Introduction to the optimization theory
(A FEW WORDS ABOUT OPTIMIZATION)

Shpilev Petr Valerievich
Faculty of Mathematics and Mechanics, SPbU

September, 2025

Санкт-Петербургский государственный университет

Comments

This is an introductory lecture on the course **Mathematical Modeling and Optimization**.

We will briefly discuss the main sections of the course and talk about how optimization methods have developed, as well as consider several well-known optimization problems of the past.

Part 1: Classical Optimization

- Optimality conditions, Taylor approximations
- Descent algorithms (gradient, Newton-type, trust-region)
- Constrained optimization, Karush–Kuhn–Tucker conditions
- Linear and quadratic programming

Part 2: Numerical & Heuristic Methods

- Methods for problems without derivatives
- Stochastic algorithms (SGD, simulated annealing)
- Population-based algorithms (genetic, particle swarm)
- Discrete optimization methods

Part 3: Model Construction

- Regression analysis: building estimates of model parameters
- Optimal experimental design to increase accuracy

Comments

I'd like to present my course, "Mathematical Modeling and Optimization", to you. It consists of 3 parts:

The first part focuses on the classical theory of optimization. We begin with smooth unconstrained problems—studying optimality conditions, Taylor-based approximations, and descent algorithms such as gradient methods and Newton-type methods. We then move to trust-region strategies and conjugate gradient methods, which are especially useful in large-scale problems. After that, we study constrained optimization, introducing the Karush–Kuhn–Tucker conditions, second-order theory, and duality. Finally, we explore linear and quadratic programming in depth, including modern active-set strategies and some alternatives. This part lays the theoretical and algorithmic foundation of classical optimization approaches.

The second part is devoted to numerical and heuristic optimization methods. When derivatives are unavailable or unreliable — or when the objective is noisy or nonconvex — we turn to alternative strategies. We begin with direct search methods, then move to stochastic techniques like Stochastic Gradient Descent, simulated annealing, and the cross-entropy method. Next, we explore population-based algorithms, including genetic algorithms, differential evolution, and particle swarm optimization. Finally, we study discrete optimization problems using methods like branch and bound, dynamic programming, and ant colony optimization. This part equips you with practical tools for solving challenging real-world problems.

The third part introduces the principles of regression modeling and optimal experimental design. Many optimization problems rely on expensive function evaluations. To address this, we use surrogate models: smooth, low-cost approximations of the true objective function. These models are fast to evaluate and optimize, guiding us toward the true optimum efficiently. By some sampling of the real function, we refine the surrogate, improving its accuracy over time. Beyond optimization, this approach is powerful for descriptive regression, where we study how variables interact. It helps uncover relationships between inputs (e.g., material properties or hyperparameters) and outputs (e.g., performance or efficiency)—without strict assumptions about data distributions. So the third part of our course dives into regression analysis basics and optimal experiment design – core techniques for building and refining models.

However, in today's introductory lecture, I'd like to take a step back and explore how the field of optimization emerged in mathematics—to shed some light on its historical context. We'll consider a few famous optimization problems that captivated the minds of past mathematicians.

Origins of Optimization

- **Euclid (c. 300 BCE):** Developed early geometric algorithms for optimality.
- **Archimedes (287–212 BCE):** Solved extremal problems in geometry and physics.
- **Key Themes:**
  - Finding the shortest path (Euclid)
  - Maximizing area with fixed perimeter (Archimedes)

Applications

- Architecture and land measurement
- Astronomy and early physics
- Engineering problems (e.g., levers and machines)

Comments

In life, we often need to make the best possible decision—what some people call the "optimal" choice. Many of these problems come up in economics and engineering. When that happens, math can be really helpful. In math, the study of finding the highest and lowest values—called maximum and minimum—started a long time ago, about 25 centuries back. Let's take a brief look at how optimization methods grew, from their early days to today's ideas, including heuristic and metaheuristic methods.

If we talk about the European tradition, the story of optimization begins in Ancient Greece with mathematicians like Euclid (around 300 BCE) and Archimedes (287–212 BCE). They started solving these problems using geometry.

• Euclid worked on things like finding the shortest path between two points, which was an early version of optimization ideas.

• Archimedes showed that a circle gives the biggest area for a set perimeter, connecting optimization to shapes in nature. Back then, their methods were simple and based on drawing shapes. There wasn't one big way to solve everything—each problem was figured out on its own, often for practical stuff like splitting land or building things.

17th Century: Calculus Breakthrough

- **Isaac Newton (1643–1727)** and **Gottfried Leibniz (1646–1716)** developed differential calculus independently.
- Enabled solving extremum problems (maxima and minima).
- Key idea: slopes of tangents used to locate optimal values.

18th Century: Variational Calculus

- **Leonhard Euler (1707–1783)** and **Joseph-Louis Lagrange (1736–1813)** introduced calculus of variations.
- Euler–Lagrange equation (1750s) solved the brachistochrone problem.
- Applications:
  - **Physics:** e.g., planetary motion
  - **Engineering:** e.g., optimal structural design

Comments

The next big leap in optimization came in the 17th century with the invention of calculus. Two great mathematicians, Isaac Newton and Gottfried Leibniz, independently developed the foundations of differential and integral calculus. One of their key contributions was the ability to analyze rates of change, which is essential for solving extremum problems—finding maximum and minimum values of functions. For example, if you want to find where a function reaches its highest or lowest value, you examine the slope of its tangent line. When the slope is zero, the function may be at a local maximum or minimum. This idea remains fundamental in modern optimization theory.

In the 18th century, calculus evolved further into a new branch called variational calculus. This branch focuses on optimizing entire functions or curves, rather than just individual points.

Two names dominate this field: Leonhard Euler and Joseph-Louis Lagrange. In the 1750s, they formulated what is now known as the Euler–Lagrange equation, which helps solve problems where the goal is to find a path or shape that minimizes (or maximizes) a certain quantity.

One famous example is the brachistochrone problem—which curve gives the shortest descent time for a particle moving under gravity? This question was asked by the famous mathematician Johann Bernoulli in 1696. The solution involves a deep understanding of both geometry and dynamics.

These tools found powerful applications: in physics, for predicting planetary motion; and in engineering, for designing optimal structures.

Linear Programming (1940s)

- **George Dantzig (1914–2005)** introduced the Simplex Method in 1947.
- Solves problems with linear objectives and linear constraints.
- Widely used for **resource allocation**, scheduling, logistics.

Convex Analysis (1930s–1950s)

- **John von Neumann (1903–1957)** and others developed the theory of convex sets and functions.
- Convexity ensures that every local minimum is a **global minimum**.
- Applications:
  - **Military and logistics:** e.g., optimal planning
  - **Economics:** e.g., production and cost optimization

Comments

The 20th century marked a turning point in optimization with the development of linear programming and convex analysis.

Let's start with linear programming. In 1947, George Dantzig introduced one of the most important tools in applied mathematics—the Simplex Method. It is used to find the best outcome for problems with linear objective functions under linear constraints. For example, how to allocate limited resources—like money, time, or materials—in a way that maximizes profit or minimizes cost.

The power of the Simplex Method is that it works efficiently in practice, even for large-scale industrial problems. Linear programming quickly became central in fields like operations research, logistics, and military planning.

Meanwhile, between the 1930s and 1950s, researchers including John von Neumann laid the foundation for convex analysis. This branch of mathematics studies convex sets and convex functions. A key property of convexity is that any local minimum is also a global minimum—a crucial advantage when solving optimization problems.

Thanks to this property, convex analysis provides a reliable theoretical framework for designing algorithms that are both efficient and robust. It underpins modern optimization theory and appears everywhere—from economic modeling to machine learning.

Together, linear programming and convex analysis form the backbone of modern optimization. They are the basis for many algorithms still used today.

Optimal Control Theory (1950s–1960s)

- Richard Bellman (1920–1984): Developed dynamic programming (1953).
- Lev Pontryagin (1908–1988): Formulated the Maximum Principle (1956).
- Applications: Aerospace (e.g., optimal rocket trajectories).

Heuristics and Metaheuristics

- Heuristics (Late 20th Century): Simple, problem-specific rules (e.g., greedy algorithms).
- Metaheuristics (1980s–Present):
  - Genetic Algorithms — John Holland (1975)
  - Simulated Annealing (1983)
  - Ant Colony Optimization (1992)
  - Particle Swarm Optimization (1995)
- Applications: AI, operations research, big data optimization.

Comments

This slide summarizes key developments in optimization methods in the second half of the 20th century. First, in the 1950s and 60s, Optimal Control Theory emerged. Richard Bellman introduced Dynamic Programming in 1953 — a method that breaks a complex problem into simpler subproblems, solving each only once and storing the results. This laid the foundation for many recursive algorithms in control and planning.

Meanwhile, Lev Pontryagin formulated the Maximum Principle in 1956 — a necessary condition for optimality in dynamic systems, playing a similar role to the Euler–Lagrange equations in the calculus of variations. These methods became central in aerospace engineering, for instance in the computation of optimal rocket trajectories.

Later, as exact methods became impractical for large or combinatorial problems, Heuristics and Metaheuristics gained popularity. Heuristics are simple, often greedy rules tailored to specific problems. Metaheuristics, developed from the 1980s onward, are more general frameworks that aim to escape local minima. Remarkable examples include Genetic Algorithms developed by John Holland in 1975, Simulated Annealing introduced in 1983, Ant Colony Optimization proposed in 1992, and Particle Swarm Optimization from 1995. These methods are widely used in artificial intelligence, operations research, and big data optimization.

Heuristics are quick tricks made for specific problems. For example: A greedy algorithm picks the best choice at every step, which often works okay for things like the traveling salesman problem.

Optimal Control Theory (1950s–1960s)

- Richard Bellman (1920–1984): Developed dynamic programming (1953).
- Lev Pontryagin (1908–1988): Formulated the Maximum Principle (1956).
- Applications: Aerospace (e.g., optimal rocket trajectories).

Heuristics and Metaheuristics

- Heuristics (Late 20th Century): Simple, problem-specific rules (e.g., greedy algorithms).
- Metaheuristics (1980s–Present):
  - Genetic Algorithms — John Holland (1975)
  - Simulated Annealing (1983)
  - Ant Colony Optimization (1992)
  - Particle Swarm Optimization (1995)
- Applications: AI, operations research, big data optimization.

Comments

Metaheuristics are universal strategies that guide heuristics for more efficient exploration of the solution space. They help to avoid local optima and are applicable to a wide range of tasks. Among the most famous:

• Genetic Algorithms (GA): They copy how nature evolves, improving a group of solutions by mixing and changing them. Great for tricky combination problems.

