Chapter 3

The Variational Principle

While Newton was still a student at Cambridge University, and before he had discovered his laws of particle motion, the great French amateur mathematician Pierre de Fermat (1601?- 1665) proposed a startlingly different explanation of motion. Fermat's explanation was not for the motion of *particles*, however, but for *light rays*. In this chapter we explore Fermat's approach, and then go on to introduce techniques in variational calculus used to implement this approach, and to solve a number of interesting problems. We then show how Einstein's special relativity and principle of equivalence help us show how the variational calculus can be used to understand the motion of particles. All this is to set the stage for applying variational techniques to general mechanics problems in the following chapter.

3.1 Fermat's principle

Imagine that a ray of light leaves a light source at point a and travels to some other given point b. Fermat proposed that out of all the infinite number of paths that the ray might take between the two points, it actually travels by the path of *least time*. For example, if there is nothing but vacuum between a and b, light traveling at constant speed c takes the path that minimizes the travel time, which of course is a straight line. Or suppose a piece of glass is inserted into part of an otherwise air-filled space between a and b. In any medium with index of refraction n, light has speed $v = c/n$, the minimum-time path in that case is no longer a straight line: Fermat's principle of least time predicts that the ray will bend at the air-glass interface by an angle given by Snell's law, as shown in Figure 3.1.

More generally, a light ray might be traveling through a medium with index

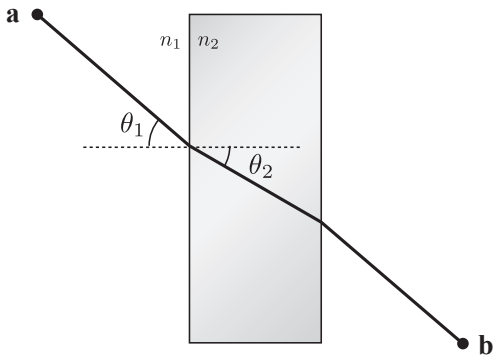

Figure 3.1: Light traveling by the least-time path between a and b, in which it moves partly through air and partly through a piece of glass. At the interface the relationship between the angle θ_1 in air, with index of refraction n_1, and the angle θ_2 in glass, with index of refraction n_2, is $n_1 \sin\theta_1 = n_2 \sin\theta_2$, known as **Snell's law**. This phenomenon is readily verified by experiment.

of refraction n that can be a continuous function of position, $n(x, y, z) \equiv n(\mathbf{r})$. In that case the time it takes the ray to travel an infinitesimal distance ds is

$$dt = \frac{\text{distance}}{\text{speed}} = \frac{ds}{c/n(\mathbf{r})}, \tag{3.1}$$

so the total time to travel from a to b by a particular path s is the integral

$$t = \int dt = \frac{1}{c}\int n(\mathbf{r})\, ds. \tag{3.2}$$

The value of the integral depends upon the path chosen, so out of the infinite number of possible paths the ray might take, we are faced with the problem of finding the particular path for which the integral is a minimum. If we can find it, Fermat assures us that it is the path the light ray actually takes between a and b.

Fermat's Principle raises many questions, not least of which is: *how does the ray "know" that of all the paths it might take, it should pick out the least-time path?* In fact, a contemporary of Fermat named Claude Cierselier, who was an expert in optics, wrote

... Fermat's principle can not be the cause, for otherwise we would be attributing knowledge to nature: and here, by nature, we understand only that order and

lawfulness in the world, such as it is, which acts without foreknowledge, without choice, but by a necessary determination.

In Chapter 12 we will explain the deep reason *why* a light ray follows the minimum-time path. But in the meantime we can state that there are similar minimizing principles for the motion of classical *particles*, so it will be important to understand how to find the path that minimizes some integral, analogous to the integral in (3.2). The technique is called the **calculus of variations**, or **functional calculus**, and that is the primary topic of this chapter.

3.2 The calculus of variations

The general methods of the calculus of variations were first worked out in the 1750's by the French mathematician Joseph-Louis Lagrange and the Swiss mathematician Leonard Euler, a century after Fermat proposed his principle. As an example of setting up these methods, return to the problem of finding the minimum-time path for a light ray traveling in a two-dimensional plane. Suppose that a light ray from a star enters Earth's upper atmosphere and travels all the way to the ground, as depicted in Figure 3.2. The density of air varies with altitude, so the index of refraction $n = n(y)$ varies with altitude as well, where y is the vertical coordinate. The time to travel by any path is

$$t = \int \frac{ds}{c/n(y)} = \frac{1}{c}\int dsn(y) \tag{3.3}$$

where ds is the distance between two infinitely nearby points along the path and c is the speed of light in vacuum. So in this case, the problem of finding the minimum-*time* path is the same as the problem of finding the minimum-*distance* path. From the Pythagorean theorem we know that $ds = \sqrt{dx^2 + dy^2}$, where x and y are Cartesian coordinates in the plane. A path could then be specified by $y(x)$. The time to travel by any path $y(x)$ is therefore

$$\begin{aligned} t &= \frac{1}{c}\int n(y)\sqrt{dx^2+dy^2} = \frac{1}{c}\int n(y)\sqrt{1+\left(\frac{dy}{dx}\right)^2}\,dx \\ &\equiv \frac{1}{c}\int n(y)\sqrt{1+y'^2}\,dx\ . \end{aligned} \tag{3.4}$$

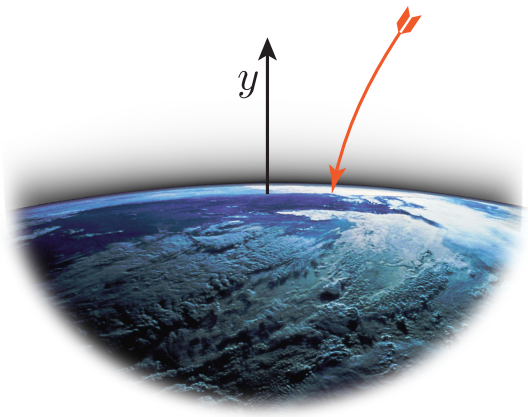

Figure 3.2: A light ray from a star travels down through Earth's atmosphere on its way to the ground.

In this case the integrand depends on *both* the path $y(x)$ and its slope $y'(x)$. It is easy to imagine that the index of refraction n might also depend upon a horizontal coordinate x (the density of air might vary somewhat horizontally as well as vertically), in which case the time for the ray to reach the ground would be

$$t = \frac{1}{c}\int n(x,y)\sqrt{1+y'^2}\,dx \equiv \int F(x,\ y(x),\ y'(x))\,dx \tag{3.5}$$

where the integrand $F(x,\ y(x),\ y'(x)) = (1/c)n(x,y)\sqrt{1+y'^2}$ depends upon all three variables $x, y(x)$, and $y'(x)$. The calculus of variations again shows us how to find the particular path $y(x)$ that minimizes this integral.

Finding the least-time path is only one example of a problem in the calculus of variations. More generally, Euler and Lagrange consider some arbitrary integral I of the form

$$I = \int F(x,\ y(x),\ y'(x))\,dx, \tag{3.6}$$

and the problem they want to solve is to find not only paths $y(x)$ that *minimize* I, but also paths that *maximize* I, or otherwise make I *stationary*. It is also possible to have a stationary path that is neither a maximum nor a minimum, as we shall see.

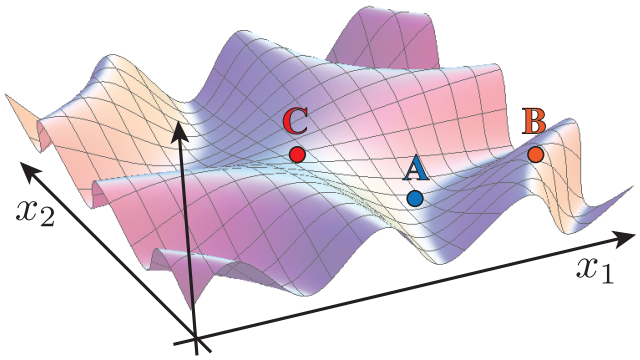

Figure 3.3: A function of two variables $f(x_1, x_2)$ with a local minimum at point A, a local maximum at point B, and a saddle point at C.

How do we go about making I stationary? Let us revisit the more familiar problem of making an ordinary function stationary. Say we are given a function $f(x_1, x_2, \cdots) \equiv f(x_i)$ of a number of independent variables x_i with $i = 1 \ldots N$, and we are asked to find the stationary points of this function. For the simpler case of a function with only two variables, we can visualize the problem as shown in Figure 3.3: we have a curved surface $f(x_1, x_2)$ over the x_1-x_2 plane, and we are looking for special points (x_1, x_2) where the surface is "locally flat". These can correspond to minima, maxima, or saddle points, as shown in the figure. Algebraically, we can phrase the general problem as follows. For every point $(x_1, x_2, \cdots)$, we move away by a small arbitrary distance δx_i

$$x_i \to x_i + \delta x_i \ . \tag{3.7}$$

We then seek a special point $(x_1, x_2, \cdots)$ where this shift does not change the function $f(x_i)$ to *linear order* in the small shifts δx_i. This is to be our intuitive meaning of being 'locally flat'. Using a Taylor expansion, we can then write

$$f(x_i) \to f(x_i + \delta x_i) = f(x_i) + \frac{\partial f}{\partial x_j}\delta x_j + \frac{1}{2!}\frac{\partial^2 f}{\partial x_j x_k}\delta x_j \delta x_k + \cdots \tag{3.8}$$

Note that the j and k indices are repeated and hence summed over, using again the Einstein summation convention of Chapter 2. If the function is to

remain constant up to linear order in δx_i, we then need N conditions

$$\frac{\partial f}{\partial x_j} = 0; \tag{3.9}$$

i.e., the slopes in all directions must vanish at the special point where the surface flattens out. This is because the δx_js are arbitrary and independent, and the only way for $(\partial f/\partial x_j)\delta x_j$ to vanish for an *arbitrary* δx_j is to have all the derivatives $\partial f/\partial x_j$ vanish.

