconds, 64 feet; and after 3 seconds, 144 feet. You can see that every additional second adds additional amount of distance because the speed has increased.

Now, let's go back to the situation where we impart an initial velocity on the ball. So, now we go back to this situation here where we start with a ball and we throw the ball in an upward direction at a given speed. So, we throw the ball straight up and let's try to analyze where the ball is at every given time. Well if our initial velocity is 48 feet per second, and we just throw it straight up, and its initial position we think of as 0, then we can do the same analysis as before—namely, we can say what the velocity will be at every time t seconds after we throw the ball up. Namely, it will be 48—the initial velocity—minus $32t$ feet per second; that will be how fast it's going. Notice that when t is small, like 1 second, this will still be an upward, that is, a positive, velocity; whereas after, say, 2 seconds, then 2×32 is 64, so $48 - 64 = -16$ feet per second; it's falling back down again after 2 seconds.

The position function is given by just taking our velocity function and integrating, that is, finding a function whose derivative is this velocity function. Because we know how to do that, we know that this function here, $48t - 16t^2$, is a function whose derivative would give this velocity function, because the derivative of $48t$ is just 48; the derivative of $-16t^2$ is—we bring down the 2—-$32t$ feet per second. This is the position function at every time t. And, of course, $p0$ is 0. We assume we're measuring where where we threw the ball was viewed as point 0. There is a description of the height of the ball at every moment of time t seconds after it leaves our hand.

Let's now turn to baseball. When you're playing a game of baseball, if the batter swings the bat and hits the ball, the ball travels in a trajectory that has both a horizontal component and a vertical component. So, we can separate those two components of velocity in the following way, and let's imagine that we have a specific example of a hit ball where the ball is hit at the following speeds: it leaves the bat in such a way that its vertical change in height is 48 feet per second. In other words, the rate at which it is rising is 48 feet per second. At the same time, it's going out into the field at 100 feet per second. So, it's a bit of a line drive. It heads out at an angle where its rise is 48 feet per second and its horizontal distance is 100 feet per second. What we're going to try to understand about this is where will the ball land? Can we compute where the ball will land? Well, here's how we can do it.

Assuming that we make our measurements so that the ball is being hit basically 0 feet from the ground; it's actually a couple feet higher, but let's just say 0 feet from the ground. Then we know that it will land when the height of the ball is 0 again. Well, when will the height be 0 again? The height is given to us by the analysis that we did just previously about a ball just going straight up and down. In other words, the horizontal component has no bearing on the height of the ball as it travels. So, the height of the ball at any time t is $48t - 16t^2$; and we want to know at what time that ball will land; that is to say, when will that be 0 again? Well, if we just do this little bit of algebra, factoring out the t, we have $48 - 16t \times t = 0$. And that will be equal to 0 at two times: one when $t = 0$—that's recording the fact that when you hit the ball it's at level 0; and it will also be 0 when $48 - 16t = 0$, which happens when $t = 3$; that is, 3 seconds later it will land, and it will land 300 feet away from home plate where it was hit. So, this little analysis will tell us the distance away from home plate at which it will land.

There is a very interesting question that you might ask if you were intending to play baseball and be an outfielder, and the question that you might ask is the following: If you were an outfielder and you look at the batter and the batter hits the ball and the ball is going up in the air, how do you know where to stand to get the ball? Well, you know how to turn to the right or the left because you'd want the ball to be coming straight toward you, so there's no question about how far to the right or the left to go. The question is, how do you know how far to stand away from the plate. It's really rather remarkable to me that outfielders are able to make that kind of an analysis so quickly that they go to the place where the ball is going to arrive. How do they know that? How do they know that?

What we're going to do is do a little bit of analysis that involves calculus that will tell us how to compute where the ball will land. So, here we go. This is a picture of the path of the ball as it follows its trajectory and lands in the field. You see, the outfielder is standing right out here someplace, what the outfielder sees about the ball—the outfielder is not able to detect the distance to the ball. What the outfielder is able to do is to see how far up the sight of the ball is as it rises off the bat. In other words, the outfielder's looking at the ball; the ball is appearing to rise against the background of whatever is behind the field of vision. So, the ball is just going up or down in the person's line of sight. That's what it is that the outfielder can actually detect about the ball. The outfielder can't tell how quickly the ball is growing in size at that kind of distance.

So, let's just do an analysis of where the ball is. At every time t, the ball is at position $100t$ feet away from the plate, and its height is $48t - 16t^2$. Those are the two coordinates of the ball's position as it flies through the air. Now, if you are standing at this point in this example 300 feet away from the plate—the plate we already computed is where it will land. If you happen to be standing right there, the line of sight that sees the ball will be going up at a particular angle. If we are looking at a certain angle, we can compute the slope of that line because you can compute the slope of the line just by knowing what the angle is. In fact, it has a name, it's the tangent of the angle; it's the rise divided by the distance away. That's something that is a function of the angle. So, you can compute the ratio of the rise over the run, which in this case is $48t - 16t^2$, divided by the distance from where the ball actually is to where we're thinking of standing, the place where the ball will land. That ratio is $48t - 16t^2$, that's the vertical height, divided by 300 $100t$, that's the distance from where the ball is to the outfielder standing at the correct position. If we just do the little bit of algebra, dividing by 16 and factoring out a t, dividing by 100; the $3 - t$'s cancel here, and 4 goes into 16 four times; 4 goes into 100 twenty-five times. We see that this value is 4/25 times t. Now what that means is that if you're standing in the correct place, then the slope of the line—that is to say, the tangent of that angle, the slope of the line, rise over run—at every time t will be $4/25 \times$ the time t.

See what that means. That means that the derivative of the slope of the line of sight is a constant. You see? In other words, the slope of the line of sight is $4/25 \times t$. It's getting bigger as the time proceeds, and it's changing at a constant rate. The slope is changing at a constant rate. So, if you're standing at a place where the slope of the line of sight is a constant, then that's telling you that you're going to be at the place where the ball will land. If you're too close, then the slope of the line of sight is going to change. It's not going to change at a constant rate. If you're too far away so that the ball actually lands in front of you, the slope of the line of sight will not change at a constant rate. It's only if you are at exactly the right position that the ball will change at exactly a constant rate. So, the derivative of the slope of the line of sight is a constant, and that's what every outfielder must compute every time that outfielder makes a decision of where to stand to get the ball.

Now let me give you a illustration of what happened in real life. In the 1954 World Series, we had a situation where the New York Giants were playing the Cleveland Indians, and in the 8th inning of this import World Series game Willie Mays was out in center field. The batter hit the ball—and those of you who are ancient will remember this—that he hit the ball and Willie Mays had been playing in close for a closer kind of a hit and the ball went over his head. Willie Mays ran back, ran back, reached out his glove over his head, leaning back to see the ball, and the ball landed right in his glove. He not only made the catch, but threw it back and made a wonderful play, and won the game, the World Series, for the New York Giants. What you don't ordinarily see in the highlights of that picture are the calculus computations that he was doing all the time in order to make this incredible play. So, outfielders really know calculus, and particularly Willie Mays. So, that was a wonderful example.

Let me show you that this concept of the trajectory of moving bodies was something that was not clearly known back in the days that people were working on these issues. Let me show you some pictures of what were supposed to be the trajectories of cannon fire. This was in various textbooks about military matters. So, this is an example of the way people thought that the trajectory of a cannonball, when it was shot, it would go straight up and then just go straight down where it landed. So, that was an interesting picture of the supposed trajectory. Here was another one. It went straight up and then straight down into the water. So, these were examples back in the days before they could describe the motion of these bodies as the parabolic equation that we actually computed.

I want to tell you about one other way that the outfielder could make the play, and this is a little bit theoretical. Suppose that we were playing baseball, the same baseball game, except that instead of the reality of the Earth, suppose that the entire mass of the Earth were centered right at a spot in the center of the Earth. You follow me? The whole mass of the Earth is just right at the center of the Earth. But, then we have the same baseball game, and they were somehow—I don't know how these baseball players were standing since there's no Earth there, but there up there near what's now the Earth, but this is just theoretical. The batter swings the bat and the ball goes off into the air. Now the wise outfielder in this game of baseball doesn't compute where the ball will land because there's no land in this theoretical world that we're talking about. Instead, that person knows that any beginning pattern of a projectile moving will attain an elliptical orbit around this other mass.

So, what will happen is the outfielder simply lets the ball go by; the ball goes up; it goes in a very tight elliptical orbit, meaning that it almost falls straight down toward the center of the Earth—remember, we're assuming the Earth itself is not there, but all the mass is right at the center. So, it doesn't hit anything. The ball goes down, accelerates, whips around that center of the Earth where all the mass is, and comes back up. Then it arrives exactly at the point where the batter hit the ball. So, the outfielder simply strolls up to home plate, looks at where the batter hit the ball, waits right there, turns the glove around backward in the same direction the batter hit the ball, holds his glove right there, and waits. Now, all he has to do is wait about a half an hour, by the way, would be the actual amount of time it would take for the ball to fall all the way down toward the center of the Earth and come back up, and it would be coming right back into his glove if he held it exactly the point of impact of the bat with the ball, it would come right back into his glove half an hour later going at exactly the same direction it left the bat at exactly the same speed that it left the bat, and then he would have got the batter out. So, that's why baseball is probably not played in Earths where there is no earth.

It was a great pleasure to talk about baseball in this lecture. In the next lecture, we'll see how to compute other features of objects that are moving in space, such as, for example, the distance that objects travel. I look forward to seeing you then.

Getting off the Line—Motion in Space

Lecture 18

The idea of calculus, and, in fact, mathematics in general, is that we develop concepts in some simple setting, and then we can extend those ideas and apply them in different settings.

Cars do not really drive in straight lines. They move about on roads that turn. A mosquito flying through the air has a position at each moment that is varying in all three spatial dimensions. Planets move around the Sun. After developing the ideas of calculus for cars moving in a straight line, we have gained enough experience and expertise to apply the same methods of reasoning to things moving around anywhere in space. We can ask all the same questions that we considered when talking about the simple case of cars on straight roads. We can talk about their instantaneous velocity and we can ask how far they travel. One of the strengths of calculus is its versatility. Ideas developed in simple settings can be modified and expanded to apply in more complicated settings.

How can we describe motion in space? At each moment in time, a flying mosquito is at some specific position in space. That position can be described by three spatial coordinates. The position of the moving mosquito is an example in which each time yields a more complicated quantity than just a single number. Each time yields all three spatial coordinates.

Let's consider an example of a bug (VW, that is) moving around a plane. Notice that the positions are time-stamped, so we know where the bug was at each moment. Specifically, we know its x- and y-coordinates at each time. Here are some questions we can consider: In what direction is the bug going at each moment? How fast is the bug moving at each moment? The combination of these two pieces of information—the direction the bug is moving and the velocity with which it's moving—is the velocity **vector** of the bug. How far did the bug travel? This last question seeks the length of a curved path. These are all calculus questions.

Let's revisit the scenario of the trajectory of a baseball. Recall that the position of the moving ball can be given by stating at each time how far it has moved horizontally and how high it is. Therefore, this is a *vector-valued function*. Specifically, suppose the ball leaves the bat with a horizontal velocity of 100 feet per second and a vertical velocity of 48 ft/sec. As we learned in Lecture 17, at any time t seconds after being hit, the ball will be located $48t - 16t^2$ feet up in the air and $100t$ feet horizontally from the plate. At any time t seconds after being hit, the ball will have vertical velocity $48 - 32t$, because gravity is changing the vertical velocity by 32 ft/sec downward during each second of travel. Its horizontal velocity will continue to be 100 ft/sec. From this information, we can compute how fast the ball is actually traveling at every moment and in what direction. To find that out, we find derivatives, one of which will give us the rate at which the ball is moving horizontally and one of which will give us the rate at which the ball is moving vertically. We combine those two rates to find the actual velocity and direction of the ball.

> **By adding up the small increments of distance traveled in each hundredth of a minute, we obtain an approximation of the total distance the car traveled from time 0 to time 3.**

To combine the motion in the two directions, we make use of a right triangle and the Pythagorean Theorem. We have one component of velocity in the horizontal direction (v_x) and another in the vertical direction (v_y). To find the diagonal length of the triangle, which represents the speed at which the ball is moving at any time and in any direction, we take the square root of $v_{[x]}(t)^2 + v_{[y]}(t)^2$. We can compute this value for a variety of times and vertical velocities.

We found the velocity and direction of travel of a car; however, we have not yet figured out how far it travels over time. In this case, the car's x coordinate is the horizontal component $p_{[x]}(t) = t^3/2 - t$. Its y coordinate is $p_{[y]}(t) = t^2$. We consider t as measured in minutes and p as measured in miles. The distance traveled is the same as the length of a curved path.

Let's obtain an approximation of the distance the car traveled from time 0 to time 3. At each time, we can find the velocity at which we are traveling, as described before, by taking derivatives in each coordinate and using the Pythagorean Theorem. We can approximate the distance traveled in a short amount of time by multiplying the velocity times a small increment of time. We can choose small increments of time between 0 and 3, such as increments of a hundredth of a minute. By adding up the small increments of distance traveled in each hundredth of a minute, we obtain an approximation of the total distance the car traveled from time 0 to time 3. This whole process is naturally represented as an integral, because the distance traveled is approximated by taking sums that are in the form of the integral. Finding the lengths of curves is generally accomplished by using an integral.

Let's close with a word about the Pythagorean Theorem. We have used it several times, but we have not proven it. We can do so in a number of ways. The 12th-century Indian mathematician Bhaskara provided an exceptionally elegant demonstration of the truth of the Pythagorean Theorem that we can understand best with a simple diagram. ■

Important Term

vector: An arrow indicating direction and magnitude (usually of motion in two-dimensional or three-dimensional space).

Suggested Reading

Any standard calculus textbook, sections on vector calculus.

Questions to Consider

1. Explain why the positions of yourself at every moment during your life form an example of a vector-valued function.

2. In English, explain the relationship between the velocity and position of the motion of the second hand versus the motion of a minute hand on a standard clock.

Getting off the Line—Motion in Space

Lecture 18—Transcript

Welcome back to *Change and Motion: Calculus Made Clear***.** Today, in fact, we're going to be talking about motion, so this is an appropriate title. Notice that cars, in fact, do not always drive on straight lines. When we introduced the concepts of derivative and integral, we talked about a car moving on a straight line. But, in fact, cars don't drive just on straight lines, cars turn around and they move in various interesting ways, and that's the way the cars move. And, in fact, other things move in more interesting ways. For example, a mosquito flying around the room, at every moment it's changing its position; not only in just one straight line, but it's changing in three different spatial dimensions. So, that's an example of a motion that goes in all sorts of strange ways; in three different dimensions. Planets move around the Sun. They move around in this beautiful elliptical orbit.

The idea of calculus, and, in fact, mathematics in general, is that we develop concepts in some simple setting, and then we can extend those ideas and apply them in different settings. In the case of calculus, we developed the concepts of the derivative and integral about a car moving in a straight line, but now that we've developed the expertise and the experience of that car moving in a straight line, we can now talk about more complicated cases. We can ask questions about instantaneous velocity and we can ask questions about how far a car has traveled when the car is not just traveling on a straight line. So, the challenge for today is to take the ideas that were developed in this simple setting and now extend them to a more complicated situation of motion that has more than one degree of freedom.

Let's think about what kinds of questions that we're interested in asking about motion of a car or a mosquito as it's moving around in different ways. Well, there are different kinds of questions that we can ask about it. One is we can ask the question, first of all, where is the moving object at each moment of time? Notice that the position of, for example, a moving mosquito. At each moment of time, you look at your watch and you say, "What time is it?" Then you say the position of that mosquito is determined by three spatial coordinates. If you think about having some sort of a base point, a point of origin for your coordinate system, you can specify the position of this moving mosquito by saying, “Well, how far over is it this way? How far this way? And how far up?” Three different numbers are associated with each time. Or, if we’re just moving on a plane. When we’re moving on a flat plane, and you think of having your flat plane have two coordinates, then the position at every time is determined by how far over the moving car is and how far up it is; two coordinates.

In other words, we’re having a situation where one single varying number, you can think of time, is giving rise to two different varying numbers. There are two different functions involved; the x-coordinate that’s changing and then the y-coordinate is also changing. So, every position can be thought of as being time-stamped to know where it is that we’re moving at each different moment of time.

Now let’s go back to the question what are the questions that we want to consider about these moving vehicles? And, they’re similar to the questions that we asked about the car moving on the straight road. One question we can ask is: In what direction is this moving car or moving mosquito going at every instant of time? Because, you see, its direction is changing. That’s one question, what direction it’s going.

Second: How fast is it going? How fast is it going? Well, it’s not just going horizontally or vertically, it’s going in some diagonal direction. How fast is that going? The combination of the question in which direction is something traveling and how fast is it traveling, that’s the definition of the velocity. So, velocity has two features to it. It has how fast and what direction? We can summarize those two questions that I talked about before by saying, “What is the velocity of the moving object at each moment of time?”

Another basic question that we can ask about a moving car or a mosquito is how far did the bug travel during a given interval of time. So, if you’re moving around the road, you might ask the question, “Well, how far total did the car go?” The reason you might want to know that is because you might want to know how much gas you’re likely to use. You use gas according to how far the car went. So, you’d be interested in knowing the length of the curved path—potentially curved path—that you traverse as you drive around in the city.

So, all of these questions are susceptible to calculus. These are susceptible to the kinds of analysis that we have become accustomed to using in this strategy of calculus. So, let's begin, then, by looking at the situation that we visited in the last lecture, talking about a baseball flying through the air after it's been hit because you notice that that is an example where we already had dealt with two different varying quantities. So, let's go ahead and now put it in the context of this—what we're really discussing now is a vector-valued function, meaning for every time t, we get two different values, its x-coordinate—how far over it is—and its y-coordinate. Since we have two different values associated with one time value that is a vector-value function. So, here we go. Here's an example of in the baseball example that we saw before.

Suppose that we have our baseball player once again striking the ball, and the ball heads out in the horizontal direction at the rate of 100 feet per second. And, then, in the vertical direction, it was going at a rate of 48 feet per second. Now, recall that we could compute what the position was at its y-direction, that is to say, its height was $48t - 16t^2$. So, we computed this before. That was telling us how high off the ground the ball was at every time, and it was just associated with the idea that if you throw a ball up in the air, it's pushed down by gravity, it's slowed down by gravity, and that this formula, $48t - 16t^2$ would tell us the exact height of the ball at every time t.

Of course, since the ball is just going horizontally at 100 feet per second, then after any time t, its position in the horizontal direction is $100t$. So, you notice that every point—that is the position of this moving ball—has two coordinates; it has an x-coordinate and a y-coordinate. Now, one thing we can ask about this is how fast is the ball traveling? We know how fast the ball is traveling in the upward direction by just taking the derivative of its upward position. In other words, since at every time t it's at height $48t - 16t^2$, then we know that the derivative of that position, if we just think of height as a vertical road, then we know that its derivative is telling us how fast the ball is moving up or down at that moment of time, $48 - 32t$. And, its velocity in the horizontal direction is always 100. But, notice that when the ball is going at a diagonal direction, it's going at a diagonal direction, then the actual speed has to be a combination of how fast it's going this way and how fast it's going up. It's faster than either of those individually because it's both going up and to the side at the same time.

Well in fact, it's very easy to geometrically see what the value of the velocity is in this diagonal direction because we have a component of velocity in the horizontal direction and another component of velocity in the vertical direction. So, we have a little triangle here. Our goal, then, would be to say what is the diagonal length of this triangle whose horizontal length is v_x, the velocity in the x-direction, and whose vertical direction is v_y in the vertical direction? v_x is its velocity in the horizontal direction, and v_y is its velocity in the vertical direction. By using the Pythagorean Theorem, we see that the velocity, that is to say, the rate at which it is proceeding— its speed—is the square root of $v_x t^2 + v_y t^2$ at every time t. Now, this is the speed because it is telling us the rate at which it is going. The velocity would include the fact that it's going in the direction of horizontally $100t$, in the vertical direction $48 - 32t$. So, the velocity would include the direction as well as the magnitude of its speed.

Here are some examples of just showing how fast the ball is actually traveling in the direction that it's traveling, that is taking into account at every moment it's going in a different kind of direction. It starts off—when it's first hit by the bat—it starts off always going, of course, at 100 feet per second in the horizontal direction and its initial vertical component of velocity was 48 feet per second. Combining those by taking the square root of the sum of the squares, using the Pythagorean Theorem, we see that its actual velocity is 111 feet per second in this diagonal direction. And, these different speeds that we can see at different times, for example, here at 1.5 seconds after striking the ball, the ball is at its apex; and, at its apex, it doesn't have any vertical component at all, but it does have its horizontal component of 100 feet per second. So, this is a chart that gives us some examples of computing this diagonal-looking speed at various representative moments of time.

We'd like to apply this same kind of analysis to a car moving on a road in a particular path to try to analyze the question of how far that car went. What we're going to do is imagine a car traveling along in the plane according to very specific function. In other words, at every moment in time I will tell you where its horizontal component is, that is the x-coordinate, is $t^3/3 - t$.

By the way, since we're thinking of a car moving on a road, we'll go back to thinking of t as being measured in minutes and the position being measured in miles from the origin. So, this is a car moving where its position is $t^3/3 - t$. Its y-coordinate is t^2.

So, let's just think why it is that that's describing the motion of a car. Well, because at $t = 0$, we know where the car is on the plane. Its x-coordinate is and its y-coordinate is 0. It starts right here. When our watch says 1 minute, where is the car? Well, we plug in 1 to this equation here to see where it is horizontally. It's at the point $1/3 - 1$. That's $-2/3$. And, what is its y-coordinate? Its y coordinate is 1^2, that's 1. At the time 1 minute, the position is right here—this dot—it's –2/3 and +1. Likewise, what is its position at time 2? Well, we plug in 2 here, we've got $8/3 - 2$. Well, let's see, 8/3 is 2 2/3 – 2 just gives a total of 2/3;so, its x-coordinate is now +2/3, and what's its y-coordinate? It's 2^2, which is 4. So, this dot is at the point x-coordinate 2/3; y-coordinate 4; and likewise, you can compute the position at time 3 the same way.

Well, this is an attractive curve, and what we would like to ask about it is, first of all, how fast is the car going at any given time? Well, we've seen the strategy for figuring out how fast the car is going; we just say that at each moment of time, wherever on the curve it is, we know that it has a velocity, a horizontal velocity, and a vertical velocity. Where do we get the horizontal velocity and the vertical velocity? Well, we get those by simply taking the derivative of the horizontal component of the position and the vertical component of the position individually. Taking those derivatives, we see that the horizontal velocity is t^2 1 and the vertical velocity is $2t$; that is the vertical component of the velocity. So, that means that the speed at time t is obtained from the Pythagorean Theorem by taking the square root of the sum of the squares of the x-coordinate of velocity and the y-coordinate of velocity. In this case, we see that the horizontal component—the x-coordinate of velocity—is $t^2 - 1$; and, consequently we put $t^2 - 1$ and square it, and then we add the y-coordinate of velocity, which is $2t$, we square that; so, by taking the square root of this term, we get the velocity, that is the speed, of the car at each moment of time.

Now here we've computed that velocity at various points of time: At t = 0, the x-coordinate of the velocity is going in the left-hand direction, that is, at –1; it has no vertical component; and, so, its total velocity is 1. If we look at time t = 1, at this point the x-coordinate has 0. It either was moving backward or forward. In fact, it's going to go up, and it's going up at the rate of 2 miles per minute, so its velocity is 2 miles per minute, by plugging into this, taking the square root of the sum of the squares. At time t = 2, once again, we see what its horizontal component of velocity is, its vertical component of velocity, taking the sum of the squares, we get 5. So, this is giving us a chart to tell us what the velocities are at various times. These are actually, of course, the speeds; the velocity includes direction.

Okay. Now, let's think about the question of how to figure out how far the car will have traveled during the first 3 minutes of its travels. So, during its first 3 minutes—it goes like this, by the way, it starts at 0; it smoothly turns around and comes up here and ends up at its final destination at the point (6, 9). So, at the time t = 3, it's at the point 6 in the horizontal direction and 9 in the vertical direction.

Our question is that we are now going to pose ourselves is: How far has the car traveled during those 3 minutes? Well, this is a difficult question because the car is traveling along a curved path. How are we going to figure out how long a curved path is? Our strategy is to use the principles of calculus; and the principles of calculus are to break up the question into small pieces, approximate how far the car will have gone in each of those small, small intervals of time, and then add them up. In other words, what we're going to do is we're going to say at time 0 the car is here; at some nearby time the car is somewhere else on the curve; at the nearby time it's somewhere else on the curve; nearby time it's somewhere else on the curve. Between any of those two consecutive positions on the curve we'll say suppose the car just went straight to that next position; and then straight to the next position; and straight to the next position; and then we added up those straight lines. Well, we would get an answer that was actually a little bit less than the actual distance that the car traveled. The reason that it's less is that in all cases, our approximation is a straight line between two points when the actual path of the car was a curved line, which would be, of course, longer. So here we go, let's just see if we can compute the distance traveled. How would we do it?