• Simulated Annealing (SA): Based on cooling metal, it sometimes picks worse options to escape small traps and find better answers.

• Particle Swarm Optimization (PSO): Copies how birds or fish move together, with particles learning from their own and others' experiences.

• Ant Colony Optimization (ACO): Inspired by ants leaving trails, it finds good paths by following "pheromone" clues.

These methods don't promise the perfect answer, but they're super useful. They help with supply chains, tuning AI networks, and planning schedules.

The history of optimization goes from ancient shape-based guesses to today's strong computer algorithms. It started with Euclid and Archimedes, grew through variational calculus, linear programming, and optimal control, and now uses heuristics and metaheuristics. These modern tools mix math with computing power to solve the toughest 21st-century problems—like artificial intelligence and global shipping. Optimization is still a key tool in science, engineering, and economics, and it keeps growing with new challenges.

Why Do We Solve Maximum and Minimum Problems?

Philosophical Motivation

- "Nothing takes place in the world whose meaning is not that of some maximum or minimum."

  — L. Euler

- "Most practical questions can be reduced to problems of largest and smallest magnitudes ... and it is only by solving these problems that we can satisfy the requirements of practice which always seeks the best, the most convenient."

  — P. L. Cebysev

Key Idea

Optimization is at the heart of science, engineering, and daily life: we seek to maximize efficiency, minimize cost, and make optimal decisions under constraints.

Comments

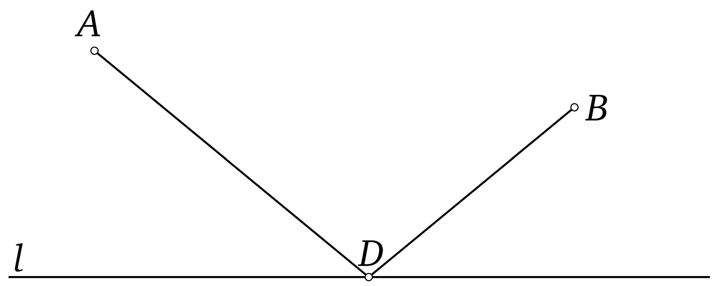
Leonhard Euler famously stated that "nothing takes place in the world whose meaning is not that of some maximum or minimum." This reflects a profound truth: nature tends toward extremal behavior — whether it's the shortest path, the least energy configuration, or the most stable state.

Similarly, Pafnuty Chebyshev emphasized that "most practical questions can be reduced to problems of largest and smallest magnitudes," pointing to the ubiquity of optimization in applied settings — from engineering design to economic policy. In practice, we always seek the best solution under given circumstances — whether it's the most efficient algorithm, the cheapest production plan, or the fastest trajectory.

This is why optimization is more than just a mathematical curiosity — it is a universal principle that governs both natural and artificial systems.

Problem Statement (Heron of Alexandria, 1st century CE)

Given two points A and B on the same side of a line, find a point D on the line such that the total distance AD + DB is minimized.



Comments

Now let's talk a little bit about the famous optimization problems of the past. We are not considering these tasks so that you can learn some new methods. I want you to feel the "spirit of optimization" that allowed scientists of the past to move forward, developing more and more effective approaches to solving a key problem: how to achieve your goal in the most effective way, since this is the idea that underlies any "optimization". In fact, we can say that any science is devoted to comprehending this idea in a certain aspect. And mathematics in this sense can be defined as the "art of optimization." That is why one of the greatest European philosophers, Immanuel Kant, argued that there is exactly as much science in every science as there is mathematics in it.

Thus, having comprehended this basic principle, you will be able to achieve success in any field of activity, since optimization methods are just a technique for implementing this principle. And with a deep understanding of it, you can work out the appropriate techniques yourself.

In our course, we mainly consider the named techniques, which can also be very useful in solving specific applied problems. However, the main purpose of our course is to provide you with an understanding of the fundamentals, an understanding of the essence of the "optimization principle", i.e. the principle of "effective goal achievement".

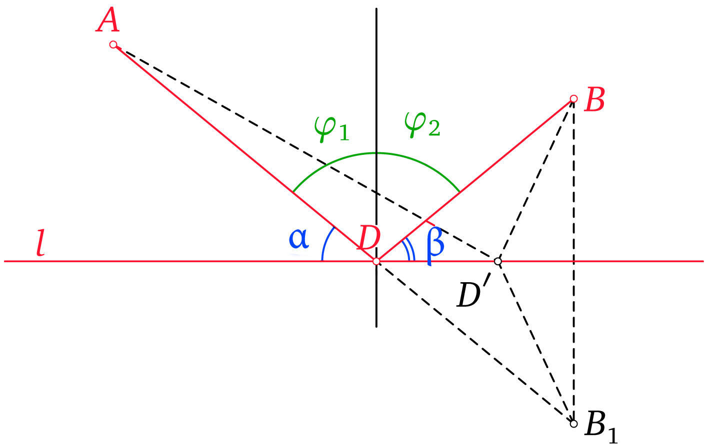
So, taking into account the above, let's look at some of the most famous optimization problems of the past. Let's start with one well-known task that you might have encountered at school. A and B are two given points on the same side of a line. Find a point D on such that the sum of the distances from A to D and from D to B is a minimum.

The presumed author is the famous ancient mathematician Heron of Alexandria. We all know about Heron through the formula for the area of a triangle that bears his name. The book containing this problem is titled «On mirrors». Scholars disagree as to when this book was written, but most believe that it was written in the first century A.D. Although Heron's book has disappeared, we know about it from later commentaries.

$$|AD'| + |D'B| = |AD'| + |D'B_1| > |AB_1| = |AD| + |DB_1|. \quad (1)$$

The symmetry properties imply:

$$|DB| = |DB_1|, \ |D'B| = |D'B_1|,$$
$$|AD'| + |D'B_1| > |AB_1|.$$



Comments

Let's recall the solution of Heron's problem. Let B1 be the point symmetric to B with respect to the line l. Join A to B_1. The required point D is the point of intersection of AB_1 and l. Indeed, if D' is a point other than D, then

(1)

$$[AD'| + |D'B| = |AD'| + |D'B_1| > |AB_1| = |AD| + |DB|$$

.

In establishing (1) we made use of symmetry properties that imply the equalities $|DB| = |DB_1|$, $|D'B| = |D'B_1|$, and the triangle inequality |AD'| +|D'B1| > |AB1| . This completes the solution of the problem.

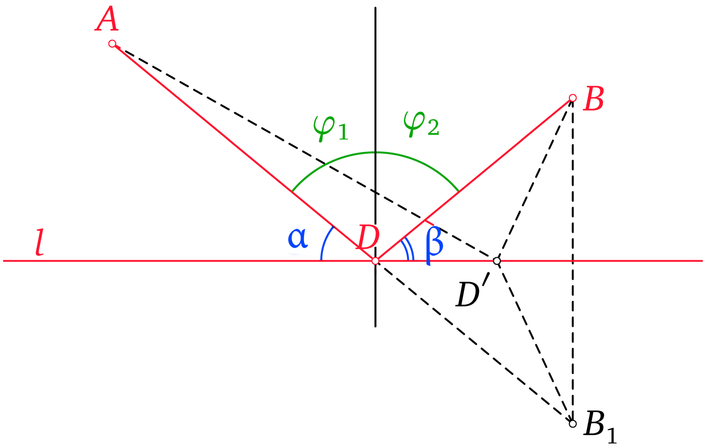
We note that the required point D has the property that the angle α is equal to the angle β. Also, the angle φ_1 is equal to the angle φ_2, or, as is usually said, the angle of incidence is equal to the angle of reflection.

In Heron's time scholars tried to comprehend the laws of nature by speculation and logical arguments, without recourse to experiment. One of the first great experimenter in the history of science was Galileo Galilei, who lived in the seventeenth century. In contrast to Galileo, Heron tried to base his explanations of the laws of reflection on logical foundations. He seems to have assumed that nature pursues the shortest path. Damianus (sixth century A.D.), a commentator on Heron, says that Heron ... showed that lines inclined at equal angles are the smallest of all intermediate ones inclined on the same side of a single line. Proving this, he says that if nature does not want a ray of light to meander to no purpose, then it breaks it at equal angles.

$$|AD'| + |D'B| = |AD'| + |D'B_1| > |AB_1| = |AD| + |DB_1|. \quad (1)$$

The symmetry properties imply:

$$|DB| = |DB_1|, \ |D'B| = |D'B_1|,$$
$$|AD'| + |D'B_1| > |AB_1|.$$



Comments

Historians of science see in this the first hint of the thought that nature is guided by extremal principles. Heron's idea was developed further by Fermat (we will have more to say about this further). Fermat deduced the law of refraction of light (established earlier experimentally by Snel) from the assumption that what characterizes the trajectory of a light ray moving from one point to another in a nonhomogeneous medium is that it is traversed in a minimum of time. From that point on, the idea of the extremal character of natural phenomena became the guiding light of science. This is confirmed by Euler's words which we quoted earlier.

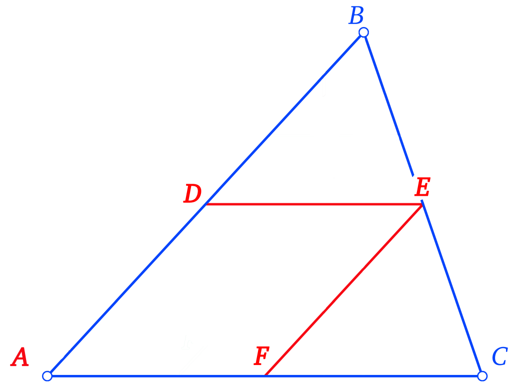
Nevertheless, what distinguishes the trajectories of light and radio waves, the motions of pendulums and planets, the flows of liquids and gases, as well as many other motions, is that they are all solutions of problems of maxima and minima. This fact provides a fruitful means of creating a mathematical description of nature.

Thus, to a first approximation, this is the basic reason that motivates us to solve problems about maxima and minima and develop a theory of extreme problems.

But the main reason, as we have already said, is something else. At all times, people have sought to minimize efforts to achieve their goals. But what if the goal is to "minimize your own efforts"? This is how mathematics actually came into being: its subject is not optimizing the achievement of some "external goals", but increasing your own skill, your own effectiveness. Therefore, studying and mastering mathematics is the most important step towards any person's main goal: self-knowledge. For, as the most famous inscription on the wall of the ancient Greek temple of Apollo at Delphi says, where the Delphic Oracle was located: "Know yourself, and you will know the gods and the universe."