The second-derivative terms involving $\partial^2 f/\partial x_j \partial x_k$ tell us how the surface curves away from the local plateau, whether the point is a minimum, maximum, or saddle point. Equation (3.9) typically yields a set of algebraic equations that can be solved for x_i, identifying the point(s) of interest. Formally, we write the condition to make a function $f(x)$ stationary as

$$\delta f \equiv f(x_i + \delta x_i) - f(x_i) = 0 \text{ to linear order in } \delta x_i \Rightarrow \frac{\partial f}{\partial x_i} = 0 \ . \tag{3.10}$$

But we already knew all this. The real problem now is to make stationary not just any old function, but the integral I as given by (3.6). The quantity I is different from a regular function as follows: A function $f(x_1, x_2, \cdots)$ takes as input a set of numbers $(x_1, x_2, \cdots)$, and gives back a number. The quantity I, however, takes as input *an entire function* $y(x)$, and gives back a single number. Take the function $f(x) = x^2$, for example: if $x = 3$, then $f(3) = 9$: one number in gives one number out. But to calculate a value for an integral $I = \int_a^b F(x, y(x), y'(x))\ dx$ with (say) $F(x, y(x), y'(x)) = y(x)^2$ and $a = 0, b = 1$, we have to substitute into it not a single number, but an entire path $y(x)$. If for example $y(x) = 5\,x$ (hence with the boundary conditions $y(0) = 0$ and $y(1) = 5$), we would write

$$I = \int_0^1 (5\,x)^2 dx = \left.\frac{25}{3}x^3\right|_0^1 = \frac{25}{3} \ . \tag{3.11}$$

The argument for I is then a path, an entire function $y(x)$. To make this explicit, we instead write I as

$$I[y(x)] = \int_a^b F(x,\ y(x),\ y'(x))\ dx \ . \tag{3.12}$$

with square brackets around the argument: I is not a function, but is called a **functional**.

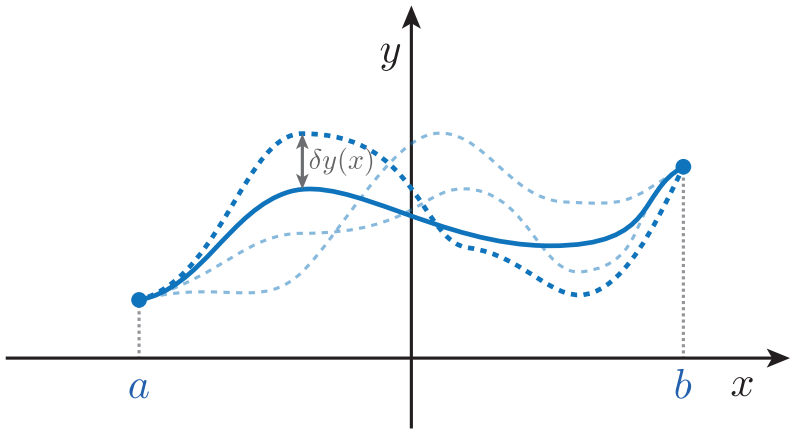

Figure 3.4: Various paths $y(x)$ that can be used as input to the functional $I[f(x)]$. We look for that special path from which an arbitrary small displacement $\delta y(x)$ leaves the functional unchanged to linear order in $\delta y(x)$. Note that $\delta y(a) = \delta y(b) = 0$.

In general, a functional may take as argument several functions, not just one. But for now let us focus on the case of a functional depending on a single function. The question we want to address is then: how do we make such a functional stationary? This means we are looking for conditions that identify a set of paths $y(x)$ where the functional $I[y(x)]$ is stationary or 'locally flat'. To do this, we can build upon the simpler example of making stationary a function. For any path $y(x)$, we look at a shifted path

$$y(x) \to y(x) + \delta y(x) \tag{3.13}$$

where $\delta y(x)$ is a function that is small everywhere, but is otherwise arbitrary. However, we require that at the endpoints of the integration in (3.12), the shifts vanish; *i.e.*, $\delta y(a) = \delta y(b) = 0$. This means that we do not perturb the boundary conditions on trial paths that are fed into $I[y(x)]$, because we only need to find the path that makes stationary the functional amongst the subset of all possible paths that satisfy the given fixed boundary conditions at the endpoints. We illustrate this in Figure 3.4. In this restricted set of trial paths, our functional extremization condition now looks very much like (3.10)

$$\delta I[y(x)] \equiv I[y(x) + \delta y(x)] - I[y(x)] = 0 \tag{3.14}$$

to linear order in $\delta y(x)$. We say: "the variation of the functional I is zero." For a function $f(x_1, x_2, \cdots)$, the condition amounted to setting all first derivatives of f to zero. Hence, we need to figure out how to differentiate a functional! Alternatively, we need to expand the functional $I[y(x) + \delta y(x)]$ in $\delta y(x)$ to linear order to identify its 'first derivative'.

Fortunately, we can deduce all operations of functional calculus by thinking of a functional in the following way. Imagine that the input to the functional, the path $y(x)$, is evaluated only on a finite discrete set of points:

$$a < x < b \to x = a + n\,\epsilon \leq b \tag{3.15}$$

for n a non-negative integer and ϵ small. In the limit $\epsilon \to 0$ and $n \to \infty$, we recover the original continuum problem. The functional is now simply a function of a finite number of variables $y(a), y(a+\epsilon), y(a+2\,\epsilon), \cdots$. In the limit $\epsilon \to 0$, the set becomes infinitely dense. You can hence view a functional as a function of an *infinite* number of variables. We can perform all needed operations on I in the discretized regime where I is treated as a function; and then take the $\epsilon \to 0$ limit at the end of the day.

Basically, we may think of x in $y(x)$ as a discrete index y_x. We then have $I[y(x)] \to I(y_x)$, a function with a large but finite number of variables y_x, with $x \in \{a \cdots b\}$ a finite set. A functional then becomes a much more familiar animal: a function. The integral I may also depend upon $y'(x)$, which can be written in our discrete way as $y'(x) \to (y_x - y_{x-\epsilon})/\epsilon$ by the definition of the derivative operation. We write it in shorthand as $y'(x) \to y'_x$; and the integration in (3.12) becomes an infinite sum: $\int dx \to \sum_x \epsilon$. To summarize, we now have a discretized form of our original functional, which has the form

$$I = \sum_x F(x, y_x, y'_x)\,\epsilon\ . \tag{3.16}$$

We can now apply the shifts $y_x \to y_x + \delta y_x$, which also implies $y'_x \to y'_x + \delta y'_x$, where $\delta y'_x = (\delta y_x - (\delta y)_{x-\epsilon})/\epsilon = d(\delta y_x)/dx$. We then need the analogue to

$$\delta f = \frac{\partial f}{\partial x_i}\delta x_i = 0 \tag{3.17}$$

with $f \to I$, and $x_i \to y_x$. Starting from (3.16), we then have

$$\delta I = \sum_x \left(\frac{\partial F}{\partial y_x}\delta y_x + \frac{\partial F}{\partial y'_x}\delta y'_x\right)\epsilon = 0\ . \tag{3.18}$$

In the $\epsilon \to 0$ limit we retrieve the integral form

$$\delta I[y(x)] = \int_a^b \left(\frac{\partial F}{\partial y(x)} \delta y(x) + \frac{\partial F}{\partial y'(x)} \frac{d}{dx} \left(\delta y(x) \right) \right) dx = 0. \tag{3.19}$$

Integrating the second term by parts, we get

$$\int_a^b \frac{\partial F}{\partial y'(x)} \frac{d}{dx} \left(\delta y(x) \right) = \delta y(x) \frac{\partial F}{\partial y'(x)} \bigg|_a^b - \int_a^b \delta y(x) \frac{d}{dx} \left(\frac{\partial F}{\partial y'(x)} \right) dx \tag{3.20}$$

where the first term on the right vanishes because we have fixed the endpoints so that $\delta y(a) = \delta y(b) = 0$. Therefore equation (3.19) becomes

$$\delta I[y(x)] = \int_a^b \left(\frac{\partial F}{\partial y(x)} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'(x)} \right) \right) \delta y(x) \, dx = 0. \tag{3.21}$$

This integral might be zero because the integrand is zero for all x, or because there are positive and negative portions that cancel one another out. However, since *arbitrary* smooth deviation functions $\delta y(x)$ are permitted, the first alternative has to be the right one. For example, if $a < x_0 < b$ and the integrand happens to be positive from a to x_0 and negative from x_0 to b so that by cancellation the overall integral is zero, the deviation function $\delta y(x)$ could be changed so that $\delta y(x) = 0$ from x_0 to b, which would force the integral to be positive. Therefore the requirement that the integral vanish for *arbitrary* smooth functions $\delta y(x)$ requires that

$$\frac{\partial F}{\partial y(x)} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'(x)} \right) = 0 \ , \tag{3.22}$$

which is known as **Euler's equation**. This equation was worked out by both Euler and Lagrange about the same time, but we will call it simply "Euler's equation", because we will reserve the term "Lagrange equations" for essentially the same equation when used in classical mechanics, as we will see in Chapter 4.