Here's the way that we think about it. We say at each time t, we can compute the speed that the car is going by taking the square root of the x component of the velocity squared plus the y component of the velocity squared; that is the speed with which the car is traveling. Now, if we multiply the speed times a short interval of time, Δt, we would have obtained how far the car would travel if it just continued at exactly that same speed for an interval of time Δt.

If we want to approximate the total distance traveled by the car between time $t = 0$ and time $t = 3$, here's the way we would do it. We'd say, how fast was the car going at time $t = 0$? We'll multiply that by Δt to see how far, approximately, the car would go in that first Δt interval of time. Then what we'd do is we'd say between the time Δt and $2\Delta t$, let's assume that the car just went at the same velocity that it was going at time Δt, and then multiply that by Δt to see how far that next little straight line interval would be; and then the velocity at $2\Delta t \times \Delta t$. So, the velocity at this moment of time just $2\Delta t$ after 0, we see how fast it's going then by taking the square root of the sum of the squares of the two components of velocity; multiplying by Δt, this is the distance traveled if the car traveled in a straight line for that interval of Δt amount of time.

If we add up all of these small increments of time, as Δt, we're thinking of breaking the interval from 0 to 3 into many small increments of time, each of width Δt, then you can see that this long summation would be an approximation to how far the car went during that amount of time. When we look at that long summation thing, we have to be reminded of the integral. This is an integral. It's the integral of the velocity, the actual speed dt, between t going from 0 to 3. So, this integral is the integral from 0 to 3 of the square root of the x-component of the velocity squared, plus the y-component of the velocity squared, dt. So, we can parse this integral to see why it's equal to the total distance traveled.

Let's go ahead and actually do it in the example that we have, of $t^3/3 - t$, and it's y-position t^2. How would we actually do it? Well, we can compute the velocity, the x-component of velocity is $t^2 - 1$; the y-component of the velocity is $2t$. The total velocity or the speed at every time t is the square root of the sum of squares, which is $t^2 + 1$.

So, if we wish to find the total distance traveled between time $t = 0$ and $t = 3$, we're interested in taking the integral of the speed at each time, that integral is the integral of $t^2 + 1dt$. The integral of $t^2 + 1dt$ is obtained by using the Fundamental Theorem of Calculus. That is to say, we find some function whose derivative is $t^2 + 1$. Well, that function is $t^3/3 + t$. That is an antiderivative of $t^2 + 1$ because if we take the derivative of this, you can see we bring down the 3, the 3's cancel, we get t^2. The derivative of t is just 1. So, sure enough, this is a function whose derivative is $t^2 + 1$. And we plug in the value 3 at the top; we just plug it in to get $27/3 + 3$; we plug in 0 and subtract; that's what the Fundamental Theorem says that we can do to find an integral, and we find that the answer is 12 miles. So, in the first 3 minutes it went 12 miles. We were really moving.

Notice that in doing this entire operation of finding the distance traveled, that what we actually used is everything that we've learned about calculus so far. In other words, we had to know the definition of the integral, the concept of breaking up that path into small pieces and just saying I'll take approximating straight lines for each of the small pieces, and then by taking finer and finer approximations, I will get this long sum of products. We recognize that as, in fact, an integral. And then, when we actually took the integral, the method that we used to take the integral was to use the Fundamental Theorem of Calculus. That is, we took an antiderivative and we plugged in the top value, subtracted the bottom value, to get the actual value of the integral.

So, this is a great example of using calculus in all of the steps that we've seen. In general, if we have a curve that's given, that is to say, we're given a curve and you can think of it as you wish as a car moving along a path where its x- and y-coordinates are given at every time between time A and time B. We can see the analysis that we've already done that the integral from A to B of the square root of the derivative of x with respect to t^2, plus the derivative of y with respect to t^2, times dt, is going to equal the total length of that path, or, if you prefer, the distance traveled by car moving along that path and those positions at every time.

Now, I'd like to take the last couple of minutes of this lecture to fill in a gap that we have had. We've used the Pythagorean Theorem several times. We used it in this lecture to resolve those two velocities into one, and we've used the Pythagorean Theorem in some previous lectures as well. So, I think it's important that we prove theorems such as the Pythagorean Theorem when we use them. So, let's begin by giving a very elegant proof of the Pythagorean Theorem. So, here is why the Pythagorean Theorem is true.

Suppose that we have a right triangle. All four of these triangles are copies of this one right triangle. We could assemble this right triangle on a square in the following way. At each side of the square is in length the sum of the shorter plus the longer sides of our right triangle. So, here we have four copies of our right triangle, and they fit neatly in this green square. Now notice that the remaining area of the green square is a square; you see, it's a square? And, it's the square whose lengths of whose sides is the hypotenuse of this right triangle. So, the remaining area of the original green square minus these four copies of the right triangle is the square of the hypotenuse.

Now, let me just reassemble these pieces in the following way. I reassembled them on the same green square, and now notice that the remaining area of the green square—minus the same triangles that we had before—is just the square on the longer side of the two sides of the right triangle plus the square on the shorter side. So, this is a proof of the Pythagorean Theorem.

Another proof, I can't resist, giving one more very quick proof of the Pythagorean Theorem, which is this, it uses the same four copies of the right triangle, and we just assemble them into a square on the hypotenuse. So, here we have the square on the hypotenuse by assembling four of these copies of the right triangle plus one little square in between.

Now, what we can do is just take this same amount of area, which is the square on the hypotenuse. And just moving two of these triangles, as you see, we get this L-shaped figure; and then if we think about just cutting this L-shaped figure into two right here, we have the square on the shorter side plus the square on the longer side. So, the object that originally had area equal the square on the hypotenuse now is seen to be the sum of the square on the longer side plus the square on the shorter side.

This proof, by the way, is a proof by the Indian mathematician from the 12th century, Bhaskara, who proved this theorem. To me, it indicates something about the fact that mathematics spans not only the ages, but also cultures around the world. So, we'll end this lecture on that note, and in the next lecture we'll extend the ideas of calculus in a different way to a different domain, namely where we're looking at terrains, such as mountainous terrains. I'll look forward to seeing you then.

Mountain Slopes and Tangent Planes

Lecture 19

In this lecture we're going to explore the situation where one value is dependent on two or more independently varying values. This kind of situation happens all the time in the real world, so this is very common.

The volume of a box depends on its length, width, and height. In other words, the volume of the box is a quantity that depends on several features, each of which can vary independently. This situation of a single value being dependent on several others is another arena in which to explore the strategies of thought that make up the world of calculus. Here, the representation of the dependency produces intriguing pictures in several dimensions. The derivative measures how much the change in one varying quantity will affect another dependent quantity. We also analyze how changes in each of several factors individually influence the whole result. The idea of a **partial derivative** is essential to the study of relationships involving several varying quantities.

In this lecture, we will study situations in which one value is dependent on two or more independently varying values. This concept is called a *function of several variables*. We will take the concepts of calculus, particularly the derivative, and see how to adapt them to functions of several variables.

Measuring the altitude at different places is a good example of a quantity (the altitude) that varies according to position, which is given by two coordinates (latitude and longitude). Imagine that we are standing on a mountain. The mountain has varying steepness depending on where we are and which direction we turn. The steepness refers to change in altitude with respect to a change in position, but here, the position can change in many ways, unlike the case of a car moving on a straight road. This leads to an idea of measuring the change in any direction we consider—a **directional derivative**. Think about expanding the location on the side of the mountain. In a small enough region, it is flat, a plane.

Let's explore the area of a rectangle. A rectangle is a simple shape. Its area is length times width ($A = l \times w$). Its area depends on two choices; thus, we can ask: How fast is the area changing with respect to a change in the length? How fast is the area changing with respect to a change in the width? If we have a rectangle 2 units wide × 5 units long, we could add a small amount to its length and see that the area would increase slightly, as well. In this case, we could ask, "What is the rate at which the area is increasing for a unit increase in length?" or we could ask, "What is the rate at which the area is increasing for a unit increase in width?"

These two rates of change in the total when the variables are altered individually are called *partial derivatives*. When we change the length by Δl, how much additional area do we get? We see that we get Δl times the width. If we divide that incremented area by the change in the length, we see that the rate of change in the area per unit change in length is w. If we change the length of our rectangle by 1 unit, we see that we will increase the area by only 2 units per additional unit of length. Likewise, if we change the width of our rectangle by 1 unit, we see that we will increase the area by 5 units per additional unit of width.

A tangent plane is that plane that contains the tangent line in the *x* direction and the tangent line in the *y* direction—lines whose slopes are determined by the partial derivatives.

Suppose we alter both the length and the width at once. Suppose we simultaneously increase the length and width the same amount, 1 unit each? We can calculate the rate at which the area of the rectangle will increase for such a change. This is an example of a *directional derivative*.

Let us now look at the function $f(x, y) = -(x^2 + y^2)$. The analog of the tangent line is the tangent plane. When we had a graph, it was natural to magnify it and see that it became more like a straight line. When we have two variables, when we magnify the graph, it becomes more like a plane. This gives the idea of a tangent plane. We can interpret the partial derivatives with respect to x and with respect to y of this function. For the function $f(x, y)$, at any given point, if we stand at the point (x, y) and we move in the positive x direction, we are fixing the y

coordinate, so the *y* direction is a constant. Thus, we are seeing the rate at which we are rising and falling. Computing that rate would be the partial derivative with respect to *x*. Likewise, if we stand at the point (*x*, *y*) and we move in the positive *y* direction, we are fixing the *x* coordinate, so the *x* direction is a constant. In this case, we are computing the partial derivative with respect to *y*.

These partial derivatives with respect to *x* and with respect to *y* tell us geometric features about the graph of the terrain. A tangent plane is that plane that contains the tangent line in the *x* direction and the tangent line in the *y* direction—lines whose slopes are determined by the partial derivatives. Finding where tangent planes are horizontal locates places where the values might be maximum or **minimum**. Such places would indicate a peak or a downward-facing peak. In other words, if we find a place where both the partial derivatives with respect to *x* and with respect to *y* are 0, we would say that is a candidate for a maximum or minimum point.

Real-world situations often involve several variables. Our weight is dependent on how much we eat, what we eat, genetic factors, and how much we exercise. Our bank balances are the result of two variables, how much we earn and how much we spend. Many complicated situations are dependent on many variables. The methods of calculus give a strategy for analyzing such situations. ■

Important Terms

directional derivative: The rate of change of a function of several variables in the direction of a given vector.

minimum: The smallest value of the outputs of a function. The *y*-value of the lowest point on the graph of a function. It does not always exist.

partial derivative: The rate of change of a quantity relative to the change of one of several quantities that are influencing its value when the other varying quantities remain fixed.

Suggested Reading

Any standard calculus textbook, chapters on functions of several variables.

Questions to Consider

1. Choose an everyday phenomenon and describe how it can be viewed as a function of several variables. For example, the price of a stock may depend on performance of the company, global money supply, fads, and so on.

2. How could we visualize (as a graph does) a function of three variables?

Mountain Slopes and Tangent Planes

Lecture 19—Transcript

Welcome back to *Change and Motion: Calculus Made Clear***.** In the last lecture, you recall, we considered the situation where one varying quantity—we were thinking of time—determined more than one quantity—namely the *x*-coordinate and the *y*-coordinate of a car or a mosquito that was moving around in the world. In this lecture we're going to turn things around. In this lecture we're going to explore the situation where one value is dependent on two or more independently varying values. This kind of situation happens all the time in the real world, so this is very common. For example, I'll give you an example, if you have the volume of a box. Well, the volume of a box is dependent on the width of the box, the length of the box, and the height of the box. So, you can have three varying quantities that determine one thing, namely the volume of the box. So, this is an example of what's called a function of several variables—there are several variables involved in producing one value.

Well this kind of situation—that is, where a single value is dependent on several variables—is a whole arena in which we can explore the strategies of thought that we've developed in the world of calculus. So, in this case, what we're going to try to do is to take the concepts of calculus that we have explored before, particularly the derivative, and then we're going to see what are the analogous features of that could be adapted to this situation of dealing with functions of several variables. So, we'll have to adapt the ideas to a new setting, just as in the last lecture we adapted the idea of one variable—like time—determining more than one thing—like the *x*-coordinate and the *y*-coordinate of a moving car.

So let's go back and think about a very specific case of a function of two variables that makes this, hopefully, clear. We are accustomed to the concept of altitude of a terrain. So, let's think about a mountainous terrain. Imagine yourself walking around on some mountainous terrain. Now, I'm not thinking of the whole Earth, so I'm just thinking of a little place, like maybe a state or something, some bounded area I can think of. On a map we have a flat plane, and at every point in that flat plane on the map, if we're actually walking on the terrain, we'd be at some altitude. So, different points on the map are associated with different altitudes. And if you are literally walking on this terrain, you would go up and down as you walked around the terrain.

Now, let's think about the kinds of issues that we have been discussing when we thought about derivatives. When we thought about derivatives of a graph of a function, we thought about when you made your varying quantity move in the positive direction, we asked the question at what rate did the value of that function alter? Well, now we're in a little bit more complicated situation, and it's sort of interesting to try and ask ourselves the question what question should we ask about this new domain of investigation? Well, here's the situation. When you're actually standing on a terrain, when you're standing there, there are different directions you can go, and depending on which direction you go, you're going to be either going uphill or downhill or level, and those are the kinds of questions that we're going to try to investigate in this topic of functions of more than one variable.

So, if you're thinking about the altitude function on a map, where on each point on the map you get the altitude, in other words, by looking at the latitude and the longitude of a point—that's two variables—it determines a particular value, namely the altitude at that point. Now, when we're standing there we need to be more sophisticated about our questioning of whether the mountain is rising or falling because it depends what direction we go. If we go this direction, it may be rising. If we turn around, it will be falling. If we go to the side, it may be level. So, in fact, in every different direction that we could point from a given point, we'd have a different slope; a different slope. So, this leads to a whole new kind of thinking about directional derivatives. That is, derivatives in all the different directions we can think of.

So this is the kind of thing that we're going to be investigating during this lecture. Another aspect of derivatives that we met in the functions of one variable, that is, just a regular graph of a function—remember that we saw that the graph of the function, of a differentiable function, looked like a straight line when we looked at it very close. If we magnified the graph of a function and we looked at it very, very closely—or is it close? I don't know. It's one of the two—we saw that it had a slope to it, and the slope was telling us how quickly the function was rising.

Well, what's the analogous situation for this terrain? The analogous situation is that at a given point, we're going to see that we have a concept of a tangent plane because it's a two-dimensional object; we're thinking of sort of a smooth hill and if you look at one point on the smooth hill and you magnify it a great deal, it will look like a flat plane, but oriented at some way. These are the kinds of questions that we are going to be investigating today in the topic of functions of more than one variable.

So to ground our discussion, let's begin with a very specific example of where one value is determined by, in this case, two specific varying quantities. Namely, let's think about the area of a rectangle where we're given the length and the width of the rectangle. So we know that if we have any rectangle whose length is l and whose width is w, then the area of that rectangle is $l \times w$. Then we can ask ourselves—the questions that are sort of natural at this point are to ask: How fast is the area changing with respect to a change in the length? How fast is the area changing with respect to a change in the width? So in other words, we're asking the following kinds of questions: suppose that this is our rectangle here, and we may as well to ground our discussion, let's assume that the rectangle is 2 units in width and 5 units in length. Now we ask ourselves what happens if we alter the length by adding a little bit additional amount to the length? We get a bigger area. And we can ask ourselves what is the rate in which the area is increasing for a unit increase in the length?

So, it's the same question as we've been asking many times about derivatives, the rate of change; in this case, the rate of change of the area with respect to a change in just one of the two varying quantities, you see? When we just talk about the change in the area with respect to a change in just one of the variables—in this case l—what we are really asking is the term for the change in the area with respect to one of the variables is the partial derivative of the area with respect to the variable, in this case l. So, the partial derivative of the area with respect to l at a given point we can compute, and this is what we'll discuss right now.

So suppose that we have this equation. The area is $l \times w$. That is for any l and w rectangle, the area is $l \times w$. When we change the length by a certain amount, Δl, how much additional area do we get? Well, we get Δl times the width. And then if we divide that incremented area by the change in the length, we would see the added area, the amount of additional area we get per unit change in the length is just *w*. At a particular point, for example, the point (5,2). If we ask ourselves how much additional area—at what rate is the area increasing—given a change in the length, we would see that the answer is 2 units per unit change in the length. Every unit change of the length by 1 gives two additional units of area in the area of the rectangle.

Well, now we can ask the same question about what happens when we augment the width. Well, when we augment the width, we're asking the question: What happens when we change the width by a certain amount? At what rate is the area of the rectangle growing? The answer is: Well, it's growing at the rate of, if we're at the point (5,2)—in other words, we're imagining a rectangle of length 5, width 2, and then we add a small amount, Δw, to the width rectangle, how much additional area do we add? At what rate do we add the additional area? The answer is: We add 5 Δw area; when we divide by Δw we get exactly 5. So, that means that the partial derivative of the area with respect to the width at the point (5,2) is equal to 5.

These are the kinds of things where we're investigating how sensitive the final answer is to the altering of one of the variables or the other. In this case, we've asked the question independently. Suppose we alter just the width; suppose we alter just the length. Now let's be a little more sophisticated and ask the question: Suppose we altered both the length and the width at the same time? In other words, at this point we're going to augment the area of the rectangle by moving out in this direction. The question that we ask is the same question as before. At what rate will the area of the rectangle increase for a unit change in moving out in that diagonal direction? You see we could ask that same question for every change of a certain amount of length and a certain amount of width.

Right now, we're going to ask the question where we say suppose we simultaneously change the length and the width at the same rate, what rate will the area of the rectangle change? Well, we can draw this little diagram to indicate what it means. We're going to have Δv be the diagonal distance that we're adding. If we notice that the vertical distance is $\left(\frac{\Delta v}{\sqrt{2}}\right)$, and the horizontal distance is $\left(\frac{\Delta v}{\sqrt{2}}\right)$ because, by the Pythagorean Theorem, this amount, $\left(\frac{\Delta v}{\sqrt{2}}\right)^2$, plus this vertical amount, $\left(\frac{\Delta v}{\sqrt{2}}\right)^2$, would be $\Delta v^2/2 + \Delta v^2/2$; so, dividing by 2, adding them up, we get $\sqrt{\Delta v^2}$, which is just Δv; so that's equal to the length of the hypotenuse.

Now we have to ask ourselves the question how much additional area are we getting? Let's just figure it out. In order to see how much additional area we're getting, we need to ask ourselves how much area is there in this additional area of the rectangle if we added at the rate of 1 unit of change in this diagonal direction? We would say that we have $\left(5+\frac{\Delta v}{\sqrt{2}}\right)\times\left(2+\frac{\Delta v}{\sqrt{2}}\right)$. We would multiply those two numbers together to get the area of the augmented rectangle. We would subtract the area of the original rectangle, and divide by Δv to get the rate at which the area of the rectangle is changing per unit change in the diagonal direction.

Doing the actual difference quotient for finding the rate at which the area is changing, we simply multiply $\left(5+\frac{\Delta v}{\sqrt{2}}\right)\times\left(2+\frac{\Delta v}{\sqrt{2}}\right)$. We subtract the area of this rectangle, which is 10; and doing that subtraction, factoring out the Δv, dividing by Δv because we're interested in the rate at which the area is changing, we see that such a change would result in a rate of change of 7/2 area increment per unit change in the diagonal direction. That would be the rate at which the area is changing for a change in the diagonal direction. So, that is an example of a directional derivative. We were asking: How fast is the area changing in the direction of a change in the same direction as the x and the y?

So far we have talked about the question of the partial derivatives with respect to x and with respect to y. Let's go on and ask a question about looking at a particular other function besides—we've now looked at the function that tells us the area of a rectangle is the product of the length times the width. Let's now look at another function. This function is a function of two variables that is $f(x)y = -x^2 + y^2$. Now it's often difficult to imagine what the graph of a function of two variables looks like, but I've selected one where we can actually see it visually. So let me see if I can help us to understand what the graph looks like. $x^2 + y^2$, remember, is the distance squared away from the origin. If you have a point, in other words, (x,y) in the plane, then the distance from the origin, by the Pythagorean Theorem, is $\sqrt{x^2} + y^2$. So, this $x^2 + y^2$ is just telling us the distance away from the origin squared of the point x and y. So, what it's saying is the $x^2 + y^2$, if you take a circle around the origin, then the value along that circle will always be the same, for all points on a circle around the origin of any size. And if we have a small circle around the origin, then the value of $x^2 + y^2$ is just a small number and we're taking the negative so that it's upward facing rather than downward facing. What this is is this: At the origin—in 3-dimensional space—at the origin we just have a point and then circles coming down like this. In other words, it's like a cone—a point, like a peak of a mountain—and going symmetrically in all directions.

This is the function $f(x)y$, and we can see here the graphic that does show this function. Now, let's just interpret the partial derivatives with respect to x and with respect to y of this function. Notice that for the function (x,y) at any given point—if we stand at the point (x,y)—and we move in the positive x-direction, then what are we doing? We're fixing the y-coordinate. We're not moving in the y-direction at all. So the y-direction is a constant, and instead we're just moving in the x-direction and we're seeing what's the rate at which we are rising or falling? So, to compute that—that's called the partial derivative with respect to x—what we're doing is we're just taking this expression, saying, "Suppose that y is just a constant; what would the derivative be for a change in the x-direction?" The answer would be just $-2x$. Likewise, what's the change if we keep x fixed and just move in the y-direction. It's the derivative of this expression, interpreting the x as fixed because it would be fixed if we're just move in the y-direction. So, its answer is also $-2y$.

So, at a given point we can compute the partial derivative with respect to x, so let's look at a particular point. Here's the point (1,1). Let's just see what is the partial derivative with respect to x? At the point (1,1), it means that when you go in the direction of the positive x-axis, you'll be declining—you'll be going downward—at the slope of –2. So, that is one step forward in the x-direction will lead to a –2 value in our function. From the point (1,1), if you go in the direction of the positive y-axis, you will find the same thing happens. That is, 1 unit of movement in the y-direction will make you descend at the rate of –2 units in the y-direction

If we take a different point; for example, here's the point (–1,1). That is, we're at the point where x is –1 and y is 1; we're at that point, and then we're asking ourselves what happens if we step 1 unit in the positive x-direction from there? Well, the partial derivative with respect to x is $-2x$. So, at the point (–1,1), the partial derivative of f with respect to x is 2; meaning that from that point, when you go forward, you're going upward. You see? So, we have this ball here, and at various points when we go in the positive x-direction—I'm thinking of the positive x-direction as toward me. So, I have this cone-shaped object and I have a point back here, you see, because here's the y-axis and this is –1 back in the x-direction. If I go this direction, I have to go upward; whereas, if I'm on this side, the cone is going down, I'm going downward for a change in the x-direction. Whereas, back at this point, you see here's the cone, back at this point, when I change in the y-direction—I'm thinking of this as the positive y-direction—in that direction I go downward because I'm on the slope of this mountain and I'm heading downward to the right. So, the partial derivative with respect to y, algebraically it's $-2y$; and, in fact, you can see that you do descend as you move in the positive y-direction.

What I'm trying to emphasize is the partial derivative with respect to x; the partial derivative with respect to y; are telling us geometric features about the graph of the terrain that are perfectly natural to think about. You're standing on the terrain and you're asking yourself if I walk in the direction of x, am I going up or down and how steeply? If I walk in the direction of y, am I going up or down and how steeply? That's what these partial derivatives are capturing.

Now let's talk about tangent planes. If we look at this graph, at the very peak of the graph, the very top, if you stood right at the very top of that tangent plane—so you've triumphed, you've climbed the mountain and you're standing at the summit to plant your flag at the summit of the thing, it would look—you would be comfortable there, because if you magnified the area around you, it would just become a flat plane. Just like any curved thing, we've talked about this before, that if you have a smoothly curving line, like a circle, if you look at it very closely, it looks like a straight line; and, if you're at the very peak of that circle, it looks like a flat straight line. Both the partial derivatives with respect to *x* and the partial derivatives with respect to *y*, they're both 0.

So that you, standing on the peak, if you magnified—so you just became a little tiny ant—it would look like a flat plane to you. Whereas, if you take our other points that we discussed before, such as the point—let's do this point here—(–1,1). Now I'm going to use these sticks to demonstrate this. So, (–1,1); this says that at this point—I'm thinking of the positive *x*-axis as coming toward me—and when I go in the positive *x*-axis I am going forward and upward. I'm going upward at slope 2. When I go in the *y*-axis, it's –2. So, these are the two directions; if I were walking locally from this point, and I magnified it a lot, I would walk steeply upward, if I went in this direction; and I would walk steeply downward if I went in this direction. Well, there's only one plane that's determined by those two lines that contains both of those straight lines. Namely, the plane that would just fit flat against these straight lines, you see, and that is called the tangent plane; the tangent plane at the point. You can see that the tangent plane is that plane that contains those two tangent lines; the tangent line going in the *x*-direction and the tangent line in the *y*-direction, whose slopes are determined by the partial derivatives.

This is sort of neat. I love these tangent planes; and you can see how the tangent planes move around on the surface.

