Chapter 1

Introduction

1.1. The brachistochrone

The calculus of variations has a clear starting point. In June of 1696, John (also known as Johann or Jean) Bernoulli challenged the greatest mathematicians of the world to solve the following new problem (Bernoulli, 1696; Goldstine, 1980):

> Given points A and B in a vertical plane to find the path AMB down which a movable point M must, by virtue of its weight, proceed from A to B in the shortest possible time.

Imagine a particle M of mass m, in a vertical gravitational field of strength g, that moves along the curve $y = y(x)$ between the two points $A = (a, y_a)$ and $B = (b, y_b)$ (see Figure 1.1). The time of descent T of the particle is

$$T = \int_0^T dt = \int_0^L \frac{dt}{ds}\, ds = \int_0^L \frac{1}{v}\, ds = \int_a^b \frac{1}{v}\sqrt{1 + y'^2}\, dx\,, \qquad (1.1)$$

where s is arc length, L is the length of the curve, and v is the speed of the particle.

If our particle moves without friction, the law of conservation of mechanical energy guarantees that the sum of the particle's kinetic

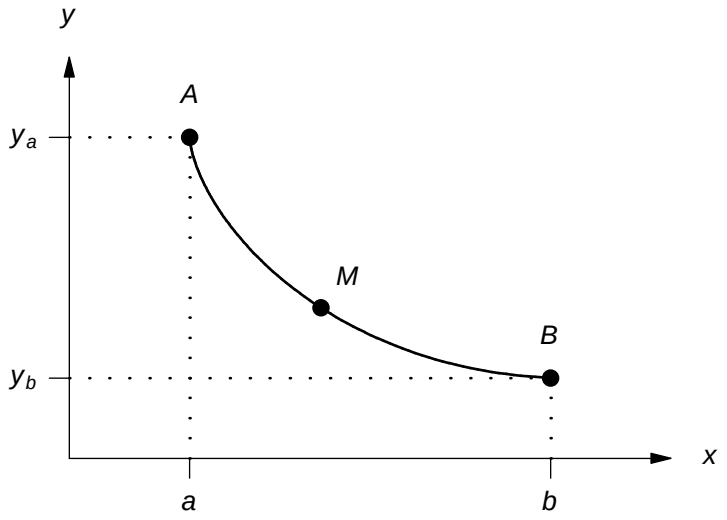

Figure 1.1. Curve of descent

energy and potential energy remains constant. If our particle starts from rest, we may thus write

$$\frac{1}{2}mv^2 + mgy \;=\; mgy_a\,. \tag{1.2}$$

The particle's speed is then

$$v \;=\; \sqrt{2g(y_a - y)}\,. \tag{1.3}$$

We now wish to find the *brachistochrone* (from $\beta\rho\alpha\chi\iota\sigma\tau\text{o}\varsigma$, shortest, and $\chi\rho\text{o}\nu\text{o}\varsigma$, time; John Bernoulli originally, but erroneously, wrote brachystochrone). That is, we wish to find the curve

$$y \;=\; y(x) \;\le\; y_a \tag{1.4}$$

that minimizes the integral

$$T \;=\; \frac{1}{\sqrt{2g}} \int_a^b \sqrt{\frac{1+y'^2}{y_a - y}}\; dx\,. \tag{1.5}$$

Several famous mathematicians responded to John Bernoulli's challenge. Solutions were submitted by Gottfried Wilhelm Leibniz

(1697), Isaac Newton (1695–7, 1697), John Bernoulli (1697a), James (or Jakob) Bernoulli (1697), and Guillaume l'Hôpital (1697).

Leibniz provided a geometrical solution. He derived the differential equation for the brachistochrone but did not specify the resulting curve (Goldstine, 1980). Leibniz also suggested that the brachistochrone be called the *tachystoptotam* (from $\tau\alpha\chi\iota\sigma\tau\mathrm{o}\varsigma$, swiftest, and $\pi\iota\pi\tau\epsilon\iota\nu$, to fall). Mercifully, this suggestion was ignored.

Newton's anonymous solution was published in the *Philosophical Transactions*; it was then reprinted in the *Acta Eruditorum*. Newton provided the correct answer but gave no clue to his method. Despite Newton's anonymity, John Bernoulli recognized that the work was "ex ungue Leonem" (from the claw of the Lion) and the *Acta Eruditorum* listed Newton in its index of authors.

John Bernoulli provided two solutions. The first solution relied on an analogy between the mechanical brachistochrone and light. Bernoulli (1697a) was quite taken with Fermat's principle of least time for light and argued that the brachistochrone "is the curve that a light ray would follow on its way through a medium whose density is inversely proportional to the velocity that a heavy body acquires during its fall." He broke up the optical medium into thin horizontal layers, chose an appropriate index of refraction, and used Snell's law of refraction and calculus to determine the shape of the brachistochrone. John Bernoulli (1718) described his second solution many years later. This second solution received little attention at the time but is now viewed as the first sufficiency proof in the calculus of variations.

James Bernoulli's solution was not as elegant as that of his younger brother, but it contained the key idea of varying only one value of the solution curve at a time. This idea provided the basis for further work in the calculus of variations. James Bernoulli called his solution an *oligochrone* (from $\mathrm{o}\lambda\iota\gamma\mathrm{o}\varsigma$, little, and $\chi\rho\mathrm{o}\nu\mathrm{o}\varsigma$, time).

We shall see that the brachistochrone is the inverted cycloid

$$x(\phi) \,=\, a + R(\phi - \sin\,\phi)\,, \quad y(\phi) \,=\, y_a - R\,(1 - \cos\,\phi)\,, \qquad (1.6)$$

where the parameter R is uniquely determined by the initial and terminal points. This cycloid is the curve traced by a point on the

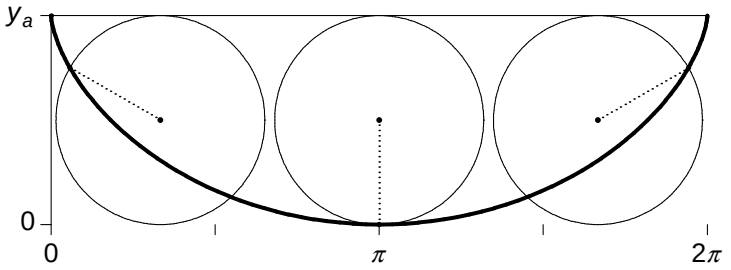

Figure 1.2. Cycloid for $R = \frac{1}{2}y_a$ and $a = 0$

circumference of a circle of radius R rolling along the bottom of the horizontal line $y = y_a$ (see Figure 1.2).

Huygens (1673, 1986) had previously shown that an inverted cycloid is a tautochrone (from $\tau\alpha\upsilon\tau$o or τo $\alpha\upsilon\tau$o, the same, and $\chi\rho$oνoς, time): the time for a heavy particle to fall to the bottom of this curve is independent of the upper starting point. To John Bernoulli's astonishment, the brachistochrone was Huygens' tautochrone.

The brachistochrone is one of many problems where we wish to determine a function, $y(x)$, that minimizes or maximizes the integral

$$J[y(x)] \;=\; \int_a^b f(x, y(x), y'(x))\; dx\,. \tag{1.7}$$

Leonhard Euler first devised a systematic method for solving such problems.

In the remainder of this chapter, we will examine three other problems that involve minimizing or maximizing integrals. We will first look at another brachistochrone problem, for travel *through* the earth. We will then look at the problem of finding the shortest path between two points on some general surface. Finally, we will look at the "soap-film problem," the problem of minimizing the surface area of a surface of revolution. All of these problems can be attacked using the calculus of variations.

1.2. The terrestrial brachistochrone

History repeats itself. In August of 1965, *Scientific American* published an article on "High-Speed Tube Transportation" (Edwards, 1965). Edwards proposed tube trains that would fall through the earth, pulled by gravity and helped along by pneumatic propulsion. The advantages cited by Edwards included:

> (1) It brings most of the tunnel down into deep bedrock, where the cost of tunneling — by blasting or by boring — is reduced and incidental earth shifts are minimized; the rock is more homogeneous in consistency and there is less likelihood of water inflow.
>
> (2) The nuisance to property owners decreases with depth, so the cost of easements should be lower.
>
> (3) A deep tunnel does not interfere with subways, building foundations, utilities, or water wells....
>
> (4) The pendulum ride is uniquely comfortable for the passenger....

Lest you think this pure fantasy, a pneumatic train was constructed in New York City, under Broadway, from Warren Street to Murray Street, in 1870 by Alfred Ely Beach (an early owner of *Scientific American*). This was New York City's first subway (Roess and Sansone, 2013). You can see a drawing of the pneumatic train on the wallpaper in older Subway Sandwich shops.

Cooper (1966a) then pointed out that straight-line chords lead to needlessly long trips through the earth. He used the calculus of variations to derive a differential equation for the fastest tunnels through the earth and integrated this equation numerically. Venezian (1966), Mallett (1966), Laslett (1966), and Patel (1967) then found first integrals and analytic solutions for this problem. See Cooper (1966b) for a summary.

Let us take a closer look at this *terrestrial brachistochrone* problem. Assume that the earth is a homogeneous sphere of radius R. Consider a section through the earth with polar coordinates centered at the heart of the earth (see Figure 1.3). Imagine a particle of mass

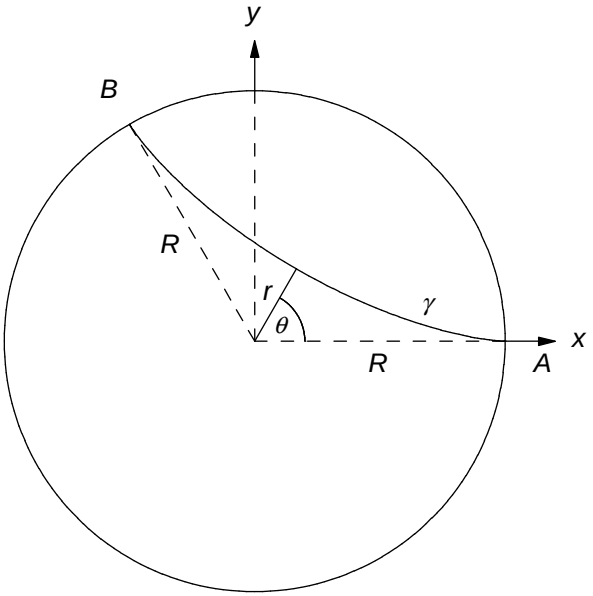

Figure 1.3. Path through the earth

m that moves between two points, $A = (r_a, \theta_a)$ and $B = (r_b, \theta_b)$, on or near the surface of the earth. We now wish to find the planar curve γ that minimizes the travel time

$$T = \int_0^T dt = \int_\gamma \frac{dt}{ds}\, ds = \int_\gamma \frac{1}{v}\, ds = \int_\gamma \frac{1}{v}\, \sqrt{dr^2 + r^2\, d\theta^2} \qquad (1.8)$$

between A and B, where s is arc length and v is the speed of the particle.

