1

Introduction

1.1 Introduction

The calculus of variations is concerned with finding extrema and, in this sense, it can be considered a branch of optimization. The problems and techniques in this branch, however, differ markedly from those involving the extrema of functions of several variables owing to the nature of the domain on the quantity to be optimized. A **functional** is a mapping from a set of functions to the real numbers. The calculus of variations deals with finding extrema for functionals as opposed to functions. The candidates in the competition for an extremum are thus functions as opposed to vectors in $\mathbb{R}^n$, and this gives the subject a distinct character. The functionals are generally defined by definite integrals; the sets of functions are often defined by boundary conditions and smoothness requirements, which arise in the formulation of the problem/model.

The calculus of variations is nearly as old as the calculus, and the two subjects were developed somewhat in parallel. In 1927 Forsyth [27] noted that the subject "attracted a rather fickle attention at more or less isolated intervals in its growth." In the eighteenth century, the Bernoulli brothers, Newton, Leibniz, Euler, Lagrange, and Legendre contributed to the subject, and their work was extended significantly in the next century by Jacobi and Weierstraß. Hilbert [38], in his renowned 1900 lecture to the International Congress of Mathematicians, outlined 23 (now famous) problems for mathematicians. His 23rd problem is entitled *Further development of the methods of the calculus of variations*. Immediately before describing the problem, he remarks:

> ... I should like to close with a general problem, namely with the indication of a branch of mathematics repeatedly mentioned in this lecture—which, in spite of the considerable advancement lately given it by Weierstraß, does not receive the general appreciation which in my opinion it is due—I mean the calculus of variations.

Hilbert's lecture perhaps struck a chord with mathematicians.[1] In the early twentieth century Hilbert, Noether, Tonelli, Lebesgue, and Hadamard among others made significant contributions to the field. Although by Forsyth's time the subject may have "attracted rather fickle attention," many of those who did pay attention are numbered among the leading mathematicians of the last three centuries. The reader is directed to Goldstine [36] for an in-depth account of the history of the subject up to the late nineteenth century.

The enduring interest in the calculus of variations is in part due to its applications. Of particular note is the relationship of the subject with classical mechanics, where it crosses the boundary from being merely a mathematical tool to encompassing a general philosophy. Variational principles abound in physics and particularly in mechanics. The application of these principles usually entails finding functions that minimize definite integrals (e.g., energy integrals) and hence the calculus of variations comes naturally to the fore. Hamilton's Principle in classical mechanics is a prominent example. An earlier example is Fermat's Principle of Minimum Time in geometrical optics. The development of the calculus of variations in the eighteenth and nineteenth centuries was motivated largely by problems in mechanics. Most textbooks on classical mechanics (old and new) discuss the calculus of variations in some depth. Conversely, many books on the calculus of variations discuss applications to classical mechanics in detail. In the introduction of Carathéodory's book [21] he states:

> I have never lost sight of the fact that the calculus of variations, as it is presented in Part II, should above all be a servant of mechanics.

Certainly there is an intimate relationship between mechanics and the calculus of variations, but this should not completely overshadow other fields where the calculus of variations also has applications. Aside from applications in traditional fields of continuum mechanics and electromagnetism, the calculus of variations has found applications in economics, urban planning, and a host of other "nontraditional fields." Indeed, the theory of optimal control is centred largely around the calculus of variations.

Finally it should be noted the calculus of variations does not exist in a mathematical vacuum or as a closed chapter of classical analysis. Historically, this field has always intersected with geometry and differential equations, and continues to do so. In 1974, Stampacchia [17], writing on Hilbert's 23rd problem, summed up the situation:

> One might infer that the interest in this branch of Analysis is weakening and that the Calculus of Variations is a Chapter of Classical Analysis. In fact this inference would be quite wrong since new problems like those in control theory are closely related to the problems of

[1] His nineteenth and twentieth problems were also devoted to the calculus of variations.

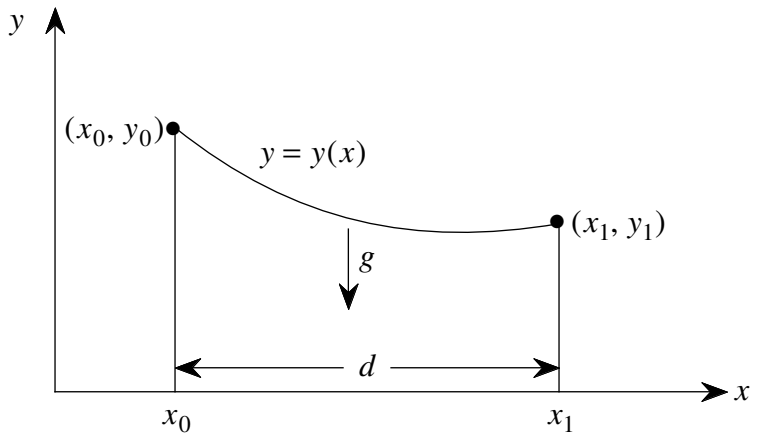

Fig. 1.1.

> the Calculus of Variations while classical theories, like that of boundary value problems for partial differential equations, have been deeply affected by the development of the Calculus of Variations. Moreover, the natural development of the Calculus of Variations has produced new branches of mathematics which have assumed different aspects and appear quite different from the Calculus of Variations.

The field is far from dead and it continues to attract new researchers.

In the remainder of this chapter we discuss some typical problems in the calculus of variations that are easy to model (although perhaps not so easy to solve). These problems illustrate the above comments and give the reader a taste of the subject. We return to most of these examples later in the book as the mathematics to solve them develops.

1.2 The Catenary and Brachystochrone Problems

1.2.1 The Catenary

Consider a thin heavy uniform flexible cable suspended from the top of two poles of height y_0 and y_1 spaced a distance d apart (figure 1.1). At the base of each pole the cable is assumed to be coiled. The cable follows up the pole to the top, runs through a pulley, and then spans the distance d to the next pole. The problem is to determine the shape of the cable between the two poles.

The cable will assume the shape that makes the potential energy minimum. The potential energy associated with the vertical parts of the cable will be the same for any configuration of the cable and hence we may ignore this component. If m denotes the mass per unit length of the cable and g the gravitational constant, the potential energy of the cable between the poles is

$$W_p(y) = \int_0^L mgy(s)\,ds, \tag{1.1}$$

where s denotes arclength, and $y(s)$ denotes the height of the cable above the ground s units in length along the cable from the top of the pole at (x_0, y_0). The number L denotes the arclength of the cable from (x_0, y_0) to (x_1, y_1). Unfortunately, we do not know L in this formulation. We can, however, recast the above expression for W_p in terms of Cartesian coördinates since we do know the coördinates of the pole tops. The differential arclength element in Cartesian coördinates is given by $ds = \sqrt{1 + y'^2}$, and this leads to the following expression for W_p,

$$W_p(y) = \int_{x_0}^{x_1} mgy(x)\sqrt{1 + y'^2(x)}\,dx. \tag{1.2}$$

Note that unlike our first expression for W_p, the above one involves the derivative of y. We have implicitly assumed here that the solution curve can be represented by a function $y : [x_0, x_1] \to \mathbb{R}$ and that this function is continuous and at least piecewise differentiable. Given the nature of the problem these seem reasonable assumptions.

