1

Introduction

The idea of writing a book on the principle of least action came to us after many conversations over coffee, while we pondered ways of communicating to students the ideas of mechanics with an historical flavor. We chose the principle of least action because we think that its importance and aesthetic value as a unifying idea in physics is not sufficiently emphasized in regular courses. To the general public, even to those interested in science at a popular level, the beautiful notion that the fundamental laws of physics can be expressed as the minimum (or an extremum) of something often seems foreign. Nature loves extremes. Soap films seek to minimize their surface area, and adopt a spherical shape; a large piece of matter tends to maximize the gravitational attraction between its parts, and as a result the planets are also spherical; light rays refracting in a glass window bend and follow the path of least time; the orbits of the planets are those that minimize something called the "action;" and the path that a relativistic particle chooses to follow between two events in space-time is the one that maximizes the time measured by a clock on the particle.

Our initial intention was to write a popular book, but the project morphed into a more technical presentation. Nevertheless we have tried to keep sophisticated mathematics to a minimum: nothing more than freshman calculus is needed for most of the book, and a good part of the book requires only high school algebra. Some familiarity with differential equations would be useful in certain sections. While the different sections have various levels of difficulty, the book does not need to be read in a linear fashion. It is quite feasible to browse through this book, as most of the chapters and many of the sections are relatively self-contained. Sections and subsections that are a bit more technical and that can easily be omitted on a first reading include 1.1, 2.5, 3.2.2, 3.2.5, 4.3, 5.7, 6.2 to 6.6, 7.7 and 8.8. These are marked in the text with an asterisk.

The gold standards on the topic of our book are *The Variational Principles of Mechanics* by Cornelius Lanczos and *Variational Principles in Dynamics and*

Quantum Theory by Wolfgang Yourgrau and Stanley Mandelstam. Our book can be regarded as a supplement to these two masterpieces, with expositions that follow the historical development of minimum principles, some elementary examples, an invitation to read the primary sources, and to appreciate science, in the words of Isidor Isaac Rabi, as a "human endeavor in its historic context, . . . as an intellectual pursuit rather than as a body of tricks."

The metaphysical roots of the least action principle are in Aristotle's statement from *De caelo* and *Politics*: "Nature does nothing in vain." If there is a purpose in Nature, she should follow a minimum path. At least that is the notion pursued by Hero of Alexandria in the first century AD to deduce the law of reflection: light follows the path that minimizes the travel time. Later, in 1657, Pierre de Fermat extended this idea to the refraction of light rays. "There is nothing as probable or apparent," says Fermat, "as the assumption that Nature always acts by the easiest means, which is to say either along the shortest lines when time is not a consideration, or in any case by the shortest time." The Arabic astronomer, Ibn al-Haytham, also uses the principle of "the simplest way" to explain refraction. Galileo, in postulating the uniform acceleration of freely falling bodies, in 1638, also echoes Aristotle: "we have been led by the hand to the investigation of naturally accelerated motion by consideration of the custom and procedure of Nature herself in all her other works, in the performance of which she habitually employs the first, simplest, and easiest means." In 1746, Pierre Louis Moreau de Maupertuis postulated the principle of least action. His proposal, based on metaphysical and religious views, reflected his adherence to notions of simplicity that had guided Fermat and Galileo: "Nature, in the production of its effects," he wrote, "does so always by the simplest means." More specifically: "in Nature, the amount of action (*la quantité d'action*) necessary for change is the smallest possible. Action is the product of the mass of a body times its velocity times the distance it moves." His formulation was vague, but, in the hands of Leonard Euler, it later became a well-formulated principle. Gottfried Leibniz used similar (but not identical) ideas to study refraction of light. Leibniz's idea is of a "most determined" path and this reflects "God's intentions to create the best of all possible worlds." "This principle of Nature," he says in his *Tentamen Anagogicum*, "is purely architectonic," and then he adds: "Assume the case that Nature were obliged in general to construct a triangle and that, for this purpose, only the perimeter or the sum were given and nothing else; then Nature would construct an equilateral triangle."