Let's think about another feature of derivatives that we have explored in some of the previous lectures. One of the features that we explored was how do we find a maximum or a minimum point when we're trying to optimize some quantity? Remember our strategy; the strategy for optimization of a graph, which we said the analysis was this. We said, "If we consider all possibilities for this curve or this thing that we're trying to optimize, we consider all possibilities; we would say the maximum occurs at a place where the derivative of analyzing all of these possibilities where the derivative is 0." Where do you think the maximum of a function of two variables is going to be? Where is it going to be? It's going to be at a mountain peak, just like we saw. It's going to be at a place where the tangent plane is horizontal. That is to say, where both the partial derivative with respect to *x* and the partial derivative with respect to *y* are both 0; and that will indicate either a peak or a downward-facing peak. You see we don't know whether it's a maximum or a minimum, just as previously we didn't know whether we were at a maximum or a minimum point when we found a point whose derivative is 0. But we found if we find a place where both the partial with respect to x and with respect to y are 0, we will say that's a candidate for our maximum point.

Let's look at an example of a function and see if we can explore it. Here's another function, $f(x)y$ = sine x × sine y. So, this is a function. Just looking at that, it would be hard to envision what it looks like in your mind. The sine of x is going up and down; the sine of y is going up and down; and then you multiply them together; it's not clear what this looks like, but we'll see in a minute what this looks like. However, we can take the partial derivative with respect to x by simply doing the following. We say, "Okay, let's make y be a constant, and then take the derivative of the varying part, x." What would we get? We would get the cosine of x—that's the derivative of sine x—× the sine of y, which is staying constant because we're saying what happens when we just move in the x-direction and we don't move in the y-direction. Similarly, we get the corresponding partial derivative of f with respect to y is sine x × the cosine of y, fixing the sine x—because it will be a constant—× the derivative of the sine of y, which is cosine of y.

We know that the cosine of x is 0 at various points of x, just by knowing what the cosine function is doing. We know when the sine of x is 0; it's 0 at π, and $-\pi$, and 0; every π units along it's 0. Let's examine the point ($\pi/2$, $\pi/2$); and notice that both the partial of f with respect to x and the partial of f with respect to y are both 0 at that point. Why? Because the cosine of $\pi/2$—this is the x-coordinate—the cosine of $\pi/2$ is 0. So 0 times anything is 0, so the partial of f with respect to x is 0.

So, there is a maximum and a minimum at the point ($\pi/2$, $\pi/2$) because when the x-coordinate is $\pi/2$, the cosine of x is equal to 0; and consequently, the partial derivative of f with respect to x is equal to 0. Likewise, when the y-coordinate is equal to $\pi/2$, the cosine of $\pi/2$ is 0; consequently, the partial of f with respect to y is 0. Therefore, we find that there has to be a maximum at the point ($\pi/2$, $\pi/2$).

Let's look graphically and see where these maxima or minima are. This is the graph of that product function, the function $f(x)y$ = (sine x)(sine y). We can see that at the point ($\pi/2$, $\pi/2$), we have a place where the derivatives, the partial derivatives, both with respect to x and with respect to y are 0. Consequently, the tangent plane is horizontal, so it's a candidate for being a maximum point or a minimum point. We would find many points, by the way, where both partial derivatives are 0; some of them are minima, such as down here, and some of them are maxima, such as up here.

Well, many real-world situations involve several variables. Our weight, for example, is dependent on how much we eat and is dependent on genetic factors and how much we exercise. Those are independent variables that determine our weight. Our bank balances—while they're the result of two variables, how much we earn and how much we spend. Many complicated situations are dependent on many variables, and we can often analyze them by asking how the final result, the final product, would change when we alter each of the variables independently. So, this idea of analyzing the variables independently often gives us an insight into how they behave in concert.

The methods of calculus, as we've seen, give us the strategies for looking at how change in these functions of two variables is the result of changing them independently. So, in this lecture, we have seen how the ideas of derivatives and tangent lines are generalized to produce ideas about partial derivatives, directional derivatives, and tangent planes in the domain of functions of more than one variable.

In the next lecture we're going to see how the idea of the integral can be expanded to deal with situations that have 3-dimensional objects involved. I look forward to seeing you then.

Several Variables—Volumes Galore

Lecture 20

In the last lecture we were talking about an extension of the ideas of calculus to functions of more than one variable, when we talked about mountain slopes, and functions of two variables. In particular, we were talking about analogs of the derivative in that new situation. Today we're going to talk about the analog of the integral in that new situation.

We have seen that integration is a way to measure area and volume. In this lecture, we extend our techniques to measure volumes and surface areas of solids obtained by a process of revolution, and we construct a theoretical solid that has finite volume but infinite surface area—that is, we have a full can of paint that does not hold enough paint to paint its own walls! Solids of revolution are nice and regular, but how can we measure volumes of stranger objects? To do this, we introduce integration in several variables, thus extending the second main idea of calculus from a line to space. Both the derivative and the integral have rich applications in settings of higher dimensions.

We have extended the ideas of calculus to functions of more than one variable, including analogs of the derivative. Now, we will discuss computing the volumes of *solids of revolution*. Suppose that we rotate around the x-axis the area under a curve $f(x)$ between points a and b. The specific function we will look at is $f(x) = x^2$. The radius of each disk at a point x is then $f(x)$ or x^2, so the volume of a thin slab is just the area of a disk of radius x^2 slightly thickened. The volume of the thin hockey puck is $\pi (x^2)^2 \Delta x$.

Consequently, the volume of our solid is approximated by adding up those volumes. In the limit, that sum is the integral:

$$V = \int_a^b \pi\big(f(x)\big)^2\, dx = \int_a^b \pi(x^2)^2\, dx\,.$$

Now, suppose we have a weirdly shaped but symmetrical vase, with its shape given by the function $f(x) = x^2$ between $a = 0$ and $b = 2$. At each point x, its radius is just x^2, so its volume is $V = \int_0^2 \pi (x^2)^2 \, dx$, which with a little bit of work, evaluates to $\pi(32/5)$. The same volume can be computed by viewing the volume as the sum of cylinders. In this case, for each y between 0 and 4, we have a boundary of a "tuna can" of radius y and height $2 - \sqrt{y}$. Thickening up that cylinder slightly, by thickness Δy, at location y, gives a volume of $2\pi y(2 - \sqrt{y})\Delta y$. Adding up those pieces of volume and passing to the limit gives us an integral whose value is the volume of that solid of revolution, namely, $\int_0^4 2\pi y(2 - \sqrt{y}) dy$, which as expected, has the same value as the previous integral, $\pi(32/5)$.

Solids of revolution also provide us with contradictory objects, including one that has finite volume but infinite surface area. This apparently nonexistent object involves the use of infinity. The object is an infinite horn obtained by revolving a curve around the x-axis. We can estimate the area and the volume of the infinite horn. Using the Fundamental Theorem of Calculus, we conclude that the total volume is finite; in fact, we discover that the volume is π cubic units. To see that there is infinite area, we divide the surface of the horn into pieces. Each piece is larger than a corresponding cylinder. Each of the infinitely many cylinders has area π; hence, the horn has infinite area. Therefore, this horn, if put upright, would be a can that would hold only π gallons of paint, yet we could not paint its walls! Integrals can be used to compute the volume and area of this infinite horn. This solid with infinite surface area but finite volume shows that infinity can be a tricky place to sell paint.

Solids of revolution also provide us with contradictory objects, including one that has finite volume but infinite surface area. This apparently nonexistent object involves the use of infinity.

We now look at how to extend the concept of integration to functions of two variables. Recall that the integral for a regular function gave the area under the curve. For a function of two variables, the integral gives the volume under the surface. A surface is the graph of a function with two variables. We can consider them in analogy to the familiar case of only one variable. Our strategy is to divide the volume into slices that run parallel to one of the axes, compute the volume of each slice using an integral, and add those together, again using an integral. Thus, we use a *double integral*. In other words, computing a double integral requires us to perform an iterated process of integration.

The notation is really neat—a double integral symbol, for example:

$$\iint f(x,y)dxdy.$$

Here's an example: Suppose we want to compute the volume contained beneath the graph of the function of two variables over a particular rectangle in the plane where the x values are between 0 and 2 and the y values are between 0 and 3; that is, $\iint_R x^2 ydxdy$, where $R=[0,2]\times[0,3]$. Our double integral becomes a matter of taking a single integral with respect to y, where at each value of y, the value we want to add is itself computed by taking an integral. Here is the equation:

$$\iint_R x^2 ydxdy = \int_0^3\left(\int_0^2 x^2 ydx\right)dy.$$

Alternatively, we can take a single integral with respect to y:

$$\iint_R x^2 ydxdy = \int_0^2\left(\int_0^3 x^2 ydy\right)dx.$$

Through our calculations, we arrive at the same answer for both equations: 12. We can think of our volume like a loaf of bread and compute the double integral by adding up thickened slices. We can conclude that calculus is the best thing since sliced bread! ■

Suggested Reading

Any standard calculus textbook, chapters on solids of revolution and functions of several variables.

Questions to Consider

1. The integral $\int_0^2 \pi x^2 dx$ represents the volume of what geometric object?

   Can you compute it?

2. Do you think it's possible to construct a horn with finite surface area but infinite volume?

3. The temperature of a rectangular plate at each point (x,y) is given by the function $T(x, y)$. How would you compute the average temperature of the plate?

Several Variables—Volumes Galore
Lecture 20—Transcript

Welcome back. In the last lecture we were talking about an extension of the ideas of calculus to functions of more than one variable, when we talked about mountain slopes, and functions of two variables. In particular, we were talking about analogs of the derivative in that new situation. Today we're going to talk about the analog of the integral in that new situation. But, before we get there, I want to talk about the integral and involving other solid 3-dimensional shapes and talk about how we find the volumes of interesting shapes using integration.

Now, we've already talked about how integrals can be used to find the volumes of various objects, but today we're going to be talking about how to find the volumes of a particular class of objects that are created by spinning an area around an axis. This is a method of creating a volume that is very common, and it produces very attractive types of shapes because they have this kind of rotational symmetry. Many objects that we see around us in everyday life have this kind of method of being creative. For example, anything that you can rotate around like this is something that you can view as being created by taking a certain area and then rotating it around a line. In fact, the way you produce such things as this, I presume, is they would put it on a rotating, spinning axis and then form it to have this kind of rotational symmetry.

So, the question we're going to ask is how can we take objects such as this or some complicated object that has several bumps in it, but that has rotational symmetry, how can we take these kinds of objects and find the volume of these objects? All of these are going to be examples of the basic idea of the integral. I hope that one of the things you get from this course is that the integral is a process of summing, of breaking things into tiny pieces and adding them up. In this case, what we're going to do is take these solids of revolution, chop them into pieces, and add them up.

Let's begin, and we'll begin by taking a particular example of a function, and taking an area under the graph of that function in the *xy* plane, and revolving it around the *x*-axis to produce our solid. So, let's consider the specific function $f(x) = x^2$. This function, as you know, has a graph that is this parabola-looking shape and it goes smoothly from the point (0,0) and it will go all the way up to where $x = 2$. So, that is, it rises up to where $y = 4$. If we take the area underneath that parabolic curve and we spin it around the x-axis, we're going to get something that looks a little bit like this—here, I'll hold it so as you look at it, you'll see the x-axis going off in this direction. It wouldn't be straight like this, because this would just be the equation of the straight line $y = x$, or $f(x) = x$, but instead it's going to be curved and bell-shaped, like the end of a trumpet—it comes out in a bell-shaped kind of a curve, because on the top we're talking about the function $y = x^2$. So, taking that and then spinning it around would give an object that looks like this cone, except that it had a curve to it. Now, our goal is to find: What is the volume that is obtained by—when we spin that area around the x-axis we get a solid—and we want to know what the volume is inside that solid.

Well, our strategy is really quite simple. What we're going to do is think about slicing that volume by taking vertical lines through vertical planes, like this, slicing it up into pieces, and then adding up those slices. Each of the slices is going to be like a hockey puck, you see, because it's going to be a disk, and then slightly thickened up because we're adding up little pieces of hockey pucks, going from one end to the other, to add up to the volume of the entire object. So, here we go. Let's think about it in the following way:

For each value x between 0 and 2, we think about taking a slice; and that slice is going to have radius equal to the distance to the x-axis up to the graph of our function, which, in this case, is x^2. So, the radius is x^2. The radius is x^2. So, when we revolve this line segment around, we get a disk, and the radius of the disk is x^2, and we think of thickening it up by a Δx thickness so that when we add up these disk, disk, disk, disk, disk, or hockey puck, hockey puck, hockey puck, hockey puck, we would get an approximation to the value of the volume of the entire solid of revolution. Since we have a formula for the area of each of these slices—namely, the area of a circle is π times the radius squared, so that's $\pi \times x^2$—that gives us a value of the area of the face of this slice, and then we thicken it up by Δx to think about getting that little increment of volume, and we add those up to, as x varies between 0 and 2, to get the total volume of the solid of revolution.

So, in other words, as we're thinking about this object, we should be able to take this integral and parse it, in the sense of saying what does each item in this integral mean? What does it mean? Well, the πx^2 refers to an area, an area of a circle. What circle? The circle that we get by looking at the point x between 0 and 2, and revolving a radius x^2 around the x-axis; that is the area, and then we thicken it up by dx, that's the Δx becomes dx in the limit; and remember that the integral is a process of summing up those little pieces as the Δx gets smaller and smaller, and the limit goes to 0.

Well, this $\pi\ x^2$ is, of course, just πx^4, and we know how to take an integral. Remember, we use the Fundamental Theorem of Calculus. So, the Fundamental Theorem of Calculus says we find some function whose derivative is πx^4. Well, that function is $\pi x^5/5$, because, remember, to take the derivative of $x^5/5$, we bring down the 5 and reduce the exponent; the 5's cancel with the over 5, and we just get πx^4. So, this is an antiderivative of this function; we plug in the upper limit of integration 2, and subtract the lower integration of 0—but, of course, we get 0 when we plug in 0, so our final answer is just $\pi\ 2^5/5$, which is $32/5 \times \pi$. And that is a way of using an integral to find this volume.

Let's look at the same object and compute its volume in a different way. By the way, I just love these things. I have to tell you, the idea of looking at integrals as adding up little pieces, and looking at a geometric object, and then seeing how you can divide it up into little pieces who sum up to the whole volume, and then adding them up with an integral—I think it just is very neat. So, let's do the same object once again. Here is the object—we have the same object, we have x^2, x going between 0 and 2, it sort of lays up like a trumpet horn, spinning it around; well, now we're going to think about dividing it up into pieces in a different way. Instead of slicing this way, we're going to take a bunch of concentric cylinders that I'm going to slice through in this way. A thin one that would go all the way from the front to the back, and then, as I expand it out, it captures a part of this solid that goes from whatever the radius is all the way to the back. Let me show it to you in the graphics, which I think will be clearer.

In this case, we're imagining taking our solid of revolution and dividing it up into cylinders that look like the boundary of a tuna can, you see? For every y between 0 and 4, we're going to say, "What would happen if I took the cylinder, this circle of radius y around the x-axis and I sliced through our object?" Well, if I slice through the object, I know what the radius is of that cylinder—it's y; and at this end it hits the solid at the point $(2,y)$ because it's at height y and the end of our solid ends at $x = 2$, and then "Where does it enter—if I'm thinking, moving in this direction—where does it enter that curve?" Well, it enters the curve at the x value, where the y value is y. So, in other words, what number squared equals y? Well, $\sqrt{y}$. So, if I see myself slicing this object by this cylindrical method, I see that I'm going to get a cylinder whose radius is y, and whose height is the difference between 2 and $\sqrt{y}$. These things are going to be added up—see, here's another part here for a different choice of y, and as y varies between 0 and 4, I can fill up my solid by these concentric rings.

Since I can write down an expression for the area of each of those boundaries of a tuna can, namely it's $2\pi\, y$, that's the circumference of a circle of radius y; times the height of that tuna can, which is $2 - \sqrt{y}$; I thicken it up by dy; and add those incremental pieces of volume up as y varies between 0 and 4, and this is an integral expression for that volume. Once again, I can actually compute the integral, using the Fundamental Theorem of Calculus, and, of course, I get the same answer, $\pi \times 32/5$, because it's the same volume that we're computing. I think this is a rather neat example of using the concept of integral, which is adding up pieces, to get an entire object.

Now, I wanted to tell you about another solid of revolution that has some interesting properties, and this is a mathematical anomaly. When we go to infinity—if we take an object that goes to infinity—we can create an object that has a very interesting property, and the interesting property is that it has finite volume, but it has infinite surface area. So, this is something that, by the way, doesn't occur in real life because we're talking about infinity here, but it actually is an object that we can construct and evaluate, and here it is. We take the function $f(x) = 1/x$. Now, you can see that such a function, $1/x$, as you choose large numbers for x, it becomes closer and closer to the x-axis. Let's start at $x = 1$ and go all the way to infinity, and just spin that area underneath the curve $f(x) = 1/x$ around the x-axis and look at the volume enclosed by that object. Well, we know how to do it; we'll just go back to our method of slicing vertically, slicing into disks, and we add up those disks.

Well, we can do that. It's the integral of 1—as x varies from 1 to infinity, of $\pi \times 1/x^2$, because at every point x the radius is $1/x$, so $\pi(1/x)^2$ is the area of a circle, or radius $1/x$; we thicken it up by dx; and add it up from x going from 1 all the way to infinity.

Well, we just take an antiderivative—by the way, an antiderivative of $1/x^2$ is $-1/x$, so $\pi \times -1/x$ is the antiderivative of $\pi \times 1/x^2$; we plug in—well, plugging in infinity doesn't really work, we just take a number k, as k gets larger, and let k go to infinity—this is a little technicality, but it doesn't matter anyway because it becomes 0. The point is, when we do this integration, the value we get is just π, meaning that the total volume contained inside that infinite horn—if you think of it, I'll do it in the direction facing you—it's an infinite horn going out in this direction, like this, and just think of it as a trumpet end at this end and getting tinier and tinier. The total volume is finite; it's just the number π—π cubic units.

Let's now look at the surface area of this. The surface area is more complicated, so instead we're going to look at the surface area of something smaller, and the way we're going to think of the surface area of something smaller is the following: We'll just say, look at the region in this sort of infinite trumpet horn between $x = 1$ and $x = 2$. If we, instead, just look at this straight cylinder, whose radius is 1/2—that's the value, when $x = =2$, the radius is 1/2—if we just go straight across rather than going up in the way the actual horn-shaped object goes, we'll get an area that is less than the part of the horn shape between $x = 1$ and $x = 2$. How big is the cylinder? This cylinder has radius 1/2; we multiply by 2π to get the circumference of a circle of radius 1/2, the radius $\times\ 2\pi$ is the circumference of the circle; and then we multiply by the height of the cylinder—well, the height is just equal to 1 for this first cylinder. So, the surface area for this piece is equal to π square units. That's the area of this cylinder.

Now, let's go to another cylinder that's twice as tall—that is, it goes between $x = 2$ and $x = 4$; and it's inside this trumpet-shaped object again, because it starts here at $x = 4$, the radius is 1/4. Since the radius is 1/4, the circumference of a circle of radius 1/4 is $1/4 \times 2\pi$; but this cylinder has height 2, because it goes from $x = 2$ to $x = 4$. So, 2, the height, $\times\ 1/4 \times 2\pi$, the circumference, is the surface area of this cylinder. Once again, if we look at these things, it's just 2 × 2 is 4/ 4 = 1—this area is π . So, this cylinder is, again, π ; and that cylinder is, of course, it has smaller surface area than this horn-shaped object between 2 and 4, because the horn-shaped object goes up and is bigger.

Now we're going to repeat the process, but this time between $x = 4$ and $x = 8$, because, once again, when we have a cylinder of radius 1/8, its circumference is: 1/8—that's the radius—× 2π—that's the way we get circumference; and then when we multiply by 4, which is the distance between $x = 4$ and $x = 8$, we get the area of this long cylinder; but the area of this is, once again, π —4 × 1/8 × 2 × π is just π . The next cylinder will go from 8 to 16, and so on and so on and so on. You see that the areas of these cylinders are just π + π + π + π—forever. So, the area of that object is infinity. We have gotten a solid of revolution, and if we put it up this way, we would have a paint can where we could pour the paint in and fill it with three units of paint, and yet the paint would not be sufficient to paint the walls. Now, that is an interesting object. So, if you're going to open a paint store in the world of infinity, you've got to be a little bit careful what walls you're going to paint. Okay, so this is the idea of using integrals to find an interesting object.

Now what I'd like to do is return to the question of how do we extend the concept of integration to functions of two variables? Remember, functions of two variables are like when we have a map, and at each point in the map we talk about the altitude of the point at that point. So, if we have a terrain, at each point we see how high that mountain is at that point, that is an example of a function of two variables, and that's what we talked about last time.

Right now, I want to say, "How are we going to deal with the concept of integration in this domain of a function of two variables?" Well, the idea, of course, is to take an analogy to a function of one variable. Remember that when we took an integral of a function of one variable, what we noticed was that integral was equal to the area underneath that curve and above the x-axis. By analogy, we want an integral associated with the function of two variables to give us the volume underneath the graph of a function of two variables. So, in other words, if we think of it this way, that we have a terrain, and there are mounds of coal on this terrain, and we want to know: What is the volume of coal that lies above the base level and underneath our surface?

That's the kind of question we would want to investigate when we're talking about taking an integral of a function of two variables.

What is the strategy for taking such an integral? Here we have a model of a surface that could be viewed as a graph of a function of two variables. So, you see, we're imagining that we have some part of the plane, which is the domain, and for every point there we have some value, how high the graph is; and we get this surface, which is the graph of this function of two variables. Our strategy for finding the volume underneath such a surface is actually quite a simple one. The idea is that we take our volume—this is a solid—and we're saying to ourselves, "Well, I could chop up this solid into slices, and these slices go parallel to one of the axis." For example, if this is the x-axis and this is the y-axis—and I'm describing them that way so you can see this graph more clearly—if this is the x-axis and this is the y-axis, it's not normal to have the y-axis going away from you, but for purposes of illustration we'll think of it that way. So, when we have this solid underneath this graph, one way to think about computing what the volume of that solid is is to simply break up our solid into slices, and then for each slice, if we can figure out the volume of each slice, we could add them together using an integral.

What is each slice? Here, I'm removing one slice. So, this is a slice at some particular value of y; you see, if I'm imagining different values of y, I take a particular value of y and I'm taking a slice out of it. This slice—our goal is to find out what the volume is of this slice. Well, we're faced with a problem of finding a volume of something that looks like a function of the variable x. We know how to find the area under the graph of the function of variable x, that's an integral. So, in order to find the value of an integral underneath a surface of a graph of a function of two variables, what we're going to do is break up that solid that we're trying to compute the volume of; we're trying to break it up into slices, and in order to find out what the volume is of each slice, we're going to use an integral again. So, it's actually a double integral. Let's actually do an example and then I think you'll see how it works.

Let's consider the specific function of two variables x^2y. Now, this is a function of two variables; what it means is that for every point in the plane, for every point x and y—for example, if $x = 1$ and $y = 2$, then we take 1^2, which is 1 × 2, which is 2, and that's the value of that function at the point (1,2). So, the function $f(x)y = x^2y$ gives us, if we graphed it above the plane, it would give us a surface, because every point has a certain height; so it gives us like the terrain of a mountain.

Suppose we want to know what is the volume contained underneath the graph of that function of two variables over a particular rectangle in the plane. Let's take a rectangle where the x values go between 0 and 2, and the y values go between 0 and 3. So, we have this rectangle in the plane, and then every point in the rectangle has a value, so we have this surface. If we want to imagine this to be our rectangle, then at every point we have a value of the surface, so we have the surface over the rectangle. Our challenge is: Can we find the value of the volume underneath that graph? And we're going to use this method of chopping, chopping into these pieces. So, how do we chop? Well, we say for every number y between 0 and 3, we're going to say what is the amount of area that we get when we slice that solid at the point y? We slice it, and we get an area, and then we thicken it up by dy to get a little piece of volume. Then we add those up as y varies between, namely 0 and 3. So, here we go; we're saying for each y between 0 and 3, so the way we read this kind of double integral is that we're saying, first of all, this double integral here is just asking the question for the function of two variables, x^2y × $dxdy$ over the region r from 0 to 3—x varying between 0 and 2; y varying between 0 and 3—this interpretation of this whole double integral is what is the volume underneath that graph over this region?

Now we're saying that this is equal to chopping it up into slices. That is, for every y between 0 and t3, we want to know what the area is of that slice, at the point y. Well, at the point y, that slice looks like a parabola. If we chop it up right there at that x, we see that the x function looks like a parabola. It's x^2y; and x is varying between 0 and 2. So, to find the area of the slice at that point, we're faced with an integration problem; but it's an integration problem we know how to do. We know how to find the area under a parabola, and particularly under the parabola x^2y. By the way, we think of y now as just the constant, because we're saying at the particular point y, at this particular point y, we're saying what is the area under that part of the curve. In other words, under this red line here, as x varies between 0 and 2, what is the area under that? Well, it's just the integral from as x varies between 0 and 2 of x^2 × the constant y.