When a particle is outside a uniform spherical shell, the shell exerts a gravitational force equal to that of an identical point mass at the center of the shell. A particle inside the shell feels no force (see Exercise 1.6.3). By integrating over spherical shells of different radii (Exercise 1.6.4), one can show that the gravitational potential energy within a spherical and homogeneous earth can be written

$$V(r) = \frac{1}{2}\frac{mg}{R} r^2, \qquad (1.9)$$

where g is the magnitude of the gravitational acceleration at the surface of the earth.

For a particle starting at rest at the surface of the earth, conservation of energy now implies that

$$\frac{1}{2}mv^2 + \frac{1}{2}\frac{mg}{R}r^2 \;=\; \frac{1}{2}mgR \tag{1.10}$$

so that

$$v \;=\; \frac{\sqrt{g(R^2 - r^2)}}{\sqrt{R}}\,. \tag{1.11}$$

It follows that the total travel time is

$$T \;=\; \sqrt{\frac{R}{g}}\int_{\theta_a}^{\theta_b}\sqrt{\frac{\left(\frac{dr}{d\theta}\right)^2 + r^2}{R^2 - r^2}}\;d\theta\,. \tag{1.12}$$

We will look at this problem in greater detail later. We shall see that the terrestrial brachistochrone is a *hypocycloid,* the curve traced by a point on the circumference of a circle of radius either $[R - (S_{AB}/\pi)]$ (see Figure 1.4) or of radius S_{AB}/π (see Figure 1.5), where S_{AB} is the arc length along the surface of the earth between A and B, as it rolls inside a circle of radius R.

The fastest Amtrak train makes the 400 mile trip between Boston and Washington, D.C., in six and a half hours. A tube train moving along a straight-line chord between Boston and Washington would penetrate 5 miles into the earth and take 42 minutes. The fastest tube train along a hypocycloid would, in turn, penetrate 125 miles into the earth and take 10.7 minutes.

1.3. Geodesics

I do not want to give the impression that the calculus of variations is only brachistochrones. In this and the next section, we will look at two other classic problems.

A line is the shortest path between two points in a plane. We also wish to find shortest paths between pairs of points on other, more general, surfaces. To find these geodesics, we must minimize arc length.

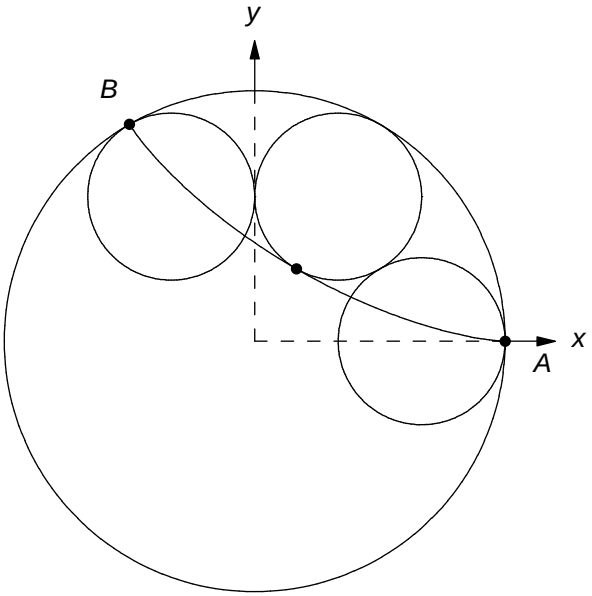

Figure 1.4. Hypocycloid with inner radius $\left(R - \frac{S_{AB}}{\pi}\right)$

The simplest case arises when the surface is a level set for one of the coordinates in a system of orthogonal curvilinear coordinates. The arc length can then be written using the scale factors of the coordinate system.

Consider, for example, two points, A and B, on a sphere of radius R centered at the origin. We wish to join A and B by the shortest, continuously differentiable curve lying on the sphere. We start by specifying position,

$$\mathbf{r}(x, y, z) \;=\; x\;\mathbf{i} \;+\; y\;\mathbf{j} \;+\; z\;\mathbf{k}\,, \tag{1.13}$$

using the Cartesian coordinates x, y, and z and Cartesian basis vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$. For points on the surface of a sphere, we now switch to the spherical coordinates r, θ, and ϕ (see Figure 1.6). Since

$$x \;=\; r\,\sin\theta\,\cos\phi\,,\quad y \;=\; r\,\sin\theta\,\sin\phi\,,\quad z \;=\; r\,\cos\theta\,, \tag{1.14}$$

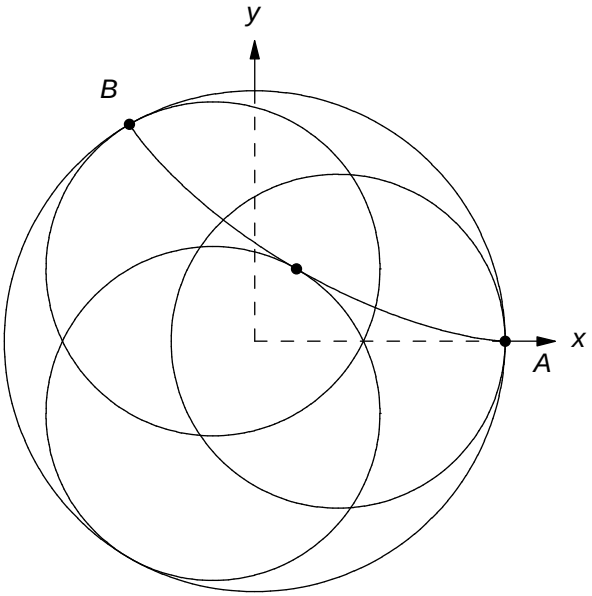

Figure 1.5. Hypocycloid with inner radius $\frac{S_{AB}}{\pi}$

the position vector $\mathbf{r}$ now takes the form

$$\mathbf{r}(r,\theta,\phi) \;=\; r\,\sin\,\theta\,\cos\,\phi\;\mathbf{i} \;+\; r\,\sin\,\theta\,\sin\,\phi\;\mathbf{j} \;+\; r\,\cos\,\theta\;\mathbf{k}\,. \quad (1.15)$$

Since this position vector depends on r, θ, and ϕ,

$$d\mathbf{r} \;=\; \frac{\partial \mathbf{r}}{\partial r}\,dr + \frac{\partial \mathbf{r}}{\partial \theta}\,d\theta + \frac{\partial \mathbf{r}}{\partial \phi}\,d\phi\,. \quad (1.16)$$

The three partial derivatives on the right-hand side of this equation are vectors tangent to motions in the r, θ, and ϕ directions. Thus

$$d\mathbf{r} \;=\; h_r\,dr\,\hat{e}_r \;+\; h_\theta\,d\theta\,\hat{e}_\theta \;+\; h_\phi\,d\phi\,\hat{e}_\phi\,, \quad (1.17)$$

where $\hat{e}_r$, $\hat{e}_\theta$, and $\hat{e}_\phi$ are unit vectors in the r, θ, and ϕ directions and

$$h_r = \left|\left|\frac{\partial \mathbf{r}}{\partial r}\right|\right| = 1\,,\;\; h_\theta = \left|\left|\frac{\partial \mathbf{r}}{\partial \theta}\right|\right| = r\,,\;\; h_\phi = \left|\left|\frac{\partial \mathbf{r}}{\partial \phi}\right|\right| = r\,\sin\,\theta \quad (1.18)$$

are the scale factors for spherical coordinates.

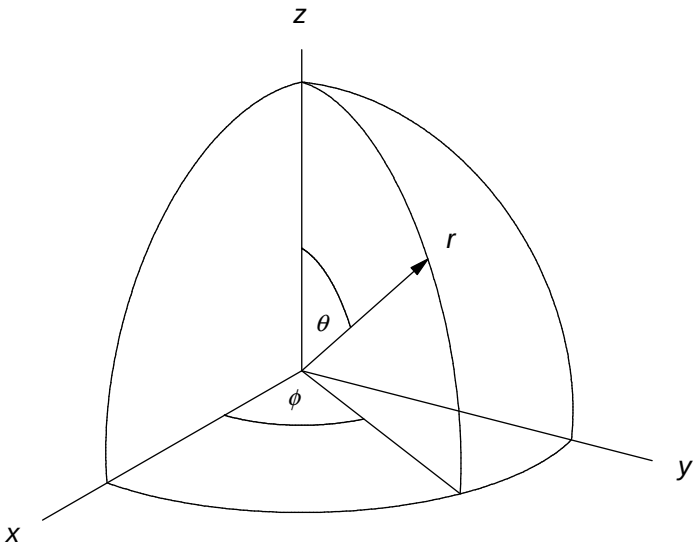

Figure 1.6. Spherical coordinates

The element of arc length in spherical coordinates is given by

$$\begin{aligned} ds &= \sqrt{d\mathbf{r}\cdot d\mathbf{r}} = \sqrt{h_r^2\,dr^2 + h_\theta^2\,d\theta^2 + h_\phi^2\,d\phi^2} \\ &= \sqrt{dr^2 + r^2\,d\theta^2 + r^2\,\sin^2\theta\,d\phi^2}\,. \end{aligned} \tag{1.19}$$

For a sphere of radius $r = R$, this element reduces to

$$ds = R\sqrt{d\theta^2 + \sin^2\theta\,d\phi^2}\,. \tag{1.20}$$

If we assume that $\phi = \phi(\theta)$, finding the curve that minimizes the arc length between the points $A = (\theta_a, \phi_a)$ and $B = (\theta_b, \phi_b)$ simplifies to finding the function $\phi(\theta)$ that minimizes the integral

$$s = \int_A^B ds = R\int_{\theta_A}^{\theta_B} \sqrt{1 + \sin^2\theta\,(d\phi/d\theta)^2}\;d\theta \tag{1.21}$$

subject to the boundary conditions

$$\phi(\theta_a) = \phi_a\,, \quad \phi(\theta_b) = \phi_b\,. \tag{1.22}$$

We will see, later, that the shortest paths on a sphere are arcs of great circles.