The formulation of mechanics in terms of minimum principles originates in the optical mechanical analogy first used by John Bernoulli to solve the "brachistochrone problem:" what path between two fixed points in a vertical plane does a particle follow in order to minimize the time taken? Bernoulli maps the problem to that of a light ray refracting in a medium of varying index of refraction,

where light follows the path of least time. The mapping between mechanics and optics becomes an isomorphism with Maupertuis's formulation. The minimization of action for a particle and the minimization of time for a light ray become the same mathematical problem provided the index of refraction is identified with the momentum of the particle: the paths are isomorphic. The principle of least action then may be viewed as an alternative and equivalent formulation to Newton's laws of motion.

In the Age of Enlightenment, Newton's ideas were extended to incorporate constraints in mechanical systems. The key figures are James Bernoulli, Jean le Rond d'Alembert, and Joseph-Louis Lagrange. The central concept for these developments is the principle of virtual work, which establishes the conditions of static equilibrium and its extension to dynamics. The work of Lagrange, starting in 1760, is of supreme importance. For a constrained system with r degrees of freedom (for example, a particle constrained to move on the surface of a sphere has $r = 2$ since at a given position it can move in only two directions), he is able to express the dynamics in terms of a single function L (the Lagrangian) through r equations identical in structure. Lagrange's equations can be derived from a minimum principle, giving rise to an expanded version of the principle of least action: Maupertuis's minimum principle gives the path between two points in space for a fixed value of the energy, while Lagrange's integral gives the path that takes a given time t between two fixed points in space. Lagrange's ideas were extended, starting in the 1820s, by William Rowan Hamilton (and also by Carl Jacobi). Hamilton and Jacobi put the optical mechanical analogy in a broader conceptual frame: the end points of paths that emanate from a given origin at $t = 0$ (each path being a minimum of the Lagrangian action) create, at a later time t, a "wave-front" that propagates. This wave-front is a surface that intersects the particle trajectories (just like a wave-front for light is perpendicular to the light rays) but does not include interference or diffraction effects peculiar to waves. However, it invites a "natural" question: if light rays are the small wavelength limit of wave optics, what is the wave theory of particles whose small wavelength limit gives the particle trajectories? Hamilton did not have an experimental reason to entertain the question in the mid-nineteenth century, but the answer came in the 1920s with Louis de Broglie's and Erwin Schrödinger's quantum theory of wave mechanics. In 1923 de Broglie wrote: "Dynamics must undergo the same evolution that optics has undergone when undulations took the place of purely geometrical optics," and in 1926 Schrödinger considered the "general correspondence which exists between the Hamilton-Jacobi differential equation and the 'allied' *wave equation*." In 1942 Richard Feynman established an even deeper connection between least action and quantum physics: a quantum particle, in propagating between two fixed points in space and time, does not follow a single path but all possible paths "at the same

time." The contribution of each path to the total propagation is the (complex) exponential of Hamilton's action.

The fact that many fundamental laws of physics can be expressed in terms of the least action principle (with the appropriate action) led Max Planck to say that, "Among the more or less general laws which manifest the achievements of physical science in the course of the last centuries, the principle of least action is probably the one which, as regards form and content, may claim to come nearest to that final ideal goal of theoretical research." And Arthur Eddington, in 1920, wrote: "the law of gravitation, the laws of mechanics, and the laws of electromagnetic fields have all been summed up in a principle of least action. ... Action is one of the two terms in pre-relativity physics which survive unmodified in a description of the absolute world. The only other survival is entropy." Although Einstein didn't follow a least action approach in his theories of relativity (special and general), Max Planck, in 1907, in the first relativity paper not written by Einstein, formulated the dynamics of the special theory in terms of the least action principle. One of the most interesting applications of the least action principle is the derivation, by David Hilbert, of the field equations of general relativity. Hilbert knew, from Einstein, that the relativistic theory of gravitation had to involve the curvature of a four-dimensional space-time. Einstein had struggled for eight years and he had eventually arrived at the solution by analyzing the properties of the field equations themselves. Hilbert followed the approach of the least action principle, guessed the "most natural" Lagrangian and, in 1915, derived the field equations before Einstein.