This value here, this integral inside, when you think of y as the constant and x as the variable, is something that we've done before. It's just an integral where we can compute the integral by taking the antiderivative of x^2y—this is with respect to x; we take some function of x whose derivative is x^2y. Well, an example of such a function is $x^3/3 \times y$, because if we take the derivative with respect to x of $x^3/3 \times y$, we bring down the 3, the 3's cancel, we get $x^2 \times y$, and that is equal to x^2y, so this is an antiderivative. We evaluate it as x goes from 2 to 0. In other words, when $x = 0$, it's 0, so we just plug in the two and we get $2^3/3y$. That is, for each slice y, we know exactly what the area is there. When we thicken it up by *dy,* we then have we're faced with a problem of finding an integral with respect to y. In other words, we think of taking all the different slices; for each slice we know what the area is thickened up; adding them up will give us the volume. We can actually compute this because we just have this $8/3 \times y$ is easy to take the integral of; we find some function now of y—because it's y that's the variable – some function of y whose derivative is equal to $8/3 \times y$, and that is $y^2/2$; we plug in the 3, by the Fundamental Theorem of Calculus, and we get that the volume is exactly 12.

Alternatively, we could do exactly the same thing by slicing the other way; by slicing first with respect to x, and then for every x, taking an integral with respect to y. This slide here demonstrates that doing the—this is called an iterated integral—first with respect to y this time, as y varies between 0 and 3 because we're slicing at every fixed point x, we can see that we, in fact, get the same answer again. You can slice this way or you can slice this way.

Well, the moral of the story is that with all this slicing going on, we can see that we could actually find the volume of a loaf of bread in the same way. And, in fact, we often do. When we eat a loaf of bread, because we say what is the volume of this loaf of bread? Well, we take each individual slice, we see what its volume is, and we generally eat it. Adding them all up, we see how much the total volume of the bread is. So, we can conclude that calculus is the best thing since sliced bread.

I'll see you next time.

The Fundamental Theorem Extended

Lecture 21

Now we've extended the idea of derivative and the idea of integral to apply to more complicated situations, so let's now extend the Fundamental Theorem of Calculus to apply to our extended notions of derivatives and integrals, and talking about curved paths and multiple functions of several variables.

You will recall that after we introduced the derivative and integral, our next step was to describe their relationship via the Fundamental Theorem of Calculus. In this lecture, we'll talk about extensions of the Fundamental Theorem of Calculus that apply to our extended notions of derivatives and integrals in curved paths and several dimensions. In particular, we will study the *Fundamental Theorem for Line Integrals*, and we will look into the background of George Green, the brilliant but largely self-taught man who developed the theorem that bears his name.

Recall that the Fundamental Theorem of Calculus connects the derivative with the integral. Basically, it observes that the accumulation of small changes (the integral) is equal to the net effect. We first saw the Fundamental Theorem of Calculus illustrated in the motion of a car. Suppose a car is moving on a straight road and its position at each time t is given by $p(t)$. We noticed that there are two ways of finding the net distance the car traveled between time a and time b. On the one hand, the distance is $p(b) - p(a)$. On the other hand, that distance equals the integral of the velocity. And the velocity is the derivative of the position function. Thus, the Fundamental Theorem of Calculus states that $\int_a^b v(t)dt = \int_a^b p'(t)dt = p(b) - p(a)$. In general then, the Fundamental Theorem of Calculus states that for any differentiable function $F(x)$, $\int_a^b F'(x)dx = F(b) - F(a)$. The Fundamental Theorem of Calculus can be rephrased as: "Accumulated change equals net effect."

With regard to a mountain slope, we considered functions of two variables. That is, for each point in the plane, we associate a number. Here, we examine what variations on the Fundamental Theorem might apply to this setting. To ground our discussion in a real situation, suppose we have a map that tells us the height of the terrain. At each point on the map, we could associate the altitude or elevation at that point, such that the map acts as a function of two variables, the x and y coordinates of that point on the map. Surveyors used to start at a point at sea level, then go a short distance and measure the rise or fall at that point, then go another short distance and repeat the measurement, and so forth. This method is suggestive of an integral because it accumulates the change in altitude. Consider the following scenario. We start at one point and draw some circuitous path on the map.

As always, the Fundamental Theorem compares the accumulation of change to a net effect.

There are two ways to compute the net change in altitude from start to finish. If we could measure our altitude at the beginning and the end, then of course, we could simply subtract. Another method would be to see how steep it is at every step, then compute how much we rise or fall with each step on the map and add up those changes to get the total change. This method of adding up the increases and decreases in altitude is reminiscent of the method that surveyors used to measure the altitude of mountains and the whole terrain. We can formulate this insight to give us a variation on the Fundamental Theorem of Calculus.

Suppose we have a function of two variables $f(x, y)$. Think about the function that tells the altitude at each point on a map. Now we draw a path on the plane or the map. The value at the end minus the value at the beginning (the net change in altitude) is equal to the integral taken along the path of the incline at each point. The incline at each point just means the steepness at each point in the direction that we are walking. The partial derivative with respect to x of $f(x, y)$ at a particular point on the map tells us how steep the terrain is if we took a step in the positive x direction. Likewise, the partial derivative with respect to y of $f(x, y)$ at a particular point on the map tells us how steep the terrain is if we took a step in the positive y direction. If we take a step diagonally, c units in the x direction and d units in the y direction, we can estimate how high we rise by multiplying the partial derivative with respect to x times c units and adding it to the partial derivative with respect to y times d units. In other words, if we take that incline or slope and multiply by the length of each step on the map, we are really taking an integral along the path—the integral of the function that tells us the rate of change of the altitude at each point, that is, the slope or steepness at each point. That integral is called a *line integral* because we are doing our familiar process of adding function value times small length, but we're doing it along a path or a potentially curvy line rather than just along the x-axis.

We can phrase a variation of the Fundamental Theorem of Calculus in this setting by saying that the line integral of the function that tells the steepness at each point along a path from one time to another equals the difference of the altitude at the end minus the altitude at the beginning of the path. Notice that taking this line integral along *any* path between two points will result in the same value.

The technical notation for the Fundamental Theorem for Line Integrals

$$\int_C \nabla f \cdot dp = f(p(b)) - f(p(a)).$$

Calculus is used to model and describe many physical phenomena. One example is the behavior of fluids and gases. For instance, how would we describe the situation of the wind blowing?We can measure its velocity and direction at each point. This gives a velocity **vector field** at each moment in time. Now let's think about a region of space bounded by a surface. For example, we can think about space bounded by a sphere. There is a mathematical statement that is a variation of the Fundamental Theorem of Calculus that helps us understand this situation. This version of the Fundamental Theorem of Calculus is called the *Divergence Theorem*. As always, the Fundamental Theorem compares the accumulation of change to a net effect. In this case, we need to describe what change we are talking about and what net effect we are talking about. The change refers to a local analysis at each point in the region of space that describes the rate at which the gas is expanding or contracting at that point. The net effect refers to the volume of gas that is crossing the containing surface per unit time.

Two other variations of the Fundamental Theorem, *Green's Theorem* and *Stokes' Theorem*, have critical applications in fluid dynamics, electricity, and magnetism. Both measure different properties accumulated over a surface compared to going around the bounding curve. Green's Theorem is named after its discoverer, **George Green** (1793–1841). George Green, who had little formal education, worked in his father's bakery for most of his life and taught himself mathematics from books. He proved his famous theorem in a privately published book that he wrote to describe electricity and magnetism. He did not attend college until he was 40. He had seven children, all with the same woman, Jane Smith, but never married. This factor became important when he secured a position in college that required that he be unmarried! He never knew the importance of his work, but it and its consequences have been described as "leading to the mathematical theories of electricity underlying 20th-century industry."

The Fundamental Theorem for Line Integrals, the Divergence Theorem (also known as **Gauss**'s Theorem), Green's Theorem, and Stokes' Theorem are all variations of the Fundamental Theorem of Calculus. Each of these variations can be predicted to be true by understanding the philosophy of the Fundamental Theorem of Calculus in its most basic form: "Accumulated change equals net effect." People are adept at generalizing and extending ideas. Those generalizations and extensions are important. In these examples, those generalizations are central to the way we understand and describe various physical phenomena, from fluid flow to electricity and magnetism. ■

Names to Know

Gauss, Carl Friedrich (1777–1855). German mathematician; commonly considered the world's greatest mathematician, hence known as the Prince of Mathematicians. He was professor of astronomy and director of the observatory at Göttingen. Gauss provided the first complete proof of the Fundamental Theorem of Algebra and made substantial contributions to geometry, algebra, number theory, and applied mathematics. He established mathematical rigor as the standard of proof. His work on the differential geometry of curved surfaces formed an essential base for Einstein's general theory of relativity.

Green, George (1793–1841). Most famous for his theorem known as *Green's Theorem*. He worked in his father's bakery for most of his life and taught himself mathematics from books. He proved his famous theorem in a privately published book that he wrote to describe electricity and magnetism. Green did not attend college until he was 40. He had seven children (all with the same woman, Jane Smith) but never married. He never knew the importance of his work, but it and its consequences have been described as "…leading to the mathematical theories of electricity underlying 20^{th}-century industry."

Important Term

vector field: A field of arrows associating a vector to each point (x,y) in the two-dimensional plane; usually represented graphically.

Suggested Reading

Any standard vector calculus book.

Schey, H. M. *Div, Grad, Curl, and All That: An Informal Text on Vector Calculus*.

Questions to Consider

1. Explain how surveying, together with global positioning technology, describes the meaning of the Fundamental Theorem for Line Integrals.

2. Suppose a function of three variables described the exact temperature at every point in a large volume of space. What are two ways to measure the difference in temperature from the starting point to the ending point of a path that a flying object might take in this volume? State a version of the Fundamental Theorem for Line Integrals for paths in space.

The Fundamental Theorem Extended

Lecture 21—Transcript

Welcome back. Let's take a few minutes to step back and look at what we've done so far in the whole course. We started the whole course and introduced calculus by looking at a car, just a car moving down a straight road, and at every moment of time the car was in some position along that straight road. You may remember, we had this car with a mileage marker, and every mileage marker was just a single number, so we had the situation where one variable, namely the time, was producing one number, the position. So, that was the most fundamental object that we were thinking about when we introduced the idea of derivative. Notice that if we think of it in terms of function terminology, the position function of the car moving on this straight road is what you call a real valued function of one real variable because each time, which is one number, gives one number, namely the position of the car. So, we had this idea at each moment the car was some place, and then we used that idea to develop the ideas of the derivative and the integral. So, we thought of a derivative and an integral for these single-valued functions of a single variable.

Then we developed those ideas somewhat, and we talked about their interpretation and this way and that, but then we pointed out that those ideas, the ideas of the derivative and the integral were so rich that we could extend them to deal with other kinds of functions and other kinds of situations. In the last several lectures, what we did is we introduced extensions of the ideas of the derivative and extensions of the idea of the integral to these different settings.

For example, we talked about motion in the plane; remember when we talked about a car moving on the road, and we spoke about a fly flying in the air, and how we could deal with that kind of a situation. So, in that situation, at every moment, if we knew where the car was as it moves in a plane, then, you see, every moment, that is a time—that's one variable, the time—was producing, in the case of the car on the plane, it was producing two numbers, namely the horizontal position of the car and also the vertical position of the car as it moved around the plane. So, it was a function where we had one variable, time, producing two numbers, the horizontal position and vertical position. Then we saw how we could develop ideas, or extensions, of the derivative and integral to describe that kind of motion, and also to compute value, such as the distance that we travel along a curved path.

Then we went on. We developed other extensions of the derivative and the integral to analyze different kinds of situations, where more than one variable was contributing to one dependent value. So, for example, in the lecture talking about mountain slopes, we said, where we have two different values—namely, you can think of it as two coordinates on a map, maybe the latitude and the longitude—two values giving one value, namely the elevation of the point on the map. So, there's an example of a function of several variables; two values giving one value.

Another example of a function of several variables was we talked about the area of a rectangle. The area of a rectangle is determined by two numbers, the length and the width. In that setting, we talked about the extensions of the derivative, namely we talked about partial derivatives. Remember, partial derivatives were telling us, when we had this function of two variables, it told us about the slope on this terrain, and we'll talk more about that today.

But then we also developed, in the last lecture, the concept of multiple integrals, double integrals that were telling us the volume under this surface, and these were all extensions of the concepts of the derivative and the integral. But, now let's go back and remember what we did right after we introduced the derivative and the integral originally. The very next step is that we described their relationship; and their relationship was described as using the Fundamental Theorem of Calculus.

So, now what we're going to do is we're going to say we took derivatives and integrals, and we saw the relationship via the Fundamental Theorem of Calculus long ago. Now we've extended the idea of derivative and the idea of integral to apply to more complicated situations, so let's now extend the Fundamental Theorem of Calculus to apply to our extended notions of derivatives and integrals, and talking about curved paths and multiple functions of several variables.

What we're going to do in the first example, it is going to involve hiking up a steep mountain slope. But, before we do that, let's go back and really recall the Fundamental Theorem of Calculus so that we have that down pat. Remember in the very first lectures we talked about a position function, $p(t)$, and its derivative was the velocity function, $v(t)$, and we saw that the distance traveled between one time a and another time b would be given, first of all, as the integral of the velocity function. And that was the way we defined the integral, because that's what it meant, adding up these little velocity times increments of time, adding them up gave us the distance traveled.

But then we also saw if we could find a function whose derivative was that velocity function, in other words, a position function that would generate that velocity function; then there was a simpler way to find the distance traveled, namely just plug in where we are, that is, what our position is at time b, and subtract our position at time a, and that difference would be the same thing, would give the net change that we accomplished, that we could also deduce by adding up all of the little distances that we traveled using the integral. So, this was the Fundamental Theorem of Calculus. Of course, we could extend it; it doesn't have to be a car moving on a straight road, it could be any function. So, if we had any function $F(x)$ whose derivative is $f(x)$, then we saw that the integral from a to b of $f(x)$, that is the derivative of $F(x)$ is just equal to the function F at b minus the function F at a.

I'd like to sort of have us remember that the Fundamental Theorem of Calculus can be phrased, sort of an English phrase for it, that accumulated change equals net effect. The integral is the accumulation of the change, the derivative—you see derivative is talking about rate of change. So, the accumulation of change can be reinterpreted as the net effect. That was the Fundamental Theorem of Calculus.

What we're going to do now is look at an extension of the Fundamental Theorem of Calculus, where we're talking about a somewhat more complicated situation. Instead of thinking about just a car moving on a straight road, what we're going to do is imagine a mountain slope, so here is the situation: We imagine that we have this mountain slope, and when we're thinking about mountains, we should be thinking about functions of two variables because every point on our map—our map is just a flat piece of paper—for every point on the map, and in order to locate a point on a map, you say, ”Where is the point? How far to the right is it? How far up is that point?” And at that point our map could tell us what the elevation is of the terrain at that point. So, a map that tells us the height of the terrain is actually a function of two variables. That is, it gives one number, the elevation, for every two numbers we give it, namely the x- and y-coordinates of that point on the map. By the way, you've seen these topographical maps; here's an example of a topographical map. It tells us the height at each point—that is, the elevation at each point—above a particular point on the terrain.

Let's think about the challenge of finding the elevation at different points on a mountain. If you think about it, back in the old days—the old days being the days before these GPS systems, Global Positioning Systems; GPS systems now can find the elevation just by taking your little hand-held GPS thing and satellites look at it and it will tell you exactly what the elevation is at that point—but back in the days before that technology was available, how were you going to find the altitude of a mountain, or any other place? Well, this is the way you do it—you do surveying. The way the surveying works is the following: You start someplace, at sea level, and then you go a short distance and you very carefully measure whether in that short distance that you were rising, and how much you were rising; or falling, how much you were falling; and that gives you the altitude at that next place. Then you take a few more steps and you see how far you've risen or fallen in that next step; and then you take the next step; and you take the next step. In that fashion, surveyors went from step to step to step to find the elevation at all of the different points on the terrain.

This is a perfectly good way to find elevation, and it is a way that is suggestive of an integral. That's what we're going to get to in a few minutes, that that method is an integral, because it's accumulating the change, the change in altitude, as you walk along a path. But let's think of it, and there are several different ways of finding the difference in altitude, or elevation, from one point to another if we're walking along on a path. The two different ways are the following:

The first way is very simple. Suppose that we actually had our topographical map, and we could just point to the point on the map and say, "Oh well, it tells me that the place I'm going to has this elevation, and the place I am now has this other elevation, so I just subtract the two to see what the change in elevation is between those two points." So, that would be a very simple way to do it if we already were told what the elevation is at each point on the map.

But the other possible way to find how much change there would be in elevation if you took a walk would be at each step to say how steep is it during that part of the walk, and then compute how much we have risen or fallen, and then go to the next step and see how much we have risen or fallen, and so on, until we add up all these rises and falls as we walk along a path, and when we get to the end, we've accumulated the change in our altitude, and we could say, "Well, we've changed this amount in altitude."

Those are two different ways, and notice that these two different strategies are very reminiscent of the Fundamental Theorem of Calculus; that is, the two different strategies of the Fundamental Theorem of Calculus were: 1) If you know where you were at then end, where you were at the beginning, just subtracting the two is one way to tell the net distance you've traveled. Another way is to say if you knew what your velocity was at each moment; you could add up those incremental velocity times times to see how far you've traveled. It's exactly the same philosophy, and all of these extensions of the Fundamental Theorem, you'll see, are all based on the same underlying concept that accumulated change will give you the same answer as the net effect.

So, let's actually do an example here. Let's consider a function that's telling us the altitude, or the elevation, of a mountain terrain over a map. That's a function of two variables, we'll call it $f(x,y)$, and remember that the partial derivative of $f(x,y)$ at a particular point on the map, think of a point on a map, it is telling us how steep the terrain is if we took a step in the x-direction. In other words, suppose that I'm standing on this terrain, and I'm looking in the direction where the map is pointing, in the x-direction, and at this particular point, if I take a step, how steep is the terrain at that point? That is what is measured by the partial derivative of f with respect to x. It's a simple idea; it's just, if you go in this direction, how steep is it?

Likewise, the partial derivative of f with respect to y at a point is just how steep is it if you go in the positive y-direction. So, imagine that you're standing on that same hill, and one direction it's going uphill, the direction in the opposite, right-angle direction may be downhill, it may also be uphill, and that's what the partial derivative with respect to *y* is.

But, now let's look at an example where you step in some direction other than in the *x*-direction or the *y*-direction; and, in fact, suppose that we imagine the following thing: Suppose that this is the *x*-direction on our map, and we're imagining we're on a terrain—we're on a mountain—and remember some fact about our concept of calculus, that if we look at a function of two variables very closely, it will look like a flat plane. That's one of the concepts of calculus—if we look at things closely, they look like a straight line or a flat plane. So, if we look at just a small part of our graph of our function of two variables, the altitude, or if we imagine that we're on mountain, we're looking at a part that's so small that it's like a flat plane. Now, suppose that we go a certain distance in the *x*-direction, and a certain distance in the *y*-direction; in other words, we're really going in some diagonal kind of direction, and I want to ask the question: How high would we rise? How much would we go up in altitude if we take a path that goes at a diagonal direction by going a certain distance, *c* units in the *x*-direction and *d* units in the *y*-direction. In other words, I'm going to go *c* units in this direction, *d* units in the *y*-direction, and resolve those to a—actually it's to a vector that is going—it points in some diagonal direction, and I'm going to ask the question: How high would we rise if we're starting at a particular point?

Well, the partial derivative with respect to *x* tells us the slope in the direction of the positive *x*-axis, and so if we're going *c* units in this direction along the map, it's *c* units along the map, and we know that the partial derivative with respect to *x* is saying how much change there is with respect to a unit change, then the product of *c*—the number of units in that direction—times the partial derivative of *f* with respect to *x* at the point (x,y) will tell us how far we've risen if we just went along this way.

Likewise, if we just go along the *y* direction, and we go distance *d* along the *y* direction, then that distance *d*, times the slope—which is the partial derivative of *f* with respect to *y*—will tell us how far we've risen here. Now, if we go in this diagonal direction that goes *c* units this way and *d* units this way, then the slope will add this amount of increase from here, and then we go from here to here, which will add the *d* times the partial of *f* with respect to *y* additional increase to get to here, so we have the little formula here that the amount you will ascend if you go (*c*,*d*) distance—that is, *c* units in the *x*-direction, *d* distance in the *y*-direction, from a point (*x*,*y*), you can compute this by taking the partial derivative with respect to $x \times c$ plus the partial derivative with respect to $y \times d$. Just adding them together gives you how far you've ascended.

What that tells us is an interesting fact that really is another version of the Fundamental Theorem of Calculus, and it is called the Fundamental Theorem for Line Integrals. Here it is, and let's interpret every symbol in this equation. Suppose that you take a path on the map, you take a path from one point and you go along that path to another point on the map, and you want to know how far you have risen during that path traversal, from the beginning point to the end point. Here's what you do: If the path is given by *u*(*t*), *v*(*t*)—in other words, at each point, each time *t*, we say where we are in this direction, that's *u*(*t*) here, *v*(*t*) here, it's a point on the map, so we know where we are on the map, and then, of course, the *f* function is giving us the altitude. We want to know what the change in altitude is if we traverse a path. We say that the change in the altitude is going to be the rate at which we're going in the *x*-direction at each time $t \times$ the partial of *f* with respect to *x*, that's going to say how much we rise in the *x*-direction, plus the partial of *f* with respect to $y \times$ the $v \times t$, *dt*, it's going to tell us how much we rise in the *y*-direction for each little time Δt. So, let me say that again: The integral from *a* to *b* of the partial of *f* with respect to $x \times u'(t) +$ the partial of *f* with respect to $y \times v'(t)$ *dt*, the way to interpret it is to move the *dt* into each—sort of distribute it—and realize that $u'(t)$, *dt*, $u'(t)$ is the rate at which you're changing on the map your *x*-coordinate. So, at times *dt* is going to give you the distance on the map that you're going in that small increment of time, and when you multiply by the partial of *f* with respect to *x* at the point that you are at that moment, it's going to tell you how much you've risen in that direction. Then you add how much you're risen in the *y*-direction to get the total amount you've risen when you go in the direction of the path.

If you sum up—which is what the integral does, of course, it's just summing up—if you sum up those values, you will get the same thing as taking the function value, that's the elevation at your terminal point, *p*(*b*) minus the function value at your beginning point, $p(a)$, regardless of what path you take to get from a to b, of course. So, this is the Fundamental Theorem for Line Integrals.

Let me go ahead and do a specific example, but we won't be able to take the time to do it in great detail, but suppose we have a function, $f(x,y)$ is this complicated function that tells us the altitude at every point, and it's actually a nice function, because if we want to climb to the peak of a mountain, which is, of course our goal, then the way that we would do it is we would take the partial derivatives with respect to x and with respect to y and set each of them equal to 0. Remember, that's the way we know we are at a peak, or a valley, but would be a place where when we go in the x-direction we don't rise or fall and when we go in the y-direction we don't rise or fall, we're at the peak; we're at the summit. Well, if we have written an equation in such a way that when we take the partial derivatives and set them to 0, we see that it occurs at the point (6,9). Now, I'm imagining a map where, therefore, the map is 6 miles over and 9 miles this way, that at that point is where the summit of the mountain is. You can see by plugging in these values that we've chosen it to have a summit that's 27,400 feet high—which would make it, by the way, the sixth highest mountain in the world. If we had a summit of that height, it would be the sixth highest—we might call it Mount Teaching Company, I'm not sure.

Here's the question: suppose we start at the point (0,0), which, incidentally, if we look back at the equation, has height 4,000 feet; so here it is, height 4,000 feet, and we know that if we took a path from the point (0,0) to the point (6,9), we would have risen a total of—the height at the peak of the mountain, 27,400, minus our original height, we would have risen 23,400 feet. Well, there are different paths that we could take to get there, and any path should give us the same answer.

Let's take two paths. One path is if we went just straight along the x-axis from 0 to 6, and then we went straight in the y-direction from 0 to 9. If we do that, in the first path, since we're just going in the x-direction, really it's just an integral with respect to x; and when we go in the -y-direction, it's just an integral with respect to y; adding them up we see that we do, in fact, get the distance, which is 23,400 change in altitude.

If we take this complicated path that we looked at from a previous lecture—remember when we took a path that went where $u(t)$ is $t^3/3 - t$ and $v(t)$ was t^3, went from (0,0) to (6,9) in this sort of curving path, we can actually write down the integrals that accumulate the steepness at every point in this diagonal direction that we're going on the curved path, and add them up and, in fact, this whole thing, although it looks a little bit complicated, it is really just following what we've discussed, and, sure enough, we get exactly the same answer again.

So, this is the fact that any path from the original point to the end point, if we see how much the accumulated ascents and descents add up to be, will be the same thing. And that is the Fundamental Theorem for Line Integrals. It has this fancy looking symbol when it's written in the technical notation.

Calculus is used to model and describe many physical phenomena, and one example of the kinds of things we want to model is the behavior of fluids and gases. So, we can think about a question, which way is the wind blowing? You're standing here and the wind is blowing in all sorts of directions, how would we describe that situation of the wind blowing? Well, one way to do it is to say at each point, which direction and at what speed is the molecule going at each moment? That gives a picture that's called a "velocity vector field." At each point in the domain that we're talking about with the wind blowing, we say at this point the wind is going in this direction at this speed, and we can draw it by drawing arrows—at each point, we can draw an arrow whose length tells us the speed, how fast it's going; and whose direction tells us which direction it's going. So, it's called a "velocity vector field," because at each point you have these vectors, which are these arrows, which are pointing in a certain direction.