Note two important features of Euler's equation:

1. The derivatives with respect to y and y' are *partial*, but the derivative with respect to x is *total*. Suppose, for example, that $F(x, y(x), y'(x)) = x\, y\, (y')^2$. Then $\partial F/\partial y = (y')^2\, x$ and $\partial F/\partial y' = 2\, x\, y\, y'$, so Euler's equation becomes

$$x\, (y')^2 - \frac{d}{dx}(2\, x\, y\, y') = x\, (y')^2 - [2\, y\, y' + 2\, x\, (y')^2 + 2\, x\, y\, y''] = 0. \tag{3.23}$$

This is an ordinary differential equation whose solution $y(x)$ is the path we are looking for. That is, in the calculus of variations, *Euler's equation converts the problem of finding which path makes a particular integral stationary into a differential equation for the path, whose solution gives the path we want.*

2. The variables x and y in Euler's equation do not have to represent Cartesian coordinates. The mathematics has no idea what x and y represent, as long as they are independent of one another. So if an integral I has the form of (3.12), but with x and y replaced by different symbols, the corresponding Euler's equation still holds. The total derivative occurring in the equation is always with respect to whatever variable of integration is chosen in the problem, which is called the independent variable. For example, if the integral to be made stationary has the form

$$I[q(t)] = \int F(t, q(t), q'(t))dt \tag{3.24}$$

then the corresponding Euler equation is

$$\frac{\partial F}{\partial q(t)} - \frac{d}{dt}\left(\frac{\partial F}{\partial q'(t)}\right) = 0. \tag{3.25}$$

t is then the independent variable while $q(t)$ is referred to as the dependent variable; and $q'(t) \equiv dq/dt$.

3.3 Geodesics

The calculus of variations is best learned through examples. Let us proceed to a sequence of explicit cases where these techniques can come in handy. One application is to find **geodesics**, which are the stationary (usually the shortest) paths between two points on a given surface.

EXAMPLE 3-1: Geodesics on a plane

We have in effect already set up the problem of finding the geodesic paths on a plane. The appropriate integral in that case is

$$s = \int \sqrt{dx^2 + dy^2} = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx \equiv \int_a^b \sqrt{1 + y'^2}\, dx. \quad (3.26)$$

We then have $F = \sqrt{1 + y'^2}$ in equation (3.6). Note that the integrand does not depend upon either x or $y(x)$ explicitly, so $\partial F/\partial y = 0$. Euler's equation (3.22) then becomes simply

$$\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0 \quad (3.27)$$

and so

$$\frac{\partial F}{\partial y'} = \frac{y'}{\sqrt{1 + (y')^2}} = k, \quad (3.28)$$

where k is a constant. Solving for y',

$$y' = \frac{\pm k}{\sqrt{1 - k^2}} \equiv m_1, \quad (3.29)$$

which defines the constant m_1 in terms of the constant k. The integral of this equation is $y = m_1\, x + m_2$, where m_2 is a constant of integration. That is, the shortest distance on a plane between two points is a straight line (!), where the slope m_1 and y-intercept m_2 may be identified by requiring the line to pass through the endpoints $a = (x_a, y_a)$ and $b = (x_b, y_b)$.

Using the calculus of variations, we have shown that among all smooth paths it is a straight line that makes the distance stationary. In this case stationary means minimum, because all nearby paths are longer. We showed earlier that minimizing the travel time of a light ray moving from a to b through a vacuum is equivalent to minimizing the distance traveled, so we have now also (no surprise) found that the minimum travel time path for a light ray is a straight line in this case.

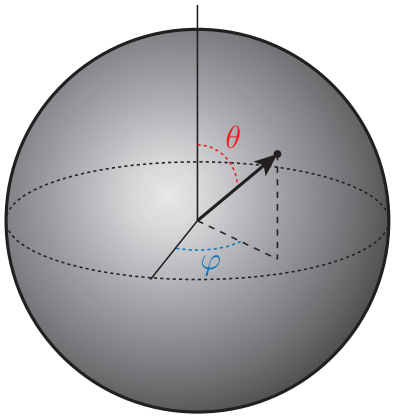

Figure 3.5: The coordinates θ and φ on a sphere.

EXAMPLE 3-2: Geodesics on a sphere

Consider the problem of finding the shortest distance between two points on the surface of a sphere, as illustrated in Figure 3.5. We can use the polar angle θ and azimuthal angle φ as the coordinates on a sphere. If R is the radius of the sphere, an infinitesimal distance in the θ direction is $ds_\theta = R\,d\theta$ and an infinitesimal distance in the φ direction is $ds_\varphi = R\sin\theta d\varphi$. These two distances are perpendicular to one another, so the distance squared between any two nearby points is the sum of squares,

$$ds^2 = R^2 d\theta^2 + R^2 \sin^2\theta d\varphi^2. \tag{3.30}$$

There are two ways to write the total distance between two points, depending upon whether we use φ or θ as the variable of integration. If we use φ, then

$$s = R\int_a^b \sqrt{\theta'^2 + \sin^2\theta}\; d\varphi, \tag{3.31}$$

where $\theta' = d\theta/d\varphi$. The corresponding Euler equation is

$$\frac{\partial F}{\partial \theta} - \frac{d}{d\varphi}\frac{\partial F}{\partial \theta'} = 0 \tag{3.32}$$

where $F = \sqrt{\theta'^2 + \sin^2\theta}$. Alternatively, we can write

$$s = R\int_a^b \sqrt{1 + \sin^2\theta\ \varphi'^2}\ d\theta, \tag{3.33}$$

where $\varphi' = d\varphi/d\theta$ with the corresponding Euler equation

$$\frac{\partial F}{\partial \varphi} - \frac{d}{d\theta}\frac{\partial F}{\partial \varphi'} = 0, \tag{3.34}$$

and where in this case $F = \sqrt{1 + \sin^2\theta\ \varphi'^2}$. Both Euler equations are correct. Is one easier to use than the other?

In the first alternative, equation (3.32) results in a second-order differential equation, since the first term $\partial F/\partial\theta \neq 0$ and by the time all the derivatives are taken the second term includes a second derivative θ''. The second alternative (3.34) is much easier to use, because in that case $F = \sqrt{1 + \sin^2\theta\ \varphi'^2}$ is not an *explicit* function of φ, so the first term in Euler's equation vanishes. The quantity $\partial F/\partial\varphi'$ must therefore be constant in θ, since its total derivative is zero. This leaves us with only a first-order differential equation

$$\frac{\partial F}{\partial \varphi'} = \frac{\sin^2\theta\ \varphi'}{\sqrt{1 + \sin^2\theta\varphi'^2}} = k, \tag{3.35}$$

for some constant k, which can be solved for φ' and rearranged to give

$$\varphi' = \pm\frac{k\csc^2\theta}{\sqrt{1 - k^2\csc^2\theta}}. \tag{3.36}$$

Using the identity $\csc^2\theta = 1 + \cot^2\theta$ and substituting $q = \alpha\cot\theta$, where $\alpha = k/\sqrt{1-k^2}$, gives

$$\varphi = \alpha\int \frac{dq}{\sqrt{1-q^2}} = \alpha\sin^{-1}q + \beta, \tag{3.37}$$

where $\alpha = \pm(\sqrt{1-k^2})/k$ and β is a constant of integration. Therefore the equation relating θ and φ is

$$\sin(\varphi - \beta) = q = \alpha\cot\theta. \tag{3.38}$$

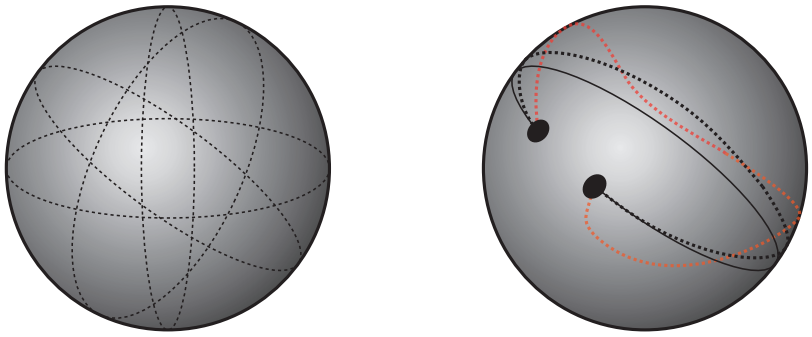

Figure 3.6: (a) Great circles on a sphere are geodesics; (b) Two paths nearby the longer of the two great-circle routes of a path.

We can better understand the meaning of this result by multiplying through by $R\cos\theta$ and using the identity $\sin(\varphi-\beta) = \sin\varphi\cos\beta - \cos\varphi\sin\beta$, which gives

$$(\cos\beta)y - (\sin\beta)x = \alpha z \tag{3.39}$$

where $x = R\sin\theta\cos\varphi$, $y = R\sin\theta\sin\varphi$, and $z = R\cos\theta$, which are the Cartesian coordinates on the sphere. Equation (3.39) is the equation of a plane passing through the center of the sphere, which slices through the sphere in a **great circle**. So we have found that the solutions of Euler's equation are great-circle routes, as illustrated in 3.6(a).