Our purpose in writing this book is to tell the above stories with some mathematical rigor while staying as close as possible to the sources. Chapter 2 visits some ancient incarnations of minimum principles before moving on to Galileo's curve of swiftest descent and Fermat's precalculus ideas. We also include Newton's calculation of the solid of least resistance, which anticipates the calculus of variations used in the principle of least action. In chapter 3 we take an excursion to Newton's *Principia,* even though this work is not directly related to variational principles. We do so for two reasons: the monumental importance of this work on mechanics, and the fact that Newton's ideas are crucial in the development of the principle of least action. Chapter 4 tells the story of the optical mechanical analogy and the true beginnings of variational principles. In chapter 5 we visit the principle of virtual work and Lagrange's equations. Here we point out that the principle of least action fails to give the dynamics of nonholonomic systems, where the constraints are expressed in terms of the possible motions rather than in terms of the possible configurations. Chapter 5 and the ones that follow require familiarity with calculus. In writing chapter 6, we decided to follow Hamilton's crucial papers as closely as possible, making some sections of this chapter perhaps less accessible

than the material found in previous chapters. We included a discussion on kinetic foci, which emphasizes the fact that the principle of least action should perhaps be called the principle of extremal (or *critical*) action since action is not always least. For classical (non-quantum) systems, the action is an extremum that can never be a maximum; that leaves us with a minimum or a saddle point, and both are possible. Chapter 7 is an overview of relativity in connection with least action. Finally, in Chapter 8 we narrate the development of quantum theory and, in particular, we trace its connection with the optical mechanical analogy and the principle of least action.

2

Prehistory of Variational Principles

2.1 Queen Dido and the Isoperimetric Problem

Consider a loop of thread lying on a table. How can we distort the loop, without stretching the thread, so that it encloses the maximum area?

The problem appears in a story told by the Roman poet, Virgil, in his epic poem, *The Aeneid* (Virgil, 19 BC):

> They sailed to the place where today you'll see
> Stone walls going higher and the citadel
> Of Carthage, the new town. They bought the land,
> Called Drumskin [Byrsa] from the bargain made, a tract
> They could enclose with one bull's hide.

These verses refer to the legend of Queen Dido who fled her home because her brother, Pygmalion, had killed her husband and was plotting to steal all her money. She ended up on the north coast of Africa, where she was given permission to rule over whatever area of land she was able to enclose using the hide of only one bull. She cut the hide into thin strips, tying them together to form the longest loop she could make, in order to enclose the largest possible kingdom. Queen Dido seems to have discovered how to use this loop to maximize the area of her kingdom: using straight coastline as her side border, she enclosed the largest area of land possible by placing the loop in the shape of a semi-circle.

Queen Dido's story is now the emblem of the so-called isoperimetric problem: for a fixed perimeter, determine the shape of the closed, planar curve that encloses the maximum area. The answer is the circle. Aristotle, in *De caelo*, while discussing the motion of the heavens, displays some knowledge or intuition of this result (Aristotle, 350 BC/1922, Book II):

> Again, if the motion of the heavens is the measure of all movement ... and the minimum movement is the swiftest, then, clearly, the movement of the heavens must be the swiftest of all movements. Now of the lines which return upon themselves the line which bounds the circle is the shortest; and that movement is the swiftest which follows the shortest line.

However, a common assumption in ancient times was that the area of a figure is determined entirely by its perimeter (Gandz, 1940). For example, Thucydides, the great ancient historian, estimated the size of Sicily from its circumnavigation time which is proportional to the perimeter (Thucydides, 431 BC, Book VI):

> For the voyage round Sicily in a merchantman is not far short of eight days; and yet, large as the island is, there are only two miles of sea to prevent its being mainland.

The confusion persisted even up to the times of Galileo, who expresses the problem in Sagredo's voice (Galilei, 1638/1974, p. 61):

> people who lack knowledge of geometry . . . make the error when speaking of surfaces; for in determining the size of different cities, they often imagine that everything is known when the lengths [*quantità*] of the city boundaries are given, not knowing that one boundary might be equal to another, while the area contained by one be much greater than that in the other.