One of the extensions of the Fundamental Theorem of Calculus involves thinking of space bounded by a surface. For example, we can think of space bounded by a ball, and this variation on the Fundamental Theorem of Calculus that refers to this situation is called the Divergence Theorem, and that will help us understand the situation. As always, the Fundamental Theorem of Calculus is going to compare the accumulation of change to a net effect. In this case, we need to describe what change we're talking about and what net effect we're talking about when we're talking about the wind.

Well, the change refers to a local analysis in each point in the region of space that describes the rate at which the gas is expanding or contracting; and then the net effect refers to the volume of gas that's crossing this containing surface per unit time. That version of the Fundamental Theorem is called the Divergence Theorem or Gauss's Theorem. It's a complicated thing, but it's rather interesting.

There are many variations on the Fundamental Theorem of Calculus. Other ones include Green's Theorem and Stokes's Theorem, and these have critical applications in fluid dynamics and electricity and magnetism. I wanted to tell you just a little story about George Green, whose bakery was the—his baking background was an unusual place, I think, to think about electricity and magnetism. George Green was a person who had no formal education to speak of—I think he went to elementary school for one year, but he was very much interested in thinking about mathematics, and was, apparently, very good at it. He lived from 1793 to 1841, and while working in his father's bakery, where he worked most of his life, he taught himself mathematics from books, and he proved this very famous theorem, that's called Green's Theorem, and privately published it by himself, and he wrote it to describe electricity and magnetism. He didn't attend college until he was 40—by the way, as an aside, he had seven children, seven children all with the same woman, whose name was Jane Smith, but he never married. This was actually important because when he went to college, he got a position in the college that required him to be unmarried. So, the fact that at that time he had six children didn't seem to deter them from giving him that position.

Anyway, the work that he did, he didn't realize the importance of his work, but its consequences after it was developed really led to some of the great ideas, particularly the mathematical theories that really underlie—the mathematical theories of electricity—that really underlie the 20th-century industrial development.

The Fundamental Theorem for Line Integrals, the Divergence Theorem—which is also, by the way, known as Gauss's Theorem, Green's Theorem—and Stokes's Theorem, these are all variations of the Fundamental Theorem of Calculus; and each of these variations can be predicted to be true when we understand the philosophy of the Fundamental Theorem of Calculus in its most basic form, namely accumulated change is equal to net effect. People are very good at generalizing and extending ideas, and in this case, these generalizations and extensions are incredibly important. They are very central to our understanding of a lot of physics, including the fluid flow and electricity and magnetism.

In our next lectures, we're going to be talking about other applications and strategies for applying calculus ideas to many domains in life by studying differential equations. I'll look forward to seeing you then.

Fields of Arrows—Differential Equations
Lecture 22

Mathematics in general—and calculus in particular—are incredibly powerful tools for allowing us to understand and manipulate the world.

In this lecture, we introduce the idea of a **differential equation**, an equation that involves not only a function but also its derivatives. Differential equations govern numerous scientific, biological, and economic processes. The solution of a differential equation is not a number but a function or, formally speaking, a family of functions. Because algebraic solutions are rather difficult to obtain, we use a geometric approach to understand the behavior of the solutions. We will see these ideas applied to real life when we examine illustrative models of population growth and the behavior of springs and pendulums.

Mathematics and calculus are incredibly powerful at allowing us to understand and manipulate the world. For example, suppose we want to know how much money will be in our savings account if we deposit $1,000 today in an account that pays 5% per year compounded continuously. The rate at which the account is growing is really information about the rate of change of the function that tells us how much money is in the account at every moment. *Rate of change* means derivative. We can write down an equation that captures this information: $dm/dt = 0.05m(t)$, where m = money and t = time. We see that the function of the amount of money in our account will become increasingly steep as money accumulates. This equation is a differential equation. It talks about a function, in this case $m(t)$, and relates the function's values with the value of its derivative. Thus, the answer to a differential equation is not a number; it is a function.

Differential equations arise in applications to many areas. Newton and Leibniz wrote down the first differential equations; they came about at the very dawn of calculus. Differential equations have central applications in all areas of physics, as well as biology, economics, the social sciences, and basically any area you can think of. Let's look at a basic example of a differential equation in physics. Suppose you know from observations that a dropped object after t seconds will travel at $-32t$ feet per second. You might want to know how far it has fallen after t seconds. The information given presents us with a differential equation, namely, $dp/dt = -32t$. Our challenge is to find a function $p(t)$ whose derivative is $-32t$. We know how to find such an antiderivative: $p(t) = -16t^2 + C$. (where C = the position of the ball at time 0). When we have a differential equation, its solution often gives us a *family of functions*. Differential equations are equations involving functions *and their derivatives*, and their solutions are not numbers but, rather, *functions*.

Differential equations have central applications in all areas of physics, as well as biology, economics, the social sciences, and basically any area you can think of.

Physics is abundant with situations modeled by differential equations. For example, Newton's law, $F = ma$, is a differential equation, if we remember that acceleration is the second derivative of position. If you have an *initial value problem* for a differential equation, it means you are given the initial conditions; thus, we get one solution equation instead of a family of equations. Often when we are dealing with differential equations, we may want to model some particular behavior. In this case, we refine the model and add more conditions to make it more realistic. For example, when we drop an object in the air, we need to account for air resistance, which is proportional to the velocity. Therefore, the more accurate equation for measuring the speed of a falling body becomes:

$$m\frac{d^2 p}{dt^2} = -32m + cm\frac{dp}{dt}.$$

That is, the force, and hence acceleration, is a function of time, position, and velocity! Another physical situation that is modeled by a differential equation is that of the pendulum, if we wish to know what the angle is at each time. For a spring of mass m stretched a distance x from equilibrium, Hooke's Law asserts that the force is kx. Thus, the differential equation is

$$m\frac{d^2x}{dt^2} = -kx \, .$$

As we see, differential equations involve discovering that a certain physical process is governed by a differential equation, then finding a method to solve that equation. Many mathematicians contributed to the growth of this field in the 17th and 18th centuries. Newton and Leibniz, having invented the notion of derivatives, were the first to write down differential equations. The Bernoulli brothers, Jakob and Johann, and Johann's son, Daniel, solved many problems in mechanics and applied differential equations to complex physics problems, such as fluid dynamics. Their jealous and quarrelsome nature did not deter progress in calculus in the late 17th and early 18th centuries. Significant contributions toward algebraic solutions of differential equations were developed by **Joseph Louis Lagrange**, whose 1788 work *Méchanique Analytique* is an elegant and comprehensive treatise on Newtonian mechanics. Another important contributor to the solutions of differential equations was **Pierre Simon de Laplace**, an expert in celestial mechanics; his method of transforming differential equations into algebraic ones, the *Laplace Transform*, is a key tool in solving differential equations, even today.

We can use differential equations and **direction fields** to study population dynamics. The simplest model, attributable to British economist Thomas Malthus in 1798, is the exponential growth model that we have already seen: $dy/dt = ry$, where r is the growth rate and y is the population. The problem with Malthus's model is that it assumes an infinite supply of resources. Because resources are limited, the population cannot grow forever.

Before we continue, we need to look at a simple geometric approach to solving such a differential equation without finding a formula for the solution. We consider an entire plane viewed as having a horizontal time axis (t) and a vertical population axis (y). Although we don't know what the function is that satisfies the differential equation, we do know that the derivative of the function y is some constant r times the value of the function, and we know that the derivative is the slope of the tangent line. Thus, to obtain a direction field, at each point (t, y), we draw a little arrow that captures the slope of the tangent line. Then, if we start at a given point, that is, at some initial condition, we simply follow the arrows to trace out our function.

An alternative to Malthus's model is the *logistic model*, or *Verhulst model*. Once the population reaches the saturation level, or carrying capacity K, it requires too many resources and begins to decline. This means that for $y > K$, $dy/dt < 0$. The differential equation is: $\frac{dy}{dt} = ry\left(1 - \frac{y}{K}\right)$. Notice that for small y, we essentially have exponential growth, and for $y > K$, dy/dt is negative, as needed. When we look at the direction field, we notice that no matter what the initial condition is, as time goes on, all solutions approach the carrying capacity K.

Another alternative to Malthus's model is the *critical threshold model*. We assume that unless the population is above some critical threshold T, it cannot grow at all. For example, if you don't have enough people in the population, the population will not grow! This means that for $y < T$, $dy/dt < 0$. The differential equation is:

$$\frac{dy}{dt} = ry\left(1 - \frac{T}{y}\right).$$

We can put together the two models. The combined model satisfies both assumptions: that there is a minimum threshold for population growth and limited resources. The direction field shows that if the initial population is less than T, then as time goes on, it declines to zero. Otherwise, it approaches the carrying capacity K. We can use computers to put in as many points in the direction field as we wish to get very accurate estimations of the actual answer. A variation of this model describes the passenger pigeon population in the United States. The pigeons had a very high critical threshold T, so when they were hunted down to such small numbers in the late 1800s, the population could no longer grow, and they became extinct in 1914. Differential equations also appear in a variety of other contexts, including the spread of epidemics, hormone balance, chemical reactions, and planetary motion. ■

Names to Know

Lagrange, Joseph Louis (1736–1813). French mathematician; professor at the Royal Artillery School in Turin, at the Ecole Normale and the Ecole Polytechnique in France, and at the Berlin Academy of Sciences. He studied algebra, number theory, and differential equations and unified the theory of general mechanics. He developed classical results in the theory of infinite series and contributed to the analytical foundation of the calculus of variations.

Laplace, Pierre Simon de (1749–1827). French mathematician and astronomer; professor at the Ecole Normale and the Ecole Polytechnique. Laplace was the author of the influential *Mécanique Céleste* that summarized all contributions to the theory of gravitation (without credit to its contributors). He developed potential theory, important in the study of physics, and partial differential equations. He made great contributions to probability theory based on techniques from calculus.

Important Terms

differential equation: An equation involving a function and its derivatives; a solution of a differential equation is a family of functions.

direction field: A two-dimensional field of arrows indicating the slope of the tangent line at each given point for a curve that is a solution of a differential equation.

Suggested Reading

Any standard calculus textbook, chapters on differential equations. For population dynamics, see Boyce and DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 8th ed., Section 2.6.

Questions to Consider

1. In the Verhulst model, at what population level is the population growing fastest?

2. In a fishery, a fraction E of the population is culled every year. Assuming the Verhulst model, what would be the differential equation for this fishery? If $E > r$, what is the long-term behavior? If $E < r$, what is the long-term behavior?

Fields of Arrows—Differential Equations

Lecture 22—Transcript

Welcome back. Mathematics in general—and calculus in particular—are incredibly powerful tools for allowing us to understand and manipulate the world. Often what that means—you know, what does it mean to understand and manipulate the world—often what it means is we devise a mathematical model that reflects some aspect of the world. So, the mathematical model constructs some mathematical relationships that describe or predict some features of the world. Let's ground this with an example—let's look at an example. Suppose we want to know how much money will be in a savings account if we deposit $1,000 today and that account earns 5 percent per year compounded continuously. We want to know what happens, if we wait for a few years, how much money will be in the account? Well, such a question is inviting us to write a mathematical function of time, where the value of the function at each time would equal the amount of money in the account at that time. Let's try to figure out what's involved in thinking of what mathematical function will reflect the reality of our savings account.

When we see the advertisement for a savings account, what does it tell us and what does it not tell us? Here's what it tells us: What it tells us is the rate of interest that the account is going to earn, that's what it tells us. It's going to tell us the rate at which our money is going to grow. So, it tells us how quickly. If we have a particular amount of money in the account, it says that the rate at which it's going to grow is 5 percent per year. It doesn't tell us, for example, you don't see a big bulletin board saying put your money in this account and you'll earn $500 next year, because it depends how much money you put in the account. If you put a lot of money in the account, you're going to earn a lot of money; if you put in less money, you'll earn less. So, the information that is told to us about the account is the rate at which the account is going to grow.

Really, the information about the rate of change of a function is information about the derivative. You see? Rate of change, remember, is what is captured by derivative. So, the information that we're given about our interest rate is really information about the derivative of the function that we really want to know, namely, the function that is going to tell us how much money is going to be in our account at every time.

So, what we can write down when we see this advertisement, your account, your money, will earn 5 percent per year compounded continuously, what we can really write down is an equation that captures the information about the amount of money in our account relative to how quickly the money is going to change, the derivative of that function. So, let's go ahead and see what kind of an equation is associated with that information.

Well, 5 percent interest compounded continuously means that the derivative of the function that tells us how much money in our account is going to be .05 times the amount of money. So, if we open an account and we put in $1,000 in the account, then what that says is that the slope of the tangent line of the graph of the function that tells us how much money will be in our account at every moment, the tangent line is going to have slope 50 when there is $1,000 in the account. But, when we have $2,000 in the account, then the slope of the tangent line will still be 5 percent, but it will be 5 percent of $2,000. That is to say, it's going to be $100; 100, the slope will be 100, which refers to $100 per year.

So, if we think about how much money will be in our account over time, we can easily see that the amount of money will increase, and the slope will also increase as we get more money in the account. Because, remember, the slope is equal to the change in dollars per unit time per year. If we have more money in the account, we get more dollars of interest in that next year. Well, the equation that relates the amount of money in our account, and then its rate of change, is a differential equation; it's an equation that talks about a function. In other words, in this case, the function that we're talking about is the amount of money at each time in the account, and that's what we may not know. But, what it does is it relates that function, the amount of money in our account at each time, it relates it to something about the derivative of that same function. Then our challenge becomes to take that differential equation, that equation that is relating an unknown function with something about how it's changing, and then our goal is to find a function that satisfies that relation. So, the answer to a differential equation is not a number, the answer is a function. We're seeking an unknown function.

Well, the example that we were just talking about, about money growing in a savings account, is a differential equation, and it is a simple example of a differential equation where the change is equal to just a constant times the amount of money in the bank account. But, differential equations in general are applied to many, many different areas of science and also in mathematics as well, but all sorts of areas of science, social science, and many different things. Newton and Leibniz were the first ones to write down differential equations, so differential equations really came about at the very dawn of calculus, and they really are the central applications of calculus to all sorts of areas. As I say, certainly physics, biology, economics, social sciences—really, essentially, any area you can think if, differential equations are the method of describing these kinds of issues.

Let's begin by talking about some basic physics examples that can be described in terms of differential equations, and then we'll get used to the idea of what we're going to do with these differential equations. So, let's suppose that we consider taking an object, like this ball, and we drop it. When we drop it, suppose we have just made observations, and we realize that the ball increases in its speed as time goes on, and by taking measurements we learn that the rate at which it is falling at every time t is $-32t$ feet per second after we've let it go. In other words, after 1 second, we've let it go, it will be dropping at 32 feet per second. And the negative just means that it's falling downward. Well, the question that we might want to know is how far will the ball have fallen after t seconds? If the information we're given is just telling us how fast it will be going at each moment of time, then we're really presented with a differential equation. So, here we go, we drop the ball, we measure its rate of falling at each moment of time, and then we have a differential equation. The differential equations is the following: If the derivative of the function that is going to tell us how far it has fallen, we'll call that function p, $p(t)$, because p is going to be the function that we're after to tell us how far it will have fallen after time t seconds. What we know about that function is something about its derivative, namely its derivative, which is the velocity, is going to be equal to $-32t$, and our question is, what is our function? Well, we've done this kind of thing before.

Our function is the function $p(t) = -16t^2$ + a constant. Any function of that sort, any position function that tells us where a falling ball is at each time t, if it's $-16t^2 + c$, then the velocity of that falling object will be exactly $-32t$, because, you see, the derivative of $-16t^2$, we bring down the 2, it's -–2 t^1. What is the c? Let's talk about the c for one minute, just to get it out of the way. The c is the following: that when you drop a ball, the initial position of the ball will tell you where the ball is, but the speed of the ball will not be affected by whether you think about having your zero point on how you're measuring the position of the ball; it doesn't matter whether you think you dropped it at point 0 or if you dropped it at the point you call 100. It doesn't matter; the speed will still be going at exactly $-32t$ feet per second after you dropped it, after time 0. So, c in this case is just the position of the ball at time 0.

In general, when we have a differential equation, the solution to the differential equation often gives us a family of functions rather than just one particular function. One of the prime examples of a differential equation is a differential equation that involves a second derivative. When we drop a ball, the rate at which the speed of a ball, the velocity of the ball increases; that is, its acceleration is –32—that's the acceleration due to gravity at the surface of the Earth. So, the differential equation, the second derivative of p, dt—this is the way you write down the second derivative, it's $\frac{d^2p}{dt^2}$—the second derivative of the unknown function p is –32, meaning that it's a constant acceleration. Now our goal is to say, "Well, what is the function?"

Well, the way that we find the function is we really take the integral twice. We first integrate to find the family of velocity functions, and then we integrate again to find the position functions. Here we've done this, and we realize that any function of the form $-16t^2$ + a constant × t + another constant will have the property that its second derivative is equal to –32. Here we've just taken the derivative; we've taken the first derivative here, and when we took the first derivative, the first derivative is $-32t + a$; and then when we took the derivative of that, which is the second derivative, we get –32, the constant –32. So, any function of this form is a solution to the differential equation, the second derivative is equal to –32. What are a and t? Well, a is going to be the velocity at time 0, and b is going to be the position at time 0.

One of the workhorses for differential equations really comes about in Newton's Second Law of Physics, which is force is equal to mass times acceleration. You see, acceleration is the second derivative of position. So often we end up with differential equations when we're using Newton's Law.

If we're told where we dropped it, that is, where the position was at time 0, and what the speed was at what we call time 0, if we threw the ball downward, its position is at a given place, but it may have an initial velocity, that those are called initial conditions for the differential equation, and if you have an initial value problem for a differential equation, it means that you're given the initial equations, and then knowing the initial conditions, we can get one particular solution equation instead of a whole family of equations.

Often when we're dealing with differential equations, we want to model some particular behavior, and when we model the behavior, we may find that it is convenient to refine the model, as we learn more and we want to put more conditions, to make it more realistic. If we actually drop an object in the air, it doesn't really accelerate at the prescribed rate of –32 feet per second per second. It doesn't quite do that because a real object is impeded by air resistance. You see, there's an upward force owing to the accumulation of air, and as a body is falling faster, that upward resistance of the air causes an upward force that resists the downward force of gravity to some extent.

Knowing that kind of fact about the world, we can conclude that there is a different differential equation that more accurately will describe the speed of a falling body; namely, if we're trying to find a function $p(t)$ that tells us where a dropped body will be, or a thrown body will be, its second derivative, the acceleration, is going to be equal to –32, the force of gravity, but then there's going to be another constant times the mass times the rate at which it's falling. In other words, if you're falling faster, there's more air resistance. So, we've got, then, a more complicated differential equation that we would try to deal with in order to describe the motion of a falling object more robustly.

Let's just look at some quick examples. If we wanted to talk about a pendulum, and ask what the angle is at each time, we could describe that by a differential equation. If we're talking about a law for a spring—Hook's Law for a Spring—if we take a spring and we expand it beyond its resting distance, and then we let go, we want to know where the spring will be at different times, that's another example of Newton's Second Law, where the force, the force on the spring, is spring constant times the amount of displacement. If you displace the spring more, it pulls harder. That force is equal to the mass of the spring times the second derivative, the acceleration. So, this is f, the force, is equal to $m \times a$, the acceleration, owing to this force. So, we have, automatically, a differential equation. So, these are all examples of differential equations in physics. A lot of physical processes are governed by describing the relationship between how a function is changing and the function itself. That's a differential equation.

Well, many mathematicians contributed to the growth of this field during the 17th and the 18th centuries. Leibniz and Newton were the first. They invented the notation for derivatives, and then they also wrote down the first differential equations. The Bernoulli brothers, Jakob and Johann and Johann's son Daniel, all these—a whole collection of Bernoullis—they solved a whole bunch of problems in mechanics, applied differential equations, and they applied these differential equations to complex physics problems, including fluid dynamics; we talked a little bit about fluid in the last lecture, and they used differential equations to describe these. The Bernoullis were famous, by the way, for being extremely jealous and quarrelsome. They fought with each other and wrote nasty things to each other. But, nevertheless, they contributed a great deal to the development of calculus, and particularly differential equations, in the late 17th and the early 18th centuries. Other significant contributions toward algebraically solving differential equations, which is a very difficult task, by the way, were developed by Joseph-Louis Legrange, whose 1788 work *Méchanique Analytique*—it's an elegant and comprehensive treatise on Newtonian mechanics—where $f = ma$ leads to differential equations.

Another big contributor to the solutions of differential equations was Pierre-Simone de la Place. He was an expert in celestial mechanics. He had a famous method for transforming differential equations into algebraic equations. It's called the *Laplace Transform*, and it's a key tool in solving differential equations even today.

Well, solving differential equations in the sense of actually getting a numerical solution is often a very difficult task. But we can get a sense of differential equations in a different method. So, that's what we're going to talk about for the rest of this lecture. We're going to particularly discuss an idea in the context of models of population growth. So, to ground our next discussion, I'm going to be talking about the idea of trying to say how many of a particular population of some animal will there be at a certain time, as certain time passes.

Such models automatically are very naturally and conveniently expressed in terms of differential equations. So, let's begin with the first model. This is Malthus's model of population growth, and it's a very simple model. It's exactly, by the way, like the model of the money growing in an account. It just says the following: That if you want to know the number of people in a population, the number of animals in a population, at time t, we'll think of that as, we'll think of that as $y(t)$ as the population at time t. Well, the change in the population is going to be equal to a constant times what the population is to begin with. Let me just remind you that we often say the rate of population growth was 1 percent. Now, what that means is that one year after today there will be 1 percent more people in the Earth than there are now. So, Malthus's equation says the rate of change to the population is just this rate, r, times the population itself. Now, by the way, the derivative of the population tells us how many new individuals there are per unit time. That's what the derivative means. Notice that if the population is growing at 1 percent per year, what that means is that it's 1 percent of the population that you have. So, if you had a population of animals where you had, say, 100,000 animals, the rate at which it's growing, if it's 1 percent, 1,000 new animals would be living after one year—the change of living animals would be 101,000,—1,000 new animals. Whereas, if you had 100 million animals in the population, you'd expect the population to be 1 million bigger than it was before. So, in other words, the population increase is related to the population that you have.

Well, that's a perfectly good model and very sensible; the only problem with it is it is totally unrealistic if you extend this time over any long period of time at all. The reason is that if you have that kind of exponential growth, which is what this population model models, very soon you have more individuals than there are atoms in the universe. It can't just keep growing forever. So, this is an example of a population growth model that has a fundamental flaw. The fundamental flaw is it can't go on forever this way, even though it may be a good model for population growth when there are plenty of resources around.

In looking at this differential equation, though, before we go on to refine it to take into account the maximum population growth, I wanted to tell you about a method for analyzing the solution for this differential equation without actually finding a formula for the solution. The method is the following: What we'll do is to consider the entire plane, and the plane is to be viewed as having a time axis, that's the horizontal axis, and then the vertical axis is the population axis. Remember, we don't know what the function is that satisfies our differential equation. Our goal always in differential equations is to find a solution to the differential equation, meaning to find a function that satisfies that relationship. Well, in the case of Malthus's model, we know that the derivative of the function y is some constant r times the value of the function. So, what we could do is go to every single point in the plane and draw an arrow that tells us, that captures the idea that we know the slope of the tangent line to our unknown function at each point, you see? That's what we know. We know the slope. The derivative is the slope of the tangent line. So, at each point we can say, "If I know what t is and I know what y is, I can look at my differential equation, plug in for my value of y here, and find out what the slope is." So, I simply draw that slope at every single point individually. I don't draw my function, I just draw the slopes at every single point.

Now, if I know an initial condition; if I know what the population was at time 0, for example, I can start to draw the solution equation because I know what direction it's going; I know what the slope of the tangent line is. So, I can draw a little bit of it. But, then I'm in a new place and I know what the slope is there, so I can draw some more; I can draw some more; I can draw some more; and, in fact, I can literally extend the line on and draw the solution to my differential equation by looking at this direction field, as it's called, arrows at ever single point. Well, this is a great strategy.

Let's go ahead and apply it to a couple of other refinements of the population growth model. Remember that the problem with the Malthus model was that every population that you started with would instantly become so great that it would use up all the resources in the world. So, in order to refine the model, the idea is to use the fact that there is a maximum capacity, a maximum population capacity, for the environment. That is, a carrying capacity, which we'll call k for the environment. If the population exceeds that maximum carrying population, then the population will decline. What we do is we write a differential equation that captures that concept of the reality, namely, that if the population comes close to the carrying capacity, then the rate at which the population will grow will decline; and if the population ever exceeds the carrying capacity, then the population will actually—the population itself will decline.

Here is a revised population model that's called the Verhulst model, also sometimes called the logistic equation, which is this. It says that the derivative of the population with respect to time is certainly going to be dependent on the population that you have. If you have more people in the population, it will tend to grow faster, but we put in this expression, we multiply by this expression, which has the following effect: suppose the population, which is y, at a given time, suppose that population is greater than the carrying capacity k of the environment. They y/k is a number bigger than 1. So, 1 minus a number bigger than 1 is a negative number. So, this is a negative number, which means that the population will decline, because now we have a negative number times something else, these are positive numbers, so it's a negative growth, which means decline. In other words, the slope is going down as time progresses.