Unfortunately, we cannot expect every interesting surface to be the level set for some common coordinate. We may, however, hope to represent our surface parametrically. We may prescribe the x, y, and z coordinates of points on the surface using the parameters u and v and write our surface in the vector form

$$\mathbf{r}(u,v) \;=\; x(u,v)\,\mathbf{i} \;+\; y(u,v)\,\mathbf{j} \;+\; z(u,v)\,\mathbf{k}\,. \tag{1.23}$$

We can now specify a curve on this surface by prescribing u and v in terms of a single parameter — call it t — so that

$$u \;=\; u(t)\,, \quad v \;=\; v(t)\,. \tag{1.24}$$

The vector

$$\dot{\mathbf{r}} \;\equiv\; \frac{d\mathbf{r}}{dt} \;=\; \frac{\partial \mathbf{r}}{\partial u}\,\dot{u} + \frac{\partial \mathbf{r}}{\partial v}\,\dot{v} \tag{1.25}$$

is tangent to both the curve and the surface. We find the square of the distance between two points on a curve by integrating

$$ds^2 \;=\; d\mathbf{r}\cdot d\mathbf{r} \;=\; \left(\frac{\partial \mathbf{r}}{\partial u}\,du + \frac{\partial \mathbf{r}}{\partial v}\,dv\right)\cdot\left(\frac{\partial \mathbf{r}}{\partial u}\,du + \frac{\partial \mathbf{r}}{\partial v}\,dv\right) \tag{1.26}$$

along the curve. Equation (1.26) is often written

$$ds^2 \;=\; E\,du^2 + 2\,F\,du\,dv + G\,dv^2\,, \tag{1.27}$$

where

$$E \;=\; \frac{\partial \mathbf{r}}{\partial u}\cdot\frac{\partial \mathbf{r}}{\partial u}\,, \quad F \;=\; \frac{\partial \mathbf{r}}{\partial u}\cdot\frac{\partial \mathbf{r}}{\partial v}\,, \quad G \;=\; \frac{\partial \mathbf{r}}{\partial v}\cdot\frac{\partial \mathbf{r}}{\partial v}\,. \tag{1.28}$$

The right-hand side of equation (1.27) is called the *first fundamental form* of the surface. The coefficients $E(u,v)$, $F(u,v)$, and $G(u,v)$ have many names. They are sometimes called first-order fundamental magnitudes or quantities. Other times, they are simply called the coefficients of the first fundamental form.

The distance between the two points $A = (u_a, v_a)$ and $B = (u_b, v_b)$ on the curve $u = u(t)$, $v = v(t)$ may now be written

$$s \;=\; \int_{t_a}^{t_b} \sqrt{E\left(\frac{du}{dt}\right)^2 + 2\,F\frac{du}{dt}\frac{dv}{dt} + G\left(\frac{dv}{dt}\right)^2}\;dt\,, \tag{1.29}$$

with

$$u(t_a) = u_a\,, \quad v(t_a) = v_a\,, \quad u(t_b) = u_b\,, \quad v(t_b) = v_b\,. \tag{1.30}$$

In this formulation, we have two dependent variables, $u(t)$ and $v(t)$, and one independent variable, t. If v can be written as a function of u, $v = v(u)$, we can instead rewrite our integral as

$$s = \int_{u_a}^{u_b} \sqrt{E + 2\,F\left(\frac{dv}{du}\right) + G\left(\frac{dv}{du}\right)^2}\; du \tag{1.31}$$

with

$$v(u_a) = v_a, \quad v(u_b) = v_b\,. \tag{1.32}$$

This is now a problem with one dependent variable and one independent variable.

To make all this concrete, let us take, as an example, the *pseudosphere* (see Figure 1.7), half of the surface of revolution generated by rotating a tractrix about its asymptote. If the asymptote is the z-axis, we can write the equation for a pseudosphere, parametrically, as

$$\begin{aligned}\mathbf{r}(u,v) &= a\,\sin u\,\cos v\;\mathbf{i} + a\,\sin u\,\sin v\;\mathbf{j} \\ &\quad + a\left(\cos u + \ln\tan\frac{u}{2}\right)\;\mathbf{k}\,.\end{aligned} \tag{1.33}$$

Since

$$\begin{aligned}\mathbf{r}_u &= \frac{\partial \mathbf{r}}{\partial u} \\ &= (a\,\cos u\,\cos v,\, a\,\cos u\,\sin v, -a\,\sin u + a\,\csc u)\end{aligned} \tag{1.34}$$

and

$$\mathbf{r}_v = \frac{\partial \mathbf{r}}{\partial v} = (-a\,\sin u\,\sin v,\, a\,\sin u\,\cos v,\, 0)\,, \tag{1.35}$$

the first-order fundamental quantities reduce to

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = a^2\,\cot^2 u\,, \tag{1.36}$$

$$F = \mathbf{r}_u \cdot \mathbf{r}_v = 0\,, \tag{1.37}$$

$$G = \mathbf{r}_v \cdot \mathbf{r}_v = a^2\,\sin^2 u\,. \tag{1.38}$$

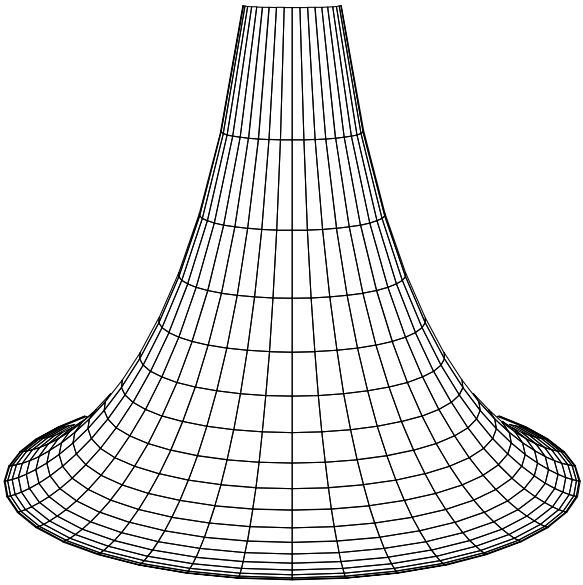

Figure 1.7. Pseudosphere

To determine a geodesic on the pseudosphere, we must thus find a curve, $u = u(t)$ and $v = v(t)$, that minimizes the arc-length integral

$$s \;=\; a \int_{t_a}^{t_b} \sqrt{\cot^2 u\, \dot{u}^2 + \sin^2 u\, \dot{v}^2}\; dt \tag{1.39}$$

subject to the boundary conditions

$$u(t_a) \;=\; u_a\,, \quad v(t_a) \;=\; v_a\,, \quad u(t_b) \;=\; u_b\,, \quad v(t_b) \;=\; v_b\,. \tag{1.40}$$

Alternatively, we may look for a curve, $v = v(u)$, that minimizes the integral

$$s \;=\; a \int_{u_a}^{u_b} \sqrt{\cot^2 u + \sin^2 u \left(\frac{dv}{du}\right)^2}\; du \tag{1.41}$$

subject to the boundary conditions

$$v(u_a) \;=\; v_a, \quad v(u_b) \;=\; v_b\,. \tag{1.42}$$

For other examples, see Exercise 1.6.6.

John Bernoulli (1697b) posed the problem of finding geodesics on convex surfaces. In 1698, he remarked, in a letter to Leibniz, that geodesics always have osculating planes that cut the surface at right angles. (An osculating plane is the plane that passes through three nearby points on a curve as two of these points approach the third point.) This geometric property is frequently used as the definition of a geodesic curve, irrespective of whether the curve actually minimizes arc length. Later, Euler (1732) derived differential equations for geodesics on surfaces using the calculus of variations. This was Euler's earliest known use of the calculus of variations.

Finding shortest paths is easiest on simple surfaces of revolution. Geodesics on surfaces of revolution satisfy a simple first integral or "conservation law" that was first published by Clairaut (1733). Jacobi (1839), in a tour de force, succeeded in integrating the equations of geodesics for a more complicated surface, a triaxial ellipsoid.

1.4. Minimal surfaces

We may minimize areas as well as lengths. Consider two points,

$$y(a) \;=\; y_a\,, \quad y(b) \;=\; y_b\,, \tag{1.43}$$

in the plane (see Figure 1.8). We wish to join these two points by a continuously differentiable curve,

$$y \;=\; y(x) \;\geq\; 0\,, \tag{1.44}$$

in such a way that the surface of revolution, generated by rotating this curve about the x-axis, has the smallest possible area S. In other words, we wish to minimize

$$S \;=\; 2\pi \int_a^b y(x)\,\sqrt{1 \,+\, y'^2}\;dx\,. \tag{1.45}$$

Some of you will recognize this as the "soap-film problem." Suppose we wish to find the shape of a soap film that connects two wire hoops. For a soap film with constant film tension, the surface energy is proportional to the area of the film. Minimizing the surface energy of the film is thus equivalent to minimizing its surface area (Isenberg,

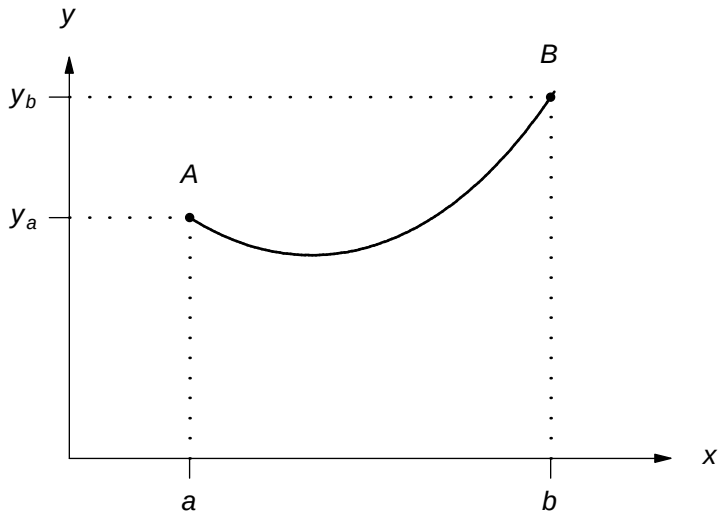

Figure 1.8. Profile curve

1992; Oprea, 2000). (For a closed soap *bubble,* without fixed boundaries, excess air pressure within the bubble prevents the surface area of the bubble from shrinking to zero.)

Euler (1744) discovered that the *catenoid,* the surface generated by a catenary or hanging chain (see Figure 1.9), minimizes surface area. As you doubtless know, however, from playing with soap films, if you pull two parallel hoops too far apart, the catenoid breaks, leaving soap film on the hoops. This was first shown analytically by Goldschmidt (1831). For two parallel, coaxial hoops of radius r, the area of a catenoid is an absolute minimum if the distance between the hoops is less than $1.056\,r$. This area is a relative minimum for distances between $1.056\,r$ and $1.325\,r$. For distances greater than $1.325\,r$, the catenoid breaks and the solution jumps to the discontinuous *Goldschmidt solution* (two disks).

Joseph Lagrange (1762) then proposed the general problem of finding a surface, $z = f(x, y)$, with a closed curve C as its boundary, that has the smallest area. That is, we now wish to minimize a *double*

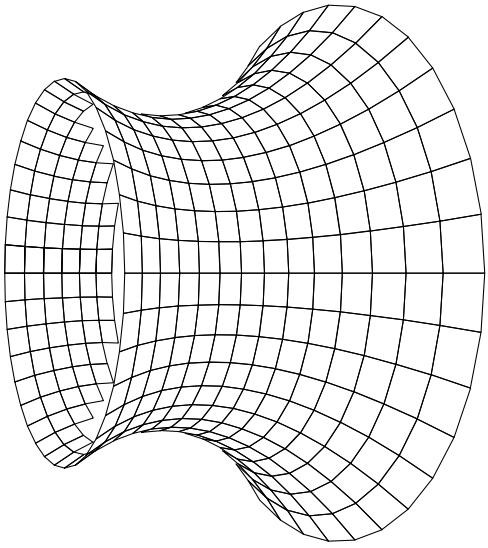

Figure 1.9. Catenoid

integral of the form

$$S \;=\; \iint\limits_{\Omega} \sqrt{1 + f_x^2 + f_y^2}\; dx\, dy \tag{1.46}$$

(see Exercise 1.6.7), where $\partial\Omega$ is the projection of the closed curve C onto the (x, y) plane and Ω is the interior of this projection. This problem has been known, starting with Lebesgue (1902), as Plateau's problem, in honor of Joseph Plateau's extensive experiments (Plateau, 1873) with soap films.