Euclid's problem. (From Euclid's Elements, 4-th century BCE)

In a given triangle ABC inscribe a parallelogram ADEF (EF||AB, DE||AC) of maximal area.

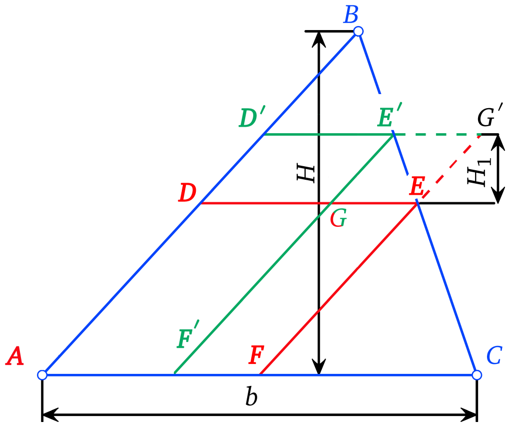


Comments

Euclid's problem. In Elements, the first scientific monograph and textbook in history (4th century BCE), Euclid posed only one maximization problem. Its modern formulation is as follows: In a given triangle ABC, inscribe a parallelogram ADEF (with corresponding sides parallel to the corresponding sides of the triangle) having the greatest possible area.

We will consider one of the possible geometric solutions of this problem; it goes back to Euclid's solution in the Elements. Specifically, we will prove that the parallelogram of maximal area is uniquely characterized by the condition that D, E, and F are the midpoints of the corresponding sides.

Let $AD'E'F'$ be a parallelogram inscribed in triangle ABC that is different from $ADEF$.
Let G' denote the point of intersection of lines $D'E'$ and EF, and G the intersection of DE and $E'F'$.

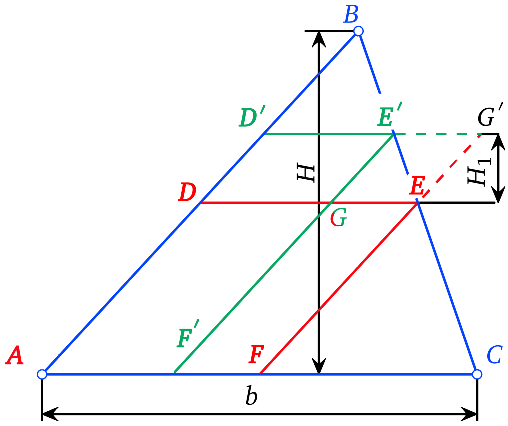


Comments

We aim to show that the area of parallelogram $AD'E'F'$ is less than that of $ADEF$ by exactly the area of parallelogram $EG'E'G$. To this end, we denote the length of the altitude drawn from vertex B of triangle ABC by H. We denote the length of the side AC by b, and the length of the altitude in triangle $GE'E$ drawn from point E' be H_1.

Triangles $\triangle GE'E$ and $\triangle ABC$ are similar, since $E'G \parallel AB$ and $GE \parallel AC$. Thus,

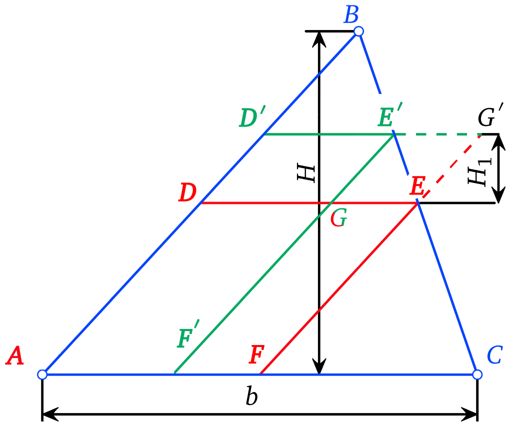
$$\frac{H_1}{|GE|} = \frac{H}{b} \quad \Leftrightarrow \quad H_1 \frac{b}{2} = |GE| \frac{H}{2}$$



Comments

In view of the similarity of the triangles $GE'E$ and ABC, we have $\frac{H_1}{|GE|} = \frac{H}{b}$, which is equivalent to an equality $H_1 \frac{b}{2} = |GE| \frac{H}{2}$

It follows that the area of parallelogram D′G′ED (with height H_1 and base b/2) is equal to that of EGF′F (with height H/2 and base |GE|).
Therefore, $S_{D'G'ED} = S_{EGF'F}$.

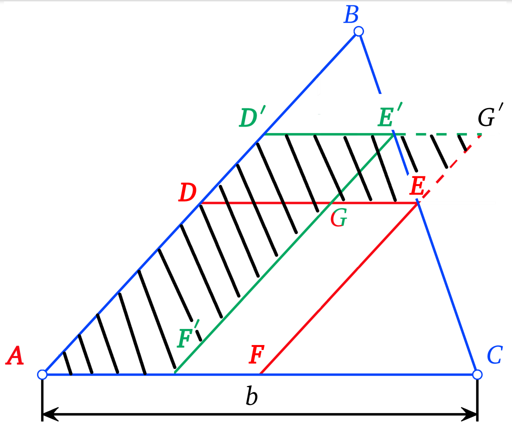


Comments

From this relation it follows that the area of parallelogram D′G′ED (with height H_1 and side DE of length b/2) equals the area of parallelogram EGF′F (with height H divided by two and side F′F of length |GE|).

The total area of ADEF equals the area of figure AD′G′EGF′, which is greater than that of AD′E′F′ by the area of parallelogram GE′G′E.

Conclusion: ADEF has greater area than any other such parallelogram ⇒ it is the maximal one.



Comments

It follows that the area of the parallelogram ADEF is equal to the area of the figure AD′G′EGF′ that is greater than the area of AD′E′F′ by the area of the parallelogram GE′G′E. This completes the solution of the problem.

A Little Immortal Poetry

There bought a space of ground, which Byrsa call'd,
From the bull's hide, they first inclos'd, and wall'd.
— The Aeneid of Vergil

Dido's problem

Among all closed plane curves of a given length, find the one that encloses the largest area.

A few comments:

- This problem also is known as the classical isoperimetric problem.
- Isoperimetric figures are those with equal perimeter.

Historical Reference:

- First rigorous proofs attributed to Hermann Amandus Schwarz.
- See also: W. Blaschke, Griechische und anschauliche Geometrie, München, 1953.

Comments

These two lines are from the Aeneid, the creation of one of the greatest poets of Ancient Rome, Publius Virgil Maron. Like all immortal creations, the Aeneid tells the story of human passions, of good and evil, of fate and suffering, of guile and love, of life and death. The quoted lines refer to an event that tradition placed in the ninth century BC.

Fleeing from persecution by her brother, the Phoenician princess Dido set off westward along the Mediterranean shore in search of a haven. A certain spot on the coast of what is now the bay of Tunis caught her fancy. Dido negotiated the sale of land with the local leader, Yarb. She asked for very little-as much as could be "encircled with a bull's hide." Dido managed to persuade Yarb, and a deal was concluded. Dido then cut a bull's hide into narrow strips, tied them together, and enclosed a large tract of land. On this land she built a fortress and, near it, the city of Carthage. There she was fated to experience unrequited love and a martyr's death. This incident suggests the question: How much land can be enclosed by a bull's hide?

To answer this question, we must pose it in a mathematically correct manner. A modem mathematician would say: Among all closed plane curves of a given length, find the one that encloses the largest area.

This question is known as Dido's problem, or the classical isoperimetric problem. (Isoperimetric figures are figures that have the same perimeter.) We will soon show that the curve that solves the classical isoperimetric problem is a circle. In describing Dido's actions, Vergil used the Latin word "circumdare" (to encircle) containing the root circus (circle). This suggests that Dido solved the classical isoperimetric problem correctly.

A Little Immortal Poetry

There bought a space of ground, which Byrsa call'd,
From the bull's hide, they first inclos'd, and wall'd.
— The Aeneid of Vergil

Dido's problem

Among all closed plane curves of a given length, find the one that encloses the largest area.

A few comments:

- This problem also is known as the classical isoperimetric problem.
- Isoperimetric figures are those with equal perimeter.

Historical Reference:

- First rigorous proofs attributed to Hermann Amandus Schwarz.
- See also: W. Blaschke, Griechische und anschauliche Geometrie, München, 1953.

Comments

Many historians are of the opinion that this was the first extremal problem discussed in the scientific literature. In addition to noting the isoperimetric property of the circle (that is, the property of the circle to enclose the largest area among all isoperimetric figures), ancient geometers also noted the isoepiphanic property of the sphere (that is, the property of the sphere to enclose the largest volume among all figures with the same surface area).

It is now impossible to tell when the thought of the maximal capacity of the circle and the sphere was first advanced. At any rate, Aristotle (4th century B.C.)-one of the most famous thinkers in European history-treats these facts as given. And who (other than Dido) did, in fact, solve the isoperimetric problem? The literature devoted to the isoperimetric property of the circle and the isoepiphanic property of the sphere is vast. One of the immense number of these works is by the German geometer Wilhelm Blaschke [2], which includes historical references.

Hermann Amandus Schwartz is thought to have given the first rigorous proofs of the maximum property of the circle and the sphere.

But in fact, Schwartz-and before him Weierstrass, and after him Blaschke himself, and numerous other mathematicians in the nineteenth and twentieth centuries-should be given credit (in connection with the isoperimetric problem) merely for shaping the ideas of their distant predecessors so as to meet the requirements of rigor of their time. The basic ways of solving the isoperimetric problem were already outlined with absolute correctness in ancient times. We will now describe one such way, due to Zenodorus, a mathematician who is thought to have lived sometime between the third century B.C. and the first century A.D.

Definition

A plane n-gon of largest area among all n-gons isoperimetric with it is called a maximal n-gon.

Zenodorus' Theorem

A maximal n-gon (if one exists) is regular.

Lemma 1

A maximal n-gon has equal sides.

Lemma 2

A maximal n-gon has equal angles.

Zenodorus' theorem follows from these two lemmas.

Comments

Zenodorus proves completely rigorously -by the standards of his time the following assertion.

STATEMENT: If there exists a plane n-gon having largest area among all n-gons of given perimeter, then it must have equal sides and equal angles.

In the interest of brevity, we will call a plane n-gon of largest area, among all n-gons isoperimetric with it, a maximal n-gon. Using this term we can state Zenodorus' theorem more briefly.