This is another model for population growth, and we can do the same strategy for drawing the direction field for that model of population growth. In other words, at every single point in the plane, where this is the time axis and this is the population axis, we can go to that point and draw an arrow that indicates which direction, what's the slope of the tangent line, of our unknown solution to that differential equation. Then, if we have a particular initial value, we can just follow the trajectory of the particular value to give us a specific solution to the differential solution that would model the growth of that population if it started at a given point. And we can see if a population begins at a number below the carrying capacity, then it will rise to the carrying capacity; and if it starts above the carrying capacity, it will fall.

There's another critical component of population growth. It has to do with a threshold model, a critical threshold model, meaning if you don't have enough people in the population, the population is not going to grow, it's going to shrink. That adds a different coefficient here to the expression that we would multiply our differential equation by, and we can combine these two different features that the population, if it grows above the carrying capacity, it's got to become a negative number; this one says that if the population is not as big as the threshold population, than this will be a negative number and the population will decline. So, we can draw a vector field that indicates this more robust population model that includes both the carrying capacity and a minimum threshold. If we're below the minimum threshold, the population goes to 0; if we're above the minimum threshold but below the carrying capacity, it will rise and become asymptotic to the carrying capacity; and if it is a population that begins above the carrying capacity and above the minimum threshold, then it will decline, the population will decline until it becomes close to the carrying capacity.

So, in drawing these direction fields, we can then sketch the answers to differential equations, and, by the way, we can get a computer to put in as many little points and direction fields as we want, so that we can get very accurate estimations of the actual answer; and, in theory, we get one specific answer to our differential equation if we have an initial condition.

By the way, a variation of that model that we just described that uses both the threshold and the carrying capacity actually does accurately describe the passenger pigeon population in the United States, back when it was alive. Those pigeons have a very high critical threshold t, so they were hunted down and became to have such small numbers in the late 1800s that the population could no longer support itself, and that population declined and, in fact, they became extinct in 1914.

Well, differential equations appear in a variety of contexts, including the spreading of epidemics, hormone balance, chemical reactions, planetary motion, and many things. In the next lecture, we're going to talk about some interesting examples involving predator-prey conditions and, also, the stringed instruments, talking about music. I'll see you then.

Owls, Rats, Waves, and Guitars

Lecture 23

In this lecture what we're going to do is to take a slightly different approach. In particular, most physical phenomena that we're trying to look at, or biological phenomena for that matter, include more than one variable. So, we have a reflection of that reality in the concept of differential equations, and that's going to be the topic of this lecture.

We have seen that differential equations describe physical and biological phenomena. We have been primarily interested in functions of one variable, namely time. We have also seen that we can understand solutions of differential equations geometrically without knowing the exact algebraic form. Most physical phenomena include more than one variable, and the topic of this lecture is differential equations in several variables.

Once again, we will attempt to develop a mathematical model that describes population dynamics, but in this case, we have two species involved that are related to one another in a predator-prey relationship. In the northern California redwoods, the woodrat provides 80% of the food for the spotted owl, which is the main predator of the woodrat. Let $w(t)$ be the population of the woodrat and $s(t)$ be the population of the spotted owl as functions of time. If there are no owls, that is, $y = 0$, we assume that the population of the woodrat grows exponentially: $dw/dt = w$. Similarly, if the are no woodrats, that is, $w = 0$, we assume that the owls die out exponentially with rate $r = 0.75$: $ds/dt = -0.75s$. (We use 0.75 for the purpose of illustration.) Whenever an owl and a woodrat meet, we assume that there is some chance that the owl eats a woodrat, hence survives and is able to reproduce, resulting in an increase of the owl population and a decrease in the woodrat population. Thus, the *system* of differential equations becomes the following:

$$\frac{dw}{dt} = w - 0.5ws$$
$$\frac{ds}{dt} = -0.75s + 0.25ws.$$

How do we analyze such a system? Consider the direction field approach again. At each point (w,s), regardless of t, we know the rates of change dw/dt and ds/dt. Thus, we know the rate of change of dw/ds, as well, and we draw an arrow in the direction (ds/dt, dw/dt). Then, if we start at a given point, that is, at some initial condition (w,s), we simply follow the arrows to trace out the evolution of w and s as functions of t.

In this system, we observe a cyclical nature for the populations of the woodrat and the spotted owl. Notice first that the populations are constant exactly when $dw/dt = 0$ and $ds/dt = 0$. This occurs at (0,0) and (3,2). The population is stable at these points. Otherwise, if we start at another point such as (3.5,1), as the population of the woodrat increases, so does the population of the spotted owl. Eventually, the population of the woodrat decreases, and this results in a decrease in the population of the spotted owl. This, in turn, fuels the growth of the population of the woodrat, and we are back to the beginning of the cycle. With different initial conditions, for example, (2,1.25), we see a different curved pattern.

These models for predator-prey relationships were pioneered by American biophysicist **Alfred Lotka** and Italian mathematician Vito Volterra in the early 20th century. They have been verified in several examples, including by the records of the Hudson Bay Company of Canada, which traded lynx and snowshoe hare pelts in 1845–1935. In these examples, the cyclical nature with a period of 9–10 years is distinctly pronounced.

Now we look at an example of differential equations that model the behavior of a vibrating string, the **wave equation**. In this case, we will find that we are seeking a function of two variables. Consider an elastic string of length L tightly stretched between two supports at the same horizontal level, for example, a guitar string. Now, pluck the string and let it vibrate freely. To describe the behavior of the string, we are interested in how each point on the string moves. That movement is captured by a function of two variables: the position along the string, x, and the time, t. At each such position x and time t, we are interested in the vertical displacement $u(x, t)$ of the string.

What do we know about a vibrating string that would allow us to figure out the function $u(x, t)$ that would capture the string's motion? If the string is exactly straight, no force is causing it to rise. But we imagine that the string is moving and we see what forces are acting on each piece of string. The forces are those that come from the tension of the string. Newton's famous law, $F = ma$, tells us how the forces pulling on a point of the string impel that point to accelerate. The derivation of the wave equation involves drawing pictures of a piece of the string and resolving the forces that are pulling at angles to tell us how the string is accelerating. We find that the wave equation is:

Once again, we will attempt to develop a mathematical model that describes population dynamics, but in this case, we have two species involved that are related to one another in a predator-prey relationship.

$$\frac{\partial^2 u(x,t)}{\partial t^2} = c^2 \frac{\partial^2 u(x,t)}{\partial x^2}.$$

These equations mean that the acceleration of a point on the string at a fixed point x as time varies is proportional to the acceleration of the height with respect to the position along the string. If we attempt to describe the vibration of a guitar string or a violin string, we have additional constraints, namely, the string is fixed at the endpoints of the string and we assume that the string is displaced in a certain way to start the vibration. That is, the string is plucked (imagine its being put into a shape and then released at time $t = 0$).

The solutions of the wave equations involve the sine function. This function is periodic, as we would expect. We can look at the solutions to the wave equation and see how the behavior of the string can be thought of as a sum of sine curves that corresponds to **integer** divisions of the total length of the string. Therefore, every integer fraction of the length of the string will give a different rate of vibration; thus, the actual notes of the string can be viewed as a combination of those sine-like functions that contribute higher pitches. We have a combination where the string will naturally vibrate up and down like a sine curve at all times; then inside it, we have the more refined numbers of vibrations. That is why the notes we hear from a guitar are so rich sounding.

Name to Know

Lotka, Alfred J. (1880–1949). An American biophysicist and the father of mathematical biology; he published the first book in this field in 1924. With Vito Volterra, he is chiefly known for the formulation of the Lotka-Volterra equations for the study of predator-prey models in population dynamics.

Important Terms

integer: A whole number (positive, zero, or negative); …–2, –1, 0, 1, 2,…

wave equation: A partial differential equation governing the displacement of a string as well as more complex phenomena, such as electromagnetic waves; it relates the second derivative of position to the second derivative of time.

Suggested Reading

Any standard differential equations book on partial differential equations and systems of differential equations.

Questions to Consider

1. What are some of the pitfalls of relying entirely on mathematical models for setting environmental (or other) policy?

2. The solutions to the wave equation for a string of length π have a factor of $\sin(k\pi x)$, where x gives the position on the string and k is a positive integer. Verify that this $\sin(k\pi x)$ term corresponds to having $(k - 1)$ equally spaced points along the string, besides the endpoints, where the string does not move up and down at all. This result was physically demonstrated in the lecture using a guitar string.

Owls, Rats, Waves, and Guitars
Lecture 23—Transcript

Welcome back. During the last lecture we talked about differential equations, and in particular we saw how differential equations can describe the physical world, like when we dropped a ball; and biological phenomena, such as population growth of a particular population. In that lecture you might have noticed that the answers that we were seeking, in other words, the solutions to the differential equations, were in each case an example of a function of one variable. In fact, in the cases we did, the variable was time. In other words, we saw that if we drop a ball, the differential equation was reflecting the fact that we know the force on the ball; we knew how the ball was accelerating as it fell, and then from that we concluded a function that told us where the ball was at each moment. Likewise, when we were talking about the population growth of a species, we were saying if we had a species of animals, then they will grow in a certain fashion. The first model was an exponential kind of fashion where they grow at a certain rate depending on how many individuals there are in that population, and then we modified that model. But, the answer was a function of one variable; namely we were seeking to know, at a given time, what will the population be?

In this lecture what we're going to do is to take a slightly different approach. In particular, most physical phenomena that we're trying to look at, or biological phenomena for that matter, include more than one variable. So, we have a reflection of that reality in the concept of differential equations, and that's going to be the topic of this lecture. In fact, remember that one of the themes of the whole course has been the following: We started out with the ideas of the derivative and the integral as applied just to activity on a straight road—you know, a car moving on a straight road—and we developed the ideas of the derivative and the integral, but those ideas were so potent that they extended to more complicated situations, such as a car moving along the plane, or a function of more than one variable. That's what we're going to do today, from the point of view of differential equations. In other words, we're going to be talking about situations where we have differential equations where the answer that we're seeking is not just a function of one variable, but a more complicated scenario. But, it's all fun, so let's just begin with our particular example and I think you'll see how this goes.

Here on the table we have a spotted owl, a northern spotted owl, and we have a collection of woodrats. What we're going to do is try to make a model that describes the population, but this time not the population just of one species, but the populations of two interacting species. In this case, the two species are interacting with each other in a predator-prey relationship. That is, the northern spotted owl, its feast of choice is the woodrat, so it eats these woodrats, and the woodrats go away, because he eats them. So, the owls eat the woodrats. In fact, by the way, for your information, 80 percent of the diet of the spotted owls are these delectable woodrats.

The question is: How can we use this somewhat more complicated situation of two species that are interacting and develop differential equations that describe how those populations will change? Well, let's just go ahead and see if we can write down something that will reflect our understanding of what's going on in the world. That's the way we make a mathematical model of anything, and in particular, in this case, of populations. Let's let $w(t)$ be the population of the woodrats at any given time t, and $s(t)$—s for spotted, be the population of the spotted owl, the predator, at any given time. Our goal is to find the two interacting functions; the function that tells us how many woodrats there are and the function that tells us how many owls there are.

We begin with our basic concept about how population grows, and we're going to simplify everything as much as possible to make it as clear as we possibly can. So, we'll start here with the assumption that if there were no owls, if no owls were in sight, then the woodrats would have plenty of resources to eat, and they would just grow exponentially. We saw in the last lecture that that's not a realistic model, because if they consume more resources than there are in the environment, then they can't continue to grow at that very fast rate. But, let's just assume that in the absence of the owls, for this level of population, there are plenty of resources for them to grow in this exponential fashion.

Now, on the other hand, let's look at the population of the owls; well the population of the owls if there were no woodrats. So, suppose that the owls were there, but the woodrats, for some reason, had all died out, then the owls, of course, would die, and they would die at a predictable rate—once again, we'll assume that it's an exponential decline, that's the negative number, and we put a coefficient .75*s* just for the purpose of illustration to say that this is the rate that their population would decline in the absence of their main food source, the woodrats.

But now we're going to look at the fact that these two populations are actually interacting. The reason that the woodrats' population doesn't just soar and become exponentially large is because the owls are around to eat them, you see. So, let's see how we can reflect that concept in an equation.

What this equation illustrates is the following: It says that we're going to assume that the population of the woodrat will change; that is, *dwdt*, that's the rate of change of the population of the woodrat. It would increase exponentially, it would just be equal to *w* in the absence of the owls, but every time that there's an interaction with an owl, the population of the woodrats will tend to decline. Now, this is somewhat of a euphemism from the point of the woodrats. You see, we're talking in this pristine world of mathematical equations and from the woodrats it's their aunts and uncles and their friends who are being eaten by the owls. They may have a different way of expressing this. But, nevertheless, as a mathematical model, we have this concept that the rate at which the population of the woodrats will change, it tends to grow because there are plenty of resources for them to eat, but on the other hand it tends to decline in proportion to the product of the number of woodrats there are and the number of owls there are. And we can think of this as reflecting the total amount of interaction between those two populations.

Similarly, let's look at the population of the owls. Well, the population of the owls, we saw, would decline. Since most of the food of an owl is the woodrats, its population would decline in the absence of woodrats. But, when there are interactions of the woodrats available, then the population of the spotted owl, the rate of growth, will tend to contribute a positive number associated with the amount of interaction, meaning the amount of eating that the owls get to have. So, the rate at which the population of the owls will change will have two features: one, its decline, which is viewed as a decline in the absence of the food source; and then, plus .25 times the amount of interaction with the woodrats, meaning the amount of food available. And you can see if there's a great deal of food available—in other words, a lot of woodrats—this number will become a bigger positive number and the population of the owls will increase.

This is a very interesting system, and it incorporates the concept that the owls are eating the woodrats. How can we understand what this model of the interaction of the woodrats and owls entails? Well, we can use the strategy that we did in the last lecture of drawing a vector field, a direction field for how these two populations are changing; *dwdt* is some value. So in other words, if we're told what the population of the woodrats is, and what the populations of the owls are, and what the time is, we can find the rate at which the population of the woodrats is changing, and similarly for the change in the population in the spotted owls.

Knowing how fast the woodrats are growing relative to the rate at which the spotted owls are growing, we can draw in a plane—we can record the relative increase in those two populations by just taking the ratio of *dw* to *ds*—the change in the woodrats to the change in the owls, basically canceling out the *dt*'s. So, in other words, if we take a certain number of—so, this axis here, by the way, is the woodrat axis, and this axis here is the spotted owl axis. That is the number of spotted owls; this is the number of woodrats. Of course, by the way, the numbers that we're talking about here are just small numbers like 2 and 3, but think of them as proportional to the populations of the woodrats and the owls.

If we had, for example, a certain number of woodrats—we had, let's say, 2 woodrats and we had 1 1/2 owls, then we could compute the *dwdt*—that would be a number, just plugging in the 2 for *w* and 1 1/2 for *s*, and we could do the same thing, plug in numbers here, and we would get *dwdt* and *dsdt*, and we could use those numbers to draw an arrow on the plane. In other words, what we do is at every single point in the plane, we can compute a number, *dwdt*. The point on the plane, the horizontal axis is the number of woodrats, *w*; the vertical point is the number of owls, *s*; and all we do is compute the *dwdt* and draw a horizontal arrow of that length, and then we compute *dsdt* and draw a vertical arrow of that length, and then we resolve those two vectors to draw an arrow that goes to the corner of those two different horizontal and vertical contributions, that change in the woodrat population compared to the change in the owl population, to get an arrow in the direction that tells us about how both populations, the owls and the woodrats, are changing.

Let me make this absolutely clear because I want to make sure that everybody actually understands here what each of the arrow in this field of arrows means. We go to a particular place—let's suppose, for example, we go to the point (4,4), and we want to know what arrow do we draw at the point (4,4)? Well, we plug in 4 for *w* and 4 for *ws,* and we get a number here; 4 − half of 16. So, it's 4 − 8, that's –4. So, we say *dwdt* is –4. So, at the point (4,4) we draw an arrow this direction of length –4—that's to the left because the positive's to the right, negative's to the left—–4. Then, here, we do the same thing. We plug in 4 for *s* and 4 for *w*. When we plug in 4 for *s*, we get –3 here, plus 4 here; –3 + 4 is +1; and so we draw an arrow at the point (4,4); we draw a vertical arrow of 1 unit length high. Then we resolve those two vectors, so that gives us a diagonal line that's slightly facing upward and to the left. So, that's why this arrow here at (4,4) is slightly up and to the left. We do that at every point; actually, we don't do that, we let a computer do that, at every point on the plane. Then things get interesting.

What gets interesting is that we can follow the populations, if we start at any population of woodrats and owls, we can follow the arrows to tell us what's going to happen to the populations of both the woodrats and the owls. So, now the owls you see are eating the woodrats; see they eat the woodrats. Suppose there are lots of owls, and if there are lots of owls—for example, up here in this area of the plane—suppose that there are lots of owls and they are eating the woodrats so much, because there's so many of them, the population of the woodrats declines. So, here come the owls, they eat the woodrats, eat the woodrats, eat the woodrats, and they eat so many woodrats that pretty soon there are just a few woodrats left, and there are so few woodrats left, that the number of owls has to start declining. You see, the number of owls is starting to decline. So, here we're following this curve, the number of owls is starting to decline. See, there go the owls, he's getting weaker.

Now, the owls get so few in number that the woodrats start to make a comeback, so that when the woodrats are down here, they start to increase in number because the owls are not around to eat them. So, here come the woodrats; they're growing; look how fast they're growing. The woodrats grow, and then after a while, since there's more food, the owl population starts to grow also. So, the owl population starts to grow, so the owl gets a little bit stronger here, and then the owl population continues to grow while the woodrat population declines, you see? And then the owl population is eating so many woodrats that the woodrat population has declined to the point where the owl population declines, and we start the circle again.

So, what this field of arrows tells us is that the populations of the owls and the woodrats do this kind of a cyclical dance with each other. If we start with an initial population of woodrats and owls, we can follow the arrows to see how the populations of the two will interact, and we'll get this kind of a curve. Now, if the populations of the woodrats were 3 and the owls were 2, then that's a stable point; that's a point where the populations would just stay exactly the same. But if we start at a different population of woodrats and owls, we'll get this cyclic pattern. And, by the way, there are some wonderful applets on the web that you can find that will let you just push the button and you'll see it draw the path of this interacting populations of the woodrats and owls circling in the numbers of them.

So, these are all different solutions to these coupled differential equations that reflect the idea that the populations of woodrats and owls are going to alternately increase and decline as they go on. In fact, we can draw this graph for the initial condition that the woodrat population at time 0 is 2, and the owl population is 1.25, in this picture we would see this kind of a curved pattern, and then in this picture where we draw the populations of the owls and the woodrats separately, we see this alternating pattern of the woodrats increasing and decreasing alternately, and the owls increasing and decreasing, shifted over a bit.

So, this is an example of a model—a predator-prey model—of the relationship of two animal populations that are interacting. These models for predator-prey relationships were pioneered by a physicist by the name of Alfred Lotka and an Italian mathematician, Vito Volterra, in the early 20th century. In fact, they've actually been verified in some real examples. The best one is from the records from the Hudson Bay Trading Company of Canada, and in the period from 1845 to 1935, the Hudson Bay Company traded in, among other things, in Canadian lynxes and snowshoe hares. This was a predator-prey relationship. The Canadian lynx had its main diet being snowshoe hares, and if we see the amount of hares and lynxes that were captured and sold by the Hudson Bay Company, we can see this alternating pattern of growth and decline; the hares here and the lynxes you can see alternating pattern such as we would predict by this predator-prey model.

This is a wonderful example of a differential equation where the answer to the differential equation gave us a curve in the plane. Let's now turn to an example of a differential equation whose answer is going to be a function of two variables. So, we'll get rid of all of the rats and the owl and this time we're going to turn to music.

When we're playing a musical instrument, such as a guitar, we might want to know how are we going to describe the movement of the string that will produce the sound? In investigating this, what we're thinking about is describing the motion of a string. Now, we have to think about this carefully to see what kind of a description would describe the motion of a string. If we think about this string as being taut and held between two points, then we can describe the motion of the string by using two different variables. So, let's think about this carefully. We have this horizontal string, and it's tightly held between the two, so there's tension on the string. At each point of the string, that string—that point—is going up and down as time passes. So, if we want to describe the entire motion of the string, we want to say for every particular point between one end of the string and the other, and every time during which we're considering this motion, we want to know how much that point is displaced from the horizontal. So, you can see that this is actually a function of two variables. The two variables—now, again are the following—so on this guitar string, we have the two ends of the string are held fixed, and we're imagining the string as being absolutely horizontal. When the string is moving and producing its sound, every point, like this point here along the string, is going up and down, so at every time t, it is displaced from the original horizontal position by some amount. We call that amount $u(xt)$. It's a function of what place—that's the x—what place along the string; and, then, the time, and as time moves it goes up and down. So, the first thing to realize then is that a description of a vibrating string is a function of two variables, and that's what we're seeking as our description of the motion of a vibrating string.

Well, how are we going to investigate this kind of motion? The way that we do it is we're going to try to think in terms of writing down a differential equation. As we did in both of the previous examples, in the last lecture and in the predator-prey model, we try to understand the world and have the understanding of the world give us a relationship between changing quantities. Those relationships are a differential equation—or in the case of the predator-prey models, two different differential equations that were related to each other. So, we're trying to analyze what's happening to this moving string in order to figure out what type of differential equation will describe that motion.

So, let's think about this string as it's vibrating, in order to try to understand how the equation that's going to describe that string will behave, what properties it will have. The first thing to realize is that as the string is vibrating, it changes its shape. It goes up and down and changes its shape in various ways. If part of the string were exactly straight, even if it's at some angle, it could be horizontal or at some angle, notice that at a point in the middle of that straight piece, there would be no upward or downward force. In other words, if you hold a string in high tension in the middle of the string, and it's exactly straight, it's not moving, it's just exactly straight, there's no force that's causing it to rise. So, if we are imagining a force making a point rise, that force is associated with the string having a curved nature at that point.

By the way, the way we produce this *wave equation*, which is what we're deducing right now, the way we think about producing the wave equation is to think about some physics. We think about a little point on that string, a little bit of the string, and we're asking: What are the forces on that string that are making it go up and down? What are the forces? Well, the only force on that little piece of string is the tension pulling on one side from one end of the string and pulling on the other side from the other end of the string.

So, we're trying to understand how this is working. As I said, if the string were exactly straight, there would be no up or downward force, but if the string were concave up, such as this, then, conceptually speaking, what that means is that this end—the tension that is coming from this side—is going to be in the upward direction compared to the tension on this side, and therefore will impel the string upward. It will make an upward force at that point.

Now, remember way back to our early lectures about derivative. Remember how we said that if you have a function that is concave up, that corresponds to the second derivative being positive. Concavity was a change in the derivative, and the derivative was the steepness of the tangent line, and so if it's concave up that meant that the derivative was increasing, and, therefore, concavity was associated with the second derivative of the position. So, that means that the force upward is associated with the second derivative of this function with respect to position.

Let's now realize that Newton's famous Law of Physics—that force is equal to mass times acceleration—tells us that the upward force on a point to satisfy this condition is associated with the second derivative with respect to time. In other words, it's changing its vertical position and it's accelerating associated with the force. Yet, we saw in the last slide, that the force is associated with the second derivative of the position part of the function. So, we have this complicated-seeming function that the second partial of the function u with respect to the time coordinate is equal to some constant times the second partial with respect to the position.

Well don't worry about it—I certainly haven't given a complete derivation of the wave equation, but let's just assume that we have, in fact, found this wave equation; that we realize a vibrating string will satisfy these conditions; that the rate at which the time at a fixed point is going up and down, the second derivative of that is going to be a constant times the second derivative with respect to the position as you move along the string. Now, when we're actually talking about a guitar string, we know that the end points of that guitar string are always going to be fixed, because they're not vibrating; they're set. So, we have an addition condition that at the beginning of the string and at the end of a string, throughout all time, its value is always going to be fixed at 0. By the way, we're going to make this guitar string exactly π units long for convenience.

Here are some solutions to the wave equation. If our guitar string is exactly π units long, then one solution would be the sine of $k \times x \times$ the cosine of $k \times ct$. It satisfies this equation because we can actually take the derivatives with respect to time and the derivatives with respect to x, and we get the same relationship that is given by the wave equation.

Now what that tells us is this equation—that's a product of a sine of some integer multiple of x times the cosine of an integer multiple of ct—satisfies the equation. What does that mean physically? It means physically that if we have our string here that's exactly length π , that when we pluck the string, one solution of the equation would be just to have the string go up and down in sort of a sine wave kind of look to it. But, it is also possible that if it were stopped in the middle, which corresponds to having the integer 2 inside of this sine function, then we would have a different kind of vibration, and you can hear that it's twice as high, it's an octave higher, than the bass note. Here's the bass note. Likewise, any integer multiple of the length is going to work, so let's go ahead and do 1/3. That note is three times the vibration level of the original note, and notice something else about this—it stopped about a point 1/3π , and it's also stopped at a point 2/3π . So, when it's vibrating, I can put my finger at 2/3π and it will not stop the note. Listen. Whereas, if I put my finger anywhere else, it stops the note; likewise at π /4.