Lagrange showed that a surface that minimizes integral (1.46) must satisfy the *minimal surface equation*

$$(1 + f_y^2)\, f_{xx} - 2\, f_x\, f_y\, f_{xy} + (1 + f_x^2)\, f_{yy} \;=\; 0\,, \tag{1.47}$$

a quasilinear, elliptic, second-order, partial differential equation. Different constraints on the function $f(x, y)$ (e.g., Exercise 1.6.10) yield different minimal surfaces.

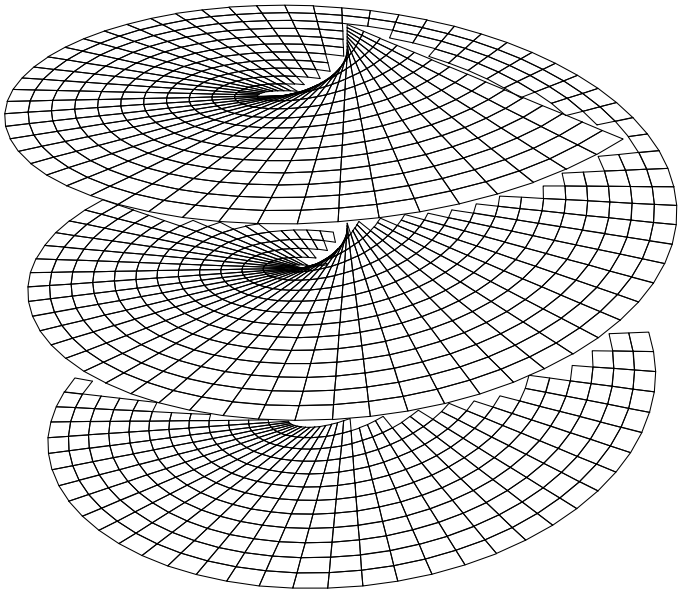

Figure 1.10. Helicoid

Jean-Baptiste-Marie-Charles Meusnier (1785) soon gave equation (1.47) a geometric interpretation. At each point P of a smooth surface, choose a vector normal to the surface, cut the surface with normal planes (that contain the normal vector but that differ in orientation), and obtain a series of plane curves. For each plane curve, determine the curvature at P. Find the minimum and maximum curvatures (from amongst all the plane curves passing through P). These are your *principal curvatures*.

Meusnier showed that the minimal surface equation implies that the *mean curvature* (the average of the principal curvatures) is zero at every point of the minimizing surface. As a result, any surface with zero mean curvature is typically referred to as a minimal surface, even if it does not provide an absolute or relative minimum for surface area. Meusnier also discovered that the catenoid and the *helicoid*, the surface formed by line segments perpendicular to the axis of a circular helix as they go through the helix (see Figure 1.10), satisfy Lagrange's

minimal surface equation. (Meusnier, like Lagrange, seemed unaware of Euler's earlier analysis of the catenoid.) The study of minimal surfaces has grown to become one of the richest areas of mathematical research.

In the remainder of this book, we will look at many other problems in the calculus of variations.

1.5. Recommended reading

Goldstine (1980), Fraser (2003), Kolmogorov and Yushkevich (1998), and Kline (1972) provide useful historical surveys of the calculus of variations.

Icaza Herrera (1994), Sussmann and Willems (1997), and Stein and Weichmann (2003) have written stimulating historical articles about the brachistochrone problem. An experimental study of the brachistochrone (using a "Hot Wheels" car) was carried out by Phelps et al. (1982).

The original 1697 solutions of John and Jacob Bernoulli can be found, translated into English, in Struik (1969). John Bernoulli's solution was recently reviewed by Erlichson (1999) and reviewed and generalized by Filobello-Nino et al. (2013).

If the endpoints A and B lie above the surface of the earth, but at vastly different heights, the gravitational field is no longer constant. One must instead determine the curve of swiftest descent in an attractive, inverse-square, gravitational field. This problem has been discovered repeatedly. Recent treatments include those of Singh and Kumar (1988), Parnovsky (1998), Tee (1999), and Hurtado (2000).

Goldstein and Bender (1986) analyzed the brachistochrone in the presence of relativistic effects and Farina (1987) showed that John Bernoulli's optical method can also be used to solve this relativistic problem. Kamath (1992) determined the relativistic tautochrone using fractional calculus.

The idea of high-speed tunnels through the earth is quite old. In Lewis Carroll's (1894) *Sylvie and Bruno Concluded*, Mein Herr describes a system of railway trains, without engines, powered by gravity:

> "Each railway is in a long tunnel, perfectly straight: so of course the *middle* of it is nearer the centre of the globe than the two ends: so every train runs half-way *down*-hill, and that gives it force enough to run the *other* half *up*-hill."

To which a protagonist replies:

> "Thank you. I understand that perfectly," said Lady Muriel. "But the velocity, in the *middle* of the tunnel, must be something *fearful!*"

You can also find a homework problem, about a tunnel-train between Minneapolis and Chicago, in Brooke and Wilcox (1929). See also Kirmser (1966).

Edwards' (1965) article reignited keen interest in gravity-powered transportation and inspired the articles by Cooper (1966a,b), Venezian (1966), Mallett (1966), Laslett (1966), and Patel (1967) on the terrestrial brachistochrone. Aravind (1981) applied John Bernoulli's optical method to the terrestrial brachistochrone and Prussing (1976), Chander (1977), McKinley (1979), and Denman (1985) pointed out that terrestrial brachistochrones are also tautochrones. Stalford and Garrett (1994) analyzed the terrestrial brachistochrone using differential geometry and optimal control theory.

Struik (1933), Carathéodory (1937), and Kline (1972) summarize the early history of the study of geodesics. Geodesics are an important topic in differential geometry (Struik, 1961; Oprea, 2007), Riemannian geometry (Berger, 2003), and geometric modeling (Patrikalakis and Maekawa, 2002). See Bliss (1902) for examples of geodesics on a toroidal anchor ring and Sneyd and Peskin (1990) for examples of geodesic trajectories on general tubular surfaces.

Isenberg (1992) and Oprea (2000) provide interesting and readable introductions to the science and mathematics of soap films. Barbosa and Colares (1986), Nitsche (1989), Fomenko (1990), and Fomenko and Tuzhilin (1991) do an excellent job of presenting the history and theory of minimal surfaces.

1.6. Exercises

1.6.1. Descent time down a cycloidal curve. Show that the descent time down the cycloidal curve

$$x(\phi) \;=\; a + R(\phi - \sin\,\phi)\,, \quad y(\phi) \;=\; y_a - R(1 - \cos\,\phi) \tag{1.48}$$

is

$$T \;=\; \sqrt{\frac{R}{g}}\,\phi_b\,, \tag{1.49}$$

where ϕ_b is the angle ϕ corresponding to the point $B = (b, y_b)$. What is the descent time to the lowest point on the cycloid?

1.6.2. Complementary curves of descent. The authors Mungan and Lipscombe (2013) recently introduced the term *complementary curves of descent* to describe curves that have identical descent times.

(a) Determine the descent time for a straight line (shown in bold in Figure 1.11).

(b) Rewrite integral (1.5) in polar coordinates assuming, for convenience, that θ increases clockwise.

(c) Determine the descent time for the lower portion of the leminiscate

$$r \;=\; 2\,c\,\sqrt{\sin\theta\,\cos\theta} \tag{1.50}$$

(shown in bold in Figure 1.11). Hint:

$$\frac{d}{d\theta}\left(\frac{\cos^{1/4}\theta}{\sin^{1/4}\theta}\right) \;=\; -\frac{1}{4}\,\cos^{-3/4}\theta\,\sin^{-5/4}\theta\,. \tag{1.51}$$

(d) Verify that the lemniscate is complementary to the straight line.

1.6.3. Potential energy due to a spherical shell. The gravitational potential energy between two point masses, M and m, separated by a distance r is

$$V(r) \;=\; -\frac{GMm}{r}\,, \tag{1.52}$$

where G is the universal gravitational constant.

Calculate the potential energy of mass m at point P due to the gravitational attraction of a thin homogeneous spherical shell of mass

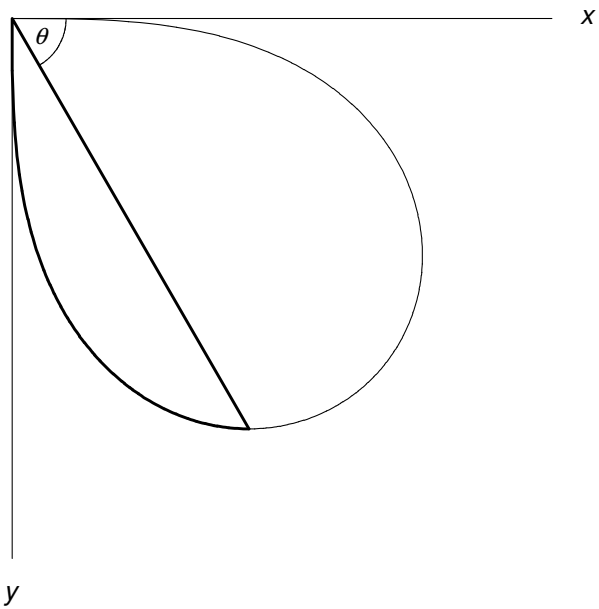

Figure 1.11. Complementary curves

M, surface (mass) density σ, and radius x by integrating over a set of ring elements. (See Figure 1.12.) Assume that point P is a distance r from the center of the shell and that y is the distance between the ring and point P. Be sure to consider the case when P is inside the shell ($r < x$) as well as outside the shell ($r > x$).

1.6.4. Potential energy inside the earth. Use your results from the last problem and integrate over shells of appropriate radii to show that the potential energy of a point mass m in a spherical and homogeneous earth can be written, to within an additive constant, as

$$V(r) \;=\; \frac{1}{2}\frac{mg}{R}\, r^2\,, \tag{1.53}$$

where R is the radius of the earth, g is the magnitude of the gravitational acceleration at the surface of the earth, r is the distance of the point mass from the center of the earth, and ρ is the (volumetric) density of the earth.