The isoperimetric problem was solved by the Greek mathematician Zenodorus (ca. 200 BC– ca. 140 BC). Although his work has been lost, we know of his proof through Pappus and Theon of Alexandria (Pappus, 1888; Heath, 1921). Zenodorus starts by shaping Queen Dido's loop into straight lines of different lengths to form an arbitrary irregular polygon. And then he shows that one can increase the area enclosed by that polygon by changing the lengths of the sides – without altering its perimeter or the number of sides – until they are all the same. In other words, he shows that, of all polygons of a given perimeter and a given number of sides, the equilateral polygon encloses the largest area. However, even if we fix the perimeter and the number of sides, there are still an infinite number of possible equilateral polygons. Zenodorus shows that, from this infinite set of equilaterals, the equiangular one encloses the largest area. And for the final piece of the proof: if we start with a regular polygon and increase its number of sides (keeping the perimeter fixed), the new polygon encloses a larger area. And this leads us naturally to the circle, which we can think of as a regular polygon with an infinite number of sides. The reader who is not interested in the details of Zenodorus's proof can skip the next section without loss of continuity.

*2.1.1 Zenodorus's Solution**

Zenodorus showed that a non-equilateral polygon can be manipulated to enclose a larger area without changing its perimeter. Pick a triangle of the irregular polygon – the shaded triangle of Figure 2.1. Keeping the base AB fixed, reshape the triangle into an isosceles triangle of the same perimeter by moving the vertex from

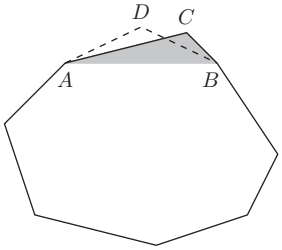

Figure 2.1 Adapted from Zenodorus. The equilateral polygon encloses a larger area than any irregular polygon with the same perimeter and the same number of sides.

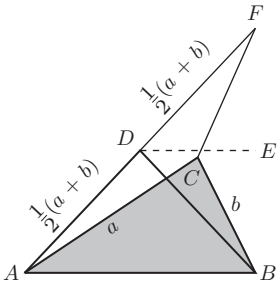

Figure 2.2 From Zenodorus. Among the triangles on a fixed base AB and fixed perimeter $AB + a + b$, the isosceles triangle ADB has the largest area. Triangle ADB, whose legs have length $(a + b)/2$, has the same perimeter as the starting (scalene) triangle ACB. Prolong AC to F so that $AD = DF$. The dashed line DE, parallel to AB, is a "line of symmetry:" the segment DB is the reflection of DF on the "mirror" DE, and all points below the line DE are closer to B than to F. Now consider the triangle ACF and use the "triangular inequality" (*in any triangle the sum of any two sides is greater than the remaining side*): $CF + a > a + b$ and the segment $CF > b$. Point C is therefore below CE; the height of the triangle ACB is therefore smaller than the height of ADB. Since both triangles have the same base, ADB has larger area.

C to D. Zenodorus shows that this reshaping increases the area. The specifics of the proof (see Figure 2.2) draw from the repertoire of "conjuring tricks" of the Greek geometers – the choreography of auxiliary lines and symmetric angles that reveal sometimes unexpected and paradoxical relations. The process can be repeated for all triangles of consecutive vertices of the polygon, allowing one to conclude that the polygon enclosing the largest area is equilateral.

The second step is to show that the maximum polygon is also equiangular. This Zenodorus proves by considering two consecutive triangles from the polygon (see Figure 2.3). Zenodorus proves that, given two non-similar isosceles triangles, if we construct, on the same bases, two *similar* triangles with the same total perimeter as the first two triangles, then the sum of the areas of the similar triangles is greater

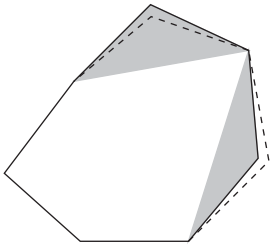

Figure 2.3 Adapted from Zenodorus. The area of a polygon can be increased by making it equiangular.