The cable will assume the shape that minimizes W_p. The constant factor mg in the expression for W_p can be ignored for the purposes of optimizing the potential energy. The essence of the problem is thus to determine a function y such that the quantity

$$J(y) = \int_{x_0}^{x_1} y\sqrt{1 + y'^2}\,dx \tag{1.3}$$

is minimum. The model requires that any candidate $\hat{y}$ for an extremum satisfies the boundary conditions

$$\hat{y}(x_0) = y_0, \quad \hat{y}(x_1) = y_1. \tag{1.4}$$

In addition, the candidates must also be continuous and at least piecewise differentiable in the interval $[x_0, x_1]$.

We find the extrema for J in Chapter 2, where we show that the shape of the cable can be described by a hyperbolic cosine function. The curve itself is called a **catenary**.[2]

The same functional J arises in a problem in geometry concerning a minimal surface of revolution, i.e., a surface of revolution having minimal surface area. Suppose that the x-axis corresponds to the axis of rotation. Any surface of revolution can be generated by a curve in the xy-plane (figure 1.2). The

[2] The name "catenary" is particularly descriptive. The name comes from the Latin word *catena* meaning chain. Catenary refers to the curve formed by a uniform chain hanging freely between two poles. Leibniz is credited with coining the term (ca. 1691).

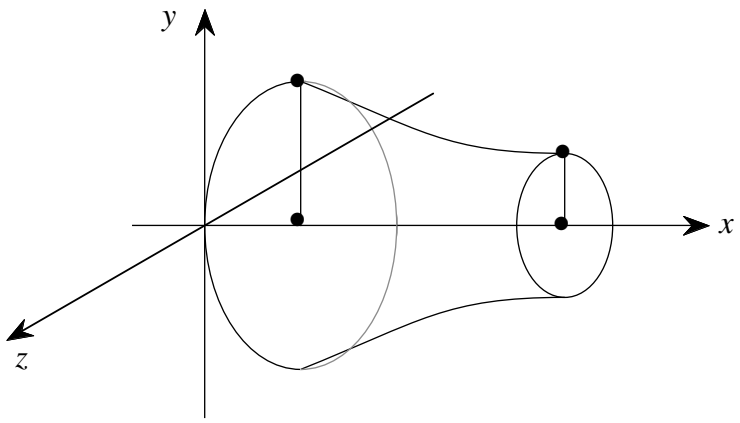

Fig. 1.2.

problem thus translates to finding the curve γ that generates the surface of revolution having the minimal surface area. As with the catenary problem, we make the assumption that γ can be described by a function $y : [x_0, x_1] \to \mathbb{R}$ that is continuous and piecewise differentiable in the interval $[x_0, x_1]$. Under these assumptions we have that the surface area of the corresponding surface of revolution is

$$A(y) = 2\pi \int_{x_0}^{x_1} |y(x)| \sqrt{1 + y'^2(x)}\, dx. \tag{1.5}$$

Here we need also make the assumption that $y(x) > 0$ for all $x \in [x_0, x_1]$.[3] The problem of finding the minimal surface thus reduces to finding the function y such that the quantity

$$J(y) = \int_{x_0}^{x_1} y \sqrt{1 + y'^2}\, dx$$

is minimum. The two problems thus produce the same functional to be minimized. The generating curve that produces the minimal surface of revolution is thus a catenary. The surface itself is called a **catenoid**.

[3] If $y = 0$ at some point $\tilde{x} \in (x_0, x_1)$ we can still generate a rotationally symmetric "object," but technically it would not be a surface. Near $(\tilde{x}, 0, 0)$ the "object" would resemble (i.e., be homeomorphic to) a double cone. The double cone fails the requirements to be a surface because any neighbourhood containing the common vertex is not homeomorphic to the plane.

Let us return to the original problem. A modification of the problem would be to first specify the length of the cable. Evidently, if L is the length of the cable we must require that

$$L \geq \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

in order that the cable span the two poles. Moreover, it is intuitively clear that in the case of equality there is only one configuration possible viz., the line segment from (x_0, y_0) to (x_1, y_1). In this case, there is no optimization to be done as there is only one candidate. We may thus restrict our attention to the case

$$L > \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}.$$

Given a cable of length L, the problem is to determine the shape the cable assumes when supported between the poles. The problem was posed by Jacob Bernoulli in 1690. By the end of 1691 the problem was solved by Leibniz, Huygens, and Jacob's younger brother Johann Bernoulli. It should be noted that Galileo had earlier considered the problem, but he thought the catenary was essentially a parabola.[4]

Since the arclength L of the cable is given, we can use expression (1.1) to look for a minimum potential energy configuration. Instead, we start with expression (1.2). The modified problem is now to find the function $y : [x_0, x_1] \to \mathbb{R}$ such that W_p is minimized subject to the arclength constraint

$$L = \int_{x_0}^{x_1} \sqrt{1 + y'^2}\, dx, \tag{1.6}$$

and the boundary conditions

$$y(x_0) = y_0, \quad y(x_1) = y_1.$$

This problem is thus an example of a constrained variational problem. The constraint (1.6) can be regarded as an integral equation (with, it is hoped, nonunique solutions). Constraints such as (1.6) are called **isoperimetric**. We discuss problems having isoperimetric constraints in Chapter 4.

Suppose that we use expression (1.1), which *prima facie* seems simpler than expression (1.2). We know L, so that the limits of the integral are known, but the parameter s is special and corresponds to arclength. We must somehow build in the requirement that s is arclength if we are to use expression (1.1). In order to do this we must use a parametric representation of the curve $(x(s), y(s))$, $s \in [0, L]$. The arclength parameter for such a curve is characterized by the differential equation

$$x'^2(s) + y'^2(s) = 1. \tag{1.7}$$

[4] There is still some dispute regarding whether Galileo thought the catenary to be the parabola. See Giaquinta and Hildebrandt [32], p. 133 for more details.

The problem thus entails finding the functions $x(s)$, $y(s)$ that minimize W_p subject to the constraint (1.7) and the boundary conditions

$$x(0) = x_0, \quad x(L) = x_1$$
$$y(0) = y_0, \quad y(L) = y_1.$$

In general, a constraint of this kind is more difficult to deal with than an isoperimetric constraint.