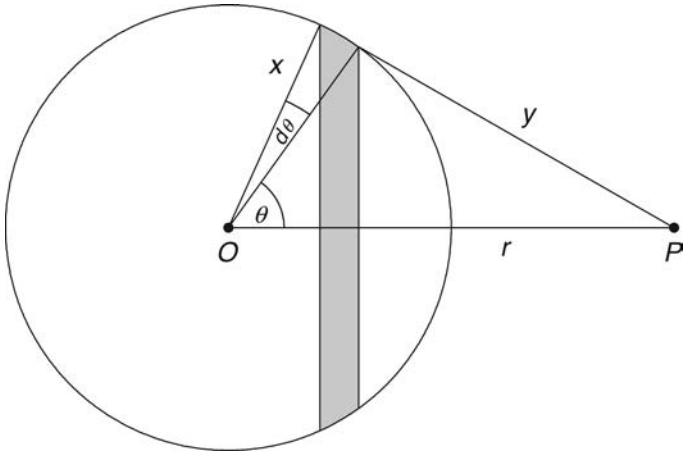

Figure 1.12. Geometry of a spherical shell

1.6.5. Gauss's law. Gauss's flux theorem for gravity states that the gravitational flux through a closed surface is proportional to the enclosed mass. Gauss's theorem can be written in differential form, using the divergence theorem, as

$$\nabla \cdot \mathbf{g} \;=\; -4\pi G\rho\,, \tag{1.54}$$

where G is the universal gravitational constant, ρ is the (volumetric) density of the enclosed mass, $\mathbf{g} = \mathbf{F}/m$ is the gravitational field intensity, m is the mass of a test point, and $\mathbf{F}$ is the force on this test mass.

(a) Use this theorem to determine the force $\mathbf{F}(r)$ acting on mass m at point P due to the gravitational attraction of a thin homogeneous spherical shell of mass M, surface density σ, and radius x. Assume that point P is a distance r from the center of the shell. Be sure to consider the case where point P is inside the shell ($r < x$) as well as outside the shell ($r > x$).

(b) Assume that $\mathbf{F}(r) = -dV/dr$, where $V(r)$ is the gravitational potential energy. Integrate the above force (starting at a reference point at infinity) to rederive the potential energy in Exercise 1.6.1.

(c) Use Gauss's flux theorem to determine the force $\mathbf{F}(r)$ acting on mass m at point P due to the gravitational attraction of a uniform solid sphere of mass M, density ρ, and radius R. Be sure to consider the case where point P is inside the shell ($r < R$) as well as outside the shell ($r > R$).

(d) Integrate the above force (starting at a reference point at infinity) to rederive the potential energy in Exercise 1.6.2.

1.6.6. First fundamental forms. Determine the first fundamental form for *three* of the following seven surfaces. The surfaces you may choose from are:

(a) the helicoid

$$x = u\cos v\,,\quad y = u\sin v\,,\quad z = a\,v\,; \tag{1.55}$$

(b) the torus

$$x = (b + a\cos u)\cos v\,,\quad x = (b + a\cos u)\cos v\,,\quad z = a\sin u\,; \tag{1.56}$$

(c) the catenoid

$$x = a\cosh\frac{u}{a}\cos v\,,\quad y = a\cosh\frac{u}{a}\sin v\,,\quad z = u\,; \tag{1.57}$$

(d) the general surface of revolution

$$x = f(u)\cos v\,,\quad y = f(u)\sin v\,,\quad z = g(u)\,; \tag{1.58}$$

(e) the sphere (with alternate parameterization)

$$x = \frac{4a^2u}{4a^2+u^2+v^2}\,,\quad y = \frac{4a^2v}{4a^2+u^2+v^2}\,, \tag{1.59}$$
$$z = a\,\frac{4a^2-u^2-v^2}{4a^2+u^2+v^2}\,;$$

(f) the ellipsoid

$$x = a\cos u\cos v\,,\quad y = b\cos u\sin v\,,\quad z = c\sin u\,; \tag{1.60}$$

(g) the hyperbolic paraboloid

$$x = a(u+v)\,,\quad y = b(u-v)\,,\quad z = uv\,. \tag{1.61}$$

1.6.7. Surface area. Consider a surface written in the vector form

$$\mathbf{r}(u,v) \;=\; x(u,v)\;\mathbf{i} \;+\; y(u,v)\;\mathbf{j} \;+\; z(u,v)\;\mathbf{k}\,, \tag{1.62}$$

where u and v are parameters.

(a) Justify or motivate the surface-area formula

$$S \;=\; \iint ||\mathbf{r}_u \times \mathbf{r}_v||\; du\, dv\,. \tag{1.63}$$

(b) Show that the above surface-area formula can also be written as

$$S \;=\; \iint \sqrt{EG - F^2}\; du\, dv\,, \tag{1.64}$$

where E, F, and G are the coefficients of the first fundamental form.

(c) Write the surface

$$z \;=\; f(x,y) \tag{1.65}$$

in vector form and show that the above formulas for area imply that

$$S \;=\; \iint_{\Omega} \sqrt{1 + f_x^2 + f_y^2}\; dx\, dy\,. \tag{1.66}$$

1.6.8. Surface area of a hyperbolic paraboloid. Consider the hyperbolic paraboloid

$$\mathbf{r}(u,v) \;=\; u\;\mathbf{i} \;+\; v\;\mathbf{j} \;+\; uv\;\mathbf{k}\,. \tag{1.67}$$

Determine the surface area for that portion of the paraboloid that is specified by values of u and v that lie in the first quadrant of (u,v) parameter space between the positive u- and v-axes and the circle

$$u^2 + v^2 \;=\; 1\,. \tag{1.68}$$

1.6.9. Surface area of a helicoid. Find the area of the portion of the helicoid

$$\mathbf{r}(u,v) \;=\; u\cos v\;\mathbf{i} \;+\; u\sin v\;\mathbf{j} \;+\; bv\;\mathbf{k} \tag{1.69}$$

that is specified by $0 \le u \le a$ and $0 \le v \le 2\pi$.

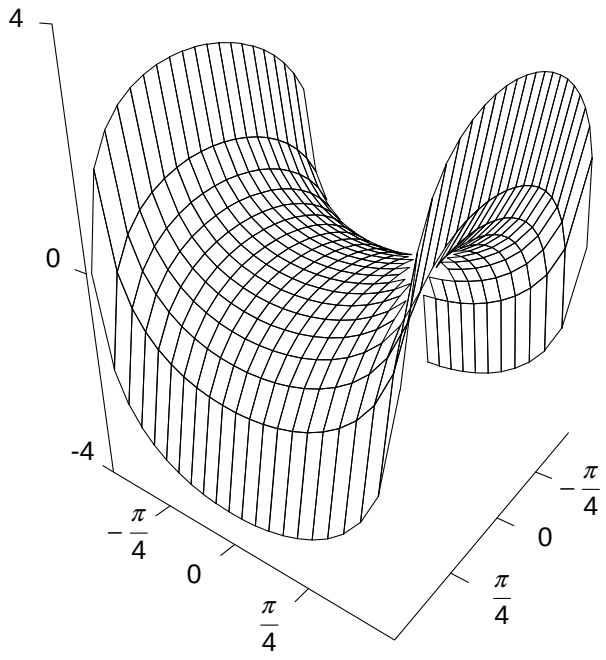

Figure 1.13. Scherk's first minimal surface

1.6.10. Scherk's minimal surface. Take the minimal surface equation, equation (1.47), and look for a solution of the form

$$f(x,y) \;=\; g(x) + h(y)\,. \tag{1.70}$$

Show that the resulting differential equation is separable. Solve for $g(x)$ and $h(y)$ to obtain *Scherk's (first) minimal surface,*

$$f(x,y) \;=\; c\,\ln\left[\frac{\cos\left(x/c\right)}{\cos\left(y/c\right)}\right]\,. \tag{1.71}$$

This surface was the first minimal surface discovered after the catenoid and the helicoid. A piece of this surface, for $c = 1$, $-\pi/2 < x < \pi/2$, and $-\pi/2 < y < \pi/2$, is shown in Figure 1.13.

Chapter 2

The First Variation

2.1. The simplest problem

Our goal is to minimize (or to maximize) a definite integral of the form

$$J[y] \;=\; \int_a^b f(x, y(x), y'(x))\; dx \tag{2.1}$$

subject to the boundary conditions

$$y(a) \;=\; y_a\,, \quad y(b) \;=\; y_b\,. \tag{2.2}$$

I wrote $J[y]$ rather than $J(y)$ to emphasize that we are dealing with *functionals* and not just functions. Our definite integral returns a real number for each function $y(x)$. A functional is an operator that maps functions to real numbers. Functional analysis was, originally, the study of functionals. The purpose of the calculus of variations is to maximize or minimize functionals.

We will encounter functionals that act on all or part of several well-known function spaces. Function spaces that occur in the calculus of variations include the following:

(a) $C[a, b]$, the space of real-valued functions that are continuous on the closed interval $[a, b]$;

(b) $C^1[a,b]$, the space of real-valued functions that are continuous and that have continuous derivatives on the closed interval $[a,b]$;

(c) $C^2[a,b]$, the space of real-valued functions that are continuous and that have continuous first and second derivatives on the closed interval $[a,b]$;

(d) $D[a,b]$, the space of real-valued functions that are piecewise continuous on the closed interval $[a,b]$; and

(e) $D^1[a,b]$, the space of real-valued functions that are continuous and that have piecewise continuous derivatives on the closed interval $[a,b]$.

A piecewise continuous function can have a finite number of jump discontinuities in the interval $[a,b]$. The right-hand and left-hand limits of the function exist at the jump discontinuities. A function that is piecewise continuously differentiable is continuous but may have a finite number of corners.

We wish to find the *extremum* of a functional. Extremum is a word that was first introduced by Paul du Bois-Reymond (1879b). Du Bois-Reymond got tired of always having to say "maximum or minimum" and so he introduced a single term, extremum, to talk about both maxima and minima. The term stuck.

We will take our lead from (ordinary) calculus. We will look for a condition analogous to setting the first derivative equal to zero in calculus. The resulting *Euler–Lagrange equation* is quite important, so much so that we will derive this equation in three ways. We will begin with Euler's heuristic derivation (Euler, 1744) and then move on to Lagrange's 1755 derivation (the traditional approach). We will then consider Paul du Bois-Reymond's modification of Lagrange's derivation (du Bois-Reymond, 1879a).