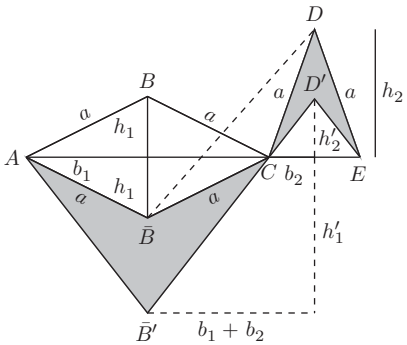

Figure 2.4 Adapted from Zenodorus. The sum of the areas of two isosceles triangles with different bases $AC = 2b_1$ and $CE = 2b_2$, but otherwise equal sides a, is smaller than the sum of the areas of two similar triangles with the same bases and equal total perimeter.

than the sum of the areas of the non-similar triangles. According to the account by Heath (1921), Zenodorus's proof is restricted. But we can show that a slight modification makes it valid in general.

Let us start with the non-similar isosceles triangles ABC and CDE (see Figure 2.4). Their bases are $AC = 2b_1$ and $CE = 2b_2$ respectively, their heights are h_1 and h_2, and all the legs are of length a. Following Zenodorus's logic, construct the triangle $A\bar{B}C$, which is the "mirror" image of triangle ABC, the mirror being along the line of the common bases.

Now construct the two similar triangles $A\bar{B}'C$ and $C\bar{D}'E$ with new heights h'_1 and h'_2, keeping the total perimeter the same:

$$B'D' = 2a. \tag{2.1}$$

Since the original triangles are not similar, the line $\bar{B}D$ joining their vertices is shorter than $2a$.[1]

$$\bar{B}D < 2a. \tag{2.2}$$

[1] This is due to the triangular inequality; the side $\bar{B}D$ is smaller than the sum of BC and CD.

Using the Pythagorean theorem

$$\begin{aligned}(\bar{B}'D')^2 &= (b_1+b_2)^2+(h_1'+h_2')^2\\ &> (\bar{B}D)^2\\ &= (b_1+b_2)^2+(h_1+h_2)^2,\end{aligned} \tag{2.3}$$

and

$$(h_1'+h_2') > (h_1+h_2). \tag{2.4}$$

This relation between the final and the initial heights is general and was obtained by Zenodorus. All we need for equation (2.4) to be valid is that $\overline{B}'D' > \overline{B}D$, and this inequality holds even for isosceles triangles of unequal legs ($BC \neq CD$).

In a slight deviation from the proof reproduced by Heath, consider the total change in area for the triangles,

$$\Delta A = b_1(h_1'-h_1)+b_2(h_2'-h_2). \tag{2.5}$$

Now take the larger of the bases (b_1 in our case) and replace it by the smaller one in equation (2.5) to obtain

$$\begin{aligned}\Delta A &> b_2(h_1'-h_1)+b_2(h_2'-h_2)\\ &= b_2(h_1'+h_2'-h_1'-h_2')\\ &> 0.\end{aligned} \tag{2.6}$$

The sum of the areas of two isosceles triangles is always increased if they are made similar (keeping the sum of the perimeters and the bases fixed): the shaded "hollow-angled (figure)" ($\kappa o\iota\lambda o\gamma\acute{\omega}\nu\iota o\nu$) $A\bar{B}'C\bar{B}$ is larger than the figure $CDED'$.

This does not mean that the maximum possible area is attained by making them similar.[2] All we have shown is that the area increases. Zenodorus proved that the largest polygon of a given number of sides is regular. The underlying argument used here is "one of a mathematician's finest weapons" (Hardy, 1967, p. 94), *reductio ad absurdum*, where an assertion is established by deriving an absurdity from its denial. We first *assume* that there exists a polygon of maximum area, but we find a way to make it even larger (by making it equilateral). We then assume that, from the set of equilateral polygons, we choose the one of maximum area. But we then show that making it equiangular makes it larger still. Thus we show that the only equilateral that cannot be further enlarged is the equiangular one. The proof "shuts all doors one after another, and leaves open only one, through which merely for that reason we must now pass" (Schopenhauer, 1969, p. 70). There is one more door to shut before we get to the circle: the proof that the area of a regular polygon is increased if we increase the number of sides (always for fixed perimeter).

[2] The maximum area, as can be proven using elementary calculus and also by a geometric method due to Steiner, is attained when $\sin\angle BCA/\sin\angle DCE = b_1/b_2$.