1.2.2 Brachystochrones

The history of the calculus of variations essentially begins with a problem posed by Johann Bernoulli (1696) as a challenge to the mathematical community and in particular to his brother Jacob. (There was significant sibling rivalry between the two brothers.) The problem is important in the history of the calculus of variations because the method developed by Johann's pupil, Euler, to solve this problem provided a sufficiently general framework to solve other variational problems.

The problem that Johann posed was to find the shape of a wire along which a bead initially at rest slides under gravity from one end to the other in minimal time. The endpoints of the wire are specified and the motion of the bead is assumed frictionless. The curve corresponding to the shape of the wire is called a **brachystochrone**[5] or a curve of fastest descent.

The problem attracted the attention of a number of mathematical luminaries including Huygens, L'Hôpital, Leibniz, and Newton, in addition of course to the Bernoulli brothers, and later Euler and Lagrange. This problem was at the cutting edge of mathematics at the turn of the eighteenth century.

Jacob was up to the challenge and solved the problem. Meanwhile (and independently) Johann and Leibniz also arrived at correct solutions. Newton was late to the party because he learned about the problem some six months later than the others. Nonetheless, he solved the problem that same evening and sent his solution anonymously the next day to Johann. Newton's cover was blown instantly. Upon looking at the solution, Johann exclaimed "Ah! I recognize the paw of the lion."

To model Bernoulli's problem we use Cartesian coördinates with the positive y-axis oriented in the direction of the gravitational force (figure 1.3). Let (x_0, y_0) and (x_1, y_1) denote the coördinates of the initial and final positions of the bead, respectively. Here, we require that $x_0 < x_1$ and $y_0 < y_1$. The Bernoulli problem consists of determining, among the curves that have (x_0, y_0) and (x_1, y_1) as endpoints, the curve on which the bead slides down from (x_0, y_0) to (x_1, y_1) in minimum time. The problem makes sense only for continuous curves. We make the additional simplifying (but reasonable) assumptions that the curve can be represented by a function $y : [x_0, x_1] \to \mathbb{R}$

[5] The word comes from the Greek words *brakhistos* meaning "shortest" and *khronos* meaning time.

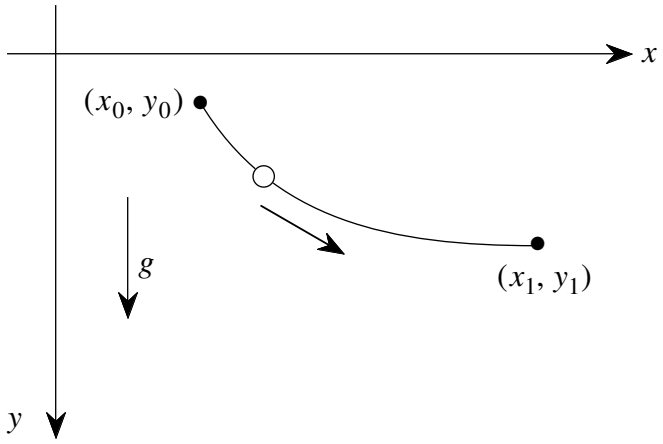

Fig. 1.3.

and that y is at least piecewise differentiable in the interval $[x_0, x_1]$. Now, the total time it takes the bead to slide down a curve is given by

$$T(y) = \int_0^L \frac{ds}{v(s)}, \tag{1.8}$$

where L denotes the arclength of the curve, s is the arclength parameter, and v is the velocity of the bead s units down the curve from (x_0, y_0). As with the catenary problem, we do not know the value of L, so we must seek an alternative formulation.

Our first job is to get an expression for the velocity in terms of the function y. We use the law of conservation of energy to achieve this. At any position $(x, y(x))$ on the curve, the sum of the potential and kinetic energies of the bead is a constant. Hence

$$\frac{1}{2}mv^2(x) + mgy(x) = c, \tag{1.9}$$

where m is the mass of the bead, v is the velocity of the bead at $(x, y(x))$, and c is a constant. Since the energy is constant along the curve, we know that

$$c = \frac{1}{2}mv^2(x_0) + mgy(x_0).$$

Solving equation (1.9) for v gives

$$v(x) = \sqrt{\frac{2c}{m} - 2gy(x)}.$$

Equation (1.8) thus implies that

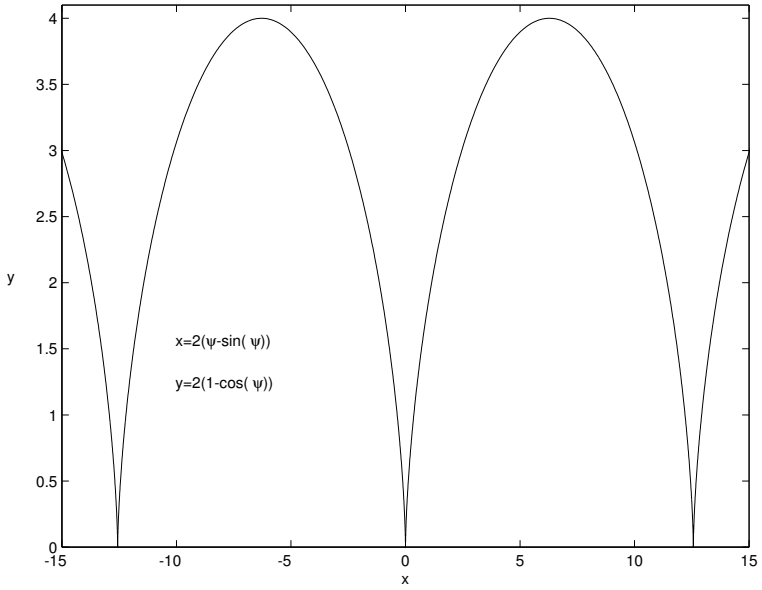

Fig. 1.4.

$$T(y) = \int_{x_0}^{x_1} \frac{\sqrt{1+y'^2}}{\sqrt{\frac{2c}{m} - 2gy(x)}}\, dx. \tag{1.10}$$

We thus seek a function y such that T is minimum and

$$y(x_0) = y_0, \quad y(x_1) = y_1.$$

Note that for the purposes of optimization T can be replaced by the functional

$$J(y) = \int_{x_0}^{x_1} \frac{\sqrt{1+w'^2}}{\sqrt{w}}\, dx, \tag{1.11}$$

and the relation

$$w(x) = \frac{1}{2g}\left(\frac{2c}{m} - 2gy(x)\right)$$

(the $\sqrt{2g}$ factor does not affect the extrema of J).