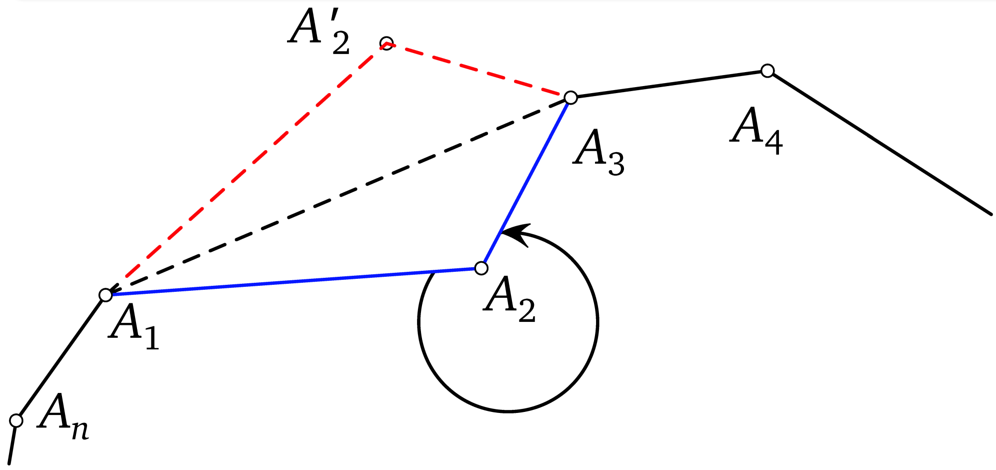
Zenodorus' Theorem: A maximal n-gon (if one exists) must be regular.

Zenodorus' theorem follows from two lemmas.

LEMMA 1. A maximal n-gon must have equal sides.

LEMMA 2. A maximal n-gon must have equal angles.

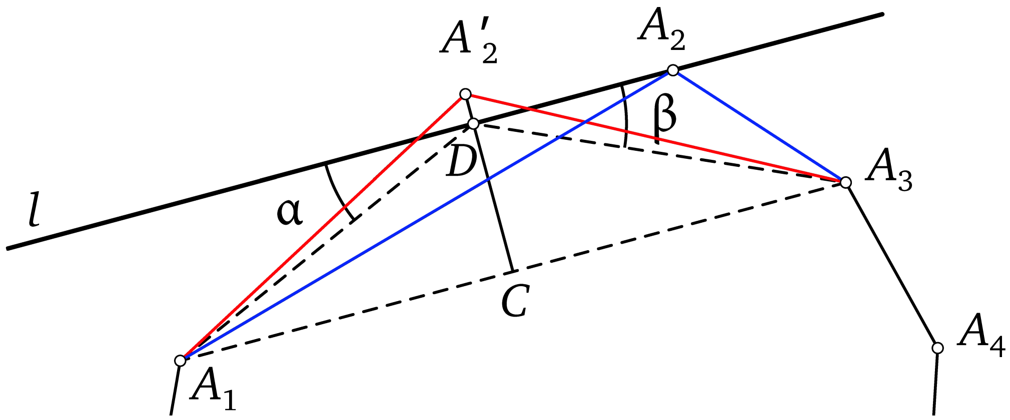
Suppose that the angle $A_1A_2A_3$, say, is larger than 180°. Let A_2' be the image of the vertex A_2 under reflection in the line A_1A_3. The polygon $A_1A_2'A_3\ldots A_n$ has greater area than the polygon $A_1A_2A_3\ldots A_n$ and is isoperimetric with it.



Comments

Before presenting the proofs, it is necessary to make an observation not mentioned by Zenodorus. As we are about to show, a nonconvex polygon cannot be maximal. Indeed, suppose that the angle A1A2A3, say, is larger than 180°. Let A'2 (A prime sub two) be the image of the vertex A2 under reflection in the line A1A3. The polygon A1A'2A3...An has greater area than the polygon A1A2A3...An and is isoperimetric with it. Now we are ready to give the proof of the lemmas.

(a) The area of $\triangle A_1DA_3$ equals the area of $\triangle A_1A_2A_3$, since they have equal altitudes and bases.
(b) The sum $A_1D + DA_3$ is less than $A_1A_2 + A_2A_3$, because D ($\neq A_2$) solves Heron's problem.
(c) The area of the isosceles $\triangle A_1A_2'A_3$, with $|A_1A_2'| + |A_2'A_3| = |A_1A_2| + |A_2A_3|$, is greater than that of $\triangle A_1A_2A_3$, since $A_2'C > DC$.



Comments

Let's prove Lemma 1: "A maximal n-gon must have equal sides.". Let $A_1A_2A_3\dots A_n$ be a maximal n-gon. As noted, it is a convex figure. We suppose that not all of its sides are equal and deduce a contradiction. Let A_1A_2 and A_2A_3 be two adjacent unequal sides. Let l be the line through A_2 parallel to A_1A_3 (see the figure). Now consider Heron's problem for the line l and the points A_1 and A_3. Recall that this is the problem of finding a point D on l that minimizes the sum of the distances $|A_1D| + |DA_3|$. As was proved earlier, the angles α and β at D must be equal. But α is equal to the angle $\angle DA_1A_3$, and β is equal to the angle $\angle DA_3A_1$ (by the property of opposite alternate angles between parallels). This means that $\triangle A_1DA_3$ is an isosceles triangle, and therefore D is different from A_2. Furthermore,

(a) the area of $\triangle A_1DA_3$ is equal to the area of $\triangle A_1A_2A_3$, since they have equal altitudes and bases; and

(b) the sum of the sides A_1D and DA_3 is less than the sum of the sides A_1A_2 and A_2A_3, since D ($\neq A_2$) is the solution of Heron's problem.

We now construct the isosceles triangle $\triangle A_1A_2'A_3$ such that $|A_1A_2'| + |A_2'A_3| = |A_1A_2| + |A_2A_3|$. Its area is, of course, larger than the area of $\triangle A_1A_2A_3$, since the altitude $A_2'C$ is larger than the altitude DC (by virtue of the fact that $|A_1A_2'|$ is longer than $|A_1D|$). But this means that the area of the polygon $A_1A_2'A_3\dots A_n$ is greater than the area of the polygon $A_1A_2A_3\dots A_n$ isoperimetric with it, a conclusion that contradicts the maximality of the last polygon.

This completes the proof of Lemma 1.

Lemma 2

A maximal n-gon has equal angles.

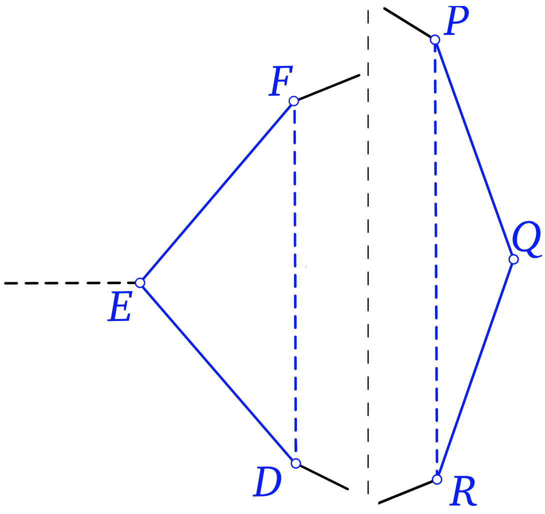
Proof: Again, let $A_1A_2A_3\dots A_n$ be a maximal polygon. Suppose that not all of its angles are equal and we will deduce a contradiction.

If the angles are not all equal, there exist two unequal adjacent angles, α and β, say. We will show this implies the existence of two unequal nonadjacent angles.

- Consider the successive angles $\alpha, \beta, \gamma, \delta, \varepsilon, \dots$ of the polygon.
- If $\gamma \neq \alpha$ or $\delta \neq \beta$, we get what we needed, since α and γ (or β and δ) are nonadjacent.
- If $\alpha = \gamma$, $\beta = \delta$, and $\alpha \neq \beta$, then the sequence is $\alpha, \beta, \alpha, \beta, \varepsilon, \dots$ — the first and fourth angles are nonadjacent, and we got what we wanted again.

Comments

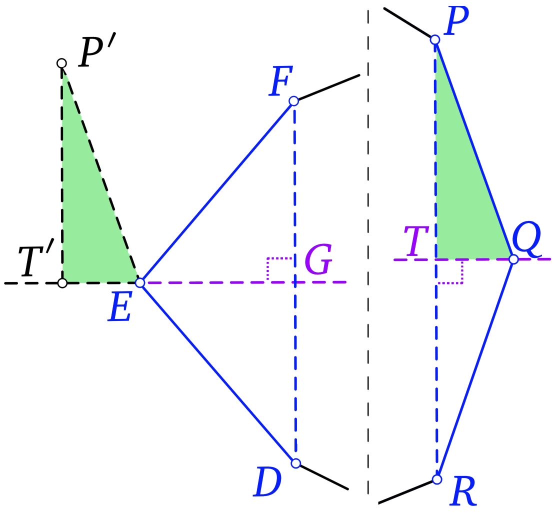
Let's prove Lemma 2: "A maximal n-gon must have equal angles.". Again, let $A_1A_2A_3\dots A_n$ be a maximal polygon. We know by now that all its sides are equal (Lemma 1) and bear in mind that it must be convex. We will suppose that not all of its angles are equal and will deduce a contradiction. If the angles are not all equal, then there must be two unequal adjacent angles, α and β, say. We will show that this implies the existence of two unequal nonadjacent angles. Consider the consecutive angles $\alpha, \beta, \alpha, \beta, \varepsilon, \dots$ (there are no fewer than five) of the polygon. If $\gamma \neq \alpha$ or $\delta \neq \beta$, then we get what we needed, since α and γ (or β and δ) are nonadjacent. If $\alpha = \gamma$, $\beta = \delta$, and $\alpha \neq \beta$, then our sequence of angles is $\alpha, \beta, \alpha, \beta, \varepsilon, \dots$ and we obtain that the first and fourth angles are nonadjacent.



- We have two disjoint triangles, DEF and PQR, from the n-gon.
- Due to our assumption, the angle $\angle E < \angle Q$.
- Since $|DE| = |EF| = |PQ| = |QR|$, the inequality of angles implies $|DF| < |PR|$.

