First of all, one observes that the reference control $u_*(\cdot)$ gives rise to a *flow*, i.e., a family of maps $\Phi_{t',t}$, defined by letting $\Phi_{t',t}(q) = \xi_{q,t}(t')$, where we use $\xi_{q,t}(\cdot)$ to denote the solution of $\dot{\xi}(\tau) = f(\xi(\tau), u(\tau), \tau)$, $\xi(t) = q$. These flow maps are of class C^1, and their differentials $D\Phi_{t',t}(q)$ can be computed by solving the variational equation $\dot{y}(\tau) = \frac{\partial f}{\partial q}(\xi_{q,t}(\tau), u(\tau), \tau) \cdot y(\tau)$. The value $D\Phi_{t',t}(q).v$ is then $y_{q,t,v}(t')$, where $y_{q,t,v}$ is the solution y of the variational equation for which $y(t) = v$.

One then makes "needle variations," in which the reference control u_* is replaced, over a finite union $[t_1, t_1+s_1] \cup \ldots [t_\nu, t_\nu+s_\nu]$ of intervals, by a control u with constant values u_i on $[t_i, t_i+s_i]$. This gives rise, for a fixed $\mathbf{t}=(t_1, \ldots, t_\nu)$ and $\mathbf{u}=(u_1, \ldots, u_\nu)$, to a ν-parameter family $\{u_{\mathbf{s}}\}_{\mathbf{s}\in\mathbb{R}^\nu_+(\sigma)}$ of controls, for some $\sigma > 0$. (Here $\mathbb{R}^\nu_+$ is the nonnegative orthant in $\mathbb{R}^\nu$, and $\mathbb{R}^\nu_+(\sigma)=\{s\in\mathbb{R}^\nu_+ : s_1+\ldots+s_\nu \leq \sigma\}$.) The $u_{\mathbf{s}}$ in turn determine a ν-parameter family $\{q_{\mathbf{s}}\}_{\mathbf{s}\in\mathbb{R}^\nu_+(\sigma)}$ of trajectories—keeping the initial condition $q_{\mathbf{s}}(a) = \bar{q}$ fixed—and then a map $\mathbf{s}\mapsto q_{\mathbf{s}}(b)=\gamma(\mathbf{s}) \in Q$.

One then computes the differentials of these maps at $\mathbf{s} = 0$, and the differential of the corresponding cost. The computation is done by means of the chain rule. For example,

$$\frac{\partial\gamma}{\partial s_i}(0) = \frac{d}{ds}\bigg|_{s=0} \Big(\Phi_{b,t_i}\Phi_{t_i,t_i+s}\Phi^{u_i}_{t_i+s,t_i}\Phi_{t_i,a}(\bar{q})\Big)\,,$$

where $\Phi^u_{t',t}$ are the flow maps corresponding to the constant control u. Since

$$\Phi_{t_i,t_i+s}\Phi^{u_i}_{t_i+s,t_i}(q) = q + s\Big(f(q,u_i,t_i) - f(q,u_*(t_i),t_i)\Big) + o(s)\,, \qquad (8.11)$$

we see that $\frac{\partial\gamma}{\partial s_i}(0)$ is obtained by solving the variational equation with initial condition $y(t_i) = f(q,u_i,t_i) - f(q,u_*(t_i),t_i)$, and evaluating the solution at time b. The justification of (8.11) when $f(q,u_i,t)$ and $f(q,u(t),t)$ are jointly continuous in q and t is trivial. In the general case, it is a much more a delicate point, and the conclusion is not valid for all possible choices of the t_i, but only for some special t_i's. However, these t_i's form a set of full measure, so the "almost everywhere" conclusion of the theorem is still valid.

With a lot of work, these basic facts lead to the versions of (MP) described above, which are essentially those of [98] and [10].

Chapter 9

Recent versions of the maximum principle: the nonsmooth connection

Stronger versions have been obtained by weakening the hypothesis of the first version, or strengthening the conclusions, or both.

9.1 The nonsmooth analysis approach

One important improvement of the classical version resulted from the use of *nonsmooth analysis* (cf. Clarke [37, 38]). Clarke replaced the continuous differentiability of f and L with respect to q by the weaker requirement that f and L be Lipschitz with respect to q. (Other than that, the hypotheses are exactly as above. Naturally, in all the integral bounds the term $||\frac{\partial f^i}{\partial q}(q, u(t), t)||$ has to be replaced by a Lipschitz constant for $f^i(\cdot, u(t), t)$ in the appropriate set K.)

This of course renders the meaning of the adjoint equation problematic, since

$$\frac{\partial H}{\partial q} = \langle p, \frac{\partial f}{\partial q} \rangle - \frac{\partial L}{\partial q}\,, \tag{9.1}$$

which in principle does not make sense unless f and L are differentiable with respect to q. Clarke, however, introduced the "generalized Jacobian" $\partial F(q)$ of a Lipschitz map F at a point q, which is a *set* of linear maps rather than a single linear map. Using this object, one can still assign a meaning to the adjoint equation in the Lipschitz case, except that one now obtains

an *adjoint differential inclusion* rather than an equation.

More recently, S. Lojasiewicz Jr., in work still unpublished, observed that in the construction of needle variations one never uses the derivatives with respect to q of the functions $f^i(q, u_j, t)$, and that the only place where these derivatives appear is in the adjoint equation, which only involves the reference control u_*. This enabled him to extend the maximum principle even further, by dispensing with the hypothesis that $f(q, u, t)$ is Lipschitz in q for all u, t, and requiring only that $f(q, u_*(t), t)$ be Lipschitz in q for each t—with a Lipschitz constant $C(t)$ integrable with respect to t—while $f(q, u, t)$ for general u, t is only required to be continuous with respect to q.