In Chapter 2 we find the extrema for J (and hence T), and show that the brachystochrone for this problem is a portion of a special type of curve called a **cycloid**. Figure 1.4 depicts a cycloid. You can visualize a cycloid in the safety of your own home by painting a white dot on a clean tyre and then rolling the tyre along a line. If you can follow the rolling dot, the curve traced out is a cycloid. Before the fabulous Bernoulli brothers came on the stage, Christiaan Huygens had already discovered a remarkable property of cycloids.

Christiaan discovered that a bead sliding down a cycloid generated by a circle of radius ρ under gravity reaches the bottom of the cycloid arch after the period $\pi\sqrt{\rho/g}$ *wherever* on the arch the bead starts from rest. This notable property of the cycloid earned it the appellation **isochrone**. The cycloid thus sports the names isochrone and brachystochrone.[6] Christiaan used the curve to good effect and designed what was then considered a remarkably accurate pendulum clock based on the laudable properties of the cycloid, which was used to govern the motion of the pendulum. The reader may find a diagram of the pendulum and further details on this interesting curve in an article by Tee [67] wherein several original references may be found.

Finally, we note that brachystochrone problems have proliferated in the three centuries following Bernoulli's challenge. Some models subjected the bead to a resisting medium whilst others changed the force field from a simple uniform gravitational field to more complicated scenarios. Research is still progressing on brachystochrones. The reader is directed to the work of Tee [67], [68], [69] for more references.

1.3 Hamilton's Principle

There are many fine books on classical (analytical) mechanics (e.g., [1], [6], [35], [48], [49], [59], and [73]) and we make no attempt here to give even a basic account of this seemingly vast subject. Nonetheless, it would be demeaning to the calculus of variations to ignore its rich heritage and fruitful interaction with classical mechanics. Moreover, many of our examples come from classical mechanics, so a few words from our sponsor seem in order.

Classical mechanics is teeming with variational principles of which Hamilton's Principle is perhaps the most important. [7] In this section we give a brief "no frills" statement of Hamilton's Principle as it applies to the motion of particles. The serious student of mechanics should consult one of the many specialized texts on this subject.

Let us first consider the motion of a single particle in $\mathbb{R}^3$. Let $\mathbf{r}(t) = (x(t), y(t), z(t))$ denote the position of the particle at time t. The **kinetic energy** of this particle is given by

$$T = \frac{1}{2} m \left(\dot{x}^2(t) + \dot{y}^2(t) + \dot{z}^2(t) \right),$$

where m is the mass of the particle and $\dot{}$ denotes d/dt. We assume that the forces on the particle can be derived from a single scalar function. Specifically, we assume there is a function V such that:

[6] It is also called a **tautochrone**, but we do not count this since the word is derived from the Greek word *tauto* meaning "same." The prefix iso comes from the Greek word *isos*, which also means "same."

[7] One need only scan through Lanczos' book [48] to find the "Principle of Virtual Work," "D'Alembert's Principle," "Gauss' Principle of Least Constraint," "Jacobi's Principle," and, of course, "Hamilton's Principle" among others.

1. V depends only on time and position; i.e., $V = V(t, x, y, z)$;
2. the force $\mathbf{f} = (f_1, f_2, f_3)$ acting on the particle has the components

$$f_1 = -\frac{\partial V}{\partial x}, \quad f_2 = -\frac{\partial V}{\partial y}, \quad f_3 = -\frac{\partial V}{\partial z}.$$

The function V is called the **potential energy**. Let

$$L = T - V.$$

The function L is called the **Lagrangian**. Suppose that the initial position of the particle $\mathbf{r}(t_0)$ and final position $\mathbf{r}(t_1)$ are specified. **Hamilton's Principle** states that the path of the particle $\mathbf{r}(t)$ in the time interval $[t_0, t_1]$ is such that the functional

$$J(\mathbf{r}) = \int_{t_0}^{t_1} L(t, \mathbf{r}, \dot{\mathbf{r}})\, dt$$

is stationary, i.e., a local extremum or a "saddle point." (We define "stationary" more precisely in Section 2.2.) In the lingo of mechanics J is called **the action integral** or simply **the action**.

Problems in mechanics often involve several particles (or spatial coördinates); moreover, Cartesian coördinates are not always the best choice. Variational principles are thus usually given in terms of **generalized coördinates**. The letter q has been universally adopted to denote generalized position coördinates. The configuration of a system at time t is thus denoted by $\mathbf{q}(t) = (q_1(t), \ldots, q_n(t))$, where the q_k are position variables. If, for example, the system consists of three free particles in $\mathbb{R}^3$ then $n = 9$.

The kinetic energy T of a system is given by a quadratic form in the generalized velocities $\dot{q}_k$,

$$T(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2} \sum_{j,k=1}^{n} C_{j,k}(\mathbf{q}) \dot{q}_j \dot{q}_k.$$

Assuming the system has a potential energy function $V(t, \mathbf{q})$, the Lagrangian is given by

$$L(t, \mathbf{q}, \dot{\mathbf{q}}) = T(\mathbf{q}, \dot{\mathbf{q}}) - V(t, \mathbf{q}).$$

In this framework Hamilton's Principle takes the following form.

Theorem 1.3.1 (Hamilton's Principle) *The motion of a system of particles* $\mathbf{q}(t)$ *from a given initial configuration* $\mathbf{q}(t_0)$ *to a given final configuration* $\mathbf{q}(t_1)$ *in the time interval* $[t_0, t_1]$ *is such that the functional*

$$J(\mathbf{q}) = \int_{t_0}^{t_1} L(t, \mathbf{q}, \dot{\mathbf{q}})\, dt$$

is stationary.

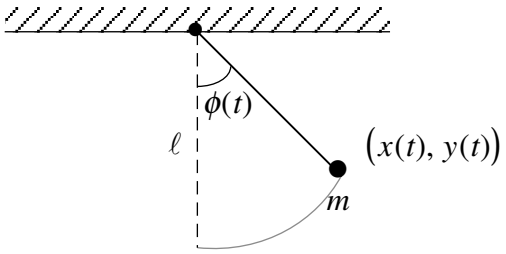

Fig. 1.5.

The dynamics of a system of particles is thus completely contained in the single scalar function L. We can derive the familiar equations of motion from Hamilton's Principle (cf. Section 3.2). The reader might rightfully question whether the motion predicted by Hamilton's Principle depends on the choice of coördinates. The variational approach would surely be of limited value were it sensitive to the observer's choice of coördinates. We show in Section 2.5 that Hamilton's Principle produces equations that are necessarily invariant with respect to coördinate choices.