Unless one endpoint is at the antipode of the other, there is a shorter distance and a longer distance along the great circle that connects them. The shorter distance is a minimum path length under small deviations in path, as is well known by airline pilots. The larger distance is a stationary path that is neither a minimum nor a maximum under all small deviations in path. Paths that oscillate around this path are generally longer than the great-circle route, while some paths pulled to one side of the great-circle route are shorter. Both kinds are sketched in 3.6(b). This behavior is fairly typical of stationary paths that are neither absolute maxima nor absolute minima relative to all neighboring paths: Some neighboring paths lead to smaller

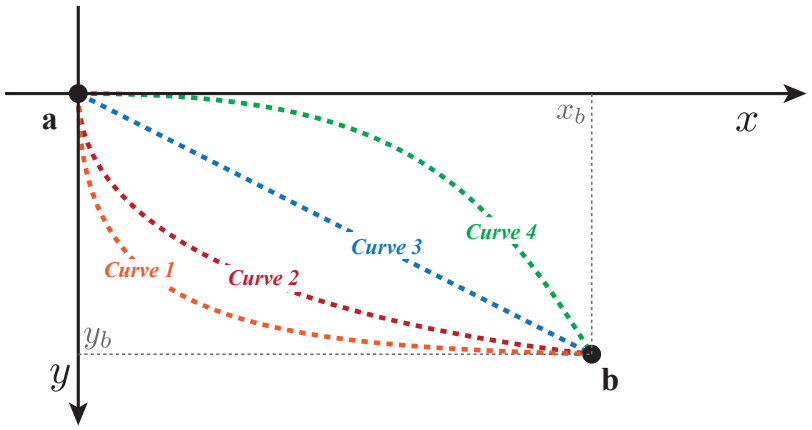

Figure 3.7: Possible least-time paths for a sliding block.

values and others lead to larger values of the integral I. In this case the set of all such paths represent a kind of saddle in a very large-dimensional space.

3.4 Brachistochrone

The **brachistochrone** ("shortest time") problem was invented and solved a half century *before* the work of Euler and Lagrange, and engaged some of the most creative people in the history of physics and mathematics. The problem is to find the shape of a frictionless track between two given points, such that a small block starting at rest at the upper point — and sliding without friction down along the track under the influence of gravity — arrives at the lower point in the shortest time. The two points a and b, and shapes of possible tracks between them, are illustrated in 3.7.

We can guess the qualitative shape of the shortest-time track by physical reasoning. Of the four curves shown in Figure 3.7, it might seem that the straight line 3 is the shortest-time path, since it is the path of shortest distance. However, curve 2 has an advantage in that the block picks up speed more quickly, so that its greater average speed may more than make up for the greater distance it has to travel. Curve 1 permits the block to pick up speed still faster, but there is a risk that the slightly increased average speed might not outweigh the greater distance involved. There is no reason to

choose curve 4, because a block will hardly get going in the first place and it also has to travel relatively far. A track whose shape is something like curve 2 should be the best choice.

To find the exact shape we choose coordinates as shown in Figure 3.7, with the origin at the release point, the positive y axis extending *downward*, and the final point designated by (x_b, y_b). The time to travel over a short distance is the distance divided by the speed, so the overall time is

$$t = \int \frac{ds}{v} \tag{3.40}$$

where v is the varying speed of the block. The infinitesimal distance is again $ds = \sqrt{dx^2 + dy^2}$. Since v changes in general along the track, we need to express it in terms of the coordinates x and y to make sense of the integral. For this, we have energy conservation which gives

$$E = \frac{1}{2}mv^2 + mg(-y) = 0, \tag{3.41}$$

since y and v are both zero initially. (We have used $-y$ in the potential energy because we are measuring y positive *downward*; *i.e.* the potential $-m\,g\,y$ decreases for larger values of y.) For any given path the time for the block to slide from beginning to end can be expressed either as

$$t = \int \frac{\sqrt{1 + y'^2}}{\sqrt{2gy}}\, dx. \tag{3.42}$$

where $y' = dy/dx$, or as

$$t = \int \frac{\sqrt{1 + x'^2}}{\sqrt{2gy}} dy \tag{3.43}$$

where $x' = dx/dy$. The Euler equation for the latter expression is

$$\frac{\partial F}{\partial x} - \frac{d}{dy}\left(\frac{\partial F}{\partial x'}\right) = 0, \tag{3.44}$$

which is the right one to use, because F is not an explicit function of x, so the first term vanishes. Therefore

$$\frac{\partial F}{\partial x'} = \frac{1}{\sqrt{2gy}} \frac{x'}{\sqrt{1 + x'^2}} = k, \tag{3.45}$$

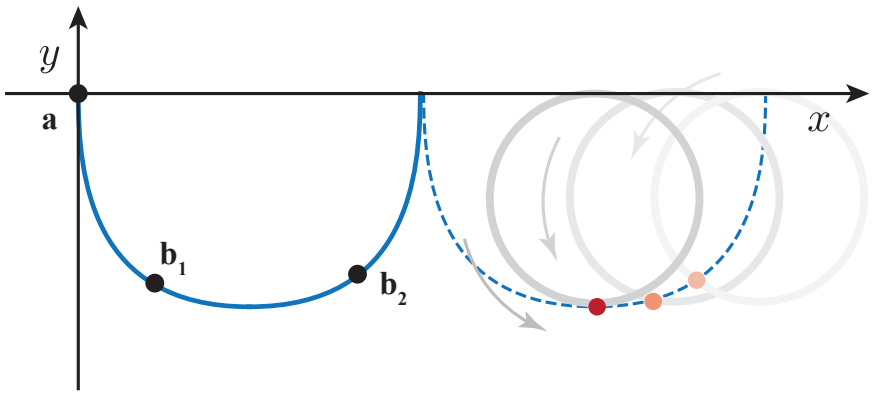

Figure 3.8: A cycloid. If in darkness you watch a wheel rolling along a level surface, with a lighted bulb attached to a point on the outer rim of the wheel, the bulb will trace out the shape of a cycloid. In the diagram the wheel is rolling along horizontally *beneath* the surface. For $x_b < (\pi/2)y_b$, the rail may look like the segment from a to b_1; for $x_b > (\pi/2)y_b$, the segment from a to b_2 would be needed.

a constant. Solving for x',

$$x' = \frac{\pm k\sqrt{2gy}}{\sqrt{1-2k^2gy}} \equiv \sqrt{\frac{y}{a-y}}, \tag{3.46}$$

choosing the plus sign and defining $a = 1/(2k^2g)$. Integrating over y,

$$x = \int dx = \int dy \sqrt{\frac{y}{a-y}}, \tag{3.47}$$

which can be evaluated using the substitution

$$y = a\sin^2\frac{\theta}{2} = \frac{a}{2}(1-\cos\theta), \tag{3.48}$$

giving the result $x = (a/2)(\theta - \sin\theta)$, where we have chosen the constant of integration so that $x = 0$ when $y = 0$ (at $\theta = 0$), which is the release point. The resulting parametric equations

$$x = \frac{a}{2}(\theta - \sin\theta) \tag{3.49}$$

$$y = \frac{a}{2}(1 - \cos\theta) \tag{3.50}$$

are the equations of a cycloid, as shown in Figure 3.8. The quantities a and the final angle parameter θ_b can be determined from the coordinates (x_b, y_b) of the final point, although this ordinarily requires the solution of a transcendental equation. Only the first cycle of the cycloid is needed; if $x_b < (\pi/2)y_b$, the minimum-time path is a piece of the left half of the cycle, as shown in Figure 3.8 (the segment from point a to point b_1); if $x_b > (\pi/2)y_b$, the right half of the cycle is needed as well (the segment from a to b_2). That is, if $x_b > (\pi/2)y_b$, the sliding particle actually descends below y_b, and then comes up to meet y_b at the end. In either case the particle begins by falling vertically when it leaves the origin, to get the maximum possible initial acceleration. The cycloid has a vertical cusp at this point.

The time required to fall to the final point can be found by returning to equation (3.43) and expressing x and y in terms of the parameter θ, according to equations (3.49) and (3.50). The result is simply

$$t = \sqrt{\frac{a}{2g}} \int_0^{\theta_f} d\theta = \sqrt{\frac{a}{2g}}\ \theta_f. \tag{3.51}$$

In particular, if $(x_b, y_b) = (\pi a/2, a)$, so that a complete half-cycle of the cycloid is needed to connect the points, then $\theta_f = \pi$ and

$$t = \pi\sqrt{\frac{a}{2g}}. \tag{3.52}$$

This is the time it would take a particle to slide from the rim to the bottom of a smooth cycloidal bowl, where a is the depth of the bowl.[1]

[1]A bit of history: On the afternoon of January 29, 1697, Sir Isaac Newton, who had left Cambridge the previous year to become Warden of the Mint in London, returned to his London home from a hard day at the Mint to find a letter from the Swiss mathematician Johann Bernoulli. The letter contained the brachistochrone problem, published the previous June. A challenge had gone forth to mathematicians to solve the problem, and they were given a time limit of six months to find the solution. Gottfried Wilhelm Leibniz, German mathematician and arch rival of Newton for recognition as the original inventor of calculus, solved the problem but asked that the deadline be extended by an additional year so that everyone would have a chance to try it. Bernoulli agreed. Although presented as a general challenge, Bernoulli specifically sent the problem to Newton, who had not seen it before, to alert him to the problem and to try to stump him, thereby showing that