In addition, a second direction developed in which (MP) is formulated not for controlled differential equations $\dot{q} = f(q, u, t)$, but for *differential inclusions* $\dot{q} \in F(q, t)$, where F is a set-valued map (cf. for example [38]).

9.2 High-order necessary conditions

While the "nonsmooth" generalizations were being developed, other authors pursued a different direction, for very smooth systems. They observed that one could get stronger results by allowing a class of variations richer than that used in the classical proof. One can then obtain "high-order necessary conditions for optimality," in which the inequality $\langle p(t), v\rangle \geq 0$ must hold not only for v of the form $f(q_*(t), u_*(t), t) - f(q_*(t), u, t)$, as in (MP), but also for other vectors resulting from more elaborate variations involving Lie brackets. (We are assuming here for simplicity that $L \equiv 1$, i.e., that we are dealing with minimum-time control.) This is a vast subject, and we refer the reader to [8, 54, 57, 73] for results and surveys of work done in the 1960s, and to [1, 2, 3, 20, 75, 77] for later, more Lie-bracket oriented, work.

9.3 The "combined systems" problem

The results discussed above are proved by different methods and cannot be combined into a single theorem.

At the risk of belaboring the obvious, we point out that ***extremality is a property which is not local with respect to time***. That is, if the domain $I = [a, b]$ of definition of a trajectory-control pair $\gamma_* = (q_*, u_*)$ can be covered by intervals I_j such that the restriction $\gamma_{*,j}$ of γ_* to each I_j is an extremal, it does not follow that γ_* itself is an extremal. The reason is, simply, that if there exist minimizing adjoint vectors p_j for each $\gamma_{*,j}$, then

these p_j need not match, so it may be impossible to use them to construct a minimizing adjoint vector for γ_*.

Remark 9.3.1 The classical calculus of variations is in this respect exceptional and misleading, because its Hamiltonian is $\langle p, u\rangle - L(q, u, t)$, so (MC) implies that $p(t) = \frac{\partial L}{\partial u}(q_*(t), u_*(t), t)$, showing that p is unique, and then the various p_j corresponding to the I_j have to match. So in this case extremality is a local property. $\diamondsuit$

Because being an extremal is not a local property, the following curious situation can arise. Suppose we are dealing with a system Σ (i.e., a dynamical law f and a Lagrangian L) which is "hybrid," in the sense that Σ satisfies the hypotheses of the classical version of (MP) in some region Q_1, while in some other region Q_2 it is very smooth, so that one can make high-order variations involving Lie brackets, and in some other region Q_3 it satisfies the conditions of Clarke's nonsmooth MP, while, finally, there is a region Q_4 where Σ is given by a differential inclusion. Suppose q_* is an optimal trajectory such that q_* visits all four regions. Then we can apply each of the four theorems to each part of q_* that stays in one region, and conclude that any such part is an extremal, but *we cannot combine them all to get one minimizing adjoint vector p_* on the whole interval of definition of q_* that satisfies in each part the corresponding conclusion.*

We will not even attempt to explain why this is so, because to do it we would have to discuss in detail the proofs of these theorems, showing that in each case one uses a different construction, and these constructions cannot be combined into a single one valid on the whole interval. But it is a fact that, due to this incompatibility of the various proofs, a single theorem covering all cases and combining them—that is, applying to "hybrid" problems as above—appeared until a few years ago to be beyond reach.

COMMENT NO. 29. JAN: the following section was written before I found the Generalized Differential Quotients and the Path-Integral Generalized Differentials. For a skecth of where I am now, see my French paper. (Item No. 58 of the "papers avaliable in the Web" link of my Web page.

9.4 A unified nonsmooth approach

Recently one of us (Sussmann [114, 115, 119, 120, 123, 124, 125]) has obtained a general version of (MP) that contains all the above results, applies

to some new cases as well, and actually covers the "hybrid" case.

The key point of this work is that the main ideas of the classical proof of [98] apply under much more general conditions, provided only that a suitable theory of "generalized differentials" is used. This theory need not be—and, in fact, *cannot be*—that of classical differentials or that of Clarke's generalized Jacobian. What is truly needed is a definition of "generalized differentials" for which there is a good calculus (and in particular a chain rule) and a nice open mapping theorem.[1]

In [114, 115, 119, 120, 123, 124, 125] it is proposed to use the theories of "semidifferentials," "multidifferentials," and "generalized differential quotients. Basically, a continuous map f is "semidifferentiable" at a point x if it can be approximated to first order near x by a Lipschitz map. (For example, the function $f(x) = |x| + |x|^{3/2} \sin \frac{1}{x}$ is semidifferentiable at 0, because $f(x) = |x| + o(|x|)$. This function does not, however, have a Clarke generalized Jacobian at 0, because it is not Lipschitz near 0, and it is clear that it does not have a classical differential at 0 either.) This theory can also be extended to define "semidifferentiability" of certain set-valued maps, needed to handle "needle variations" when $f(q, u, t)$ is only continuous in q, so that the flow maps $\Phi^u_{t',t}$ are set-valued. It turns out that, using semidifferentiable maps, one can prove a general version of (MP) that contains and combines the classical, nonsmooth and Lojasiewicz versions. In [120, 123, 124, 125, 126] we propose two even more general differentiability theories—the "multidifferentials" and the "generalized differential quotients"—which make it possible to extend (MP) farther into a version that also covers the differential inclusions case.