Example 1.3.1: Simple Pendulum
Consider a simple pendulum of mass m and length ℓ in the plane. Let $(x(t), y(t))$ denote the position of the mass at time t. Since $x^2 + y^2 = \ell^2$ we need in fact only one position variable. Rather than use x or y it is natural to use polar coördinates and characterize the position of the mass at time t by the angle $\phi(t)$ between the vertical and the string to which the mass is attached (figure 1.5). Now, the kinetic energy is

$$T = \frac{1}{2}m(\dot{x}^2(t) + \dot{y}^2(t)) = \frac{1}{2}m\ell^2\dot{\phi}^2(t),$$

and the potential energy is

$$V = mgh = mg\ell(1 - \cos\phi(t)),$$

where g is a gravitation constant. Thus,

$$L(\phi, \dot{\phi}) = \frac{1}{2}m\ell^2\dot{\phi}^2 - mg\ell(1 - \cos\phi),$$

and Hamilton's Principle implies that the motion from a given initial angle $\phi(t_0)$ to a fixed angle $\phi(t_1)$ is such that the functional

$$J(\phi) = \int_{t_0}^{t_1} \left(\frac{1}{2}m\ell^2\dot{\phi}^2 - mg\ell(1 - \cos\phi)\right) dt$$

is stationary.

Example 1.3.2: Kepler problem
The Kepler problem models planetary motion. It is one of the most heavily studied problems in classical mechanics. Keeping with our no frills approach, we consider the simplest problem of a single planet orbiting around the sun, and ignore the rest of the solar system. Assuming the sun is fixed at the origin, the kinetic energy of the planet is

$$T = \frac{1}{2}m(\dot{x}^2(t) + \dot{y}^2(t)) = \frac{1}{2}m\left(\dot{r}^2(t) + r^2(t)\dot{\theta}^2(t)\right),$$

where r and θ denote polar coördinates and m is the mass of the planet. We can deduce the potential energy function V from the gravitational law of attraction

$$f = -\frac{GmM}{r^2},$$

where f is the force (acting in the radial direction), M is the mass of the sun, and G is the universal gravitation constant. Given that

$$f = -\frac{\partial V}{\partial r},$$

we have

$$V(r) = -\int f(r)\,dr = -\frac{GmM}{r};$$

hence,

$$L(r,\theta) = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + \frac{GmM}{r}.$$

Hamilton's Principle implies that the motion of the planet from an initial observation $(r(t_0), \theta(t_0))$ to a final observation $(r(t_1), \theta(t_1))$ is such that

$$J(r,\theta) = \int_{t_0}^{t_1} \left(\frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + \frac{GmM}{r}\right) dt$$

is stationary.

The reader may be wondering about the fate of the constant of integration in the last example. Any potential energy of the form $-GmM/r + const.$ will produce the requisite force f. In the pendulum problem we tacitly assumed that the potential energy was proportional to the height of the mass above the minimum possible height. In fact, for the purposes of describing the dynamics it does not matter; i.e., $V(t, \mathbf{q})$ and $V(t, \mathbf{q}) + c_1$ produce the same results for any constant c_1. We are optimizing J and the addition of a constant in the Lagrangian simply alters the functional $J(\mathbf{q})$ to $\tilde{J}(\mathbf{q}) = J(\mathbf{q}) + const.$ If one functional is stationary at $\mathbf{q}$ the other must also be stationary at $\mathbf{q}$.

In the lore of classical mechanics there is another variational principle that is sometimes called the "Principle of Least Action" or "Maupertuis'

Principle," which predates Hamilton's Principle. This principle is sometimes confused with Hamilton's and the situation is not mitigated by the fact that Hamilton's Principle is sometimes called the Principle of Least Action. [8] Maupertuis' Principle concerns systems that are **conservative**. In a conservative system we have that the total energy of the system at any time t along the path of motion is constant. In other words, $L + V = k$, where k is a constant. For this special case $L = 2T - k$, and Hamilton's Principle leads to Maupertuis' Principle that the functional

$$K(\mathbf{q}) = \int_{t_0}^{t_1} T(\mathbf{q}, \dot{\mathbf{q}})\, dt$$

is stationary along a path of motion. Hence, Maupertuis' Principle is a special case of Hamilton's Principle. Most books on classical mechanics discuss these principles (along with others). Lanczos [48] gives a particularly complete and readable account that, in addition to mechanics, deals with the history and philosophy of these principles. The eminent scientist E. Mach [51] also writes at length about the history, significance, and philosophy underlying these principles. His perspective and sympathies are somewhat different from those of Lanczos. [9]

1.4 Some Variational Problems from Geometry

1.4.1 Dido's Problem

Dido was a Carthaginian queen (ca. 850 B.C.?) who came from a dysfunctional family. Her brother, Pygmalion, murdered her husband (who was also her uncle) and Dido, with the help of various gods, fled to the shores of North Africa with Pygmalion in pursuit. Upon landing in North Africa, legend has it that she struck a deal with a local chief to procure as much land as an oxhide could contain. She then selected an ox and cut its hide into very narrow strips, which she joined together to form a thread of oxhide more than two and a half miles long. Dido then used the oxhide thread and the North African sea coast to define the perimeter of her property. It is not clear what the immediate reaction of the chief was to this particular interpretation of the deal, but it is

[8] The translators of Landau and Lifshitz [49], p. 131, go so far as to draft a table to elucidate the different usages.

[9] Mach is not so generous with Maupertuis. In connexion with Maupertuis' Principle he writes, "It appears that Maupertuis reached this obscure expression by an unclear mingling of his ideas of *vis viva* and the principle of virtual velocities" (p. 365). In defense of Mach, we must note that Maupertuis suffered no lack of critics even in his own day. Voltaire wrote the satire *Histoire du docteur Akakia et du naif de Saint Malo* about Maupertuis. The situation at Frederick the Great's court regarding Maupertuis, König, and Voltaire is the stuff of soap operas (see Pars [59] p. 634).

clear that Dido sought to enclose the maximum area within her ox and the sea. The city of Carthage was then built within the perimeter defined by the thread and the sea coast. Dido called the place *Byrsa* meaning hide of bull.[10]

The problem that Dido faced on the shores of North Africa (aside from family difficulties) was to determine the optimal path along which to place the oxhide thread so as to provide Byrsa with the maximum amount of land. Dido did not have the luxury of waiting some 2500 years for the calculus of variations to develop and thus settled for an "intuitive solution."

Dido's problem entailed determining the curve γ of fixed length (the thread) such that the area enclosed by γ and a given curve σ (the North African shoreline) is maximum. Although this is perhaps the original version of Dido's problem, the term has been used to cover the more basic problem: among all closed curves in the plane of perimeter L determine the curve that encloses the maximum area. The problem did not escape the attention of ancient mathematicians, and as early as perhaps 200 B.C. the mathematician Zenodorus[11] is credited with a proof that the solution is a circle. Unfortunately, there were some technical loopholes in Zenodorus' proof (he compared the area of a circle with that of polygons having the same perimeter). The first complete proof of this result was given some 2000 years later by Karl Weierstraß in his Berlin lectures.