Comments

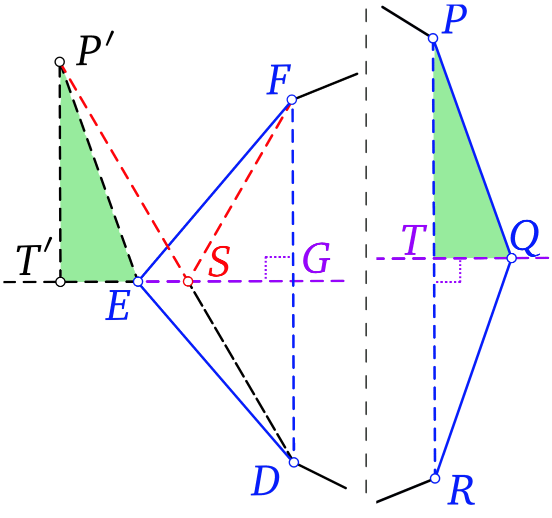
We see that our assumption justifies the conclusion that there are two triangles DEF and PQR with disjoint interiors, each of which is formed by successive vertices of our n-gon and such that angle E is smaller than angle Q. Since $|DE| = |EF| = |PQ| = |QR|$, the inequality of the angles E and F implies that $|DF| < |PR|$.



- We have two disjoint triangles, DEF and PQR, from the n-gon.
- Due to our assumption, the angle $\angle E < \angle Q$.
- Since $|DE| = |EF| = |PQ| = |QR|$, the inequality of angles implies $|DF| < |PR|$.
- From vertices E and Q, we drop perpendiculars EG and QT to the bases DF and PR respectively.
- We then extend EG and construct a triangle congruent to QTP on this extension.

Comments

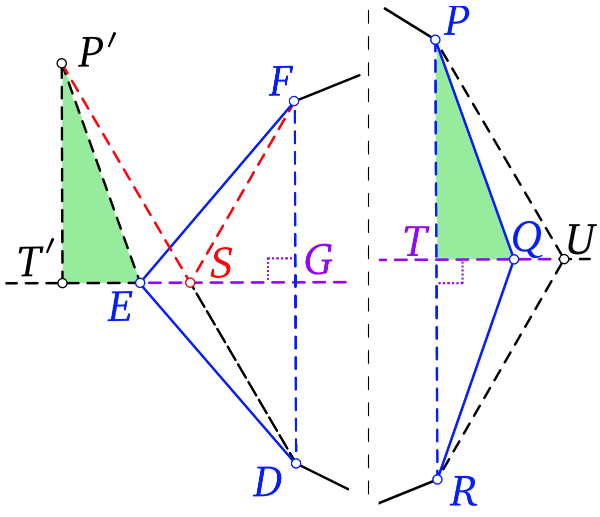
From E and Q we drop perpendiculars EG to DF and QT to PR. Next, we extend the segment EG and apply to the extension the triangle $ET'P'$ congruent to the triangle QTP (i.e. T goes over into T', P into P' and Q into E).



- We consider Heron's problem for the line segment T′G and points P′ and F.
- The solution S minimizes the sum of distances from P′ and F to a point on T′G.
- Due to our assumption $\angle P'ET' > \angle FEG$, the optimal point S does not coincide with E.
- Furthermore, S lies on the segment EG.

Comments

Now we consider Heron's problem for the line T′G and the points P′ and F. Let S be the solution of Heron's problem, that is S is a point on T′G such that the sum of the distances from P′ to S and from S to F is minimal. Since the angle P′ET′ (equals half the angle Q) is larger than the angle FEG (equals half the angle E) , the point S does not coincide with the point E (the angles P′ST′ and FSG are equal) and, furthermore, S lies on the segment EG.



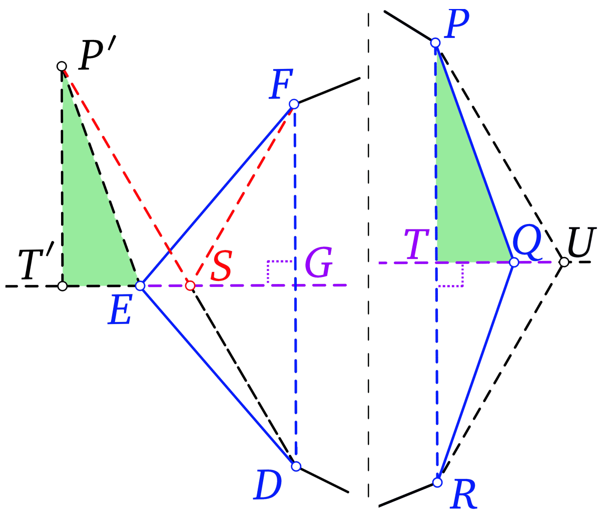
- We lay off the segment TU on line QT, with $|TU| = |T'S|$.
- The sum of the lateral sides of the new triangles ($\triangle DSF$ and $\triangle PUR$) is smaller than the original ones ($\triangle DEF$ and $\triangle PQR$).
- This is a direct result of S being the solution to Heron's problem.
- The area of $\triangle P'ES$ is larger than the area of $\triangle ESF$.
- This is because their altitudes satisfy $|P'T'| > |FG|$, as $|PR| > |DF|$.

Comments

Now we lay off on the line QT the segment TU of the same length as the segment T′S and consider the triangles $\triangle DSF$ and $\triangle PUR$. The sum of the lateral sides of these triangles is smaller than the sum of the lateral sides of the original triangles $\triangle DEF$ and $\triangle PQR$. In fact,

$$|DS|+|SF|+|PU|+|UR| = 2(|SF|+|SP'|) < 2(|FE|+|EP'|) = |DE|+|EF|+|PQ|+|QR|.$$

We have used the fact that our triangles are isosceles and that S is the solution of Heron's problem. On the other hand, the area of $\triangle P'ES$ is larger than the area of $\triangle ESF$, since their respective altitudes are $|P'T'| = \frac{1}{2}|PR|$ and $|FG| = \frac{1}{2}|DF|$ and we have shown that $|DF| < |PR|$. It follows that the sum of the areas of the triangles $\triangle DSF$ and $\triangle PUR$ is greater than the sum of the areas of the original triangles $\triangle DEF$ and $\triangle PQR$.



Denoting the area of $\triangle ABC$ by $S_{\triangle ABC}$, we have:

$S_{\triangle DSF} + S_{\triangle PUR} = (S_{\triangle DEF} - 2S_{\triangle ESF}) + (S_{\triangle PQR} + 2S_{\triangle P'ES}) > S_{\triangle DEF} + S_{\triangle PQR}$

This means the new polygon DSF...PUR... has a **smaller perimeter** and a **larger area** than the original polygon DEF...PQR...

- We can transform this new polygon to an isoperimetric one with an even **greater area**.
- This contradicts the assumption that the original polygon was maximal.
- This contradiction completes the proof of Lemma 2 and the theorem of Zenodorus.

Comments

Indeed, denoting the area of a triangle ABC by $S_{\triangle ABC}$, we have

$$S_{\triangle DSF} + S_{\triangle PUR} = S_{\triangle DEF} - 2S_{\triangle ESF} + S_{\triangle PQR} + 2S_{\triangle P'ES} > S_{\triangle DEF} + S_{\triangle PQR}.$$

This means that the polygon DSF . . . PUR . . . has a smaller perimeter and a larger area than our original polygon DEF . . . PQR . . .. Now we can treat either triangle ($\triangle DSF$ or $\triangle PUR$) as we treated triangle $\triangle A_1DA_3$ in proving Lemma 1, that is, we can raise it to obtain a polygon isoperimetric with the polygon DEF . . . PQR . . .. Since the area of the new polygon is greater than the area of the polygon DSF . . . PUR . . ., it is certainly greater than the area of the polygon DEF . . . PQR . . .. This contradicts the maximality of the polygon DEF . . . PQR . . . and completes the proof of Lemma 2 and, thereby, also of the theorem of Zenodorus.

LEMMA 3

There exists a maximal n-gon.

This and Lemmas 1 and 2 imply Zenodorus' Theorem (A maximal n-gon is regular).

THEOREM 2

The area enclosed by an arbitrary closed curve of a given length does not exceed the area enclosed by a circle of the same length.

Proof:

- Let P be the perimeter of a regular n-gon and S its area. $P = 2nR\sin(\pi/n)$ and $S = rP/2$, where $r = R\cos(\pi/n)$.
- These relations give the formula: $P^2 - 4n\tan(\pi/n)S = 0$.
- Zenodorus' Theorem implies that for any arbitrary n-gon with perimeter P and area S: $P^2 - 4n\tan(\pi/n)S \geq 0 \quad (2)$
- Using the inequality $\tan\alpha \geq \alpha$ (for $0 < \alpha \leq \pi/2$), and inequality (2), we get: $P^2 - 4\pi S \geq 0 \quad (3)$
- This inequality holds for any n-gon and all n.

Comments

We have established that if a maximal n-gon exists, it must be regular. But does such a maximal n-gon actually exist? If it doesn't, then the solution to Dido's problem collapses into dust and ashes. After all, not every function attains a maximum.

The ancient authors paid little attention to the question of whether a solution exists at all. It was only about a hundred years ago that mathematicians began to truly appreciate the importance of existence questions — and to develop rigorous methods for proving them.

For now, we simply state — without proof — a fact that seems to have been obvious to Zenodorus: Lemma 3. There exists a maximal n-gon. Combining this with Lemmas 1 and 2, we obtain Zenodorus' Theorem: A maximal n-gon is regular. From this, it remains to deduce the classical isoperimetric theorem: Theorem 2. The area enclosed by any closed curve of a given length does not exceed the area enclosed by a circle of the same length.

Let P denote the perimeter of a regular n-gon and S its area. We know from geometry that $P = 2nR\sin(\pi/n)$, where R is the radius of the circumscribed circle, and that $S = rP/2$, where r is the radius of the inscribed circle. We have $r = R\cos(\pi/n)$. All these yield the following formula linking S and P:

$$P^2 - 4n\tan(\pi/n)S = 0.$$

Zenodorus' Theorem implies that if P is the perimeter of an arbitrary n-gon and S is its area, then

$$P^2 - 4n\tan(\pi/n)S \geq 0 \tag{2}$$

The inequality $\tan\alpha \geq \alpha$ (valid for $0 < \alpha \leq \pi/2$) and (2) imply the inequality

$$P^2 - 4\pi S \geq 0, \tag{3}$$

which holds for an arbitrary n-gon and all n.

We note that for an arbitrary circle we have

$$P^2 - 4\pi S = 0,$$

where P is the circumference and S is the area.