Remark 9.4.1 The conventional wisdom of the subject holds that differential inclusions are more general than controlled dynamical systems, because a set-valued function usually does not have continuous selections. While this is true, A. Bressan has shown (cf. [18, 19, 21, 22, 24, 25]) that lower semicon-

[1] The classical open mapping theorem says that if a continuous map F from an open subset U of a finite-dimensional real linear space X to another such space Y is such that F is differentiable at a point q and the differential $DF(q)$ is a surjective linear map, then F is *open at* q, in the sense that F maps neighborhoods of q to neighborhoods of $F(q)$. What we need for a proof of the maximum principle under general conditions is a generalization of this to a large enough class of maps, including many maps that are not necessarily differentiable in the ordinary sense, and not Lipschitz or even single-valued (since, for example, flow maps generated by continuous vector fields should be in the class, and these flow maps are in general set-valued). The main point of [114, 115, 119, 120, 123, 124, 125] is to single out a class of maps which is large enough to contain all the maps that are needed for the proof of the maximum principle, and at the same time such that a concept of "generalized differential," with good properties—in particular, a chain rule and an open mapping theorem—exists for that class.

tinuous set-valued maps admit "conically continuous selections," and that these selections are as good as continuous ones as far as the properties of the corresponding flows are concerned. Adapting Bressan's theory, it is shown in [119, 120, 123] that the general differential inclusion case can be reduced to the case of a controlled dynamical system, and one ends up with a result that covers all cases. ♢

Chapter 10

The geometric connection

Both geometry and control are at least as hold as humankind and, as we shall show in this section, these two areas have been connected from their very early days by the thread of optimal control, a large component of which is curve minimization.

We have already explained what control theory is about, and at this point we may just retain from our earlier discussion the simple statement that ***the essence of control problems is that they are problems of design, in which it is desired to produce an object or a behavior so as to satisfy certain specifications***.

A similar question could, of course, be asked of "geometry." Exactly what is geometry about? Acording to E. T. Bell, in [7], p. 301,

> Attempts to obtain from professional geometers a statement of what geometry is are likely to be only nebulously successful. Beyond agreeing, more or less, on "geometry is the product of a particular way of thinking," few geometers will commit themselves to anything less hazy.

If, instead of asking geometers, we just take the simple-minded approach of looking up the word in a dictionary, we will find definitions such as

> [Geometry is] the branch of mathematics concerned with the properties and relations of points, lines, surfaces, and solids in space.

Actually, this dictionary definition is not too bad as a first approximation to what mathematicians might want to say. They would, however, supplement it in a few ways:

- ***first***, by making it clear that "lines" refers not only to "straight lines" but to more general curves (exactly as "surfaces" does not mean "planes");
- ***second***, by incorporating into the definition the revolutionary 19th century discovery that in mathematics there is no such thing as "space" but, rather, there is a multiplicity of "spaces," of which the space of classical Euclidean geometry is one but not the only one;
- ***third***, by taking into account the modern understanding that the concept of a truly general "curve" or "surface" or "solid" is much too broad, and what geometry really studies is some special classes of "curves, surfaces and solids,"
- ***fourth***, by getting rid of the limitation to three dimensions implicit in the reference to "curves, surfaces and solids," and the statement that these objects live in "space."

One would then end up with the following modified version of the standard dictionary definition:

> [Geometry is] the branch of mathematics concerned with the properties and relations of points and various special curves, surfaces, solids and other special higher-dimensional configurations and the study of the spaces where these configurations live.

The new definition is still far from transparent but, despite its evident shortcomings, it certainly represents an improvement over the previous attempt.

The main problem with the new definition, which is bound to be a problem with ***any*** definition of geometry, and is the reason for Bell's "nebulous success," is the fact that nobody has ever been able to say exactly what it is that distinguishes the special configurations studied in geometry from completely general configurations such as arbitrary subsets of "space." All that can be said is that the true feature that singles out these configurations and makes them special is that they are ... geometric. So "geometry" is, ultimately, defined in a tautological way as the study of "geometric objects."

Even though it is far from clear what it is exactly that makes an object "geometric," mathematicians feel and have always felt that they can tell one when they see it. For example, if you were to define an ellipse to be a curve that, relative to some system of Cartesian coordinates x, y in the plane has an equation of the form $ax^2 + by^2 = 1$, with $a < 0$ and $b > 0$, then this would not be regarded[1] as "geometric." If, on the other hand, you take the very same notion, and characterize it instead as "a curve which is the locus of all the points for which the sum of the distances to two fixed given points is constant," then the "geometricity" of the ellipse becomes apparent.

Although the extension of the term "geometric" appears to have been fairly clear at each particular time, it has varied with time. Since ancient times, nobody ever doubted that straight lines and circles were "geometric curves." So were ellipses, hyperbolas and parabolas, since they can be characterized "geometrically" as "loci" of points satisfying certain "geometric" conditions, as we indicated above for the case of the ellipse.

At the other extreme, the graph of a nonmeasurable function, or of a continuous nowhere differentiable function, are objects that can definitely be talked about in 20th century mathematics, but they are clearly not "geometric curves" by most standards, although occasionally some of these objects can acquire a "geometric" status if they become amenable to a particularly simple description, as may be the case nowadays with some fractal sets.

The "geometricity" of objects lying between these two extremes has been somewhat unclear, and has certainly changed through the ages. For example, Descartes only called a curve "geometric" if it was algebraic[2], thereby denying admission to the world of geometry to transcendental loci such as sinusoids, cycloids, tractrices, graphs of exponentials, clotoids, logarithmic spirals, and other natural curves. This was regrettable, because some of these curves admit rather simple "geometric" descriptions. (For example, the exponential curve C is characterized by the property that there is a straight line L for which, if P varies along C, and Q, R are the points in L for which the line PQ is tangent to C at P and the line PR is perpendicular to L, then the distance from R to Q is constant.)

With the advent of the calculus, a language became available in terms of which these curves are simple and natural. (For example, the exponential function is the solution of the differential equation $y' = y$.) So the notion of

[1] or, at least, would not have been regarded before the introduction of analytic geometry by Descartes

[2] that is, defined by a system of polynomial equations; for example, straight lines and conic sections are "geometric" in this sense, but an exponential curve is not, because $y = e^x$ is not a polynomial equation.

geometricity shifted, and the geometric world was enlarged by the admission of many new curves.