2.2. Euler's approach

Leonhard Euler was the first person to systematize the study of variational problems. His 1744 opus, *A Method for Finding Curved Lines Enjoying Properties of Maximum or Minimum, or Solution of Isoperimetric Problems in the Broadest Accepted Sense*, is a compendium of 100 special problems. The book also contains a general method for

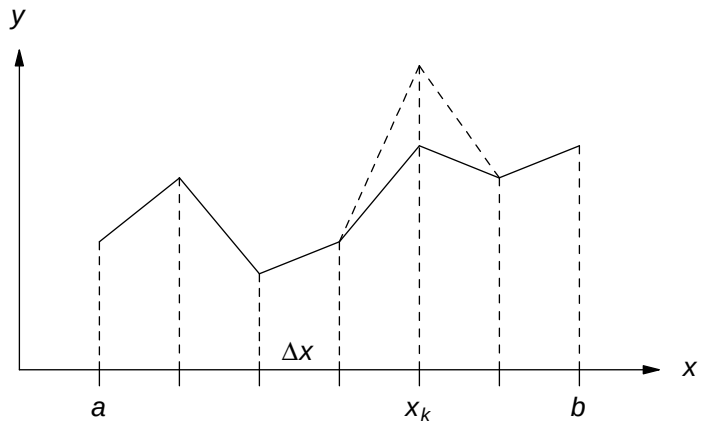

Figure 2.1. Polygonal curves

handling these problems. Euler dropped his method for Lagrange's more elegant "method of variations" after receiving Lagrange's (August 12, 1755) letter. Euler also named this subject the *calculus of variations* in Lagrange's honor.

Euler's essential idea was to first go from a variational problem to an n-dimensional problem and to then pass to the limit as $n \to \infty$. We will borrow from the modernized treatment of Euler's method found in Elsgolc (1961) and Gelfand and Fomin (1963). See Goldstine (1980) and Fraser (2003) for more on the original approach.

Let us divide the closed interval $[a, b]$ into $n+1$ equal subintervals (see Figure 2.1). We will assume that the subintervals are bounded by the points

$$x_0 \;=\; a,\, x_1, \ldots,\, x_n,\, x_{n+1} \;=\; b\,. \tag{2.3}$$

Each subinterval is of width

$$\Delta x \;=\; x_{i+1} - x_i \;=\; \frac{(b-a)}{n+1}\,. \tag{2.4}$$

We will also replace the smooth function $y(x)$ by the polygonal curve with vertices

$$(x_0, y_0),\, (x_1, y_1), \ldots,\, (x_n, y_n),\, (x_{n+1}, y_{n+1})\,. \tag{2.5}$$

Here, $y_i = y(x_i)$. We can now approximate the functional $J[y]$ by the sum

$$J(y_1, \ldots, y_n) \;\equiv\; \sum_{i=0}^{n} f\left(x_i, y_i, \frac{y_{i+1} - y_i}{\Delta x}\right) \Delta x\,, \tag{2.6}$$

a function of n variables. (Remember that $y_0 = y_a$ and $y_{n+1} = y_b$ are fixed.)

What is the effect of raising or lowering one of the free y_i? To answer this question, let us choose one of the free y_i, y_k, and take the partial derivative with respect to y_k. Since y_k appears in only two terms in our sum, the partial derivative is just

$$\begin{aligned} \frac{\partial J}{\partial y_k} \;&=\; f_y\left(x_k, y_k, \frac{y_{k+1} - y_k}{\Delta x}\right) \Delta x \\ &\quad + f_{y'}\left(x_{k-1}, y_{k-1}, \frac{y_k - y_{k-1}}{\Delta x}\right) \\ &\quad - f_{y'}\left(x_k, y_k, \frac{y_{k+1} - y_k}{\Delta x}\right). \end{aligned} \tag{2.7}$$

To find an extremum, we would ordinarily set this partial derivative equal to zero for each k. We also, however, want to take the limit as $n \to \infty$. In this limit, $\Delta x \to 0$ and the right-hand side of equation (2.7) goes to zero. The equation $0 = 0$, while true, is, sadly, not very helpful. To obtain a nontrivial result, we must first divide by Δx,

$$\begin{aligned} \frac{1}{\Delta x}\frac{\partial J}{\partial y_k} \;&=\; f_y\left(x_k, y_k, \frac{y_{k+1} - y_k}{\Delta x}\right) \\ &\quad - \frac{1}{\Delta x}\left[f_{y'}\left(x_k, y_k, \frac{y_{k+1} - y_k}{\Delta x}\right) - f_{y'}\left(x_{k-1},\, y_{k-1}, \frac{y_k - y_{k-1}}{\Delta x}\right)\right]. \end{aligned} \tag{2.8}$$

As we now let $n \to \infty$ and $\Delta x \to 0$, equation (2.8) yields the *variational derivative*

$$\frac{\delta J}{\delta y} \;=\; f_y(x, y, y') - \frac{d}{dx} f_{y'}(x, y, y')\,. \tag{2.9}$$

This variational derivative plays the same role for functionals that the partial derivative plays for multivariate functions. For a relative (or local) minimum, we expect this derivative to vanish at each point,

leaving us with the *Euler–Lagrange equation*

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) \;=\; 0\,. \tag{2.10}$$

This condition must be modified if the minimizing curve lies on the boundary rather than in the interior of the region of interest. Moreover, the Euler–Lagrange equation is only a *necessary* condition, in the same sense that $f'(x) = 0$ is a necessary, but not a sufficient, condition in calculus.

I should, perhaps, add that the above discussion is misleading to the extent that the formal notion of a variational or functional derivative was not introduced until much later, by Vito Volterra (1887), in the early stages of the development of functional analysis. See the recommended reading at the end of this chapter for more information about variational derivatives.

Example 2.1 (Shortest curve in the plane)**.**

Let's see what the Euler–Lagrange equation has to say about the shape of the shortest curve between two points, (a, y_a) and (b, y_b), in the plane. We clearly wish to minimize the arc-length functional

$$J[y] \;=\; \int_a^b ds \;=\; \int_a^b \sqrt{1 + y'^2}\; dx\,. \tag{2.11}$$

The integrand,

$$f(x, y, y') \;=\; \sqrt{1 + y'^2}\,, \tag{2.12}$$

does not depend on y and so the Euler–Lagrange equation reduces to

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{1 + y'^2}}\right) \;=\; 0\,. \tag{2.13}$$

Integrating once produces

$$\frac{y'}{\sqrt{1 + y'^2}} \;=\; \text{constant} \tag{2.14}$$

and we quickly conclude that

$$y' \;=\; c\,, \tag{2.15}$$

c a constant. If we integrate once again and set our new constant of integration to d, we conclude that

$$y \;=\; cx + d\,. \tag{2.16}$$

This is the equation of a straight line. The constants c and d can be determined from the boundary conditions.

2.3. Lagrange's approach

Let us return to the problem of minimizing or maximizing the functional

$$J[y] \;=\; \int_a^b f(x, y(x), y'(x))\; dx \tag{2.17}$$

subject to the boundary conditions

$$y(a) \;=\; y_a\,, \quad y(b) \;=\; y_b\,. \tag{2.18}$$

Euler derived the Euler–Lagrange equation by varying a single ordinate. Lagrange realized that he could derive this same equation while simultaneously varying *all* of the (free) ordinates.

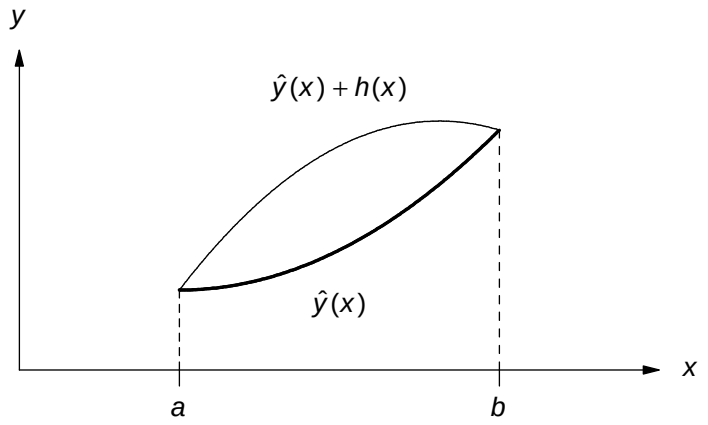

Figure 2.2. A small variation

Let us suppose that the function $y = \hat{y}(x)$ solves our problem. We now introduce $h(x)$, a small deviation or *variation* from this idealized solution,

$$y(x) = \hat{y}(x) + h(x) \tag{2.19}$$

(see Figure 2.2), that satisfies

$$h(a) = 0 \quad \text{and} \quad h(b) = 0\,. \tag{2.20}$$

At this point, we need to discuss a subtle point that escaped Lagrange but that turns out to be rather important. What exactly do we mean when we say that a variation is small? The usual way to measure the nearness of two functions is to compute the *norm* of the difference of the two functions. There are many possible norms and we will see that our conclusions about extrema (maxima and minima) are rather sensitive to which norm we use.

We will use two different norms throughout this course. They are the *weak* norm

$$||h||_w = \max_{[a,b]} |h(x)| \tag{2.21}$$

and the *strong* norm

$$||h||_s = \max_{[a,\,b]} |h(x)| + \sup_{[a,b]} |h'(x)\,|\,. \tag{2.22}$$

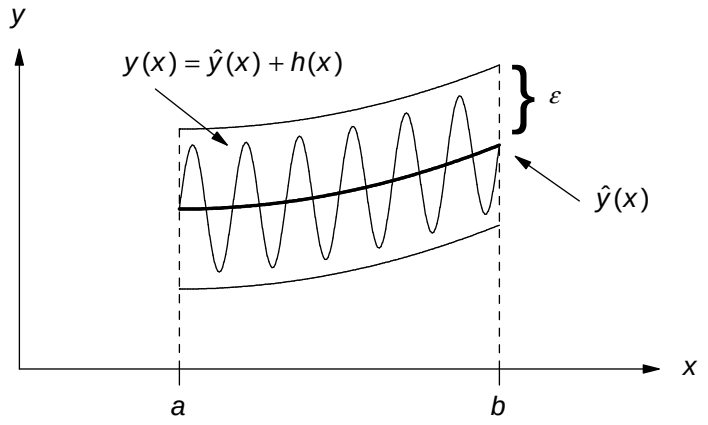

Figure 2.3. Strong variation

The supremum (or least upper bound) is here in case we are working with functions that are piecewise continuously differentiable. If our functions are continuously differentiable, the supremum can be replaced by a maximum.

We will use the weak and strong norms to establish neighborhoods in function space. Weak and strong norms permit different variations about the optimal solution. Since the weak norm does not impose any restriction upon the derivative, an ϵ-neighborhood in a weakly-normed space will include *strong variations* (see Figure 2.3) that differ significantly from the optimal solution in slope while remaining close in ordinate.

Strong variations may have arbitrarily large derivatives.

Example 2.2.

The function

$$h(x) \;=\; \epsilon \sin\left(\frac{x}{\epsilon^2}\right) \tag{2.23}$$

never exceeds ϵ and yet its derivative,

$$h'(x) \;=\; \frac{1}{\epsilon}\cos\left(\frac{x}{\epsilon^2}\right), \tag{2.24}$$

may become arbitrarily large as ϵ is made small.