Lemma 4

For every closed plane curve of length P^* enclosing an area S^* and for every $\varepsilon > 0$, there exists an n-gon of perimeter P and area S such that

$$|P - P^*| \leq \varepsilon, \quad |S - S^*| \leq \varepsilon.$$

Lemma 4 and relation (3) imply: for every $\varepsilon > 0$ there exists a polygon with perimeter P and area S such that

$$4\pi S^* \leq 4\pi S + 4\pi\varepsilon \leq P^2 + 4\pi\varepsilon$$
$$\leq (P^* + \varepsilon)^2 + 4\pi\varepsilon = P^{*2} + \varepsilon\,(2P^* + 4\pi + \varepsilon)\,.$$

Since ε is arbitrary, we conclude

$$4\pi S^* \leq P^{*2},$$

and for a circle, $4\pi S^* = P^{*2}$. This completes the proof of the isoperimetric inequality (Theorem 2). □

Comments

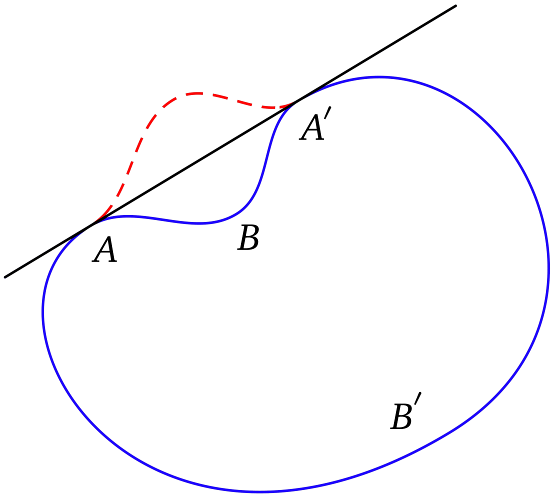
We note that for an arbitrary circle we have the obvious equality, equation four: $P^2 - 4\pi S = 0$, where P is the circumference of the circle and S is its area.

Now we state a lemma that connects all the concepts appearing in the formulation of the classical isoperimetric problem with the notion of an n-gon. The idea is that one can approximate both the length of a curve and the area it encloses by the perimeter and area of an n-gon, with arbitrarily high precision.

Lemma 4. For every closed plane curve of length P^*, enclosing an area S^*, and for every positive ε, there exists an n-gon of perimeter P and area S such that $|P - P^*| \leq \varepsilon, \quad |S - S^*| \leq \varepsilon$.

From Lemma 4 and relation three, it follows that for any positive ε there is a polygon with perimeter P and area S such that: $4\pi S^* \leq 4\pi S + 4\pi\varepsilon \leq P^2 + 4\pi\varepsilon$, which is less than or equal to $(P^* + \varepsilon)^2 + 4\pi\varepsilon = P^{*2} + \varepsilon\,(2P^* + 4\pi + \varepsilon)$.

Since epsilon is arbitrary, we arrive at the final inequality: $4\pi S^* \leq P^{*2}$. According to equation four, this becomes an equality in the case of a circle.



Steiner's Proof

- ASSERTION 1: The extremal curve is convex.
- If a curve is not convex, it has two points, A and A′, with both arcs connecting them lying on the same side of line AA′.
- Replacing one arc with the reflection of the other across line AA′ creates a new curve.
- This new curve has the same perimeter but encloses a larger area.

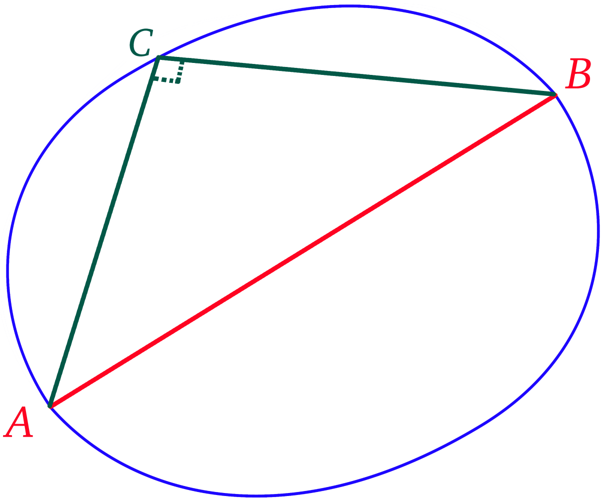
Comments

Having presented a proof based on the ideas of the ancients, it is difficult to resist presenting an outline of yet another proof, whose key thought is due to Jakob Steiner, a mathematician who enriched geometry with many remarkable ideas. A tacit assumption of Steiner's proof is the existence of the curve that solves the isoperimetric problem. (We already know that this is a justified assumption.) It remains to show that this extremal curve is a circle.

ASSERTION 1. The extremal curve is convex.

What is a convex curve? It is a curve whose interior (that is, the region bounded by the curve) includes the segment joining any two of its points.

If the curve is not convex then it must contain two points A and A' such that both arcs ABA′ and AB′A′ joining A and A′ lie on the same side of the line AA′. By replacing one of these arcs with its image under reflection in AA′, we obtain a new curve of the same length that encloses a larger area.



- ASSERTION 2: If points A and B halve the length of the extremal curve, then the chord [AB] halves the area it encloses.
- ASSERTION 3: Suppose that points A and B halve the extremal curve. If C is any point on the curve, then the angle ACB is a right angle.

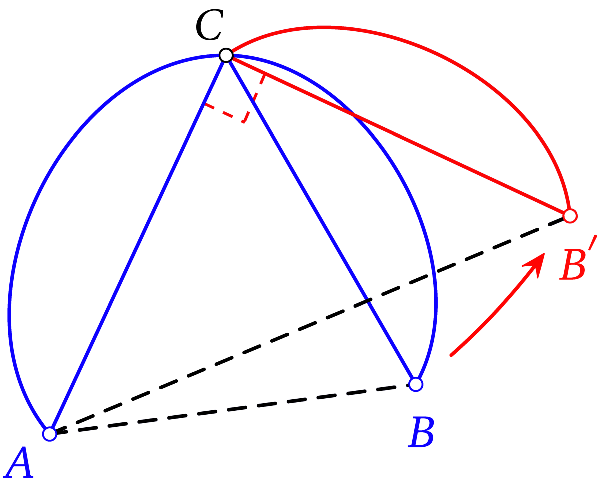
Comments

ASSERTION 2. If points A and B halve the length of the extremal curve, then the chord [AB] halves the area it encloses.

In fact, if the chord [AB] divided the area into unequal parts, then the figure consisting of the larger part and its image under reflection in the diameter AB would add up to a figure with the same length and a larger area.

ASSERTION 3. Suppose that points A and B halve the extremal curve. If C is any point on the curve, then the angle ACB is a right angle.

This is the heart of the matter. The method we will employ to prove this assertion is known as Steiner's four-hinge method.



- Assume there is a point C where the angle ACB is not a right angle.
- By "spreading" the segments at C to make a right angle, we form a new triangle.
- The right triangle has the maximal area for a given side length.
- Reflecting the new arc across the chord creates a new curve with the same perimeter but a larger area.
- This contradicts the assumption of a maximal curve.
- Thus, for any point C on the curve, the angle ACB must be a right angle, which proves the curve is a circle.

Course content
Heron's Problem
Euclid's problem
Dido's Problem
Isoperimetric Theorem
Steiner's Proof
Light's Refraction
Light's Refraction

Comments

Suppose there is a point C such that the angle $\angle ACB$ is not a right angle.

The area bounded by the arc $\overset{\frown}{ACB}$ and the diameter AB can be thought of as three parts: first, the triangle $\triangle ABC$, and second, the two circular segments adjacent to sides AC and CB.

Now, picture a hinge placed at point C that joins the two segments. We "spread" this hinge so that the new angle $\angle ACB'$ becomes exactly ninety degrees.

When we do this, the area bounded by the new arc $\overset{\frown}{ACB'}$ increases. Why? Because among all triangles with the same two lateral sides, the right-angled triangle has the largest possible area.

Indeed, the area of triangle $\triangle ABC$ equals $\frac{1}{2} \cdot AC \cdot BC \cdot \sin(\angle C)$, and since $\sin(\angle C) \leq 1$, the maximum is reached when $\angle C$ is a right angle.

Next, reflect the curve $\overset{\frown}{ACB'}$ across the chord AB'. The resulting figure has the same perimeter as the original, but a strictly larger area. That's the key point — and this establishes our claim.

So, the extremal figure is made up of all points C from which the chord that halves the curve's length is seen under a right angle. By the inscribed angle theorem, such a curve must be a circle.

At this moment, an enthusiast might exclaim: "Astounding!" A skeptic, however, will mutter: "Not so fast... you haven't proven existence... What if, when the hinge opens too far, the segments at C overlap?"

We'll ignore such grumbling for now. Yes, the argument is dazzling — but it still calls for a rigorous justification.

Literature on Isoperimetric Problems

- H. Rademacher and O. Toeplitz, The enjoyment of mathematics. Princeton Univ. Press, 1966.
- R. Courant and H. Robbins, What is mathematics?. Oxford Univ. Press, 1978.
- Bruce van Brunt, The Calculus of Variations. New York: Springer, 2004.
- Frank Morgan, Geometric Measure Theory: a beginner's guide. Academic Press, 2009.

Comments

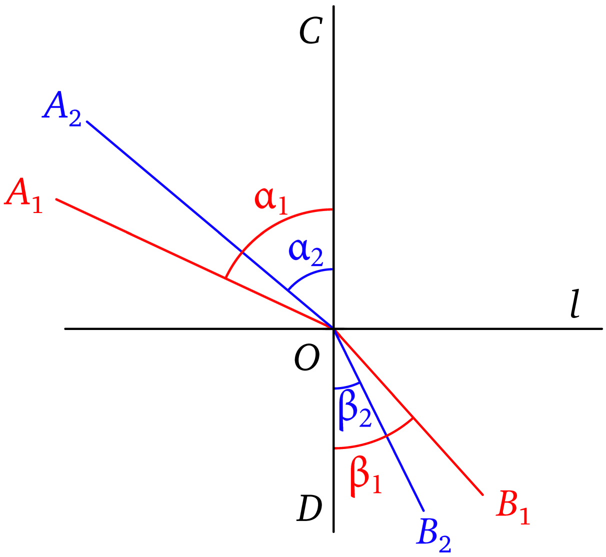
For those who are interested, here are some remarkable works dealing with the isoperimetric problem:

Richard Courant in collaboration with Herbert Robbins. What is mathematics? Oxford Univ. Press, Oxford, 1978.

Hans Rademacher and Otto Toeplitz, The enjoyment of mathematics. Princeton Univ. Press, Princeton, 1966.