EXAMPLE 3-3: Fermat again

We return to where we began the chapter, with Fermat's principle of stationary time, illustrated in Figure 3.9(a). Bringing to bear the calculus of variations, we can now find the path of a light ray in a medium like Earth's atmosphere, where the index of refraction n is a continuous function of position. If a ray of light from a star descends through the atmosphere it encounters an increasing density and an increasing index of refraction. We might therefore expect the ray to bend continuously, entering the atmosphere at some angle θ_a and reaching the ground at a steeper angle θ_b. For simplicity, take the Earth to be essentially flat over the horizontal range of the ray and assume the index of refraction $n = n(y)$ only, where y is the vertical direction. The light travel time is then

$$t = \frac{1}{c}\int n(y)\sqrt{1+y'^2}\,dx = \frac{1}{c}\int n(y)\sqrt{1+x'^2}\,dy. \tag{3.53}$$

The easiest solution comes from using the second form. Then $\partial F/\partial x = 0$, so

$$\frac{d}{dy}\left(\frac{\partial F}{\partial x'}\right) = 0, \tag{3.54}$$

and so

$$\frac{\partial F}{\partial x'} = \frac{n(y)\;x'}{\sqrt{1+x'^2}} = k, \tag{3.55}$$

a constant. The derivative $x' = dx/dy = \tan\theta$, where θ is the local angle of the ray relative to the vertical, so the quantity $x'/\sqrt{1+x'^2} = \sin\theta$. Therefore

$$n(y)\sin\theta = k, \tag{3.56}$$

---

he did not really understand calculus as well as the continental mathematicians.

Newton's niece, Catherine Barton, was living with him in London at the time. She later testified that "Sr I. N. was in the midst of the hurry of the great recoinage and did not come home till four from the Tower very much tired, but did not sleep till he had solved it wch was by 4 in the morning." Newton sent off the solution that same morning to the Royal Society, and it was published anonymously in the February issue of the *Philosophical Transactions*. Bernoulli had no doubt who was responsible, and wrote to a friend that it was "ex ungue Leonum" — "from the claws of the Lion." Aside from Newton, Leibniz, and Johann Bernoulli himself, the brachistochrone problem was solved by only two other mathematicians at that time, Bernoulli's older brother Jacob and the French mathematician de l'Hospital. All of the solutions were ad hoc, involving algorithms suited to the particular problem, but not necessarily easily generalizable to a wider class of problems.

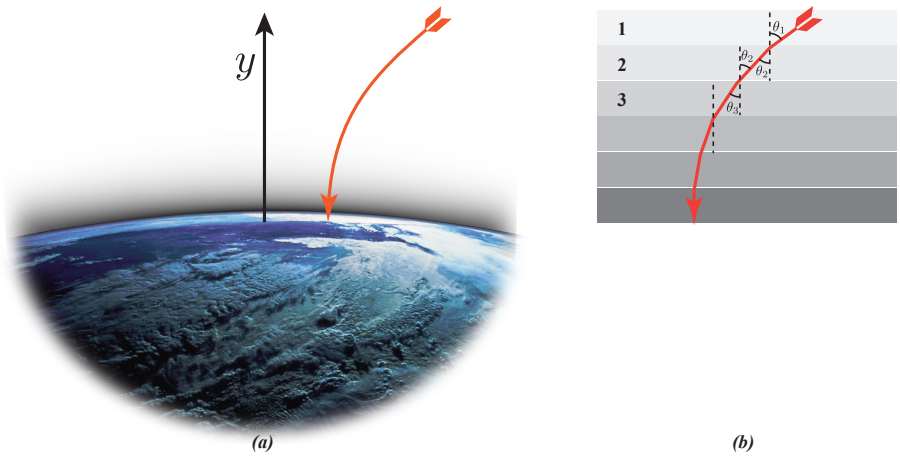

Figure 3.9: (a) A light ray passing through a stack of atmospheric layers; (b) The same problem visualized as a sequence of adjacent slabs of air of different index of refraction.

a constant everywhere along the path. This result could also have been obtained immediately from Snell's law, by modeling the atmosphere as a large number of thin horizontal layers, where n is constant within each layer, but with n increasing slightly as one passes from one layer to the layer just beneath it. Snell's law is obeyed at each boundary: for example, $n_1 \sin\theta_1 = n_2 \sin\theta_2$ as shown in Figure 3.1. However, the angle θ_2 at which the ray leaves layer 2 is the same angle at which the ray leaves layer 2 at the boundary with layer 3 (see Figure 3.9(b)). Therefore also $n_2 \sin\theta_2 = n_3 \sin\theta_3$, etc., so in the stack of layers it follows that $n(y)\sin\theta = \text{constant}$. In the limit where the stack approaches an infinite number of layers of infinitesimal thickness, we get equation (3.56). Given a function $n(y)$ we can then find the specific path shape $y(x)$ from $\theta(y)$ (See problems at the end of the chapter.)

Note that the constancy of $n\sin\theta$ allows us to predict the ray angle θ_b at the ground without knowing the detailed index of refraction $n(y)$ or the path of the ray! If we know the indices of refraction at the top of the atmosphere n_a and at the ground n_b, and the angle at which the ray enters the atmosphere θ_a (from the true location of the star) we can find the angle at the ground θ_b — which is the angle at which a telescope would observe the star — as

$$n_a \sin\theta_a = n_b \sin\theta_b = \text{constant} \tag{3.57}$$

3.5 Several Dependent Variables

We have so far considered problems with one independent variable (such as x) and one dependent variable (such as $y(x)$). There are many additional problems that require two or more dependent variables, such as both $y(x)$ and $z(x)$. For example, to find the shortest-distance path between two given points in three-dimensional space, we would need both y and z as well as x to describe an arbitrary path. Consider the more general functional

$$I[t, y_i(x), y_i'(x)] = \int_{x_1}^{x_2} F(x,\ y_1(x), ...\ y_N(x),\ y_1'(x), ...\ y_N'(x))\ dx \quad (3.58)$$

with $y_1(x)$, $y_2(x)$, ..., $y_N(x)$, we then have N dependent variables. The goal is to make I stationary under variations in *all* of the functions $y_i(x)$ with $i = 1, 2, \ldots, N$. In the preceding section, the single function $y(x)$ could be visualized as a path in the two-dimensional x, y space; in the more general case the N functions $y_i(x)$ can be visualized as together defining a path in an $N + 1$- dimensional space, with axes $x, y_1, y_2, ...y_N$.

For example, the distance between the two points (x_a, y_a, z_a) and (x_b, y_b, z_b) in three dimensions is

$$s = \int ds = \int \sqrt{dx^2 + dy^2 + dz^2} = \int_{x_a}^{x_b} \sqrt{1 + y'^2 + z'^2}\ dx \quad (3.59)$$

along a path described by $y(x)$ and $z(x)$, restricted to pass through the given endpoints. The three-dimensional path that minimizes s is a problem in the calculus of variations, and the integral is a simple case of the form written in equation (3.58).

Analogs to the Euler equations can readily be found in the $N + 1$ dimensional case. Let the shift in the paths now be

$$y_i(x) \to y_i(x) + \delta y_i(x) \qquad (i = 1, ...N) \quad (3.60)$$

Therefore the functions $\delta y_i(x)$ describe the deviations of the arbitrary path $y_i(x)$. Looking back at equation (3.19), we note that the only difference is that we simply have more than one function on which I depends. We can then immediately extend (3.19) to

$$\delta I[y(x)] = \int_a^b \left(\frac{\partial F}{\partial y_i(x)} \delta y_i(x) + \frac{\partial F}{\partial y_i'(x)} \frac{d}{dx} (\delta y_i(x)) \right) dx = 0 \quad (3.61)$$

where the index i is repeated and hence summed over. Applying the same trick of integration by parts for every i

$$\begin{aligned}\int_a^b \frac{\partial F}{\partial y_i'(x)}\frac{d}{dx}\left(\delta y_i(x)\right) &= \delta y_i(x)\frac{\partial F}{\partial y_i'(x)}\Big|_a^b \\ &- \int_a^b \delta y_i(x)\frac{d}{dx}\left(\frac{\partial F}{\partial y_i'(x)}\right)\,dx \end{aligned} \tag{3.62}$$

we find again that the first term on the right vanishes because $\delta y_i(a) = \delta y_i(b) = 0$ by construction. Therefore equation (3.61) becomes

$$\delta I[y(x)] = \int_a^b \left(\frac{\partial F}{\partial y_i(x)} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_i'(x)}\right)\right)\delta y_i(x)dx = 0 \tag{3.63}$$

from which we get N copies of the original Euler equation

$$\frac{\partial F}{\partial y_i(x)} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_i'(x)}\right) = 0 \qquad \text{for } i = 1\cdots N\ . \tag{3.64}$$

We are now equipped to handle variational problems involving for than one dependent function.

EXAMPLE 3-4: Geodesics in three dimensions

From equation (3.59), the setup for the problem of finding geodesics in three dimensions, we have $F = \sqrt{1 + y'^2 + z'^2}$ choosing x as the independent variable. Hence, we use (3.64) with $N = 2$ and we get

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'} = 0 \qquad \text{and} \qquad \frac{\partial F}{\partial z} - \frac{d}{dx}\frac{\partial F}{\partial z'} = 0, \tag{3.65}$$

which reduce to

$$\frac{\partial F}{\partial y'} = \frac{y'}{\sqrt{1 + y'^2 + z'^2}} = k_1 \qquad \text{and} \qquad \frac{\partial F}{\partial z'} = \frac{z'}{\sqrt{1 + y'^2 + z'^2}} = k_2 \tag{3.66}$$

where k_1 and k_2 are constants. The equations can be decoupled by taking the sum of the squares of these two equations to show that the denominator of each equation is constant, so that y' and z' must themselves each be constants. Therefore the minimum-length path has constant slope in both the x, y and x, z planes, corresponding to a straight line, as expected. The constants can be determined by requiring the line to pass through the given endpoints.