Every integer multiple of the length of the string will give a different kind of vibration, and, therefore, the actual notes of the string can be viewed as a combination of those integer sine-like functions. So, we have this sort of combination of where the total string is vibrating up and down like a sine curve at all times, and then inside it we have these more refined numbers of vibrations that are included in it; and that's why when we hear a guitar note it sounds very rich, because we not only hear the fundamental frequency, but we also hear multiples of that fundamental frequency. So, this is an example where, following differential equations and using calculus, we can actually understand music and the guitar.

I'll see you next time.

Calculus Everywhere

Lecture 24

It would be really difficult to exaggerate the importance of calculus and its influence on our real lives and how we live. It's so effective, because it's a tool that gives us the ability to quantitatively study change and how parts combine to make the whole.

Calculus is so effective because it is a tool for quantitatively studying change and how parts combine to create the whole. The techniques and strategies of calculus all arise from two fundamental ideas—the derivative and the integral. Further extensions of the ideas of calculus and further applications of these ideas are actively pursued to this day.

We have touched the wealth of ideas related to calculus, and each idea is the tip of the iceberg. For example, we started the course by considering the motion of a car and developing the two fundamental ideas of calculus, the derivative and the integral, in that context. But calculus is also extensively used in the inner workings of the car: the physical motion of the engine pistons, the optimization problem for optimal fuel consumption, and the differential equations guiding the heating and cooling systems. Likewise, the **heat equation** not only helps us to understand the way heat moves in the engine, but its variations are among the tools used in the analysis of the pricing of financial derivatives. As well, electric circuits are described by differential equations.

Let's remember some of the concepts we have covered in the course and how we could have expanded those concepts. We considered the problem of hydrostatic force acting on a dam. But we could have talked about any number of other optimization problems involving power generation at a dam. We have also seen how calculus helps determine optimal architectural structures, such as suspension bridges and arches. We saw an example of using calculus as part of product design when we studied how calculus helps us design soda cans that minimize the amount of aluminum used. The same optimization techniques are used to design rockets that put telecommunications satellites in orbit. The rocket consists of several propulsion stages with the satellite on top. Each stage accelerates the rocket by burning fuel. Once a stage burns all its fuel, it is simply jettisoned, and the next stage is activated. Calculus is used to minimize the mass of the rocket while achieving the desired velocity.

The potency of calculus is a testament to the power of abstraction. We have seen repeatedly that the methods of solving problems in a bewildering array of fields magically involve the same processes that are captured in the derivative, the integral, and the limit. We can view even apparently static objects as dynamically growing or changing. This perspective allows us to analyze them in a different and often effective way. The habit of recasting a problem to see it dynamically creates new insights. Calculus gives us specific tools and an underlying philosophy about how to view and interpret the world. It gives a perspective that can be applied or attempted in many settings. Often, looking at an issue from the calculus point of view can lead to productive insights.

Calculus is also extensively used in the inner workings of the car: the physical motion of the engine pistons, the optimization problem for optimal fuel consumption, and the differential equations guiding the heating and cooling systems.

From one point of view, the story of calculus is the story of the intellectual conquest of infinity. The derivative and the integral are the two fundamental ideas of calculus. We saw them arise through a commonsensical analysis of the everyday phenomenon of motion—a car on a straight road. But the derivative and the integral were processes that had amazingly rich structures. The definition of the derivative arose from the analysis of the car moving on a straight road, studying the average velocity of the car in increasingly small increments of time. The derivative is the limit of those average velocities as Δt goes to 0. The integral is a sum. It allowed us to add up little pieces—thin sections of areas or slices of volumes to create the whole from its parts. We also talked about infinite series, which is a way of taking a complicated function and dealing with it as a simple function. The sine function, for example, can be viewed as an infinite polynomial. In the last two lectures, we talked about direction fields as solutions to differential equations. If we imagine the arrows being more and more numerous, to infinity, we could find a smooth solution to a differential equation.

One of the geometric views of calculus was that if you magnified a graph of a differential function closely, it would look like a straight line. Likewise, if you have a function of two variables that is differentiable, it will look like a flat plane. In the 19th century, however, mathematicians **Weierstrass** and Bolzano both showed a function that was continuous but not differentiable. It was not smooth at all. In the world of fractals, the Koch Curve uses an infinite process to produce a curve that is not smooth-looking when viewed arbitrarily closely. The Devil's Staircase is an example of a graph of a perfectly continuous function that is flat almost everywhere. This is another example of applying the idea of infinity to different situations.

There is much more to come in calculus and beyond. Thousands of papers are written about calculus today, and hundreds of thousands of papers in engineering, physics, biology, and other arenas use calculus. Newton said, "If I have seen farther than others, it is because I have stood on the shoulders of giants." In exploring calculus, we have stood on Newton's shoulders, on Leibniz's shoulders, on Archimedes's shoulders, and those of many other great thinkers. We have communed with some truly spectacular highpoints of human thought. All of us can share an appreciation for that giant conceptual development that shaped and continues to shape history—calculus. ■

Name to Know

Weierstrass, Karl (1815–1897). German mathematician. He left the University of Bonn without a degree and taught secondary school for more than 12 years while independently studying analysis. His 1854 paper led to a professorship in Berlin. He provided a rigorous definition of the limit, thus placing calculus at last on solid mathematical ground, and he provided a precise definition of **real numbers**.

Important Terms

heat equation: A partial differential equation governing the heating and cooling of objects; it relates the second derivative of position to the first derivative of time.

real number: Any decimal number.

Suggested Reading

Any standard calculus textbook with applications.

Boyer, Carl B. *The History of the Calculus and Its Conceptual Development*.

Questions to Consider

1. What intellectual developments in history compare to calculus in their influence?

2. To what realms of the world is mathematical analysis appropriate or not appropriate? How has the answer to that question changed over time?

Calculus Everywhere
Lecture 24—Transcript

Welcome back. I'm always astounded at the amazing richness that flows from the two fundamental ideas of calculus. It would be really difficult to exaggerate the importance of calculus and its influence on our real lives and how we live. It's so effective, because it's a tool that gives us the ability to quantitatively study change and how parts combine to make the whole. The techniques and the strategies of calculus all arise from just two fundamental ideas: the derivative and the integral, and each of those ideas was motivated by just a very straightforward but extremely clear observation of everyday occurrences, a car moving on a road.

But, nevertheless, even though from some perspective it was a sort of a simple analysis, it took literally thousands of years to develop those ideas and to develop those ideas into the calculus that we know today. And, calculus, by the way, is not dead subject. It's not a dead subject today. The extensions of calculus that are further being developed, taking those ideas and applying them to many more different things, all of that activity is actively being pursued to this very day. Calculus, there's no doubt, has been one of the most fruitful and influential strategies for analyzing our world that has ever been devised.

But, when I look back, by the way, at the 23 previous lectures in the course, there are two things that strike me. First thing that strikes me is the wealth of ideas that we've touched on. There really is an enormous collection of really interesting ideas, at least from my point of view. But, the second thing that strikes me is the fact that each idea is just really the tip of the iceberg, that every single example that we did, every application that we looked at and talked about, for each one there were maybe a thousand related examples or ideas or applications that could have been explored, and are explored by different people all over the world throughout the preceding years and now.

Well, let's remember some of these. When we started the course, we considered the idea of a car moving on a straight road, and from that we developed the two fundamental ideas of calculus: the derivative and the integral. And we developed them in that context of just a car driving on a road and seeing where the car was at each time. That was the basic thing we were looking for. But, if we wanted to look at a car in calculus, we might look at it from an entirely different point of view and see the extensive use that is made of calculus in talking about the inner workings of a car, you know things like the physical motion of the engine pistons; what they're optimization problems associated with optimal fuel consumption. We have the idea of differential equations that talk about the heating and the cooling system. The heat equation is an equation that tells us something about the way heat dissipates through a medium, like a rod or through the engine block, and the heat equation helps us to understand the way that an engine works, among other things, or many other things that involve heat.

What is the heat equation? How do we go about thinking about the heat equation? Well, the heat equation is a differential equation, and we think about it just as we thought about the wave equation, that if we know something about the rate at which one point on a rod changes its heat if there's a differential amount, you know, more heat on this side, and it's cooler than that heat, then heat will flow toward it. That idea can be expressed in a differential equation that gives us this heat equation whose applications are enormous. In fact, one of the interesting aspects of the heat equation is not only is it used to describe heat, which was its original purpose, but, in fact, variations on the heat equation are used to develop these fancy mathematical models of financial instruments—you know, the derivative changing, the black skulls model of financial, evaluating the cost of puts and calls in the financial world. These things come from analysis of calculus using the heat equation, as well as other ideas at the same time.

Calculus is about change. That's what makes it so important. So, it can be applied to change in many of these different physical concepts and conceptual contexts, not only the economic and financial issues, but talking about electric circuits; electricity is described by differential equations, for example.

Let's go back and remember some of the things that we did in the course. When we thought about how much pressure there is on the wall of a dam—remember that? We talked about the hydrostatic force, the hydrostatic force, how much total force there was on the side of the dam owing to the pressure on each point pushing against the side of the dam. But we could have talked about some optimization problems associated with generating power from a dam. For example, this is the kind of question to which calculus is ideally suited, suppose that you have this dam here, and there's a power station on it, and you've got three different turbines, and each of these turbines has some unique power function that gives the amount of power that can be generated from that turbine as a function of the amount of water that flows through it. Well, you might have, if you're in charge of this dam, you might be in charge of deciding where to take the arriving water and how to partition it among the turbines to maximize the total energy production. And the water flow, that is, the volume of water per second, it might vary depending on weather conditions or power needs or things like that. Calculus could be used to solve the kind of a problem of saying, "How should we partition the water to the turbines to maximize the power that we produce?" So, this was an example where you could take one thing that we talked about and actually do many other associated ideas.

We talked about how calculus helps determine optimal architectural structures. Remember, we talked suspension bridges and we talked about making an arch that was a very strong arch. But, we could also talk about a question like this; this is a typical calculus question: Suppose that we have a log, a round log, and we want to cut from that a rectangular-shaped beam that has maximum strength. How do we go about analyzing that kind of situation? Well, we combine knowledge about the real world, about how strong the beams are, and we find that the strength of the beam is dependent on how wide that it is and how tall it is in a particular way; being tall is better, but it has to be a certain width, or the square of the height times the width is proportional to the strength of the beam; and then using that kind of understanding of the physical world, and then using techniques of calculus, allows us to optimize the shape of the beam that can be cut from the log that will give the maximum strength. So, that's an example of using a technique that has been applied to different areas, and we applied it to different areas, that can be applied to this architectural issue.

What's an example? Remember when we talked about product design? Remember the product design when we took a can and we wanted to minimize the amount of aluminum in the side of the can that would hold 12 ounces of soda. Remember that? Well, there was a technique involved, and that same technique, that same optimization technique, can literally be used in rocket science. It is rocket science. The same technique applies to maximizing ideas in putting telecommunication satellites into orbit, because of this: A rocket consists of having several propulsion stages. The satellite you really want to get in orbit is on top, and then you have these stages that are full of the burning fuel. Once all the fuel has been burned out of a stage, then you jettison it and go on to the next stage. A good calculus question is: How should you divide the fuel and how should you jettison things so that you minimize the mass of the rocket while you do achieve the desired velocity of getting your satellite into orbit?

Well, one of the reasons I want to bring these up is one of the things that makes calculus so potent is that it's a testament to the power of abstraction. The techniques that we've seen over and over again, the strategy that we saw in one arena was available for solving problems in another. We repeatedly saw that the methods of solving problems in a whole bewildering array of fields, magically involve the same processes that were captured by the derivative and the integral, and by the limit concept.

One of the things that I think is most attractive about calculus is we can take the calculus point of view and apply it to all sorts of situations, including, for example, the idea of just looking at static objects as dynamic things. We did this many times. When we look at a static object, like a cube, what could be more staid and static than a cube? But we could look at it as something that's evolving, that's growing, and it's a dynamic view that allows us, then, to analyze such things as the equation of the volume of different structures, like a cube or other things. By looking at it in a dynamic way, it gives us a way to recast our view of the world in this kind of calculus perspective.

So, calculus does several things. It gives us specific tools, and we've seen these specific tools, how to take derivatives, how to take integrals, how to maximize—optimize—things by finding places where the derivative is 0. It gives us these specific tools, but it also gives us an underlying philosophy, a way to look at the world, a way to interpret the world, and that perspective can be applied to many, many settings.

Often when you look at an issue from this calculus point of view, you say to yourself can I take this object and look at it from this calculus point of view and get a productive insight? And, very often, the answer is yes, indeed you can, that it gives you a different point of view, seeing things as evolving, as dynamically changing, and analyzing it from the point of view as change.

One thing that I wanted to say in this last lecture is from one point of view, the story of calculus can be viewed as the story of the intellectual conquest of infinity. Think about all of the basic ideas of calculus. We started with the derivative and the integral. So, these were the two fundamental ideas of calculus. Let's remember them as an example of trying to deal with infinity. When I was talking about these lectures with a friend of mine from the physics department, and I told him that I was going to explain the idea of calculus and we were going to resolve one of Zeno's Paradoxes—or several of Zeno's Paradoxes—particularly the paradox of a moving object and at each instant it's still. He said, "Oh yes, you are able to deal with $0/0$, or infinitesimals over infinitesimals." To him, that was the triumph of calculus, the idea that we could somehow deal with the infinitely small. And that was the idea of the limit.

So, let's remember. Let's take the concept of the perspective of calculus as a conquest of infinity and explore it for a few minutes. Let's look at all the basic ideas of calculus that we saw. First the derivative. Remember, the definition of the derivative arose from our analysis of a car moving on a straight road, and it arose by saying we see where the car is at a time $t + \Delta t$; we see where it was at time t. We divide by Δt to see what the average velocity was over that small integral of time. And then we took the limit as Δt goes to 0. The limit as Δt goes to zero. That is an infinite process. An infinite process, because it wasn't good enough to just take one Δt; it wasn't good enough to say: How fast were you going in 1/100 of a minute? That wasn't good enough. It wasn't good enough to say: How fast were you going in 1/1000 of a minute? The actual definition of the derivative was that it was a limit as Δt goes to 0.

So, there were infinitely many steps involved to resolve Zeno's Paradox. When we were resolving Zeno's Paradox, we had to come to grips with making sense of an infinite process. It was not solved by saying choose Δt

to be an infinitely small number, because there is no infinitely small number. It was resolved by taking individual snapshots at specific moments of time, and then combining all of that information using a limit process that went to infinity to find out what the answer was.

Now, let's look at the integral. The integral was a sum, and when we thought about an integral, and I hope by this time, when you think about an integral, you say: The integral is a sum of little pieces. And those little pieces, in the symbol of the integral—we have this integral sign with a function dx, and what that means is it's the value if $f \times dx$, a little thickness. When we saw the integral as measuring an area under a curve, we saw that what we were really doing is taking the function values, and dividing up the interval from a to b into little Δx kinds of widths, and then saying that the area of each of these rectangles, whose height is the function value and whose width is Δx, we add up all of those sums of areas of rectangles to get not the integral, but to get an approximation of the integral. It's not until we broke those pieces into yet thinner slices, we broke them up into thinner slices, and added them up; and we got something else that was not quite the integral, but it was close to the integral. It was getting closer, but we had to do an infinite process to take the limit to get to one exact value. So, in a sense, the very fundamental concepts of calculus were involved with understanding the idea of infinity; going to the limit as Δx goes to 0 is really dealing with an infinite process. But let's look at some other examples of where it is that we talked about the infinite processes.

In Lecture 14 we talked about infinite series. Let me remind you of what an infinite series is. An infinite series was a way of taking a complicated function and dealing with it as a simple function. Functions that are simple are polynomials. They're the ones, for example, that our calculators can deal with, because they only involve multiplying numbers together and dividing and adding and subtracting numbers. It's just the arithmetic operations, it's with numbers. So, a polynomial; remember, a polynomial is just something $\times$ x^n + something $\times$ $x^n - 1$, and so on. It's just a simple expression. It's a simple expression. It's one of the expressions whose derivatives we actually took, and it's very easy to take the derivative of a polynomial because you just bring down the exponent and subtract 1; we saw that. So, you can take derivatives and integrals of polynomials. But, suppose that we have functions that are much more complicated than polynomials?

We saw in Lecture Fourteen an example of the sine function. The sine function was one where there was no obvious way to take a number and figure out what the sine of that number is. How would you go about doing that? Well, the answer is the way to do it is to realize that sine function can be viewed as an infinite polynomial; as an infinite polynomial. And we saw that polynomial in Lecture 14, here it is, and what it means is when we plug in x to the sine function, if we plug it into that infinite polynomial, we'll get the exact same number, meaning that your calculator can compute that number and also meaning that, in a sense, a sine function is just a polynomial, but extended to infinity. It's something that involves going to infinity. Well, the sine function, the cosine, the trigonometric functions are not the only functions that have this representation as an infinite series. In fact, Isaac Newton himself, and Leibniz, they used infinite series approximations of functions a great deal. This was one of the basic premises that they used.

One of the functions that we talked about generally that has come up in several lectures, when we talked about exponential growth in the Malthus model of population growth, or when we talked about money growing, that these were examples of exponential functions. But if you think about taking an exponential function, what does it mean to take a number and raise it to, say, the $\sqrt{2}$ power? What does that mean? What would you multiply together to take the number 3 and raise it to the $\sqrt{2}$ power? Even thinking about that question, it's not so obvious. You can't multiply 3 by itself, the $\sqrt{2}$ times; what does that mean? So, the idea is that exponential functions can be expressed as infinite polynomials, meaning that if you plug in a particular number x and you plug it into this infinite polynomial, the values that you get, that if you added all of them together, you would get a specific value.

Now, of course, you don't actually add all of the numbers together. What you do is you add a certain finite number together and you get one answer, and you add more of them together and you get another number, and you add more together and get another number, then in the limit you get a value. So, this limiting process came up in infinite series.

Let's go back to just two lectures ago and last lecture when we talked about direction fields as solutions to differential equations. Remember how we thought about that. We said suppose that you have a differential equation, and at each point in the plane you knew what the slope was at that point. Then you could take your pencil and follow—draw a curve—that always had that slope, the slope that was given at each point. We had little arrows at each single point. We could draw a curve that had the requisite derivatives at each point—that had the requisite slope at each point. That was a way to understand differential equations.

Well, the idea is that if we imagine those arrows being more and more numerous, and instead of just starting at a point and following a straight arrow for an inch, and then seeing what direction the next arrow is and going another inch and another inch, instead of that we could just go 1/100 of an inch, and then 1/100 of an inch, and then 1/100 of an inch, then 1/1,000,000 of an inch and 1/1,000,000 of an inch. And going to infinity allows us to find an actual smooth solution to a differential equation. So, this is another example of the use of infinity as a fundamental part of calculus.

One of the geometric views that we had about calculus was if we magnified the graph of a differential function very closely, it would look like a straight line. That's one of the most fundamental aspects of calculus; that is one of the premises on which everything depends, that if we look at a differential function closely enough, it looks like a straight line. And if you plug in any function into your graphing calculator, and you say to your graphing calculator, graph this function; and then you take just a tiny part of the x-axis and a tiny part of the y-axis—in other words, you magnify it—you will find that the graph will begin to look exactly like a straight line. You just look at a small enough part of the x-axis and a small enough part of the y-axis. It'll look like a straight line.

If you have a function of two variables that's differentiable; we've talked about this before. At every point it will look like a flat plane. That is one of the basic premises on which calculus is built. But, the question is: Is that the way the world is? Is that really an accurate description of the world? That things are really straight like that? Well, one of the explorations of infinity that came about in the middle of the 19th century had to do with trying to come to grips with these basic ideas of the derivative, and see if it's true that a function that you draw, when looked at very closely, actually looks like a straight line? It was actually shown, and it was rather disturbing to people. Actually, in 1872, the first famous example was given by Weierstrass, although, in fact, Bolzano had given an example 34 years before, of a function that although it was continuous, it was nowhere differentiable, meaning there was no concept of smoothness. If you magnified it, it did not look smooth. Yet, many mathematicians in the old days thought of mathematics as sort of describing the world, and the world would naturally be smooth when looked at very closely. But that's not really right. Mathematics is about abstraction, and one of the abstractions of mathematics is that we can think about using an infinite process to produce curves that are not at all smooth-looking when looked at close; and this is a whole new realm of mathematics.

So, let me give you an example of such a function. This is in the world of fractals. Fractals are a wonderful concept that we have objects that look the same at every magnification, and they don't look like a straight line. So, here's an example of a fractal and how we go about producing it. Once again, it uses the idea of infinity. So, here we go, let's start with just a straight line as a first approximation to our fractal, and let's remove the middle third from that straight line, and build just a tent, where each edge of the tent is exactly 1/3 of the length of the line, and so we have four segments now that are constructing the second state in the construction, or the first step in the construction of this fractal. Then we proceed to do the same thing again; that is, we replace each of the segments by four segments, each 1/3 as long as the segment that we started with. So, for example, this first horizontal segment is replaced by one, two, three, four segments. This diagonal segment is replaced by one, two, three, four; and so on and so on; so that this segment here has a total of 16 smaller segments. We repeat, every one of these flat segments is replaced by four segments by removing the middle third and putting in a little tent, and you can see that we get a curve that looks like that, and then we repeat again.

Notice that as we repeat this process, we're converging down to one shape, but this shape has the property that under any magnification it does not look smooth, because look what it's going to look like. This part of the curve, of the final fractal—by the way, this fractal is called the Koch curve; K-O-

C-H, the Koch curve. Let's look at just this part of the final fractal. How is it produced? Well, it's produced by starting with a straight segment and doing this infinite process to produce it. So, in other words, it is produced exactly as the entire curve was produced, but the entire curve started with a segment this long. So, what this shows is that every little piece of this Koch curve, upon magnification, looks precisely the same as the entire Koch curve. Well, this defies the concept of calculus. This is not a curve that is amenable to the ideas of calculus because these curves are not smooth. So, this is an example of a curve that is no where differentiable; nowhere differentiable.

Here's another example of a very strange object that looks very much like the other example that we saw. This is an object that is flat almost everywhere, and yet it gets from 0 to 1. This is called "The Devil's Staircase." So, between 1/3 and 2/3, let's just put a horizontal line at the value 1/2, and then in the middle third of the first third—that is between 1/9 and 2/9—we put a horizontal line at height 1/4 and between 7/9 and 8/9 we put a horizontal line at height 3/4. Then we take each of the remaining segments and put another horizontal curve; horizontal curve; horizontal curve; horizontal curve. If you continue this process forever, you can see that these fit together to form a path, it's actually the graph of a perfectly continuous function that starts at 0, it goes to (1,1), and it's flat almost everywhere. These are examples of carrying the idea of infinity and applying it to many different kinds of situations.

Well, Isaac Newton was well aware of the idea that calculus was just one step in a long story, and he, of course, was also a great physicist. But he had the concept that is absolutely accurate, that the world continues to change; that it's not done. What he said, he had a great quote near the end of his life, he said, "I do not know what I may appear to the world, but to myself I seem to have been only like a boy playing on the seashore and diverting myself by now and then finding a smoother pebble or a prettier shell than ordinary whilst the great ocean of truth lay all undiscovered before me." This was the philosophy that I think is accurate, and one expression of the accuracy of the unbounded limits of what we have in store for us in the future is the fact that literally thousands of papers are written about calculus and their extensions today in the mathematical world, and literally hundreds of thousands of papers will use calculus in engineering, in physics, and biology, and will use it applying to all sorts of different arenas.

So, another quote from Newton says, “If I have seen further than others, it is because I have stood on the shoulders of giants.” Mathematics is an incremental exploration; and in exploring calculus ourselves, we have stood on the shoulders of giants. We stood on Newton’s shoulders, on Leibniz’s shoulders, on Archimedes’s shoulders, and many other of the great thinkers who have contributed to this wonderful spectacle of calculus. We’ve really communed, I think, with some of the truly spectacular high points of human thought, and we can all appreciate that giant conceptual development that has shaped and continues to shape our history—calculus.

Thank you.

Timeline

All dates are approximate in the sense that the mathematical activities mentioned each spanned several or many years.

540 B.C. .... Pythagoras founded his school and proved the Pythagorean Theorem.

450 B.C. .... Zeno posed his paradoxes of motion.

355 B.C. .... Eudoxus was associated with the method of exhaustion that was an integral-like process.

300 B.C. .... Euclid presented the axiomatic method in geometry in his *Elements*.

225 B.C. .... Archimedes used integral-like procedures to find formulas for various areas and volumes of geometrical figures.

225 B.C. .... Apollonius described the geometry of conic sections.

A.D. 1545 .... Tartaglia, Cardano, and Ferrari were involved with the algebraic solution of cubic and quartic equations.

1600 .... Kepler and Galileo did work on motion and planetary motion, describing those mathematically.

1629 .... Fermat developed methods of finding *maxima* and *minima* using infinitesimal methods resembling the derivative.

1635 .... Cavalieri developed a method of indivisibles.

1650.. Descartes developed connections between geometry and algebra and invented methods for finding tangents to curves.