Prior to Weierstraß, Steiner (ca. 1841) proved that *if* there exists a "figure" γ whose area is never less than that of any other "figure" of the same perimeter, then γ is a circle. Not content with one proof, Steiner gave five proofs of this result. The proofs are based on simple geometric considerations (no calculus of variations). The operative word in the statement of his result, however, is "if." Steiner's contemporary, Dirichlet, pointed out that his proofs do not actually establish the existence of such a figure. Weierstraß and his followers resolved these subtle aspects of the problem. A lively account of Dido's problem and the first of Steiner's proofs can be found in Körner [45].

Some simple geometrical arguments can be used to show that if γ is a simple closed curve solution to Dido's problem then γ is convex (cf. Körner, *op. cit.*). This means that a chord joining any two points on γ lies within γ

[10] The reader will find various bits and pieces of Dido's history scattered in Latin works by authors such as Justin and Virgil. One account of the hide story comes from the *Aeneid*, Bk. I, vs. 367. The story gets even better once Aeneas arrives on the scene. Finally, good ideas never die. It is said that the Anglo-Saxon chieftains Hengist and Horsa (ca. 449 A.D.) acquired their land by circling it with oxhide strips [37]. Beware of real estate transactions that involve an ox.

[11] The proof may have been known even earlier, but Zenodorus in any event is the author of the proof that appears in the commentary of Theon to Ptolemy's *Almagest*. Zenodorus quotes Archimedes (who died in 212 B.C.) and is quoted by Pappus (ca. 340 A.D.). Aside from these rough dates we do not know exactly when Zenodorus lived. At any rate, the solution was of little comfort to Dido's heirs as the Romans obliterated Carthage/ Byrsa in the third Punic war just after 200 B.C. and sowed salt on the scorched ground so that nothing would grow.

and the area enclosed by γ. The convexity of γ is then used to show that Dido's problem can be distilled down to the problem of finding a function $y : [x_0, x_1] \to \mathbb{R}$ such that

$$A(y) = \int_{x_0}^{x_1} y(x)\, dx$$

is maximum subject to the constraint that the arclength of the curve γ^+ described by y is $L/2$. If we assume that y is at least piecewise differentiable then this amounts to the condition

$$\frac{L}{2} = \int_{x_0}^{x_1} \sqrt{1 + y'^2}\, dx.$$

The problem with this formulation is that we do not know the limits of the integral. The geometrical character of the problem indicates that we do not need to know both x_0 and x_1 (we could always normalize the construction so that $x_0 = 0 < x_1$), but we do need to know $x_1 - x_0$. This problem is effectively the opposite of the problem we had with the first formulation of the catenary. Since we know arclength, a natural formulation to use would be one in terms of arclength.

Suppose that γ^+ is described parametrically by $(x(s), y(s))$, $s \in [0, L/2]$, where s is arclength. Suppose further that x and y are at least piecewise differentiable. Green's theorem in the plane can then be used to show that the area of the set enclosed by γ^+ and the x-axis is

$$A(y) = \frac{1}{2} \int_0^{L/2} y(s) \sqrt{1 - y'^2(s)}\, ds, \tag{1.12}$$

where we have used the relation $x'^2(s) + y'^2(s) = 1$. The basic Dido problem is thus to determine a positive function $y : [0, L/2] \to \mathbb{R}$ such that A is maximum.

1.4.2 Geodesics

Let Σ be a surface, and let p_0, p_1 be two distinct points on Σ. The geodesic problem concerns finding the curve(s) on Σ with endpoints p_0, p_1 for which the arclength is minimum. A curve having this property is called a **geodesic**. The theory of geodesics is one of the most developed subjects in differential geometry. The general theory is complicated analytically by the situation that simple, common surfaces such as the sphere require more than one vector function to describe them completely. In the language of geometry, the sphere is a manifold that requires at least two charts. We have encountered and sidestepped the analogous problem for curves, and we do so here in the interest of simplicity. We focus on the local problem and refer the reader to any general

text on differential geometry such as Stoker [66] or Willmore [75] for a more precise and in-depth treatment of geodesics.[12]

Suppose that Σ is described by the position vector function $\mathbf{r} : \sigma \to \mathbb{R}^3$, where σ is a nonempty connected open subset of $\mathbb{R}^2$, and for $(u, v) \in \sigma$,

$$\mathbf{r}(u, v) = (x(u, v), y(u, v), z(u, v)) .$$

We assume that $\mathbf{r}$ is a smooth function on σ; i.e., x, y, and z are smooth functions of (u, v), and that

$$\left|\frac{\partial \mathbf{r}}{\partial u} \wedge \frac{\partial \mathbf{r}}{\partial v}\right| \neq \mathbf{0}, \tag{1.13}$$

so that $\mathbf{r}$ is a one-to-one mapping of σ onto Σ. If γ is a curve on Σ, then there is a curve $\hat{\gamma}$ in σ that maps to γ under $\mathbf{r}$. Any curve on Σ may thus be regarded as a curve in σ. Suppose that the points p_0 and p_1 correspond to $\mathbf{r}_0 = \mathbf{r}(u_0, v_0)$ and $\mathbf{r}_1 = \mathbf{r}(u_1, v_1)$, respectively. Any curve γ from $\mathbf{r}_0$ to $\mathbf{r}_1$ maps to a curve $\hat{\gamma}$ from (u_0, v_0) to (u_1, v_1).

For the geodesic problem we restrict our attention to smooth simple curves (no self-intersections) on Σ from $\mathbf{r}_0$ to $\mathbf{r}_1$. Let Γ denote the set of all such curves. Thus, if $\gamma \in \Gamma$, then there exists a parametrization of γ of the form

$$\mathbf{R}(t) = \mathbf{r}(u(t), v(t)), \quad t \in [t_0, t_1], \tag{1.14}$$

where $\mathbf{R}(t_0) = \mathbf{r}_0$, $\mathbf{R}(t_1) = \mathbf{r}_1$, and u and v are smooth functions on the interval $[t_0, t_1]$ such that

$$u'^2(t) + v'^2(t) \neq 0 \tag{1.15}$$

for all $t \in [t_0, t_1]$. In the parameter space σ, the last condition ensures that the curve $\hat{\gamma}$ is also a smooth curve and has a well-defined unit tangent vector. The differential of arclength along γ is given by

$$\begin{aligned} ds^2 &= |\mathbf{R}'(t)|^2 \, dt^2 \\ &= \left| \frac{\partial \mathbf{r}}{\partial u} u'(t) + \frac{\partial \mathbf{r}}{\partial v} v'(t) \right|^2 dt^2 \\ &= \left(E u'^2 + 2F u' v' + G u'^2 \right) dt^2, \end{aligned}$$

where

$$E = \left| \frac{\partial \mathbf{r}}{\partial u} \right|^2, \quad F = \frac{\partial \mathbf{r}}{\partial u} \cdot \frac{\partial \mathbf{r}}{\partial v}, \quad G = \left| \frac{\partial \mathbf{r}}{\partial v} \right|^2 .$$

The functions E, F, and G are called components of the **first fundamental form** or **metric tensor**. Note that these components depend only on u and v. Note also that the identity

$$\left| \frac{\partial \mathbf{r}}{\partial u} \wedge \frac{\partial \mathbf{r}}{\partial v} \right| = EG - F^2$$

[12] A more specialized discussion can be found in Postnikov [62].

and condition (1.13) indicate that the quadratic form

$$I = Eu'^2 + 2Fu'v' + Gv'^2$$

is positive definite.