From all this we have to conclude that "geometricity" is in the eyes of the beholder, and the notion of a "geometric object" is undergoing perpetual change. What makes an object "geometric" is that ***it admits a simple characterization***, with the understanding that "simplicity" here is a time-varying concept, depending on the mathematical language available at a particular moment in history.

Now, one of the most natural ways of singling out objects in a "simple" way has been, at least since the classical era, to require that they solve some "simple" minimization or maximization problem[3]. On the other hand, an important and very common type of specification for a design problem is that something be minimized or maximized. In many design problems, what is desired is to find a curve that achieves a minimum or a maximum of some functional—that is, to solve an ***optimal control problem***—and the solution frequently turns out to be a very special curve, which becomes "geometric" by virtue of this special nature. In this way, geometry emerges from the need to solve optimal control problems.

10.1 Birds, bees, and the birth of geometry and optimal control

Homo sapiens is not the first entity that has engaged in problem-solving and arrived at "geometric" curves and surfaces as solutions of design problems. The animal kingdom already exhibits many examples of optimal control problems and their geometric solutions[4]. For example, J. L. Coolidge, in his 1940 book [43] (page 1), asked

> What is the earliest example of puposive geometrical construction?

and argued that

[3]The key word here, of course, is "simple." A 20th Century mathematician would pedantically observe that *every* curve—a Peano curve as well as a straight line segment—is the solution of *some* minimization problem.

[4]Naturally, one could even go farther in time and attribute an infinite age to our fields, based on the fact that they have existed for all eternity in God's mind. Here, we are only trying to establish the much weaker claim that these fields are at least as old as humankind, and for that extremely modest purpose we have no need of such a hypothesis.

> This is a perfectly intelligible question if we avoid metaphysical difficulties over the problem of freedom of the will. If a pigeon fly straight rather than on a curve, once it has decided the direction of its home, we are safe in saying that it has intentionally taken the shortest path and applaud its sagacity, without inquiring as to whether the real reason was that it had an impulse to to straight, and none to bend.

Whatever the true reason for the pigeon's behavior might be, what the pigeon is actually doing is doing is solving a *curve minimization problem*, and obtaining a straight line segment—the most important of all "geometric" configurations—as the solution[5]

COMMENT NO. 30. NEED LOTS OF REFERENCES HERE!!!!!

Bees have also been given credit for exceptional geometrical ability by many human observers, some of whom appeared all too eager to attribute to these industrious insects the loftiest of ideals and a profoundly superanimal wisdom. For example, Pappus of Alexandria (*ca.* 290-*ca.* 350) wrote[6] that

[5]Nowadays there exists a vast body of truly fascinating literature showing that many aspects of animal anatomy, physiology, and behavior, can be explained by assuming that "Nature" is in fact solving optimal control design problems. The "algorithm" used by Nature to solve these problems, is, of course, natural selection.

[6]Quoted from Coolidge [43], p. 1

> Presumably because they know themselves to be entrusted with the task of bringing from the gods, to the accomplished portion of mankind a share of ambrosia in this form, they do not think it proper to pour it carelessly on ground or wood, or any other irregular material; but first collecting the sweets of the most beautiful flowers which grow upon the earth, they make from them for the reception of the honey, the vessels which we call honeycombs (with cells) all equal and similar, and contiguous to one another, and hexagonal in form. And that they have contrived this by virtue of a certain geometrical forethought we may infer in this way. They would necessarily think that the figures must be such as to be contiguous to one another, that is to say, to have their sides in common in order that no foreign matter could enter into the interstices between them and so defile the purity of the produce. Now there are three rectilinear figures which are capable of fulfilling this condition, I mean regular figures which are equilateral and equiangular, for bees would have none of figures which are not uniform. . . . There beoing then three figures capable by themselves of exactly filling the space about the same point, the bees by reason of their instinctive wisdom chose for their construction the figure which has most angles, because they conceived it would hold more honey than the other two.

That is, the bees want their cells to be identical regular polygons ("for bees would have none of figures which are not uniform") and they want them to tile the plane. This leaves them with only three options, namely, triangles, squares and hexagons, and the bees cleverly choose the hexagons because they maximize area ("hold more honey"). Pappus is not entirely precise here. It is not true that any hexagon will have a larger area than any triangle or square. What is true is that, for regular polygons with n sides ***and a given perimeter*** P, the enclosed area grows with n. In particular, a hexagon of a given perimeter will "hold more honey" than a triangle or square of the same perimeter. But, if we assume that our superintelligent bees start from the natural observation that "we are M bees and we have an area A that we can tile with our cells," tben it is quite obvious that, whether they use hexagons or squares or triangles the area of each cell will be $\frac{A}{M}$. This means, of course, that if they had used a triangular tiling then the perimeter of the cells would be larger than that of the hexagonal cells. It is unclear to us exactly what process of reasoning may have led Pappus' wise bees to focus on area maximization ***subject to a fixed perimeter***.

On the other hand,

COMMENT NO. 31. Must complete this section

10.2 Line segments

With *Homo sapiens*, the new dimension of *conscious awareness and understanding* is added to the relationship between geometry and control. Many aspects of human behavior—such as walking straight to go from point A to point B—are similar to Coolidge's pigeon example, and can be explained in terms of animal-like problem-solving. But in addition there are other aspects that can only be accounted for by attributing to humans the *knowledge* that certain curves solve certain optimal control problems.

For example, in ancient Egypt and India the art of *rope-stretching* was practiced in order to produce segments and other geometrical figures, showing knowledge of the fact that a segment minimizes length. In Egypt, according to Democritus, geometric constructions were carried out with the help of specialized workers—whom Democritus describes as experts in "composing lines," and calls by the Greek word "harpenodaptai," which means "rope-strethcers"—by means of pegs and cords. Their skills were used to build altars and temples, where it was deemed necessary to produce certain geometric shapes that would obey very precise specifications, such as being made of perfectly straight segments or perfect circles. These demands often originated with the Gods, who insisted that an altar or temple be built so as to meet very strict requirements, failing which their wrath would be unleashed upon the builders and their communities. For example, the problem of doubling the cube appears in a story told in the Dialogue "Platonikos" of Erathosthenes. The Athenians asked the oracle at Delos what they should do to be liberated from aplague that had struck Athens in BCE. The god Apollo replied throught the oracle that they had to build an altar of exactly the same shape but twice the volume of the existing one.