1665–1666.............................. During these plague years, Newton devised calculus, his laws of motion, the universal law of gravitation, and works on optics.

1669.. Barrow formulated ideas leading to the Fundamental Theorem of Calculus and resigned his chair at Cambridge in favor of Newton.

1672.. Leibniz independently discovered calculus and devised notation commonly used now.

1700.. The Bernoullis were involved with the development and application of calculus on the continent.

1715.. Brook Taylor and Colin Maclaurin developed ideas of approximating functions by infinite series.

1750.. Euler developed a tremendous amount of mathematics, including applications and extensions of calculus, especially infinite series.

1788.. Lagrange developed ideas about infinite series and the calculus of variations.

1805.. Laplace worked on partial differential equations and applications of calculus to probability theory.

1822.. Fourier invented a method of approximating functions using trigonometric series.

1827 .... Cauchy developed ideas in infinite series and complex variable theory and refined the definitions of limit and continuity.

1851 .... Riemann developed a modern definition of the integral.

1854 .... Weierstrass formulated the rigorous definition of the limit used today.

1700–present day .... Innumerable mathematicians, scientists, engineers, and others have applied calculus to all areas of mathematics, science, economics, technology, and many other fields. Mathematicians continue to develop new mathematics based on the ideas of calculus.

Glossary

acceleration: Rate of change of velocity; a measure of how fast the velocity is changing. The second derivative of position. Units are distance/time^2.

antiderivative (of a function *f*): A function with a derivative equal to *f*.

Brachistochrone: A curve traced between two points (not atop one another), along which a freely falling object will reach the bottom point in the least amount of time.

continuous function: A function that has no breaks or gaps in its graph; the graph of a continuous function can be drawn without lifting the pen.

cosine: A function of angle θ giving the ratio of the length of the adjacent side to the length of the hypotenuse of a right triangle, as well as the horizontal coordinate of a point on the circle of radius 1 corresponding to the angle θ.

delta (Δ): The Greek letter capital delta is used in such expressions as Δt or Δx to denote a small change in the varying quantity. We should think of the Δ as shorthand for "difference." In the definition of the integral, Δx or Δt appears in the long sums used to define the integral. The Δ is transformed into a *dx* or *dt* in the integral symbol to remind us of the origins of the integral as a sum.

derivative: Mathematical description of (instantaneous) rate of change of a function. Characterized geometrically as the slope of the tangent line to the graph of the function. The derivative of a function $f(x)$ is written $f'(x)$ or $\frac{d}{dx}(f(x))$ and is formally defined as $\lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x}$.

differentiable function: A function whose derivative exists at every point where the function is defined; a continuous function without kinks or cusps.

differential equation: An equation involving a function and its derivatives; a solution of a differential equation is a family of functions.

direction field: A two-dimensional field of arrows indicating the slope of the tangent line at each given point for a curve that is a solution of a differential equation.

directional derivative: The rate of change of a function of several variables in the direction of a given vector.

epicycles: A circle on a circle. For centuries before Kepler, people believed that planets' orbits were circular, but because that image did not accord with observation, the planets were viewed as revolving around on little circles whose centers were going in circles. The smaller circles with centers on circles are epicycles.

exhaustion (the Greek method of exhaustion): A geometric technique by which formulas for areas of different shapes could be verified through finer and finer approximations.

Fourier Series: An infinite series of sines and cosines, typically used to approximate a function.

function: Mathematical description of dependency. A rule or correspondence that provides exactly one output value for each input value. Often written algebraically as $f(x)$ = formula involving x (for example, $f(x) = x^2$).

Fundamental Theorem of Calculus: The most important theorem in calculus. Demonstrates the reciprocal relationship between the derivative and the integral.

graph: A geometric representation of a function, showing correspondences via pairs of points (input, output) drawn on a standard Cartesian (x-y) plane.

heat equation: A partial differential equation governing the heating and cooling of objects; it relates the second derivative of position to the first derivative of time.

infinite series (also called an infinite sum): The sum of an infinite collection of numbers. Such a series can sum to a finite number or "diverge to infinity"; for example, $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+...$ sums to 1, but $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+...$ does not sum to a finite number.

integer: A whole number (positive, zero, or negative); …–2, –1, 0, 1, 2,…

integral: Denoted $\int_a^b v(t)dt$. If we think of function *v*(*t*) as measuring the velocity of a moving car at each time *t*, then the integral is a number that is equal to the distance traveled, because the integral is obtained by dividing the time from *a* to *b* into small increments and approximating the distance traveled by assuming that the car went at a steady speed during each of those small increments of time. By taking increasingly smaller increments of time, approximations converge to a single answer, the integral. This sounds complicated, but the naturality of it is the topic of Lecture Three. The integral is also equal to the area under the graph of *v*(*t*) and above the *t*-axis. The integral is related to the derivative (as an inverse procedure) via the Fundamental Theorem of Calculus. See also **antiderivative**.

Law of Large Numbers: The theorem that the ratio of successes to trials in a random process will converge to the probability of success as increasingly many trials are undertaken.

limit: The result of an infinite process that converges to a single answer. Example: the sequence of numbers $1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5},...$ converges to the number 0.

maximum: The largest value of the outputs of a function. The *y*-value of the highest point on the graph of a function. It does not always exist.

minimum: The smallest value of the outputs of a function. The *y*-value of the lowest point on the graph of a function. It does not always exist.

Newton-Raphson Method: An iterative technique for finding solutions of an equation using graphs and derivatives.

parabola: A conic section defined as the set of all points equidistant between a point and a line.

paradox: Two compelling arguments about the same situation that lead to two opposite views. Zeno's paradoxes of motion give logical reasons why motion cannot occur. On the other hand, we experience motion. The opposite conclusions deduced from Zeno's logic versus our experience compose the paradox.

partial derivative: The rate of change of a quantity relative to the change of one of several quantities that are influencing its value when the other varying quantities remain fixed.

partial differential equation: An equation involving a function of several variables and its partial derivatives.

π (pi): Greek letter denoting the value 3.1415926… equal to the ratio of the circumference of a circle to its diameter.

probability: The quantitative study of uncertainty.

real number: Any decimal number.

sine: A function of angle θ giving the ratio of the length of the opposite side to the length of the hypotenuse of a right triangle, as well as the vertical coordinate of a point on the circle of radius 1 corresponding to the angle θ.

slope (of a straight line): The ratio of distance ascended to distance traversed, sometimes known as "rise over run."

smooth function: A function that is continuous and whose first derivative, second derivative, and so forth are all continuous.

tangent line: A straight line associated to each point on a curve. Just grazing the curve, the tangent line "parallels" the curve at a point.

variable: The independent quantity in a functional relationship. For example, if position is a function of time, time is the variable.

vector: An arrow indicating direction and magnitude (usually of motion in two-dimensional or three-dimensional space).

vector field: A field of arrows associating a vector to each point (x,y) in the two-dimensional plane; usually represented graphically.

velocity: Average velocity is total distance divided by the time it took to traverse that distance; units are length/time (for example, miles per hour). Instantaneous velocity is the speed at one moment of time, approximated by average velocity for smaller and smaller time intervals; units are also length/time. The instantaneous velocity is the derivative of the position function for a moving object.

wave equation: A partial differential equation governing the displacement of a string as well as more complex phenomena, such as electromagnetic waves; it relates the second derivative of position to the second derivative of time.

Biographical Notes

Archimedes (c. 287–212 B.C.). Ancient Greek mathematician, physicist, astronomer, inventor, and prolific author of scientific treatises. He studied hydrostatics and mechanics and discovered the general principle of the lever, how to compute tangents to spirals, the volume and surface area of spheres, the volume of solids of revolution, many applications of the method of exhaustion, and an approximation of the value of π, among other work. Archimedes was killed by a Roman soldier when Syracuse was conquered during the Second Punic War.

Barrow, Isaac (1630–1677). Lucasian professor of mathematics at Cambridge. In 1669, Barrow resigned from his chair to give Newton the professorship. He contributed to the development of integral calculus, particularly through the recognition of its inverse relationship with the tangent. He published works in optics and geometry and edited the works of ancient Greek mathematicians, including Euclid and Archimedes.

Bernoulli, Daniel (1700–1782). Swiss professor of mathematics at St. Petersburg and at Basel. He is best known for work in fluid dynamics (the *Bernoulli principle* is named for him) and is also known for work in probability. He was the son of Jean Bernoulli.

Bernoulli, Jacques (often called Jakob or James) (1654–1705). Professor of mathematics at Basel and a student of Leibniz. He studied infinite series and was the first to publish on the use of polar coordinates (the *lemniscate of Bernoulli* is named for him). He formulated the Law of Large Numbers in probability theory and wrote an influential treatise on the subject. Together, Jacques and brother Jean were primarily responsible for disseminating Leibniz's calculus throughout Europe.

Bernoulli, Jean (often called Johannes or John) (1667–1748). Swiss mathematician. He was professor of mathematics at Groningen (Holland) and Basel (after the death of his brother Jacques). He was a student of Leibniz and applied techniques of calculus to many problems in geometry and mechanics. He proposed the famous Brachistochrone problem as a challenge to other mathematicians. Jean Bernoulli was the teacher of Euler and L'Hôpital (who provided Jean a regular salary in return for mathematical discoveries, including the well-known *L'Hôpital's Rule*).

Buffon, Georges Louis Leclerc, Comte de (1707–1788). French naturalist and author of *Histoire naturelle*. He translated Newton's *Method of Fluxions* into French. He formulated the Buffon's Needle problem, linking the study of probability to geometric techniques.

Cauchy, Augustin Louis (1789–1857). Prolific French mathematician and engineer. He was professor in the Ecole Polytechnique and professor of mathematical physics at Turin. He worked in number theory, algebra, astronomy, mechanics, optics, and elasticity theory and made great contributions to analysis (particularly the study of infinite series and of complex variable theory) and the calculus of variations. He improved the foundations of calculus by refining the definitions of limit and continuity.

Cavalieri, Bonaventura (1598–1647). Italian mathematician; professor at Bologna; student of Galileo. He developed the method of indivisibles that provided a transition between the Greek method of exhaustion and the modern methods of integration of Newton and Leibniz. He applied his method to solve a majority of the problems posed by Kepler.

Descartes, René (1596–1650). French mathematician and philosopher. He served in various military campaigns and tutored Princess Elizabeth (daughter of Frederick V) and Queen Christina of Sweden. Descartes developed crucial theoretical links between algebra and geometry and his own method of constructing tangents to curves. He made substantial contributions to the development of analytic geometry.

Euler, Leonhard (1707–1783). Swiss mathematician and scientist. Euler was the student of Jean Bernoulli. He was professor of medicine and physiology and later became a professor of mathematics at St. Petersburg. Euler is the most prolific mathematical author of all time, writing on mathematics, acoustics, engineering, mechanics, and astronomy. He introduced standardized notations, many now in modern use, and contributed unique ideas to all areas of analysis, especially in the study of infinite series. He lost nearly all his sight by 1771 and was the father of 13 children.

Fermat, Pierre de (1601–1665). French lawyer and judge in Toulouse; enormously talented amateur mathematician. Fermat worked in number theory, geometry, analysis, and algebra and was the first developer of analytic geometry, including the discovery of equations of lines, circles, ellipses, parabolas, and hyperbolas. He wrote *Introduction to Plane and Solid Loci* and formulated the famed *Fermat's Last Theorem* as a note in the margin of his copy of Bachet's *Diophantus*. He developed a procedure for finding maxima and minima of functions through infinitesimal analysis, essentially by the limit definition of derivative, and applied this technique to many problems, including analyzing the refraction of light.

Fourier, Jean Baptiste Joseph (1768–1830). French mathematical physicist and professor in the Ecole Polytechnique. Fourier accompanied Napoleon on his campaign to Egypt, was appointed secretary of Napoleon's Institute of Egypt, and served as prefect of Grenoble. He carried out extensive studies in heat propagation, which form the foundation of modern partial differential equations with boundary conditions. He developed the *Fourier Series*, which represents functions by infinite (trigonometric) series.

Galilei, Galileo (1564–1642). Italian mathematician and philosopher; professor of mathematics at Pisa and at Padua. He invented the telescope (after hearing of such a device) and made many astronomical discoveries, including the existence of the rings of Saturn. He established the first law of motion, laws of falling bodies, and the fact that projectiles move in parabolic curves. Galileo made great contributions to the study of dynamics, leading to consideration of infinitesimals (eventually formalized in the theory of calculus). He advocated the Copernican heliocentric model of the solar system and was subsequently placed under house arrest by the Inquisition.

Gauss, Carl Friedrich (1777–1855). German mathematician; commonly considered the world's greatest mathematician, hence known as the Prince of Mathematicians. He was professor of astronomy and director of the observatory at Göttingen. Gauss provided the first complete proof of the Fundamental Theorem of Algebra and made substantial contributions to geometry, algebra, number theory, and applied mathematics. He established mathematical rigor as the standard of proof. His work on the differential geometry of curved surfaces formed an essential base for Einstein's general theory of relativity.

Green, George (1793–1841). Most famous for his theorem known as *Green's Theorem*. He worked in his father's bakery for most of his life and taught himself mathematics from books. He proved his famous theorem in a privately published book that he wrote to describe electricity and magnetism. Green did not attend college until he was 40. He had seven children (all with the same woman, Jane Smith) but never married. He never knew the importance of his work, but it and its consequences have been described as "leading to the mathematical theories of electricity underlying 20th-century industry."

Kepler, Johannes (1571–1630). German astronomer and mathematician; mathematician and astrologer to Emperor Rudolph II (in Prague). Kepler assisted Tycho Brahe (the Danish astronomer) in compiling the best collection of astronomical observations in the pre-telescope era. He developed three laws of planetary motion and made the first attempt to justify them mathematically. They were later shown to be a consequence of the universal law of gravitation by Newton, applying the new techniques of calculus.

Lagrange, Joseph Louis (1736–1813). French mathematician; professor at the Royal Artillery School in Turin, at the Ecole Normale and the Ecole Polytechnique in France, and at the Berlin Academy of Sciences. He studied algebra, number theory, and differential equations and unified the theory of general mechanics. He developed classical results in the theory of infinite series and contributed to the analytical foundation of the calculus of variations.

Laplace, Pierre Simon de (1749–1827). French mathematician and astronomer; professor at the Ecole Normale and the Ecole Polytechnique. Laplace was the author of the influential *Mécanique Céleste* that summarized all contributions to the theory of gravitation (without credit to its contributors). He developed potential theory, important in the study of physics, and partial differential equations. He made great contributions to probability theory based on techniques from calculus.

Leibniz, Gottfried Wilhelm von (1646–1716). German diplomat, logician, politician, philosopher, linguist, and mathematician; president of the Berlin Academy. Leibniz is regarded, with Newton, as a co-inventor of calculus. He was the first to publish a theory of calculus. Leibniz's notation is used currently. He made substantial contributions to formal logic, leading to the establishment of symbolic logic as a field of study. He discovered an infinite series formula for $\pi/4$. He was accused of plagiarism by British partisans of Newton, and his supporters counterclaimed that Newton was the plagiarist. Now, he is acknowledged to have independently discovered calculus.

L'Hôpital, Guillaume François Antoine (1661–1704). French marquis, amateur mathematician, and student of Jean Bernoulli. L'Hôpital provided one of the five submitted solutions to Bernoulli's Brachistochrone problem. He was the author of the first calculus textbook (1696), written in the vernacular and based primarily on the work of Jean Bernoulli. This text went through several editions and greatly aided the spread of Leibniz's calculus on the Continent.

Lotka, Alfred J. (1880–1949). An American biophysicist and the father of mathematical biology; he published the first book in this field in 1924. With Vito Volterra, he is chiefly known for the formulation of the Lotka-Volterra equations for the study of predator-prey models in population dynamics.

Newton, Sir Isaac (1642–1727). Great English mathematician and scientist; Lucasian professor of mathematics at Cambridge. Newton was the first discoverer of differential and integral calculus. He formulated the law of universal gravitation and his three laws of motion, upon which classical physics is based. In 1687, he published his results in *Philosophiae Naturalis Principia Mathematica*. He formulated the theory of colors (in optics) and proved the binomial theorem. He is possibly the greatest genius of all time. Newton was a Member of Parliament (Cambridge), long-time president of the Royal Society, and Master of the Mint. The controversy with Leibniz over attribution of the discovery of calculus poisoned relations between British and Continental scientists, leading to the isolation of British mathematicians for much of the 18th century.

Riemann, Georg Friedrich Bernhard (1826–1866). German mathematician; professor of mathematics at Göttingen. He made great contributions to analysis, geometry, and number theory and both extended the theory of representing a function by its Fourier series and established the foundations of complex variable theory. Riemann developed the concept and theory of the Riemann integral (as taught in standard college calculus courses) and pioneered the study of the theory of functions of a real variable. He gave the most famous job talk in the history of mathematics, in which he provided a mathematical generalization of all known geometries, a field now called Riemannian geometry.

Weierstrass, Karl (1815–1897). German mathematician. He left the University of Bonn without a degree and taught secondary school for more than 12 years while independently studying analysis. His 1854 paper led to a professorship in Berlin. He provided a rigorous definition of the limit, thus placing calculus at last on solid mathematical ground, and he provided a precise definition of real numbers.

Zeno of Elea (c. 495–430 B.C.). Ancient Greek dialectician and logician. He is noted for his four paradoxes of motion. He was a student of Parmenides, whose school of philosophy rivaled that of the Pythagoreans.

Bibliography

Readings

Bardi, Jason Socrates. *The Calculus Wars: Newton, Leibniz, and the Greatest Mathematical Clash of All Time*. New York: Thunder's Mouth Press, 2006. This book tells the story of the controversy between Newton and Leibniz and their supporters over who should receive credit for the discovery or invention of calculus. The fact that a book on this topic for the general reader is published in 2006 is a testament to both the significance of calculus and the level of rancor of the dispute.

Bell, E. T. *Men of Mathematics*. New York: Simon & Schuster, 1937. A classic of mathematics history, filled with quotes and stories (often apocryphal) of famous mathematicians.

Berlinski, David. *A Tour of the Calculus*. New York: Pantheon Books, 1995. This book is written in a flowery manner and gives the nonmathematician a journey through the ideas of calculus.

Blatner, David. *The Joy of π*. New York: Walker Publishing Company, 1997. This fun little paperback is filled with gems and details about the history of the irrational number π.

Boyer, Carl B. *The History of the Calculus and Its Conceptual Development*. Mineola, NY: Dover Publications, 1959. A scholarly history of calculus from ancient times through the 19th century.

———. *A History of Mathematics*. Princeton: Princeton University Press, 1968. An extensive survey of the history of mathematics from earliest recorded history through the early 20th century. Each chapter includes a good bibliography and nice exercises.

Burger, Edward B., and Michael Starbird. *The Heart of Mathematics: An invitation to effective thinking*. Emeryville, CA: Key College Publishing,

2000. This award-winning book presents deep and fascinating mathematical ideas in a lively, accessible, and readable way. The review in the June–July 2001 issue of the *American Mathematical Monthly* says, “This is very possibly the best ‘mathematics for the non-mathematician’ book that I have seen—and that includes popular (non-textbook) books that one would find in a general bookstore.”

———. *Coincidences, Chaos, and All That Math Jazz: Making Light of Weighty Ideas*. New York: W.W. Norton & Co., 2005. This book fuses a professor’s understanding of rigorous mathematical ideas with the distorted sensibility of a stand-up comedian. It covers many beautiful topics in mathematics that are only touched on in this course, such as probability, chaos, and infinity. “Informative, intelligent, and refreshingly irreverent,” in the words of author Ian Stewart.

Cajori, Florian. *A History of Mathematics*, 5th ed. New York: Chelsea Publishing Co., 1991 (1st ed., 1893). A survey of the development of mathematics and the lives of mathematicians from ancient times through the end of World War I.

———. “History of Zeno’s Arguments on Motion.” *American Mathematical Monthly*. Vol. 22, Nos. 1–9 (1915). Cajori presents a philosophical and mathematical discussion of the meaning of Zeno’s four paradoxes of motion.

Calinger, Ronald. *A Contextual History of Mathematics*. Upper Saddle River, NJ: Prentice-Hall, 1999. This modern, readable text offers a survey of mathematics from the origin of numbers through the development of calculus and classical probability. It includes a nice section on the Bernoulli brothers.

Churchill, Winston Spencer. *My Early Life: A Roving Commission* (available through out-of-print bookstores only). Winston Churchill won the Nobel Prize in literature. The writing in this autobiography of his early life is absolutely delightful. We refer to only a few pages about his struggles with mathematics, but the whole book is a joy.

Davis, Donald M. *The Nature and Power of Mathematics*. Princeton: Princeton University Press, 1993. This wide-ranging book does not study the history of calculus in particular; rather, it describes an array of ideas from all areas of mathematics. It includes brief biographies of Gauss and Kepler, among other mathematicians.

Dunham, William. *Journey through Genius: The Great Theorems of Mathematics*. New York: John Wiley & Sons, 1990. Each of this book's 12 chapters covers a great idea or theorem and includes a brief history of the mathematicians who worked on that idea. Mathematicians discussed include Archimedes, Newton, the Bernoullis, and Euler.

Eves, Howard. *Great Moments in Mathematics (after 1650)*. Washington, DC: The Mathematical Association of America, 1981. This collection of lectures includes four entertaining chapters relevant to the development of calculus. These begin with the invention of differential calculus and conclude with a discussion of Fourier series.

Goodstein, David L., Judith R. Goodstein, and R. P. Feynman, *Feynman's Lost Lecture: The Motion of Planets around the Sun*. New York: W.W. Norton & Co., 1996. This book and CD give Feynman's geometric explanation for elliptical orbits of planets, as well as a history of the problem from Copernicus and Kepler to the present day.

Kline, Morris. *Mathematics: A Cultural Approach*. Reading, MA: Addison-Wesley Publishing Co., 1962. This survey of mathematics presents its topics in both historical and cultural settings, relating the ideas to the contexts in which they developed.

Priestley, William M. *Calculus: An Historical Approach*. New York, Heidelberg, and Berlin: Springer-Verlag, 1979. Develops the standard theory of calculus through discussions of its historical growth, emphasizing the history of ideas rather than the history of events.

Schey, H. M. *Div, Grad, Curl, and All That: An Informal Text on Vector Calculus*, 4th ed. New York: W.W. Norton, 2005. This book has been a favorite introduction to ideas of vector calculus for more than 30 years. It is short and clearly written. It presents vector calculus in the context of electrostatics, so it is especially attractive to people who have a feel for the physics of electricity and magnetism.

Simmons, George F. *Calculus with Analytic Geometry*. New York: McGraw-Hill, 1985. This college-level mathematics text provides a standard development of calculus along with appendices that include biographical notes and supplementary topics.

Thompson, Sylvanus P., and Martin Gardner. *Calculus Made Easy*, New York: St. Martin's Press, 1998. This book is a revision, by the great mathematical expositor Martin Gardner, of a classical exposition of calculus for the general public.

Standard Textbooks

There are dozens of standard calculus and differential equations textbooks, usually titled *Calculus* and *Differential Equations*. Two of them are:

Boyce, William E., and Richard C. DiPrima. *Elementary Differential Equations and Boundary Value Problems*, 8th ed. New York: John Wiley & Sons, 2005. This is the bestselling introductory differential equations textbook. It presents differential equations from the applied mathematicians' point of view and includes explanations of both the theory and applications, as well as many examples, applications, and exercises.

Stewart, James. *Calculus*, 5th ed. Belmont, CA: Brooks/Cole, 2003. The Stewart textbook is the best-selling calculus textbook in the United States. It is well-written and comprehensive and contains many worked examples, applications, and exercises.

Internet Resources

A History of the Calculus. School of Mathematics and Statistics, University of St. Andrews, Scotland. www-history.mcs.st-andrews.ac.uk/ history/ HistTopics/The_rise_of_calculus.html. This site presents a synopsis of the history of calculus from ancient times through the time of Newton and Leibniz and includes links to many more history sites.

Index of Biographies. School of Mathematics and Statistics, University of St. Andrews, Scotland. www-history.mcs.st-andrews.ac.uk/~history/BiogIndex.html. This site contains biographical articles on many of the world's mathematicians from ancient times to the present. Both chronological and alphabetical indexes are presented, as well as such categories as famous curves, history topics, and so forth.

The Math Forum @ Drexel. Drexel School of Education, Drexel University. www.mathforum.org/library/topics/svcalc/. Includes links to many other sites that contain articles and demonstrations of concepts in calculus.

Mnatsakanian, Mamikon. *Visual Calculus by Mamikon*. California Institute of Technology. www.its.caltech.edu/~mamikon/calculus.html. This site offers animations demonstrating some of Mnatsakanian's applications of his clever sweeping tangent method for solving calculus problems.

Predator-Prey Models. Department of Mathematics, Duke University. www.math.duke.edu/education/webfeats/Word2HTML/Predator.html. This site provides an interactive location for creating direction fields and solutions to differential equation predator-prey models.

Weisstein, Eric. *Wolfram MathWorld*, *The Web's Most Extensive Mathematics Resource*. mathworld.wolfram.com. This website is like a mathematical encyclopedia. If you come across a mathematical term or concept, no matter how trivial or how involved, it's likely to be described here.