Bruce van Brunt, The Calculus of Variations, New York: Springer, 2004

Frank Morgan Geometric Measure Theory: a beginner's guide, Academic Press, 2009



- Consider two rays, A_1OB_1 and A_2OB_2, that refract at point O.
- The angles α_1 and α_2 are the incidence angles.
- The angles β_1 and β_2 are the refraction angles.
- Snel's Law states that the ratio of the sine of the incidence angle to the sine of the refraction angle is a constant:

$$\frac{\sin \alpha_1}{\sin \beta_1} = \frac{\sin \alpha_2}{\sin \beta_2} = \text{constant}$$

- This constant is independent of the incidence angle.

Comments

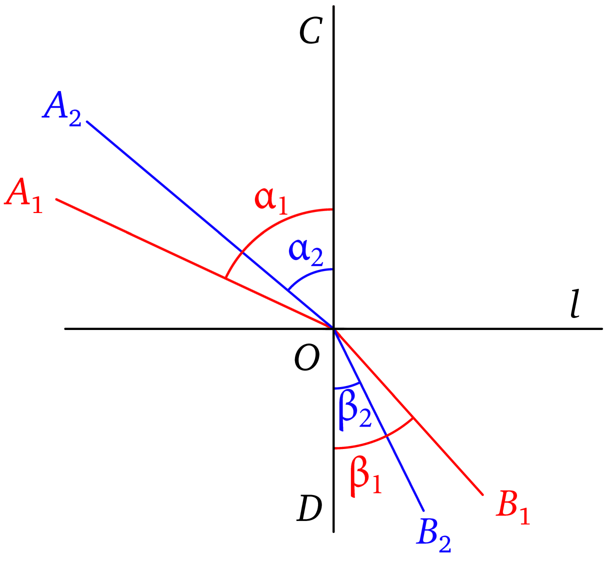
Let's begin our consideration of the next task with the dictum of a famous scientist, Carl Siegel: «According to Leibniz our world is the best possible. That is why its laws can be described by extremal principles.»

Carl Siegel, an eminent twentieth-century mathematician, obtained fundamental results in many areas of mathematics and mechanics. His quote, which we mentioned earlier, is a joke, of course, but it contains a kernel of truth. When discussing Heron's problem, we had cause to remark that nature "employs" extremal principles. For example, we said that a reflection from a flat surface "chooses" a trajectory of least length.

Heron's words quoted at the beginning of our lecture contain the germ of a fundamental idea established between the seventeenth and nineteenth centuries. During this time it became clear that nature "operates" optimally in optics, in mechanics, in thermodynamics-in fact, everywhere. The extremal principle associated with natural phenomena was clearly formulated for the first time in optics in an attempt to comprehend the law of refraction of light.

Ancient philosophers tried to discover the law of refraction. In particular, in the second century B. C. Ptolemy tried to obtain this law experimentally but failed to do so.

The law was first found by the Dutch scientist Snel. Snel's name is not as well known today as the names of his great contemporaries Descartes, Huygens, and Fermat. Nowadays Snel's fame only results from his experimental discovery of the law of refraction of light, a discovery that remained unpublished in his lifetime. But in his time Snel was very famous. Kepler regarded him as "the glory of the geometers [mathematicians] of our age."



- Consider two rays, A_1OB_1 and A_2OB_2, that refract at point O.
- The angles α_1 and α_2 are the incidence angles.
- The angles β_1 and β_2 are the refraction angles.
- Snel's Law states that the ratio of the sine of the incidence angle to the sine of the refraction angle is a constant:

$$\frac{\sin \alpha_1}{\sin \beta_1} = \frac{\sin \alpha_2}{\sin \beta_2} = \text{constant}$$

- This constant is independent of the incidence angle.

Comments

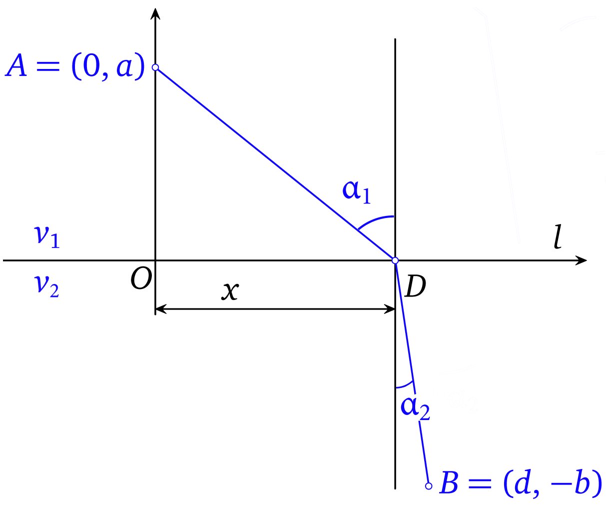
Snel's law of refraction can be stated as follows. Let A_1OB_1 and A_2OB_2 be two rays (going "from above to below") that refract at the point O. The angles α_1 and α_2 formed by the vertical OC and the respective lines A_1O and A_2O are called incidence angles (a term with which you should already be familiar). The angles β_1 and β_2 formed by the vertical OD and the respective lines B_1O and B_2O are called refraction angles.

Snel showed that

$$\frac{\sin \alpha_1}{\sin \beta_1} = \frac{\sin \alpha_2}{\sin \beta_2}$$

that is, the ratio of the sine of the incidence angle to the sine of the refraction angle is a constant that is independent of the incidence angle.

Descartes, one of the greatest French thinkers and scholars, arrived at the same law independently of Snel. He deduced the law of refraction from his conceptions of the propagation of light rays. These conceptions have not withstood the test of time, although they led later to the law of conservation of momentum. Descartes' theory implied that the speed of light is greater in a denser medium, such as water, than in a less dense medium such as air. Many other scientists doubted this. Fermat explained the law of refraction from the opposite assumption, that light moves more slowly in a denser medium. In this case Fermat turned out to be correct. Experiments showed that the denser the medium, the slower the speed of light.



- Fermat's principle allows the precise formulation of a minimum problem to derive Snel's law.
- It requires computing the minimum of the following function of one variable:

$$f(x) = \frac{\sqrt{a^2 + x^2}}{v_1} + \frac{\sqrt{b^2 + (d - x)^2}}{v_2}$$

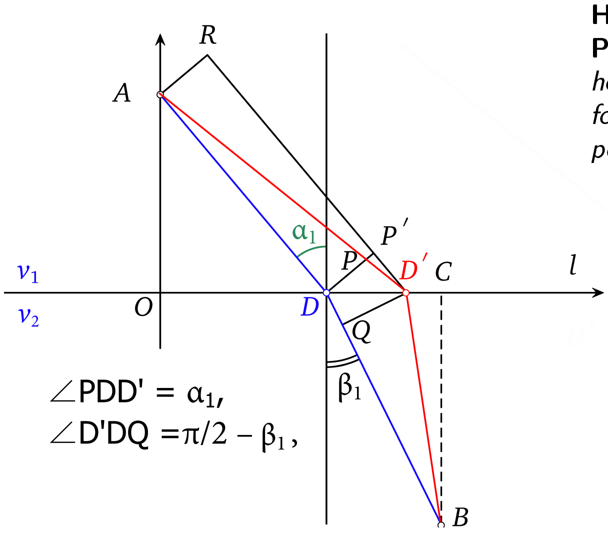
- Fermat had a method to find maxima and minima by setting the derivative to zero.
- However, he did not know how to apply it to radical expressions, which made his solution very complicated.

Comments

To explain the law of refraction of light, Fermat advanced an extremal principle for optical phenomena. It was later named for him. This principle states that, in an inhomogeneous medium, light travels from one point to another along the path requiring the shortest time. Fermat's principle allows the precise formulation and solution of a minimum problem that leads to the derivation of Snel's law. Specifically, this principle requires the computation of the minimum of the following function of one variable: f of x equals the square root of a squared plus x squared, divided by v sub one, plus the square root of b squared plus the quantity d minus x squared, divided by v sub two.

It is worth noting that, at the time that he advanced his extremal principle (approximately 1660), Fermat already had at his disposal an algorithm for finding maxima and minima of functions that was equivalent to setting the derivative equal to zero. The use of derivatives so simplifies the derivation of Snel's law that it now can be carried out by high school students. Fermat himself obtained the required result in a far more elaborate way. It is natural to ask: Why did Fermat not use his algorithm? The answer is very simple: Fermat could apply his method to polynomials-and here he actually anticipated the notion of a derivative-but he did not know how to apply it to radical expressions. That is why the deduction of Snel's law using derivatives was first accomplished by Leibniz, who introduced this concept in the very same work of 1684 in which he laid the foundations of mathematical analysis.

Thus, Fermat deduced Snel's law from his extremal principle, but his solution was very complicated. A far simpler solution, also based on Fermat's principle, was given by Huygens, yet another scientific genius of the seventeenth century and the author of the wave theory of light.



Huygens' solution

Problem: *Find a point* D *on the horizontal line* l *that minimizes the time for a light ray to travel from point* A *to point* B.

- The velocity of light is v_1 in the upper medium and v_2 in the lower medium.
- Huygens showed that the minimum time occurs at a point D where:
$$\frac{\sin\alpha_1}{\sin\beta_1} = \frac{v_1}{v_2}$$
- The proof demonstrates that for any other point $D' \neq D$, the time taken to traverse the path $AD'B$ is greater than the time for path ADB.

Comments

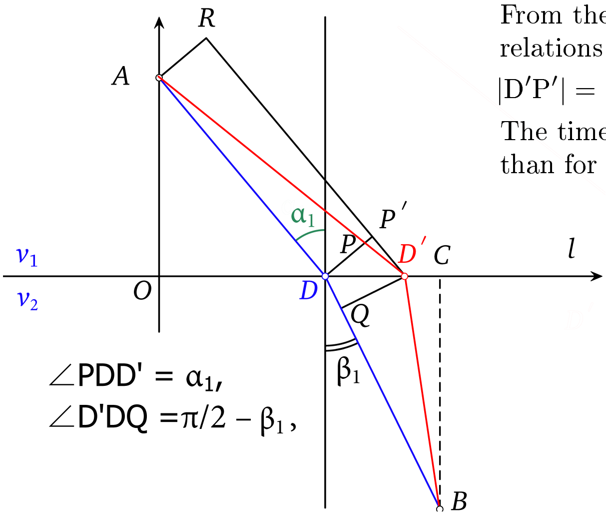
Before reproducing Huygens' solution let's state the problem precisely. Given two points A and B on either side of a horizontal line l separating two media. It is required to find a point D such that the time it takes for a light ray to traverse the path ADB is a minimum, provided that the velocity of propagation of light is v_1 in the upper medium and v_2 in the lower one. Note that the formula on the previous slide is a mathematical reformulation of this problem and that this problem is very similar to Heron's problem.