The fascinating story of the "ritual origins of geometry" is told in a series of papers (cf. [104, 105, 106]) by A. Seidenberg, and was incorporated by B. L. van der Waerden in his 1983 study [131] of geometry and algebra in ancient civilizations.

The important point for us is that

COMMENT NO. 32. Must complete this section

Curve optimization problems had been studied and solved at least since the Greeks. The oldest of all is of course the one of determining the shortest

path joining two points, whose solution—which must have been well known since very ancient times—is a straight-line segment.

10.3 The parallel postulate

Euclid (ca. 330-275 BCE) stated and used the following postulate, number V in his list:

> *If a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if prolonged indefinitely, meet on that side on which are the angles less than the two right angles.*

Euclid himself appears to have felt that this postulate was less evident than the others, as shown by the fact that he avoids using it for as long as possible. And after Euclid, according to R. Bonola ([16], p. 2),

> Even the earliest commentators on Euclid's text held that Postulate V. was not sufficiently evident to be accepted without proof, and they attempted to deduce it as a consequence of other propisitions.

In the 17th century, the Fifth Postulate attracted the interest of many mathematicians, several of whom (for example, P. A. Cataldi in 1603 and 1604, G. A. Borelli in 1658, G. Vitale in 1680, and J. Wallis in 1693, cf. [34, 35], [17],citevitale, [133]) published works entirely devoted to the subject.

A major effort to prove Euclid's postulate was by G. Saccheri (1667-1733), a contemporary of Johann Bernoulli, whose 1733 book *Euclides ab Omni Naevo Vindicatus*—"Euclid vindicated from every flaw"—is an attempt to derive the postulate by contradiction. Saccheri assumes that all the other postulates are true but the Fifth is false, and sets out to derive a contradiction. At one point, he claims to have done it, but his argument is incorrect, as it depends on extending to the point at infinity certain properties that are true at finite distances. (Specifically, Saccheri assumes that if two different lines intersect at infinity, then they cannot have a common perpendicular there, arguing that if they did then this would be "contrary to the nature of the straight line.")

10.4 Circles and the isoperimetric problem

The oldest curve optimization problem after the one one about segments is the *isoperimetric problem*, also known as *Dido's problem*, inspired by the mythical story told by Virgil (70-19 B.C) in the *Aeneid* about the foundation of Carthage (*c.* 850 B.C.): the question is (I) *to find the plane curve of a given length that encloses the largest possible area*. (It is quite easy to see that this is equivalent to the problem of (II) *to determine the shortest of all plane curves that enclose a given area*.) The solution was known by the Greeks to be the circle, although it took until the 19th century for this to be proved in a way that meets our contemporary standards of rigor. Some philosophers were convinced that *this had to mean something* in the grand design of the universe. For example, Aristotle believed that "all motion is either rectilinear or circular or a combination of the two, because these are the only 'simple movements'" ([134], p. 4). He wrote (*ibid.*) that:

> if the motion of the heavens is the measure of all movements whatever in virtue of being alone continuous and regular and eternal, and if, in such kind, the measure is the minimum, and the minimum movement is the swiftest, then, clearly, the movement of the heavens must be the swiftest of all movements. *Now of lines which return upon themselves the line which bounds the circle is the shortest*; and that movement is the swiftest which follows the shortest line. Therefore, if the heaven moves in a circle and moves more swiftly than anything else, it must necessarily be spherical. [Our Italics.]

So Aristotle thought[7] that the motion of heavenly bodies is determined by some kind of "minimum" principle, and the fact that this motion was circular—as he believed it was—had to have something to do with the isoperimetric property. (The italicized assertion in the quotation is of course not true unless we impose some extra constraint, e.g. that the area enclosed by the curve be a fixed constant, since otherwise the shortest path would be reduced to a single point. So it is reasonable to assume that Aristotle had the isoperimetric problem—in its Version II—in mind.)

[7] Aristotle was far ahead of its time: his clarity and precision are up to the most exacting standards of rigor of the very best postmodern and deconstructionist writing of the late 20th century, and the technique of using words such as "clearly" and "therefore" to justify the most unintelligible nonsequiturs is still widely used today in contemporary scientific publications.

It clear that the Greeks understood that these problems were intimately related to geometry. However, as we shall argue, one important aspect of this relationship was not really grasped until the end of the 19th century. This aspect, that minimization problems can be the source of new geometries, was not truly appreciated until the era of Gauss and Riemann. In modern terms, this point is a trivial remark, namely, that one can turn around the fact that the segment is the shortest path between two points, and make it the *defintion of segment*, using a more general prescription to measure "length." We will show how the failure to notice this point—which looks so simple today—made Johann Bernoulli miss a discovery that would have have added much to his fame and glory.

10.5 Shortest paths in geometrical optics

In the years between the flourishing of science in classical Greece and the late 17th century, interest in geometric optimization problems grew for at least two reasons: (i) because an increasing number of problems of this kind were solved, and the solutions posed nontrivial challenges to mathematicians, since in some cases they were far from obvious and some of the sharpest minds got them wrong, (ii) because new evidence supported the intuition that these problems were in some deep way related to the workings of the physical world.