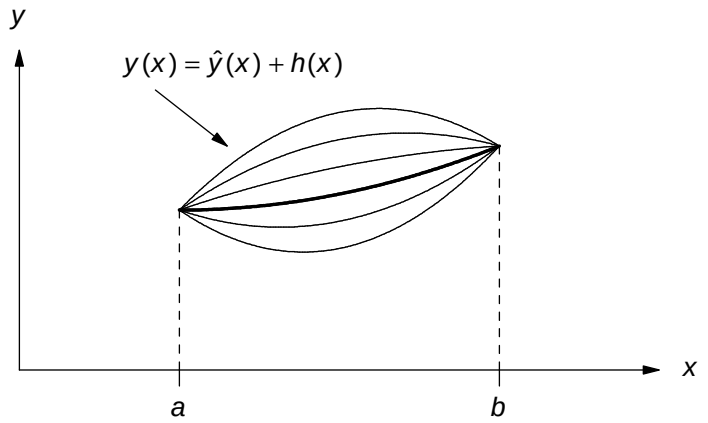

Figure 2.4. Weak variations

The strong norm, in contrast, *does* place a restriction on the size of the derivative. Stating that

$$||h||_s \;<\; \epsilon \tag{2.25}$$

implies not only

$$\max_{[a,b]} \;|h(x)| \;<\; \epsilon \tag{2.26}$$

but also

$$\sup_{[a,b]} \;|h'(x)| \;<\; \epsilon\,. \tag{2.27}$$

An ϵ-neighborhood in a strongly-normed space contains only weak variations (see Figure 2.4) that are close to the optimal solution in both ordinate *and* slope.

Since strong variations are a superset of weak variations, a function that minimizes a functional relative to nearby strong variations also minimizes that functional relative to nearby weak variations. Conversely, a necessary condition for a weak relative minimum is also a necessary condition for a strong relative minimum. Lagrange's approach uses weak variations. This is alright if we want necessary conditions but is a problem if we want sufficient conditions. In due course, we will encounter examples of functionals that have minima relative to weak variations, but not relative to strong variations.

To make Lagrange's assumption as explicit as possible, we will consider small weak variations

$$h(x) \;=\; \epsilon\,\eta(x) \tag{2.28}$$

where

$$\eta(a) \;=\; 0\,, \quad \eta(b) \;=\; 0 \tag{2.29}$$

and $h(x)$ and $h'(x)$ are of the same order of smallness. The function $\eta(x)$ is thus assumed to be independent of the parameter ϵ. As ϵ tends to zero, the variation $h(x)$ tends to zero in both ordinate and slope. For notational convenience, we will also think of the functional $J[y]$ as a function of ϵ,

$$J(\epsilon) \;\equiv\; J[\hat{y} + \epsilon\eta] \;=\; \int_a^b f(x, \hat{y} + \epsilon\eta, \hat{y}' + \epsilon\eta')\; dx\,. \tag{2.30}$$

Let us now look at the *total variation*

$$\Delta J \;=\; J(\epsilon) - J(0)\,. \tag{2.31}$$

That is,

$$\begin{aligned}\Delta J \;&=\; \int_a^b f(x, \hat{y} + \epsilon\eta,\, \hat{y}' + \epsilon\eta')\; dx - \int_a^b f(x, \hat{y}, \hat{y}')\; dx \qquad (2.32)\\ &=\; \int_a^b [f(x, \hat{y} + \epsilon\eta, \hat{y}' + \epsilon\eta') - f(x, \hat{y}, \hat{y}')]\; dx\,.\end{aligned}$$

If f has enough continuous partial derivatives — and we shall assume that it does — we may expand the total variation in a power series in ϵ. Using the usual Taylor expansion, we obtain

$$\Delta J \;=\; \delta J \,+\, \frac{1}{2}\,\delta^2 J \,+\, O(\epsilon^3)\,. \tag{2.33}$$

Here,

$$\begin{aligned}\delta J \;&=\; \left.\frac{dJ(\epsilon)}{d\epsilon}\right|_{\epsilon\,=\,0} \epsilon \qquad (2.34)\\ &= \epsilon \int_a^b [f_y(x, \hat{y}, \hat{y}')\,\eta + f_{y'}\,(x, \hat{y}, \hat{y}')\,\eta']\; dx\end{aligned}$$

is the *first variation.* Likewise,

$$\begin{aligned}\delta^2 J \;&=\; \left.\frac{d^2 J(\epsilon)}{d\epsilon^2}\right|_{\epsilon\,=\,0} \epsilon^2 \qquad (2.35)\\ &= \epsilon^2 \int_a^b [f_{yy}(x, \hat{y}, \hat{y}')\,\eta^2 + 2\, f_{yy'}\,(x, \hat{y}, \hat{y}')\,\eta\,\eta' + f_{y'y'}\,(x, \hat{y}, \hat{y}')\,\eta'^2]\; dx\end{aligned}$$

is the *second variation.* For ϵ sufficiently small, we expect that a nonvanishing first variation will dominate the right-hand side of total variation (2.33). Likewise, we expect that a nonvanishing second variation will dominate higher-order terms.

If $J[\hat{y}]$ is a relative (or local) minimum, we must have

$$\Delta J \;\geq\; 0 \tag{2.36}$$

for all sufficiently small ϵ. Since, however, the first variation is odd in ϵ, we can change its sign by changing the sign of ϵ. To prevent this change in sign, we require that

$$\delta J \;=\; 0\,. \tag{2.37}$$

For a minimum, we also require that

$$\delta^2 J \;\geq\; 0\,. \tag{2.38}$$

If we want a relative maximum, we will, in turn, require

$$\delta J \;=\; 0\,, \quad \delta^2 J \;\leq\; 0\,. \tag{2.39}$$

It is convenient, at this early stage of the course, to focus on the first variation. In light of the above arguments, we may safely say:

First variation condition:
A necessary condition for the functional $J[y]$ to have a relative (or local) minimum or maximum at $y = \hat{y}(x)$ is that the first variation of $J[y]$ vanish,

$$\delta J \;=\; 0\,, \tag{2.40}$$

for $y = \hat{y}(x)$ and for all admissible variations $\eta(x)$.

The first variation,

$$\delta J \;=\; \epsilon \int_a^b \left[f_y(x, \hat{y}, \hat{y}')\, \eta + f_{y'}\,(x, \hat{y}, \hat{y}')\, \eta' \right] dx\,, \tag{2.41}$$

is rather unwieldy as written. We will rewrite the first variation so as to factor out the dependence on the admissible variations $\eta(x)$. There are two different ways to do this. Both methods involve integration by parts. We start with Lagrange's approach.

2.3.1. Lagrange's simplification. Let us subject the second term in integrand (2.41) to integration by parts,

$$\int_a^b f_{y'}\,(x, \hat{y}, \hat{y}')\eta'\; dx \;=\; \eta(x)\; \frac{\partial f}{\partial y'}\bigg|_{x\,=\,a}^{x\,=\,b} - \int_a^b \eta \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) dx\,. \tag{2.42}$$

Since our variations from the idealized solution vanish at the endpoints of the interval,

$$\eta(a) \,=\, 0\,, \quad \eta(b) \,=\, 0\,, \tag{2.43}$$

our first necessary condition reduces to

$$\epsilon \int_a^b \eta(x) \left[\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \right]_{\hat{y},\hat{y}'} dx \;=\; 0 \tag{2.44}$$

for all admissible $\eta(x)$. The subscript in this last equation signifies that the expression in square brackets is evaluated at $y = \hat{y}(x)$ and $y' = \hat{y}'(x)$.

Let us note, right away, that our use of integration by parts, in this way, pretty much forces us to assume that $\hat{y}(x)$ is twice differentiable. The partial derivative $f_{y'}$ is generally a function of y' (as well as of y and x) and if y'' does not exist, the existence of

$$\frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \tag{2.45}$$

becomes doubtful. We shall see, momentarily, that Lagrange's simplification actually forces us to assume that $\hat{y}''(x) \in C[a,b]$ or $\hat{y}(x) \in C^2\,[a,b]$.

Lagrange claimed, without proof, that the coefficient of $\eta(x)$ in equation (2.44) must vanish, yielding the Euler–Lagrange equation,

$$\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \;=\; 0\,. \tag{2.46}$$

Euler pointed out, in a communication to Lagrange, that Lagrange's statement was not self-evident and that he really ought to *prove* that the coefficient of $\eta(x)$ must vanish. This proof was eventually supplied by du Bois-Reymond (1879a). Du Bois-Reymond's result is now known as the fundamental lemma of the calculus of variations.

Fundamental lemma of the calculus of variations:
If $M(x) \in C[a,b]$ and if

$$\int_a^b M(x)\,\eta(x)\;dx \;=\; 0 \tag{2.47}$$

for every $\eta(x) \in C^1[a,b]$ such that

$$\eta(a) \;=\; \eta(b) \;=\; 0\,, \tag{2.48}$$

then

$$M(x) \;=\; 0 \tag{2.49}$$

for all $x \in [a,b]$.

Proof. The proof is by contradiction. Suppose (without loss of generality) that $M(x)$ is positive at some point in (a,b). $M(x)$ must then, by continuity, be positive in some interval $[x_1,x_2]$ within $[a,b]$. Now (see Figure 2.5), let

$$\eta(x) \;=\; \begin{cases} (x-x_1)^2\,(x-x_2)^2\,, & x \in [x_1,x_2]\,, \\ 0\,, & x \notin [x_1,x_2]\,. \end{cases} \tag{2.50}$$

Clearly, $\eta(x) \in C^1[a,b]$. With this choice of $\eta(x)$,

$$\int_a^b M(x)\,\eta(x)\,dx \;=\; \int_{x_1}^{x_2} M(x)(x-x_1)^2(x-x_2)^2\;dx\,. \tag{2.51}$$

Since the integrand is nonnegative,

$$\int_a^b M(x)\,\eta(x)\;dx \;>\; 0\,. \tag{2.52}$$

This in contrary to our original hypothesis and it now follows that

$$M(x) \;=\; 0\,, \quad x \in (a,b)\,. \tag{2.53}$$

The continuity of $M(x)$, in turn, guarantees that $M(x)$ also vanishes at the endpoints of the interval. ♣

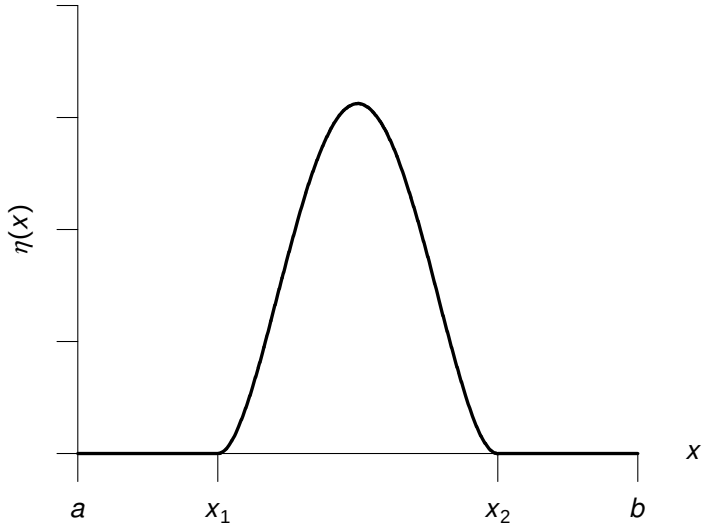

Figure 2.5. A nonnegative bump

To be able to apply this fundamental lemma of the calculus of variations, we must be sure that

$$M(x) \;=\; \frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) \tag{2.54}$$

is continuous on the closed interval $[a,\, b]$. If we apply the chain rule, we may rewrite the right-hand side of this last equation in the *ultra-differentiated* form

$$\begin{aligned} M(x) \;&=\; \frac{\partial f}{\partial y} - \frac{\partial f_{y'}}{\partial x}\,\frac{dx}{dx} - \frac{\partial f_{y'}}{\partial y}\,\frac{dy}{dx} - \frac{\partial f_{y'}}{\partial y'}\,\frac{dy'}{dx} \\ &=\; f_y - f_{y'x} - f_{y'y}\,y' - f_{y'y'}\,y''\,. \end{aligned} \tag{2.55}$$

To obtain the Euler–Lagrange equation using Lagrange's simplification, we must therefore make the additional assumption that $\hat{y}''(x) \in C[a,b]$ or that $\hat{y}(x) \in C^2[a,b]$.