3.6 Mechanics from a variational principle

Through a series of intriguing examples, in the preceding sections we were able to solve certain problems by extremizing travel time. We may now ask whether there is a *general* formulation of *mechanics* that is based entirely on a variational principle. A variational principle can lead to second-order differential equations, and so does Newton's second law. Perhaps we can cast any classical mechanics problem in the form of a statement about finding the stationary paths of some functional?

Motivated by the examples already explored, a natural starting point is to extremize *travel time.* We start with the case of a free relativistic particle, and we will require the formalism to be Lorentz invariant from the outset. After all, the variational principle — if general and fundamental — should look the same in all inertial frames. This immediately leads us to write the simple candidate functional

$$I = \int d\tau, \tag{3.67}$$

the proper time for a particle to travel between two fixed points in spacetime. We propose that extremizing this quantity leads to the trajectory of a free relativistic particle, equivalently described by

$$\frac{d}{dt}\left(\gamma\, m\mathbf{v}\right) = 0\ . \tag{3.68}$$

from (2.100). Armed with the techniques developed in the previous sections, we can check whether this statement is correct.

We write the functional in terms of the coordinate system of some inertial observer $\mathcal{O}$ using coordinates $(c\,t, x, y, z)$

$$I = \int d\tau = \int \frac{dt}{\gamma} = \int dt\, \sqrt{1 - \frac{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}{c^2}} \tag{3.69}$$

where we used the time dilation relation $dt = \gamma\, d\tau$. We need to determine three functions $x(t)$, $y(t)$, and $z(t)$ that extremize the functional I whose independent variable is t. We can imagine that the endpoints of the trajectory are fixed, and so we have a familiar variational problem. We can then use Euler's equations (3.64) with

$$F = \sqrt{1 - \frac{\dot{x}^2}{c^2} - \frac{\dot{y}^2}{c^2} - \frac{\dot{z}^2}{c^2}}\ . \tag{3.70}$$

and $N = 3$. We have three equations

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\frac{\partial F}{\partial \dot{x}} = 0 \quad , \quad \frac{\partial F}{\partial y} - \frac{d}{dt}\frac{\partial F}{\partial \dot{y}} = 0 \quad , \quad \frac{\partial F}{\partial z} - \frac{d}{dt}\frac{\partial F}{\partial \dot{z}} = 0 \; . \tag{3.71}$$

It is straightforward to show that these lead to

$$\frac{d}{dt}\left(\gamma \dot{x}\right) = 0 \quad , \quad \frac{d}{dt}\left(\gamma \dot{y}\right) = 0 \quad , \quad \frac{d}{dt}\left(\gamma \dot{z}\right) = 0 \; ; \tag{3.72}$$

That is, equation (3.68). This is already very promising: we can describe a free relativistic particle by extremizing the particle's proper time.

Let us next look at the low-velocity regime of our functional. We write (3.69) in an expanded form for $\beta \ll 1$

$$I \simeq \int dt \left(1 - \frac{1}{2}\frac{v^2}{c^2} + \cdots\right) \; . \tag{3.73}$$

The first term is a constant and does not affect a variational principle: Euler's equations involve derivatives of F and hence constant terms in F may safely be dropped. The second term is quadratic in the velocity. We rewrite our functional as

$$I \to \int dt \left(\frac{1}{2} m \, \left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right)\right) \; . \tag{3.74}$$

In addition to dropping the constant shift term, we have also multiplied I from (3.73) by $-m\,c^2$ for convenience. This is a multiplication by a constant and hence, once again, does not affect the Euler equations (3.64). It makes things a little more suggestive, however: we are now extremizing the particle's non-relativistic kinetic energy. If we now use Euler's equations (3.64) with $F = (1/2)mv^2$, we get the familiar three differential equations

$$\frac{d}{dt}\left(m\mathbf{v}\right) = 0 \; , \tag{3.75}$$

as expected. So far so good. We have the expected results for free particles! But how about problems that involve forces?

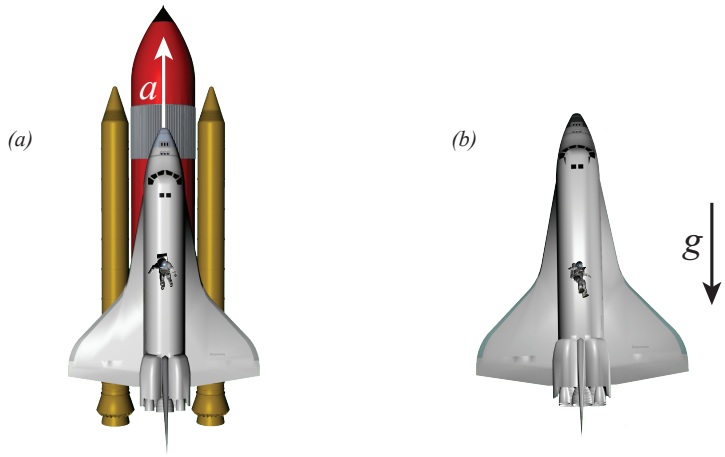

Figure 3.10: Two spaceships, one accelerating in gravity-free space (a), and the other at rest on the ground (b). Neither observers in the accelerating ship nor those in the ship at rest on the ground can find out which ship they are in on the basis of any experiments carried out solely within their ship.

3.7 Motion in a uniform gravitational field

Shortly after developing his special theory of relativity, Einstein saw a beautiful way to understand the effect of uniform gravitational forces, which he called the **principle of equivalence**. He later said that it was "the happiest thought of my life", because it was a wonderfully simple but powerful idea that became a crucial steppingstone to achieving his relativistic theory of gravity: *general relativity*.

The equivalence principle can be illustrated by experiments carried out in two spaceships, one accelerating uniformly in gravity-free empty space and one standing at rest in a uniform gravitational field, as shown in Figure 3.10. The acceleration a of the first ship is numerically equal, but opposite in direction, to the gravitational field g acting on the second ship. The equivalence principle then claims that if observers in either one of the ships carry out any experiment whatever that is confined entirely within their own ship, the results cannot be used to determine which ship they are living in: the two situations are *equivalent*. This is a statement inspired by observation — dating back to Galileo's Pisa tower experiment equating inertial and gravitational masses — which Einstein then elevated to the stature of a principle

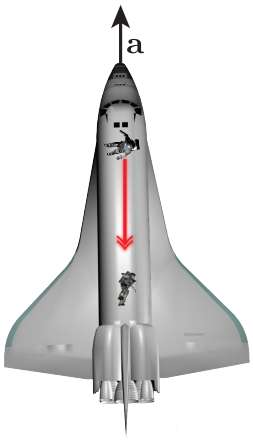

Figure 3.11: A laser beam travels from the bow to the stern of the accelerating ship.

of Nature.

We use the principle here to deduce two related effects of gravity that are not contained in Newton's theory: the gravitational frequency shift and the effect of gravity on the rate of clocks. We start by considering a particular thought experiment with light waves. An observer in the bow of the accelerating ship shines a laser beam at another observer in the stern of the ship, as shown in Figure 3.11. The laser emits monochromatic light of frequency $\nu_{\rm em}$ in the rest frame of the laser. We assume that the distance traveled by the ship while the beam is traveling is very small compared with the length h of the ship, so that the time it takes for the beam to reach the stern is essentially $t = h/c$.

During this time the stern attains a velocity $v = at = ah/c$ with respect to the velocity of the laser when the light was emitted. This velocity is small compared with the speed of light, so the ship suffers no appreciable length contraction[2]. The stern observer is moving toward the source, so will observe a *blue*shift due to the Doppler effect. The nonrelativistic Doppler formula is given by equation (2.113) approximated at small v as

$$\nu_{ob} = \nu_{\rm em}(1 + v/c) = \nu_{\rm em}(1 + ah/c^2) \tag{3.76}$$

[2]The ship's length contraction would scale as v^2/c^2. The physical effect we focus on arises from the Doppler shift, which is linear in v/c.

can then be used to compare the observed frequency with the emitted frequency.

Now according to the equivalence principle, the same result will be observed in the ship at rest in a uniform gravitational field, if we substitute the acceleration of gravity g for the rocket acceleration a. That is, if the observer at the top of the stationary ship shines light with emitted frequency $\nu_{\rm em}$ toward the observer at the bottom, the bottom observer will see a blueshifted frequency

$$\nu_{ob} = \nu_{\rm em}(1 + gy/c^2), \tag{3.77}$$

where now we have used the symbol y for the altitude of the top clock above the bottom clock. It is also true that the top observer will see a redshift if he or she looks at a light beam sent off by the bottom observer. In neither case can we blame the shift on Doppler, however, because neither observer is moving. Instead, the shift in this case is due to a difference in altitude of the two clocks at rest in a uniform gravitational field.