As we have already pointed out in §4.1, Hero of Alexandria showed that when a light ray emitted by an object is reflected by a mirror it follows a path from the object to the eye which is the shortest of all possible such paths. In Hero's setting, which involved a single medium and therefore a constant speed of light, "shortest" was equivalent to "fastest." This was no longer the case in the work of Fermat (1601-1665), who formulated the general principle that *light rays follow the fastest—i.e., minimum-time—paths*, and showed how this explained not only Hero's observation about reflection but also Snell' law of refraction. (See Scholium VII, in page 105; we shall soon see that Fermat's principle played a crucial role in Johann Bernoulli's solution of the brachistochrone problem, cf. Chapter 13 below, especially §13.4 in page 213.)

These observations were no longer speculative guesses about the distant heavens, but empirically verifiable assertions about the world around us, which in fact made nontrivial testable predictions. (For example, as we indicated §Scholium VII, Descartes had an alternative explanation for Snell' law, which required that light travel *faster* in a denser medium, whereas

Fermat's principle yielded the correct law of refraction only if light was assumed to travel more slowly in a denser medium. Huygens (1629-1695) subsequently provided a justification of Fermat's principle by deriving it from wave optics.)

There is, addition, an important "geeometric" aspect to Fermat's least time principle. To understand this, let us just retain from Hero's theory about reflection the following observation, the we shall call "Hero's geometric principle":

(HGP) *the path of the light rays is determined by "pure geometry," and light follows the shortest paths.*

Then Fermat's principle can be interpreted in two ways, as saying *either*

(1) that Hero's geometric principle no longer applies, since we now need to take the physical properties of the medium into account,

or

(2) that Hero's geometric principle still applies, and light rays still follow the shortest paths, but we now have to define "length" in terms of a new, more refined geometry, in which the physical properties of space influence the geometry, and shortest paths can be curved or bent.

Under the second interpretation, Fermat's principle would still say that light follows a shortest path—i.e., a "geodesic"—but in a different, non-Euclidean geometry. With hindsight, we now know that Interpretation 2 is the most fruitful. For example, G. Lochak writes in [87], p. 57–58 [our translation]:

> [Fermat's principle] anticipated the greatest revolution to be undergone by geometry in the future, namely, the discovery of non-Euclidean geometries, because it showed for the first time that the physical properties of a medium might cause the shortest paths not to be straight lines.

One might be tempted to suspect that authors like Lochak perhaps allow themselves to be carried away by their own 20th century hindsight about the geometric possibility, and may be putting too much of this hindsight into 17th century minds, attributing to them ideas that they were far from ready to conceive. It appears reasonable to ask

- To whom exactly did Fermat's principle "show" what Lochak says it did?
- Did it "show" it to Fermat himself?
- Did Fermat really think that his minimum-time paths were "shortest" rather than just "fastest"?

We prefer to remain agnostic as to what Fermat's actual beliefs may have been, but point out, in defense of Lochak, that Leibniz, in his analysis of the refraction problem in the 1684 *Nova mathodus* paper, definitely uses the word "shortest." This suggests to us that at least Leibniz definitely ***was*** entertaining the possibility that light rays in a medium with variable speed of light truly were "shortest paths" in some new geometry.

COMMENT NO. 33. HAVE TO CHECK LEIBNIZ' ORIGINAL LATIN TO MAKE SURE.

10.6 Johann Bernoulli misses an opportunity to solve the problem of Euclid's fifth postulate

As will be explained in Chapter 13, Johann Bernoulli's brachistochrone problem is equivalent to that of determining the light rays in a half-plane $\{(x, y) : y \geq 0\}$, where the speed of light is $\sqrt{y}$. Johann Bernoulli was eager to generalize his work on the brachistochrone by studying other problems of a similar kind. One obvious class of such problems is obtained by trying to determine the paths of light rays that obey Fermat's minimum-time principle, for a medium with a variable speed of light given by a function $c(x, y)$ other than $\sqrt{y}$.

Johann Bernoulli solved this problem for other functions c. In particular, he did it for $c(x, y) = y$ and for $c(x, y) = y^{1/3}$, and showed in both cases that one can obtain simple formulae for the solutions. When $c(x, y) = y$, he found that the solution curves are either half-circles with center in the x axis or vertical half lines.

His result for $c = y$ put him, unknowingly, within a hair's breadth from the solution to a problem that had defied the efforts of mathematicians for 2000 years, and would only be solved 172 years later, in 1868.

The problem was that of the independence from the other axioms of geometry of Euclid's famous fifth postulate about plane geometry. The postulate asserts that if L is a line and P a point not on it then there exists a

unique line through P parallel to L, and was regarded by many authors as much less believable than all the others. Since Euclid wrote his *Elements* (ca. 300 B.C.), enormous amounts of energy were spent in efforts to derive it from the other axioms. The "proofs" were usually by contradiction: one assumes that all the other postulates hold but the fifth one does not, and sees what happens, hoping that a contradiction will eventually emerge. Although all these proofs were wrong, because at some point they incorrectly stated that something was a contradiction and therefore the fifth postulate had been proved, they provided valuable knowledge. Indeed, the proofs really were extensive explorations of a non-Euclidean universe, notwithstanding the authors' firm conviction that such a universe was a logical impossibility, and their unawareness that they were in fact laying the foundations for non-Euclidean geometry. Eventually, the proofs grew longer and longer, and the strange properties of a non-Euclidean world became so familiar and started to make so much sense, that the dominant view shifted. In the 1820s, Taurinus (1794-1874) and Gauss (1777-1855) finally realized that the contradiction was nowhere to be found. Lobachevski (1793-1856) in 1829, and J. Bolyai (1802-1860) in 1832, openly advocated the view that non-Euclidean geometry was simply another geometry, free from contradiction and worth studying. This, however, did not settle the question, for merely asserting with conviction that a logical system is without contradiction does not prove that it is. The problem was finally solved in 1868 by Beltrami (1835-1900), who gave a *model* for hyperbolic geometry, a non-Euclidean geometry in which parallels are not unique. Finally, in 1882, Poincaré (1854-1912) gave another, much simpler model.