Let D be a point at which the following equality holds: the fraction of the sine of α_1 divided by the sine of β_1 equals the fraction of v_1 divided by v_2, i.e.,

$$\frac{\sin\alpha_1}{\sin\beta_1} = \frac{v_1}{v_2}.$$

We will show that for any other point $D' \neq D$ the time of traversal of the path $AD'B$ is greater than the time of traversal of the path ADB.

To this end we erect perpendiculars to the line AD at A and D, respectively. Let P be the point of intersection of AD' and the perpendicular at D. We draw a line through D' parallel to AD and denote its points of intersection with the perpendiculars (to AD) at D and A by P' and R, respectively. Finally, we drop the perpendicular $D'Q$ from D' to DB.



From the geometry of the figure, we have the relations:

$|\mathrm{D'P'}| = |\mathrm{DD'}| \sin\alpha_1$ and $|\mathrm{DQ}| = |\mathrm{DD'}|\sin\beta_1$ $(*)$

The time taken for the path AD′B is greater than for ADB:

$$\frac{|\mathrm{AD'}|}{\mathrm{v}_1} > \frac{|\mathrm{AD}|}{\mathrm{v}_1} + \frac{|\mathrm{DD'}|\sin\alpha_1}{\mathrm{v}_1}$$
$$\frac{|\mathrm{D'B}|}{\mathrm{v}_2} > \frac{|\mathrm{DB}|}{\mathrm{v}_2} - \frac{|\mathrm{DD'}|\sin\beta_1}{\mathrm{v}_2}$$

Combining these inequalities with the condition $\frac{\sin\alpha_1}{\mathrm{v}_1} = \frac{\sin\beta_1}{\mathrm{v}_2}$, we get:

$$\frac{|\mathrm{AD'}|}{\mathrm{v}_1} + \frac{|\mathrm{D'B}|}{\mathrm{v}_2} > \frac{|\mathrm{AD}|}{\mathrm{v}_1} + \frac{|\mathrm{DB}|}{\mathrm{v}_2}$$

Thus, the ratio of the sines of the angles of incidence and refraction is a constant, which proves Snel's Law.

Comments

From the figure we see that the angles ∠PDD′ and ∠D′DQ are respectively equal to α_1 and $\pi/2 - \beta_1$, respectively. Hence

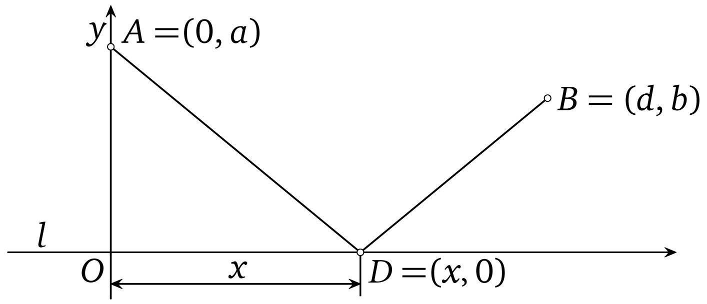
$$|\mathrm{D'P'}| = |\mathrm{DD'}|\sin\alpha_1, \quad |\mathrm{DQ}| = |\mathrm{DD'}|\sin\beta_1. \qquad (*)$$

Now we compare the traversal times along the paths ADB and AD′B. The relation (*) and the inequalities $|\mathrm{AP}| > |\mathrm{AD}|$, $|\mathrm{D'P}| > |\mathrm{D'P'}|$, and $|\mathrm{D'B}| > |\mathrm{BQ}|$ (inclined segments are longer than perpendicular ones) imply the following inequalities. The length of segment AD′ divided by v_1 is strictly greater than the length of AD divided by v_1, plus the length of DD′ multiplied by $\sin\alpha_1$ divided by v_1. And the length of D′B divided by v_2 is strictly greater than the length of DB divided by v_2, minus the length of DD′ multiplied by $\sin\beta_1$ divided by v_2. If we now add these inequalities together, and use the fact that $\sin\alpha_1/\mathrm{v}_1 = \sin\beta_1/\mathrm{v}_2$, the extra terms cancel. We then obtain the inequality:

$$\frac{|\mathrm{AD'}|}{\mathrm{v}_1} + \frac{|\mathrm{D'B}|}{\mathrm{v}_2} > \frac{|\mathrm{AD}|}{\mathrm{v}_1} + \frac{|\mathrm{DB}|}{\mathrm{v}_2}.$$

This shows that the path ADB has a smaller travel time than any nearby path AD′B. Thus the refraction point that minimizes the time of traversal of the broken path from A to B is characterized by the fact that the ratio of the sines of the angles of incidence and refraction is equal to $\mathrm{v}_1/\mathrm{v}_2$, that is, to a constant. But this is just Snel's law.

What underlies Fermat's principle is the assumption that light is propagated along certain lines. This idea ties in most readily with the corpuscular theory of light that regards light as a flow of particles. We owe to Huygens another explanation of the propagation and refraction of light, based on the notion of light as a wave whose front moves in time. In this case, it can also be shown that Snel's law holds.



- The formalization of problems, such as Heron's problem, marked a new era in mathematics.
- This allows for a precise solution by finding the minimum of a function.
- The problem can be formulated as finding the least value of the function:

$$f(x) = \sqrt{a^2 + x^2} + \sqrt{b^2 + (d - x)^2}$$

- This approach was a significant step toward modern calculus.

Comments

In fact, until the end of the 17th century, mathematicians approached each new optimization problem individually, developing original methods for solving it. The mathematical analysis apparatus that appeared by the beginning of the 18th century made it possible to develop general approaches to solving such problems. He made it possible to switch to a single form of writing the conditions of problems and use a common mathematical language in which one could talk about problems of such different content. In particular, at the beginning of the 18th century, with the work of Euler and Lagrange, variational analysis appeared, which made it possible to solve problems of optimizing functionals such as curve length or descent time (for example, the brachystochron problem). Mathematical analysis made it possible to formulate problems in a unified form using the concepts of functions, derivatives and integrals, which simplified communication between mathematicians and standardized approaches.

Extremal problems arising in mathematics, in the natural sciences, or in practical enterprises are traditionally stated first without formulas, using the terminology of the domain in which they arise. In order to be able to use a general theory, it is necessary to translate the statements of the problems from each specific language into the language of mathematics. Such a translation is called a formalization.

Let's illustrate how the formalization is carried out using the example of the Heron's problem.

We take the given line as the x-axis and draw the y-axis through the point A perpendicular to the x-axis. Let the coordinates of the points A and B be (0,a) and (d, b) , respectively. On the x-axis we take a point D with coordinates (x, 0). Then the sum of the distances from A to D and from D to B is $\sqrt{a^2 + x^2} + \sqrt{b^2 + (d - x)^2}$ This results in the following problem: Find the least value of the function f(x) for all values of x.

Formalizations discussed today

- Heron's problem: $f_1(x) = \sqrt{a^2 + x^2} + \sqrt{b^2 + (d - x)^2}$
- The problem of refraction of light: $f_2(x) = \dfrac{\sqrt{a^2 + x^2}}{v_1} + \dfrac{\sqrt{b^2 + (d - x)^2}}{v_2}$
- Euclid's problem: $f_3(x) = H/b\ x(b - x)$

Problem for practice

- Try to formalize the Kepler's planimetric problem:
  Inscribe a rectangle of maximal area in a circle of unit radius.

Source: V. M. Tikhomirov, "Stories about Maxima and Minima" - a fascinating book about the development of these ideas in mathematics.

Comments

Here are examples of the formalizations of some tasks that we discussed today. Function $f_1(x)$ for Heron's problem, $f_2(x)$ for the problem of refraction of light and $f_3(x)$ for the Euclid's problem.

As a small exercise, you can try to formalize the Kepler's planimetric problem: Inscribe a rectangle of maximal area in a circle of unit radius.

So, why have we spent time looking at these famous problems of the past — Heron's problem, Euclid's problem, the isoperimetric problem, and the problem of light refraction — and especially at their geometric solutions?

It was not just to admire elegant reasoning or historical ingenuity. The deeper reason is that they illustrate how mathematical thinking developed long before the formal optimization methods we know today.

The formalization is the truly revolutionary step — turning intuitive statements into precise mathematical formulations. This simple-sounding move opened the way to a unified theory of optimization, one that is still an active and rapidly developing field.

For you, formalization may feel natural; you've seen it in school mathematics and in many university courses. But in reality, it is one of the most powerful techniques we have. Formalization allows us to take problems from completely different domains, express them in a common mathematical language, and apply systematic methods to solve them. In practice, we often face challenges that, in their raw, unstructured form, seem intractable. Transforming them into a formal model is often the key step — and it is a skill that, unfortunately, does not always receive the attention it deserves in university training. Yet it is one of the most essential qualities of an applied researcher.

Formalizations discussed today

- Heron's problem: $f_1(x) = \sqrt{a^2 + x^2} + \sqrt{b^2 + (d - x)^2}$
- The problem of refraction of light: $f_2(x) = \dfrac{\sqrt{a^2 + x^2}}{v_1} + \dfrac{\sqrt{b^2 + (d - x)^2}}{v_2}$
- Euclid's problem: $f_3(x) = H/b\ x(b - x)$

Problem for practice

- Try to formalize the Kepler's planimetric problem:
  Inscribe a rectangle of maximal area in a circle of unit radius.

Source: V. M. Tikhomirov, "Stories about Maxima and Minima" - a fascinating book about the development of these ideas in mathematics.

Comments

My advice is this: never underestimate the power of reformulating a problem; train yourself to translate vague goals into precise mathematical terms. This habit will open the door to powerful tools and unexpected insights.

Finally, I should mention that the main material for today's lecture was drawn from the book Stories about Maxima and Minima by V. M. Tikhomirov, which contains many fascinating stories about the development of these ideas in mathematics — not only in geometry, but also in algebra and mathematical analysis.

You can still find many non-formalized problems in areas like economics, engineering, medicine, social sciences, and environmental studies. For example, in economics, deciding on a long-term investment strategy under uncertain market conditions often starts as a vague question without precise numbers or constraints. In engineering, designing a new type of bridge or vehicle may begin as a broad goal like "make it lighter but stronger," which needs careful definition before it can be solved. In medicine, choosing the best treatment plan for a patient with multiple health issues often involves unclear trade-offs that must be translated into measurable criteria. These examples show that in many fields, problems first appear as general ideas, and turning them into clear mathematical formulations is an essential step toward finding a solution.