How can we *explain* the blueshift seen by the person at the bottom of the stationary ship? If we think of the laser atoms that radiate light at the top as clocks whose rate is indicated by the frequency of their emitted light, the observer at the bottom will be forced to conclude that these top clocks are running *fast* compared to similar clocks at the bottom of the ship! For suppose a clock at the top of the ship has a luminous second-hand that emits light of frequency $\nu_{\rm em}$. In one second, the hand emits $\nu_{\rm em}$ wavelengths of light. The observer at the bottom must collect all these wavelengths, since none of them is created or destroyed in transmission. However, the frequency of the waves observed at the bottom is increased by the factor $(1 + gy/c^2)$, which means that the observer at the bottom will collect all of these waves in less that one second according to his or her own clock. That is, the second-hand of the clock at the top appears to advance by one second in less than one second to the observer at the bottom, by the exact same factor. The observer at the top agrees with this judgment. The top observer sees a redshift when looking at clocks at the bottom, so it is natural for a person at the top to believe that bottom clocks run slower than top clocks.

If atomic clocks at high altitude run faster, it is of course true that *all* clocks up high run faster, because they can be continuously compared with one another. And if all stationary clocks at high altitude run fast compared

with all stationary clocks at lower altitude, we can conclude that time itself runs fast at higher altitude: That is, for time intervals Δt,

$$\Delta t_{\text{high}} = \Delta t_{\text{low}}(1 + gy/c^2). \tag{3.78}$$

This is the time difference for two clocks at rest in a uniform gravitational field. Now suppose the lower clock remains at rest, reading time t, but the upper clock is allowed to move with a speed v that can change with time. Then in an infinitesimal time dt according to the lower clock, the upper clock advances by time

$$d\tau = dt(1 + gy/c^2)\sqrt{1 - v^2/c^2}, \tag{3.79}$$

with factors showing that it runs fast due to its altitude and slow due to its speed. For a nonrelativistic particle moving near Earth's surface, both gh/c^2 and v^2/c^2 are very small, so

$$d\tau \cong dt(1 + gy/c^2)(1 - v^2/2c^2) \cong dt(1 + gy/c^2 - v^2/2c^2) \tag{3.80}$$

using the binomial expansion to obtain the first expression and neglecting the product of two very small quantities to obtain the second expression. Therefore as the lower clock advances from some time t_a to a later time t_b, with these approximations the upper clock advances by time

$$\tau = \int_{t_a}^{t_b} dt(1 + gy/c^2 - v^2/2c^2). \tag{3.81}$$

Notice that if m is the mass of the upper clock, this can be written in the form

$$\begin{aligned} \tau &= t_b - t_a - \frac{1}{mc^2}\int_{t_a}^{t_b}\left(\frac{1}{2}mv^2 - mgy\right)\, dt \\ &= (t_b - t_a) - \frac{1}{mc^2}\int_{t_a}^{t_b} dt\ (T - U) \end{aligned} \tag{3.82}$$

where $T = (1/2)mv^2$ and $U = mgy$ are the kinetic and potential energies of the upper clock, if the lower clock is at rest and has zero potential.

The value of τ depends not only upon the initial and final times t_a and t_b, but also upon the path the clock takes in getting from the beginning

point to the end point. So looking at the problem of two clocks in a uniform gravitational field, where the lower clock is at rest and the upper clock has altitude $y(t)$ and moves with speed $v(t)$, we have shown that the proper time interval read by the upper clock's rest frame as it moves between two given points, while the lower clock advances from time t_a to time t_b, is

$$\tau = (t_b - t_a) - \frac{1}{mc^2}\int_{t_a}^{t_b} dt\ (T - U) \tag{3.83}$$

where the integrand is now the *difference* between the kinetic and potential energies of the upper clock.

Let us now find that particular path of the upper clock which extremizes the time τ as it travels between two given points in space, starting at fixed time t_a and ending at time t_b according to the lower clock. Extremizing τ in this problem is the same as *minimizing* the functional

$$I \equiv \int_{t_a}^{t_b} \left(\frac{1}{2}mv^2 - mgy\right)\ dt, \tag{3.84}$$

with the integrand

$$F = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - mgy \tag{3.85}$$

since the $t_b - t_a$ term is a constant. Euler's equations for the x, y, and z directions then give

$$m\ddot{x} = 0, \quad m\ddot{y} = -mgy, \quad \text{and} \quad m\ddot{z} = 0, \tag{3.86}$$

which are Newton's laws of motion for a particle in a uniform gravitational field! Our goal of identifying a variational principle for the motion of a particle in a uniform gravitational field has been successful. Furthermore, the form of the functional, as given in equation (3.84), is highly suggestive, a fact we will exploit in Chapter 4.

3.8 Summary

In this chapter we have shown that a variational principle — Fermat's principle of stationary time — can be used to find the paths of light rays. Such a variational principle seems totally unlike the approach of Newton to finding the paths of particles subject to forces. Yet we have shown that the

associated **calculus of variations** of **functional calculus** allows us to convert the problem of making stationary a certain integral into a differential equation of motion. We applied these techniques to solve several interesting problems.

We then went on to show that the relativistic and nonrelativistic mechanics of a free particle can be understood from a variational principle, and extended that approach, using Einstein's principle of equivalence, to find the motion of nonrelativistic particles in uniform gravitational fields. The functional

$$I \equiv \int_{t_a}^{t_b} \left(\frac{1}{2}mv^2 - mgy\right)\ dt = \int_{t_a}^{t_b} (T-U)\ dt, \tag{3.87}$$

where the integrand is the difference between the kinetic and gravitational potential energies of the particle, gives the correct differential equations of motion for a nonrelativistic particle.

Can we do something similar for *any* mechanics problem? One involving normal and tension forces? Or frictional forces? How about non-conservative forces in general, which do not have potentials? In short, can we always find the equations of motion of a particle through this program of extremizing an associated functional? These are questions for Chapter 4.

3.9 Exercises and Problems

PROBLEM 3-1 : Describe the geodesics on a right circular cylinder. That is, given two arbitrary points on the surface of a cylinder, what is the shape of the path of minimum length between them, where the path is confined to the surface? *Hint*: A cylinder can be made by rolling up a sheet of paper.

PROBLEM 3-2 : A particle falls along a cycloidal path from the origin to the final point $(x, y) = (\pi a/2, a)$; the time required is $\pi\sqrt{a/2g}$, as shown in Section 3.4. How long would it take the particle to slide along a straight-line path between the same points? Express the time for the straight-line path in the form $t_{\text{straight}} = kt_{\text{cycloid}}$, and find the numerical factor k.

PROBLEM 3-3 : A unique transport system is built between two stations 1 km apart on the surface of the Moon. A tunnel in the shape of a full cycloid cycle is dug, and the tunnel is lined with a frictionless material. If mail is dropped into the tube at one station, how much later (in seconds) does it appear at the other station? How deep is the lowest point of the tunnel? (Gravity on the Moon is about 1/6th that on Earth.)

PROBLEM 3-4 : A hollow glass tube is bent into the form of a slightly tilted rectangle, as shown below. Two small ball bearings can be introduced into the tubes at one corner; one rolls clockwise and the other counterclockwise down to the opposite corner at the bottom. The balls are started out simultaneously from rest, and note that each ball must roll the same distance to reach the destination. The question is: which ball reaches the lower corner first, or do they arrive simultaneously? Why?

PROBLEM 3-5 : Prove from Fermat's Principle that the angles of incidence and reflection are equal for light bouncing off a mirror. Use neither algebra nor calculus in your proof! (*Hint*: The result was proven by Hero of Alexandria 2000 years ago.)

PROBLEM 3-6 : An ideal converging lens focusses light from a point object onto a point image. Consider only rays that are straight lines except when crossing an air-glass boundary, such as those shown below. Relative to the ray that passes straight through the center of the lens, do the other rays require more time, less time, or the same time to go from O to I? That is, in terms of Fermat's Principle, is the central path a local minimum, maximum, or a stationary path that is neither

a minimum nor a maximum?

PROBLEM 3-7 : Light focusses onto a point I from a point O after reflecting off a surface that completely surrounds the two points, as shown in cross section below. The shape of the surface is such that all rays leaving O (excepting the single ray which returns to O) reflect to I. (a) What is the shape of the surface? (b) Pick any one of the paths. Is it a path of minimum time, maximum time, or is it stationary but of neither minimum nor maximum time for all nearby paths?

PROBLEM 3-8 : Consider the ray shown bouncing off the bottom of the surface in the preceding problem. Replace the surface at this point by the more highly-curved surface shown below in dotted lines. The ray still bounces from O to I. Is the ray now a path of minimum time, maximum time, or is it stationary but of neither minimum nor maximum time? Compare with nearby paths that bounce once but are otherwise straight. Suppose the paths must bounce once but need not be segments of straight lines. What then?

PROBLEM 3-9 : When bouncing off a flat mirror, a light ray travels by a minimum time path. (a) For what shape mirror would the paths of all bouncing light-rays take equal times? (b) Is there a shape for which a bouncing ray would take a path of greatest time, relative to nearby paths?

PROBLEM 3-10 : A hypothetical object called a straight **cosmic string** (which may have been formed in the early universe and may persist today) makes the r, θ space around it conical. That is, set an infinite straight cosmic string along the z axis; the two-dimensional space perpendicular to this, measured by the polar coordinates r and θ, then has the geometry of a cone rather than a plane. Suppose there is a cosmic string between Earth and a distant quasi-stellar object. What might we see when we look at this QSO? [Assume light travels in least-time paths (here also least-distance paths) relative to nearby paths.]