Having made (or, more honestly, having been forced into) the assumption that $\hat{y}(x) \in C^2\,[a,b]$, we can now state the following *necessary* condition for a relative maximum or minimum:

Euler–Lagrange condition:
Every $\hat{y}(x) \in C^2[a,b]$ that produces a relative extremum of the integral

$$J[y] \;=\; \int_a^b f(x,y,y')\;dx \tag{2.56}$$

satisfies the Euler–Lagrange differential equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) \;=\; 0\,. \tag{2.57}$$

Lagrange's simplification forces us to *assume* that our solutions have continuous second derivatives. Can we loosen this assumption? Let us start with the necessary condition that the first variation must vanish,

$$\delta J[\eta] \;=\; \epsilon \int_a^b [f_y(x,\hat{y},\hat{y}')\,\eta + f_{y'}(x,\hat{y},\hat{y}')\,\eta']\;dx \;=\; 0\,, \tag{2.58}$$

and try a different approach.

2.3.2. Du Bois-Reymond's simplification. Let us now assume that the functions $\hat{y}(x)$ and $\eta(x)$ are merely continuously differentiable, $\hat{y}(x)$, $\eta(x) \in C^1[a,b]$. Since $f_{y'}(x,\hat{y},\hat{y}')$ depends on $\hat{y}'(x)$, this function need not be differentiable. As a result, we cannot integrate the second term in integrand (2.41) by parts.

Let us instead integrate the first term in integrand (2.41) by parts. Doing so, we obtain

$$\int_a^b f_y(x,\hat{y},\hat{y}')\,\eta\;dx \;=\; [\eta(x)\phi(x)]_{x=a}^{x=b} - \int_a^b \phi(x)\,\eta'(x)\;dx\,, \tag{2.59}$$

where

$$\phi(x) \;=\; \int_a^x f_y(u,\hat{y}(u),\hat{y}'(u))\;du\,. \tag{2.60}$$

Since we have only assumed the continuity of $f_y(x, \hat{y}, \hat{y}')$ and of $\eta'(x)$, this integration by parts *is* legal. Since

$$\eta(a) \;=\; \eta(b) \;=\; 0\,, \tag{2.61}$$

necessary condition (2.58) now reduces to

$$\int_a^b \left(\frac{\partial f}{\partial y'} - \int_a^x \frac{\partial f}{\partial y}\, du \right)_{\hat{y},\hat{y}'} \eta'(x)\, dx \;=\; 0\,. \tag{2.62}$$

We clearly need another lemma to progress further. Here it is:

Lemma of du Bois-Reymond:
If $M(x) \in C[a, b]$ and

$$\int_a^b M(x)\, \eta'(x)\, dx \;=\; 0 \tag{2.63}$$

for all $\eta(x) \in C^1[a, b]$ such that

$$\eta(a) \;=\; \eta(b) \;=\; 0\,, \tag{2.64}$$

then

$$M(x) \;=\; c\,, \tag{2.65}$$

a constant, for all $x \in [a, b]$.

Proof. We may prove this lemma by considering one well-chosen variation $\eta(x)$. Let μ denote the mean value of $M(x)$ on the closed interval $[a, b]$,

$$\mu \;=\; \frac{1}{(b-a)} \int_a^b M(x)\, dx\,. \tag{2.66}$$

Clearly,

$$\int_a^b [M(x) - \mu]\, dx \;=\; 0\,. \tag{2.67}$$

Now, consider the variation $\eta(x)$ defined by the equation

$$\eta(x) \;=\; \int_a^x [M(u) - \mu]\; du\,. \tag{2.68}$$

It is easy to see that $\eta(x) \in C^1[a,\, b]$. The function $\eta(x)$ also vanishes at $x = a$ and $x = b$. It is clearly an admissible variation. Moreover,

$$\eta'(x) \;=\; M(x) - \mu\,. \tag{2.69}$$

By hypothesis,

$$\int_a^b M(x)\,\eta'(x)\; dx \;=\; \int_a^b M(x)\,[M(x) - \mu]\; dx \;=\; 0\,. \tag{2.70}$$

Also,

$$\int_a^b M(x)\,[M(x) - \mu]\; dx - \mu \int_a^b [M(x) - \mu]\; dx \;=\; 0\,. \tag{2.71}$$

But, this last equation may be rewritten

$$\int_a^b [M(x) - \mu]^2\; dx \;=\; 0\,. \tag{2.72}$$

Let $x_0 \in [a, b]$ be a point where $M(x)$ is continuous. If $M(x_0) \neq \mu$, then there would have to exist a subinterval about $x = x_0$ on which $M(x) \neq \mu$. But this is clearly impossible in light of our last displayed equation. Thus $M(x) = \mu$ at all points of continuity. It follows that $M(x)$ is constant for all $x \in [a, b]$. ♣

We now wish to apply this lemma to necessary condition (2.62),

$$\int_a^b \left(\frac{\partial f}{\partial y'} - \int_a^x \frac{\partial f}{\partial y}\; du \right)_{\hat{y},\hat{y}'} \eta'(x)\; dx \;=\; 0\,. \tag{2.73}$$

Note that

$$M(x) \;=\; \frac{\partial f}{\partial y'} - \int_a^x \frac{\partial f}{\partial y}\, du \tag{2.74}$$

is continuous on $[a, b]$ and that the assumptions of the lemma are satisfied. It now follows that

$$\frac{\partial f}{\partial y'} \;=\; \int_a^x \frac{\partial f}{\partial y}\, du \,+\, c \tag{2.75}$$

for all $x \in [a, b]$. This is known as the *integrated form* of the Euler–Lagrange equation.

The right-hand side of equation (2.75) is differentiable. This, in turn, implies that the left-hand side of equation (2.75) is differentiable and that $\hat{y}(x)$ satisfies the Euler–Lagrange equation,

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) \;=\; \frac{\partial f}{\partial y}\,. \tag{2.76}$$

In other words, all solutions with continuous first derivatives, not just those with continuous second derivatives, must satisfy the Euler–Lagrange equation.

The differentiability of $f_{y'}(x, \hat{y}, \hat{y}')$ can also be used to *prove* (Whittemore, 1900–1901) the existence of the second derivative $\hat{y}''(x)$ for all values of x for which $f_{y'y'}(x, \hat{y}, \hat{y}') \neq 0$.

We will see later that we can weaken the conditions on $\hat{y}(x)$ and $\eta(x)$ even further, so that they are merely piecewise continuously differentiable. One can then show that the Euler–Lagrange equation is satisfied between corners of the solution and that additional conditions must be satisfied at the corners. Determining these conditions requires additional tools, and so we will defer the topic of corners until Chapter 10. For the time being, we will focus our attention on continuously differentiable solutions.

2.4. Recommended reading

Goldstine (1980), Fraser (1994, 2005a), and Thiele (2007) analyze Euler's early contributions to the calculus of variations. Euler's idea of using a polygonal curve to approximate the solution of a variational problem was revived in the 20th century by Russian mathematicians working on *direct methods* of solution. In a direct method, you construct a sequence of approximating functions, determine the unknown values and coefficients in each function using minimization, and let the sequence of functions converge to the solution. Euler's approach suggests the *direct method of finite differences* (Elsgolc, 1961). Other direct methods include the Ritz method, which is frequently and inappropriately (Leissa, 2005) called the Rayleigh-Ritz method, the Kantorovich method, and the Galerkin method. See Forray (1968) for an introduction to these direct methods.

The theory of the differentiation of functionals has its origins in the work of Volterra (1887). See Gelfand and Fomin (1963), Hamilton and Nashed (1982, 1995), Kolmogorov and Yushkevich (1998), and Lebedev and Cloud (2003) for more on variational derivatives. There are errors in the statements and proofs of the existence of the variational derivative for the simplest problem of the calculus of variations in Volterra (1913) and Gelfand and Fomin (1963). These errors were pointed out and corrected by Bliss (1915) and Hamilton (1980).

Lagrange announced his new approach to the calculus of variations in a 1755 letter to Euler; his results appeared in print seven years later (Lagrange, 1762). Fraser (1985) reviews the lengthy correspondence between Joseph Lagrange and Leonhard Euler and traces the development of Lagrange's approach to the foundations of the calculus of variations.

It has been said that the fundamental lemma is like a watchdog that guards the entrance gates to the entire classical domain of the calculus of variations (Dresden, 1932). Many early writers took the conclusion of the fundamental lemma as self-evident while others erred in their proof of this lemma. Huke (1931) traces the long and fascinating history of the fundamental lemma and of the lemma of du Bois-Reymond.

In writing this chapter, we leaned heavily on Bolza (1973) and Sagan (1969). I encourage all students of the calculus of variations to read these two books.

2.5. Exercises

2.5.1. Euler's approach. Using Euler's approach from Section 2.2, determine polygonal approximations to the curve that minimizes

$$\int_0^2 [(y')^2 + 6x^2 y]\, dx \tag{2.77}$$

subject to

$$y(0) = 2, \quad y(2) = 4 \tag{2.78}$$

for $n = 1$, $n = 2$, and $n = 3$. Write down and solve the Euler–Lagrange equation for this problem. Compare your polygonal approximations to your solution of the Euler–Lagrange equation.

2.5.2. Another lemma. Let $M(x) \in C[a, b]$ be a continuous function on the closed interval $a \le x \le b$ that satisfies

$$\int_a^b M(x)\, \eta''(x)\, dx = 0 \tag{2.79}$$

for all $\eta(x) \in C^2[a, b]$ satisfying

$$\eta(a) = \eta(b) = \eta'(a) = \eta'(b) = 0\,. \tag{2.80}$$

Prove that

$$M(x) = c_0 + c_1 x \tag{2.81}$$

for suitable constants c_0 and c_1. What can you say about c_0 and c_1?