The arclength of γ is given by

$$L(\gamma) = \int_{t_0}^{t_1} \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}\, dt.$$

The geodesic problem is thus to find the functions u and v (i.e., the curve $\hat{\gamma}$) such that L is a minimum and

$$\begin{aligned} u(t_0) &= u_0, \quad v(t_0) = v_0 \\ u(t_1) &= u_1, \quad v(t_1) = v_1. \end{aligned}$$

Example 1.4.1: Geodesics on a Sphere
Let Σ be an octant of the unit sphere. The surface Σ can be described parametrically by

$$\mathbf{r}(u, v) = (\sin u \cos v, \sin u \sin v, \cos u)$$

for $\sigma = \{(u, v) : 0 < u < \pi/2, 0 < v < \pi/2\}$. Now,

$$\begin{aligned} E &= \left|\frac{\partial \mathbf{r}}{\partial u}\right|^2 = |\,(\cos u \cos v, \cos u \sin v, -\sin u)\,|^2 \\ &= 1, \\ F &= \frac{\partial \mathbf{r}}{\partial u} \cdot \frac{\partial \mathbf{r}}{\partial v} \\ &= (\cos u \cos v, \cos u \sin v, -\sin u) \cdot (-\sin u \sin v, \sin u \cos v, 0) \\ &= 0, \\ G &= \left|\frac{\partial \mathbf{r}}{\partial v}\right|^2 = |\,(-\sin u \sin v, \sin u \cos v, 0)\,|^2 \\ &= \sin^2 u. \end{aligned}$$

The arclength integral is thus

$$L(\gamma) = \int_{t_0}^{t_1} \sqrt{u'^2 + v'^2 sin^2 u}\, dt.$$

A feature of the basic geodesic problem described above is that it does not involve the function $\mathbf{r}$ directly. The arclength of a curve depends only on the three scalar functions E, F, and G. Geodesics are part of the **intrinsic geometry** of the surface, i.e., the geometry defined by the metric tensor. The metric tensor does not define a surface uniquely even modulo translations and

rotations. There are any number of distinct surfaces in $\mathbb{R}^3$ that have the same metric tensor. For example, a plane, a cone, and a cylinder all have the same metric tensor. If a cylinder is "unrolled" and "flattened" to form a portion of the plane, then a geodesic on the cylinder would become a geodesic on the plane.

One direction for a generalization of the above problem is to focus on the space $\sigma \subseteq \mathbb{R}^2$ and *define* the components of the metric tensor. For notational simplicity, let $u = u^1$, $v = u^2$, and $\mathbf{u} = (u, v)$. We can choose scalar functions $g_{jk} : \sigma \to \mathbb{R}$ $j, k = 1, 2$ and define the arclength element ds by

$$\begin{aligned} ds^2 &= g_{11}(du^1)^2 + g_{12}du^1du^2 + g_{21}du^2du^1 + g_{22}(du^2)^2 \\ &= g_{jk}du^j du^k, \end{aligned}$$

where the last expression uses the Einstein summation convention: summation of repeated indices when one is a superscript and the other is a subscript. Of course we must place some restrictions on the g_{jk} in order to ensure that our arclength element is positive and that the length of a curve does not depend on the choice of coördinates $\mathbf{u}$. We can take care of these concerns by requiring that the g_{jk} produce a quadratic form that is positive definite and that the g_{jk} form a second order covariant tensor. To mimic the earlier case we also impose the symmetry condition

$$g_{jk} = g_{kj},$$

so that

$$ds^2 = g_{11}(du^1)^2 + 2g_{12}du^1du^2 + g_{22}(du^2)^2. \tag{1.16}$$

In terms of the former notation, $E = g_{11}$, $F = g_{12} = g_{21}$, and $G = g_{22}$. For this case, the positive definite requirement amounts to the condition

$$g_{11}g_{22} - g_{12}^2 > 0$$

with $g_{11} > 0$. The condition that the g_{jk} form a second-order covariant tensor means that under a smooth coördinate transformation from $\mathbf{u} = (u^1, u^2)$ to $\hat{\mathbf{u}} = (\hat{u}^1, \hat{u}^2)$, the components $g_{jk}(\mathbf{u})$ transform to $\hat{g}_{lm}(\hat{\mathbf{u}})$ according to the relation

$$\hat{g}_{lm} = g_{jk}\frac{\partial u^j}{\partial \hat{u}^l}\frac{\partial u^k}{\partial \hat{u}^m}.$$

The set σ equipped with such a tensor can be viewed as defining a geometrical object in itself (as the surface Σ was). It is a special case of what is called a **Riemannian manifold**. Let $\mathcal{M}$ denote this geometrical object. A curve $\hat{\gamma}$ in σ generates a curve γ in $\mathcal{M}$, and the arclength is given by

$$L(\gamma) = \int_{t_0}^{t_1} \sqrt{g_{jk}u^{j\prime}u^{k\prime}}\, dt,$$

where $(u^1(t), u^2(t))$, $t \in [t_0, t_1]$ is a parametrization of $\hat{\gamma}$. The condition that the g_{jk} form a second-order covariant tensor ensures that $L(\gamma)$ is invariant

with respect to changes in the curvilinear coördinates $\mathbf{u}$ used to represent $\mathcal{M}$. Note also that $L(\gamma)$ is invariant with respect to orientation-preserving parametrizations of $\hat{\gamma}$.