PROBLEM 3-11 : There are definitely galaxies between ourselves and distant quasi-stellar objects. The gravity of the galaxies affects the geometry of spacetime; the effect on light rays is as though a lens of a particular shape were placed between ourselves and the QSO. (See the diagram. The effect is called gravitational lensing and has been observed.) What might the distant QSO look like?

PROBLEM 3-12 : Model Earth's atmosphere as a spherical shell 100 mi thick, with index of refraction $n = 1.00000$ at the top and $n = 1.00027$ at the bottom. Is a light rays final angle φ_f relative to the normal at the ground greater or less

than its initial angle φ_i relative to the normal at the top of the atmosphere? (Take Earth to have radius $R = 4000$ mi.)

PROBLEM 3-13 : We seek to find the path $y(x)$ that minimizes the integral $I = \int f(x, y, y')dx$. Find Euler's equation for $y(x)$ for each of the following integrands f, and then find the solutions $y(x)$ of each of the resulting differential equations if the two endpoints are $(x, y) = (0, 1)$ and $(0, 3)$ in each case. (a) $f = ax + by + cy'^2$ (b) $f = ax^2 + by^2 + cy'^2$ (c) $f = x^2y'^2$.

PROBLEM 3-14 : Find a differential equation obeyed by geodesics in a plane using polar coordinates r, θ. Integrate the equation and show that the solutions are straight lines.

PROBLEM 3-15 : Find two differential equations obeyed by geodesics in three-dimensional Euclidean space, using spherical coordinates r, θ, φ.

PROBLEM 3-16 : Two-dimensional surfaces that can be made by rolling up a sheet of paper are called *developable* surfaces. Find the geodesic equations on the following developable surfaces and solve the equations. (a) A circular cylinder of radius R, using coordinates θ and z. (b) A circular cone of half-angle α (which is the angle between the cone and the axis of symmetry) using coordinates θ and l, where l is the distance of a point on the cone from the apex. *Hint*: Find the distance ds between nearby points on the surface in terms of $l, \alpha, d\theta$, and dl.

PROBLEM 3-17 : Find the geodesic equations on the torus shown below, using the coordinates θ, φ. Show that the circles running around the outer edge with $\varphi = \pi/2$, circles running around the inner edge with $\varphi = -\pi/2$, and circles running around the torus at fixed θ, are all geodesics. Show that a circle running around the torus with fixed $\varphi = 0$ is *not* a geodesic.

PROBLEM 3-18 : Using Euler's equation for $y(x)$, prove that

$$\frac{\partial f}{\partial x} - \frac{d}{dx}(f - y'\frac{\partial f}{\partial y'}) = 0. \tag{3.88}$$

This equation provides an alternative method for solving problems in which the integrand f is not an explicit function of x, because in that case the quantity

$f - y'\partial f/\partial y'$ is constant, which is only a first-order differential equation.

PROBLEM 3-19 : A line and two points not on the line are drawn in a plane, as shown below. A smooth curve is drawn between the two points and then rotated about the given line, also as shown. Find the shape of the curve that minimizes the area generated by the rotated curve. A lampshade manufacturer might use this result to minimize the material used to produce a lampshade of given upper and lower radii.

PROBLEM 3-20 : The time required for a particle to slide from the cusp of a cycloid to the bottom was shown in Section 3.4 to be $t = \pi\sqrt{a/2g}$. Show that if the particle starts from rest at any point *other* than the bottom, it will take this same length of time to reach the bottom. The cycloid is therefore also the solution of the *tautochrone*, or equal-time problem. *Hint*: The energy equation for the particle speed in terms of y written in Section 3.4 must be modified to take into account the new starting condition. [The tautochrone result was known to the author Herman Melville. In the chapter called "The Try-Works" in *Moby-Dick*, the narrator Ishmael, on board the whaling ship Pequod, describes the great try-pots used for boiling whale blubber: "Sometimes they are polished with soapstone and sand, till they shine within like silver punchbowls. ... It was in the lefthand try-pot of the Pequod, with the soapstone diligently circling around me, that I was first indirectly struck by the remarkable fact, that in geometry all bodies gliding along the cycloid, my soapstone for example, will descend from any point in precisely the same time."]

PROBLEM 3-21 : Derive Snell's law from Fermat's Principle.

PROBLEM 3-22 : A lifeguard stands on the beach a distance l_1 from the shoreline. A swimmer calls for help, a distance l_2 directly out from the shoreline and a lateral distance h from the lifeguard. If the lifeguard can run twice as fast as she can swim, at what angle θ should she run relative to the shoreline in order to reach the swimmer as soon as possible?

PROBLEM 3-23 : Assume Earth's atmosphere is essentially flat, with index of refraction $n = 1$ at the top and $n = n(y)$ below, with y measured from the top, and the positive y direction downward. Suppose also that $n^2(y) = 1 + \alpha y$, where

α is a constant. Find the light-ray trajectory $x(y)$ in this case.

PROBLEM 3-24 : Suppose the Earth's atmosphere is as described in the preceding problem, except that $n^2(y) = 1+\alpha y+\beta y^2$, where α and β are constants. Find the light-ray trajectory $x(y)$ in this case.

PROBLEM 3-25 : Consider Earth's atmosphere to be spherically symmetric above the surface, with index of refraction $n = n(r)$, where r is measured from the center of the Earth. Using polar coordinates r, θ to describe the trajectory of a light ray entering the atmosphere from high altitudes, (a) find a first-order differential equation in the variables r and θ that governs the ray trajectory; (b) show that $n(r)r\sin\varphi =$ constant along the ray, where φ is the angle between the ray and a radial line extending outward from the center of the Earth. This is the analog of the equation $n(y)\sin\theta =$ constant for a flat atmosphere.

PROBLEM 3-26 : Using the result found in part (b) of the preceding problem, and supposing that $n^2(r) = 1 + \alpha/r^2$ (where α is a constant), find the light-ray trajectory $r(\theta)$.

PROBLEM 3-27 : According to Einstein's general theory of relativity, light rays are deflected as they pass by a massive object like the Sun. The trajectory of a ray influenced by a central, spherically symmetric object of mass M lies in a plane with coordinates r and θ (so-called *Schwarzschild coordinates*); the trajectory must be a solution of the differential equation

$$\frac{d^2u}{d\theta^2} + u = \frac{3GM}{c^2}u^2 \tag{3.89}$$

where $u = 1/r$, G is Newtons gravitational constant, and c is the constant speed of light. (a) The right-hand side of this equation is ordinarily small. In fact, the ratio of the right-hand side to the second term on the left is $3GM/rc^2$. Find the numerical value of this ratio at the surface of the Sun. The Sun's mass is 2.0×10^{30} kg and its radius is 7×10^5 km. (b) If the right-hand side of the equation is neglected, show that the trajectory is a straight line. (c) The effects of the term on the right-hand side have been observed. It is known that light bends slightly as it passes by the Sun and that the observed deflection agrees with the value calculated from the equation. Near a black hole, which may have a mass comparable to that of the Sun but a much smaller radius, the right-hand side becomes very important, and there can be large deflections. In fact, show that there is a single radius at which the trajectory of light is a circle orbiting the black hole, and find the radius r of this circle. (d) Suppose we wish to make a spherical

piece of glass with a varying index of refraction $n(r)$, such that trajectories of light rays within it will be exactly the same as the trajectories of light around a black hole. Find the index $n(r)$ required to do this.

PROBLEM 3-28 : A clock is thrown straight upward on a planet where gravity $g = 10.0$ m/s^2, and it returns to the surface exactly 20 seconds after it was thrown, according to clocks at rest on the ground. (a) Using the clock's motion as derived in Section 3.7, how much less than 20 seconds will have elapsed according to this moving clock? (b) Now suppose that instead of the freely-falling motion used in part (a), the moving clock has constant speed v_0 straight up for exactly 10 seconds according to ground clocks, and then moves straight down again at the same constant speed v_0 for another 10 seconds, according to ground clocks. How much less than 20 seconds will have elapsed according to this moving clock? (c) Show that the freely-falling clock returns reading a lower total elapsed time than the clock that moves at constant speed up and down.

PROBLEM 3-29 : As we will show in Chapter ?, in nonrelativistic mechanics the *shape* of a particle's path between points a and b can be found by making stationary the so-called **Jacobi Action**

$$J = \int_a^b \sqrt{E - U}\; ds, \tag{3.90}$$

where E and U are the particle's total energy and potential energy, respectively, and ds is the infinitesimal path length. Using this principle, find the shape $y(x)$ of the path of a particle that moves in a uniform gravitational potential $U = mgy$. *Hint*: Begin by showing that the integrand can be written $f = \sqrt{E - mgy}\,\sqrt{1 + x'^2}$.

PROBLEM 3-30 : An object of mass m can move in two dimensions in response to the simple harmonic oscillator potential $U = (1/2)kr^2$, where k is the force constant and r is the distance from the origin. Using the Jacobi action introduced in the preceding problem, find the shape of the orbits using polar coordinates r and θ; that is, find $r(\theta)$ for the orbit. Show that the shapes are ellipses and circles centered at the origin $r = 0$.

PROBLEM 3-31 : A comet of mass m moves in two dimensions in response to the central gravitational potential $U = -k/r$, where k is a constant and r is the distance from the Sun. Using the Jacobi action introduced two problems earlier, and using polar coordinates (r, θ), find the possible shapes of the comet's orbit. Show that these are (a) a parabola, if the energy of the comet is $E = 0$; (b) a

hyperbola if $E > 0$; (c) an ellipse or a circle if $E < 0$, where in each case $r = 0$ at one of the foci.