In the Poincaré model, the "plane" is the open half-plane $\{(x, y) : y > 0\}$. A "line" is a half-circle with center in the x axis, or a vertical half line. One can then easily see that given any two points A, B in the "plane" there is a unique line joining them. The distance between A and B is the "length" of the unique "line segment" joining them, except that "length" is not the standard length of a curve. *The "length" of an arc is the travel time of a particle that moves along the arc with velocity y.*

In other words, *the "lines" of the Poincaré model are none other than the solutions of one of Johann Bernoulli's modified versions of the brachistochrone problem, with $c = y$.*

The cycloids that occur in the original brachistochrone problem (cf. Chapter 13 below) have many of the incidence properties of Euclidean lines (see, for example, statement (*) in Chapter 13 below), and therefore come close to satisfying all Euclid's postulates other than the fifth. But they do

not yet provide a good model for non-Euclidean geometry because they "go to infinity" (i.e., approach the x axis) with finite length. The half-circles of the Poincaré model have all the desired incidence properties and also have infinite length, so they furnish a perfect model where all the Euclidean axioms other than the fifth are true. A simple way to understand why these two cases are so different is to observe that in both of them one is studying minimum-time paths for a control system $\dot{x} = \varphi(x,y)u$, $\dot{y} = \varphi(x,y)v$, with control constraint $u^2 + v^2 = 1$. When $\varphi(x,y) \equiv \sqrt{|y|}$ [8]—the cycloid case—the x axis consists of equilibria, but since the function φ is not Lipschitz it is reasonable to expect that one can join points in the x axis to points strictly above it in finite time. When $\varphi \equiv y$—the circle case—the function φ is Lipschitz, so there is no way to connect in finite time a point not on the x axis to a point on the x axis.

In modern terms, the circles of Johann Bernoulli's brachistochrone-like problem are the geodesics of a Riemannian metric on the upper half plane given by $ds^2 = y^{-2}(dx^2+dy^2)$. It is interesting to note that Johann Bernoulli had made significant contributions to the early stages of the development of these ideas, by studying the curves of minimum length on a surface, which later came to be known as "geodesics." (The word "geodesic" was coined much later by Liouville (1809-1892).)

We also note that Johann Bernoulli lived at a time when the problem of Euclid's parallel postulate was very much at the center of mathematical research. For example, G. Saccheri published in 1733 his book *Euclides ab omni naevo vindicatus* ("Euclid vindicated from every flaw"), where he pushed very far the development of non-Euclidean geometry, only to end up dismissing it by claiming that a particular conclusion he had reached had to be rejected because it was "repugnant to the nature of a straight line."

To those of us who know that it can take years of hard work to understand the solution of an important open problem such as Fermat's last theorem, the question of the independence of Euclid's postulate is a mathematical rarity. It is a problem that mathematicians took about 2200 years to solve, but the solution is so simple that it can be explained to any undergraduate with a moderate understanding of plane analytic geometry.

Johann Bernoulli had essentially all the tools he needed to solve the problem. Why didn't he? What he lacked was the understanding of the

[8] Since one usually requires that the state space of a control system be an open subset of $\mathbb{R}^n$ or a smooth manifold, it is more natural to regard the brachistochrone and related problems as defined in the *whole plane* rather than in a half-plane. This is why we are using the absolute value of y. We will pursue this point in 13.10, where we will study the "reflected brachistochrnoe problem."

geometric significance of the cycloids and half circles that he had found as solutions of his minimization problems. He certainly saw these curves as "geometric," but failed to see that they could serve as the "lines" of a different geometry.

10.7 The missing link: curve minimization as a rich source of new geometries

"Geometry" is, tautologically, the study of "geometric objects," but it is far from clear what it is exactly that makes an object "geometric." Since ancient times, nobody ever doubted that straight lines and circles were "geometric curves." So were ellipses, hyperbolas and parabolas, since they can be characterized "geometrically" as "loci" of points satisfying certain "geometric" conditions (e.g., for the ellipse, that the sum of the distances to two given points is a constant) or as conic sections. On the other hand, the graph of a nonmeasurable function, or of a continuous nowhere differentiable function, are clearly not "geometric curves" by most standards, although occasionally some of these objects can acquire a "geometric" status if they become amenable to a particularly simple description, as may be the case nowadays with some fractal sets. The geometricity of objects lying between these two extremes has been somewhat unclear, and has certainly changed through the ages. For example, Descartes only called a curve "geometric" if it was algebraic[9], thereby denying admission to the world of geometry to transcendental loci such as sinusoids, cycloids, tractrices, graphs of exponentials, clotoids, logarithmic spirals, and other natural curves. This was regrettable, because some of these curves admit rather simple geometric descriptions. (For example, the exponential curve C is characterized by the property that there is a straight line L for which, if P varies along C, and Q, R are the points in L for which the line PQ is tangent to C at P and the line PR is perpendicular to L, then the distance from R to Q is constant.)

With the advent of the calculus, a language became available in terms of which these curves are simple and natural. (For example, the exponential function is the solution of $y' = y$.) So the notion of geometricity shifted, and the geometric world was enlarged by the admission of many new curves.

So "geometricity" is in the eyes of the beholder, and the notion of a "geometric object" is undergoing perpetual change. What makes an object

[9]that is, defined by a system of polynomial equations; for example, straight lines and conic sections are "geometric" in this sense, but an exponential curve is not, because $y = e^x$ is not a polynomial equation.

"geometric" is that it admits a simple characterization, with the understanding that "simplicity" here is a time-varying concept, depending on the mathematical language available at a particular moment in history.

Within this framework, one of the most natural ways of singling out objects in a "simple" way has been, at least since the classical era, to require that they solve some "simple" minimization or maximization problem. (The key word here, of course, is "simple." A 20th Century mathematician would pedantically observe that *every* curve—a Peano curve as well as a straight line segment—is the solution of *some* minimization problem.)