The advantage of the above abstraction is that it can be readily modified to accommodate higher dimensions. Suppose that $\sigma \subseteq \mathbb{R}^n$ and $\mathbf{u} = (u^1, \ldots, u^n)$. We can define an n-dimensional (Riemannian) manifold $\mathcal{M}$ by introducing a metric tensor with components g_{jk} such that:

1. the quadratic form $g_{jk}du^j du^k$ is positive definite;
2. $g_{jk} = g_{kj}$ for $j, k = 1, 2, \ldots, n$;
3. under any smooth transformation $\mathbf{u} = \mathbf{u}(\hat{\mathbf{u}})$ the g_{jk} transform to $\hat{g}_{lm}$ according to the relation

$$\hat{g}_{lm} = g_{jk} \frac{\partial u^j}{\partial \hat{u}^l} \frac{\partial u^k}{\partial \hat{u}^m}.$$

A curve γ on $\mathcal{M}$ is generated by a curve $\hat{\gamma}$ in $\sigma \subseteq \mathbb{R}^n$. Suppose that $\mathbf{u}(t) = (u^1(t), \ldots, u^n(t))$, $t \in [t_0, t_1]$ is a parametrization of $\hat{\gamma}$. The arclength of γ is then defined as

$$L(\gamma) = \int_{t_0}^{t_1} \sqrt{g_{jk} u^{j\prime} u^{k\prime}}\, dt.$$

A generalization of the geodesic problem is thus to find the curve(s) $\hat{\gamma}$ in σ with specified endpoints $\mathbf{u}_0 = \mathbf{u}(t_0)$, $\mathbf{u}_1 = \mathbf{u}(t_1)$ such that $L(\gamma)$ is a minimum.

Geodesics are of interest not only in differential geometry, but also in mathematical physics and other subjects. It turns out that many problems can be interpreted as geodesic problems on a suitably defined manifold.[13] In this regard, the geodesic problem is even more important because it provides a unifying framework for many problems.

1.4.3 Minimal Surfaces

We have already encountered a special minimal surface problem in our discussion of the catenary. The rotational symmetry of the problem reduced the problem to that of finding a function y of a single variable x, the graph of which generates the surface of revolution having minimal surface area. Locally, any surface can be represented in "graphical" form,

$$\mathbf{r}(x, y) = (x, y, z(x, y)), \tag{1.17}$$

where $\mathbf{r}$ is the position function in $\mathbb{R}^3$. Unless some symmetry condition is imposed, a surface parametrization requires two independent variables. Thus the problem of finding a surface with minimal surface area involves two independent variables in contrast to the problems discussed earlier.

[13] In the theory of relativity, where differential geometry is widely used, the condition that the metric tensor be positive definite is relaxed to positive semidefinite.

Given a simple closed space curve γ, the basic minimal surface problem entails finding, among all smooth simply connected surfaces with γ as a boundary, the surface having minimal surface area. Suppose that the curve γ can be represented parametrically by $(x(t), y(t), z(t))$ for $t \in [t_0, t_1]$, and for simplicity suppose that the projection of γ on the xy-plane is also a simple closed curve; i.e., the curve $\hat{\gamma}$ described by $(x(t), y(t))$ for $t \in [t_0, t_1]$ is a simple closed curve in the xy-plane. Let Ω denote the region in the xy-plane enclosed by $\hat{\gamma}$. Suppose further that we restrict the class of surfaces under consideration to those that can be represented in the form (1.17), where z is a smooth function for $(x, y) \in \Omega$. The differential area element is given by

$$dA = \sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\, dx\, dy,$$

and the surface area is thus

$$A(z) = \int\int_{\Omega} \sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\, dx\, dy.$$

The (simplified) minimal surface problem thus concerns determining a smooth function $z : \Omega \to \mathbb{R}$ such that $z(x(t), y(t)) = z(t)$ for $t \in [t_0, t_1]$, and $A(z)$ is a minimum. There is a substantial body of information about minimal surfaces. The reader can find an overview of the subject in Osserman [58].

1.5 Optimal Harvest Strategy

Our final example in this chapter concerns a problem in economics dealing with finding a harvest strategy that maximizes profit. Here, we follow the example given by Wan [71], p. 6 and use a fishery to illustrate the model.

Let $y(t)$ denote the total tonnage of fish at time t in a region Ω of the ocean, and let y_c denote the carrying capacity of the region Ω for the fish. The growth of the fish population without any harvesting is typically modelled by a first-order differential equation

$$y'(t) = f(t, y). \tag{1.18}$$

If y is small compared to y_c, then f is often approximated by a linear function in y; i.e., $f(t, y) = ky + g(t)$, where k is a constant. More complicated models are available for a wider range of $y(t)$ such as logistic growth

$$f(t, y) = ky(t)\left(1 - \frac{y(t)}{y_c}\right).$$

The ordinary differential equation (1.18) is accompanied by an initial condition

$$y(0) = y_0 \tag{1.19}$$

that reflects the initial fish population.

Suppose now that the fish are harvested at a rate $w(t)$. Equation (1.18) for the population growth can then be modified to the relation

$$y'(t) = f(t, y) - w(t). \tag{1.20}$$

Given the function f, the problem is to determine the function w so that the profit in a given time interval T is maximum.

It is reasonable to expect that the cost of harvesting the fish depends on the season, the fish population, and the harvest rate. Let $c(t, y, w)$ denote the cost to harvest a unit of fish biomass. Suppose that the fish commands a price p per unit fish biomass and that the price is perhaps season dependent, but not dependent on the volume of fish on the market. The profit gained by harvesting the fish in a small time increment is $(p(t) - c(t, y, w))w(t)\,dt$. Given a fixed period T with which to plan the strategy, the total profit is thus

$$P(y, w) = \int_0^T (p(t) - c(t, y, w))w(t)\,dt.$$

The problem is to identify the function w so that P is maximum.

The above problem is an example of a constrained variational problem. The functional P is optimized subject to the constraint defined by the differential equation (1.20) (a nonholonomic constraint) and initial condition (1.19). We can convert the problem into an unconstrained one by simply eliminating w from the integrand defining P using equation (1.20). The problem then becomes the determination of a function y that maximizes the total profit. This approach is not necessarily desirable because we want to keep track of w, the only physical quantity we can regulate.

A feature of this problem that distinguishes it from earlier problems is the absence of a boundary condition for the fish population at time T. Although we are given the initial fish population, it is not necessarily desirable to specify the final fish population after time T. As Wan points out, the condition $y(T) = 0$, for example, is not always the best strategy: "green issues" aside, it may cost far more to harvest the last few fish than they are worth. This simple model thus provides an example of a variational problem with only one endpoint fixed in contrast to the catenary and brachystochrone.

In passing we note that economic models such as this one are generally framed in terms of "present value." A pound sterling invested earns interest, and this should be incorporated into the overall profit. If the interest is compounded continuously at a rate r, then a pound invested yields e^{rt} pounds after time t. Another way of looking at this is to view a pound of income at time t as worth e^{-rt} pounds now. Considerations of this sort lead to profit functionals of the form

$$P(y, w) = \int_0^T e^{-rt}(p(t) - c(t, y, w))w(t)\,dt.$$

http://www.springer.com/978-0-387-40247-5

The Calculus of Variations
van Brunt, B.
2004, XIV, 292 p., Hardcover
ISBN: 978-0-387-40247-5