As we pointed out above, Greek philosophers such as Aristotle had a distinct feeling that there had to be a deep meaning to the fact that the circle was the solution of the isoperimetric problem.

We now understand that even the simple statement that "a line segment is the shortest path joining its two endpoints" is rich in profound implications. All it takes is to rewrite it as "a *line segment* is the shortest distance between two points," and it becomes clear that the length minimization property can be regarded as a *definition* of line segment rather than a proposition that happens to be true about "real" line segments. Once this momentous step is taken, one can move on and study general geometries in which "length" is measured in some other way, e.g. as travel time in a medium with a variable speed of light, and one can investigate whether the "lines" in this geometry behave like Euclidean lines.

So maybe Johann Bernoulli could have solved an important 2000-year old problem, thereby succeeding in his passionate quest for fame and glory surpassing those of his brother and his son (cf. Scholium II in page 10), if only it had occurred to him to look for the hidden meaning of a trivial statement about segments, or to read it aloud with just a slight change of emphasis.

10.8 Covariance and invariance, from Lagrange to Riemann

We saw in Chapter 8 that Lagrange was proud of having freed mechanics from geometry by doing everything analytically. Little did he know that one of his discoveries anticipated Riemann's revolutionary creation of differential geometry.

Lagrange allowed the configuration of a collection of N particles to be represented in terms of completely general systems of $3N$ coordinates. This

was a decisive step in the process that would free geometry from the constraint of having to deal with 3-dimensional Euclidean space. "To appreciate how all was new in this work [the *Mécanique Analytique*], let us point out that, when Lagrange described a motion in terms of a system of Cartesian coordinates in space, this was such a nontrivial step that he devoted two pages to trying to persuade the reader" (Lochak, [87], p. 49).

He then showed that the Euler-Lagrange system (4.3) has the same form in "curvilinear" or "generalized" coordinates, so that (4.3) is invariant under *arbitrary*, nonlinear coordinate changes. This was the first example of what would later become the cornerstone of differential geometry, namely, the principle of "coordinate democracy," according to which all coordinate systems, "rectangular" as well as "curvilinear," have to be treated equally.

Lagrange saw the significance of this invariance property:[10]

> It is perhaps one of the principal advantages of our method that it expresses the equations of every problem in the most simple form relative to each set of variables and that it enables us to see beforehand which variables one should use in order to facilitate the integration as much as possible. [Quoted from [134], p. 33.]

In 1854, sixty-six years after the publication of Lagrange's treatise, Riemann gave his famous lecture, sketching the brilliant ideas that would become the basis of modern differential geometry. It turns out that one of these ideas was precisely that of *invariance under general coordinate changes*, whose importance Lagrange had presciently anticipated.

Riemann worked on spaces of any number n of dimensions, whose points are represented, locally, by n coordinates $x^1, \ldots, x^n$. In addition, he finally took the step of turning around the trivial statement about the segment being the shortest path between two points, and used the length-minimization property to *define* the concept of a "line segment," now called a "geodesic."

The length of an arbitrary curve $t \mapsto \xi(t) = (x^1(t), \ldots, x^n(t))$ is measured using the *arc-length* element $ds^2 = \sum_{ij} g_{ij}(x) dx^i dx^j$, where $(g_{ij})_{i,j=1}^n$ is an n by n matrix of smooth functions of x, assumed to be positive definite for each x.

The functions g_{ij} define the metric relative to a particular coordinate system $(x^1, \ldots, x^n)$. If we change to another coordinate system $\tilde{x}$, related

[10]For a striking illustration of what Lagrange had in mind here, think of how much simpler it is to derive the equations of motion of a body under an inverse-square attracting force if one first proves that the motion must take place in a plane, and then does the whole calculation in polar coordinates, using the invariance of the Euler-Lagrange equations.

to x by $\tilde{x} = \Phi(x)$, $x = \Psi(\tilde{x})$, then the coefficients $\tilde{g}_{\alpha\beta}$ expressing the metric in the new system are related to the g_{ij} by

$$\tilde{g}_{\alpha\beta}(\tilde{x}) = \sum_{ij} g_{ij}(x) \cdot \frac{\partial x^i}{\partial \tilde{x}^\alpha} \cdot \frac{\partial x^j}{\partial \tilde{x}^\beta} \,. \tag{10.1}$$

The metric itself should be thought of as a "coordinate free object," represented in each coordinate system by an array $\{g_{ij}\}$ of functions. These arrays are related by the transformation (10.1). More generally, *covariant quantities are "objects" that for each coordinate system are represented by an array of functions, in such a way that when one changes coordinates these arrays transform according to some specified law.* There are different types of such quantities, depending on the transformation law, and among them there is a class of quantities called *tensors*.

Remark 10.8.1 In modern differential geometry, one seeks to define covariant objects intrinsically, rather than as "arrays of functions that transform in a certain way." This is a much better and more powerful language, exactly as, say, the language of linear maps on abstract vector spaces is superior to the language of matrices. But to do justice to this idea one needs much more algebra, so we will keep our discussion simple by sticking with the old-fashioned way of talking about covariance. ♢

The metric is an example of a tensor, known as the ***Riemann metric tensor***. A "flat metric" is one whose metric tensor, relative to some coordinate system, is the identity matrix. (Ordinary Euclidean space is of course flat.) It is easy to see that if a metric is flat, and one changes coordinates, then in the new coordinate system the g matrix will in general no longer be the identity. In other words, the fact that a metric is flat may not be apparent if the metric is presented in "curvilinear" coordinates. This raises the question of how to characterize flatness intrinsically, in a coordinate-free way.

Riemann answered the question by introducing the ***curvature tensor*** R. In each coordinate system, R has n^4 components R^ℓ_{ijk}, labelled by four indices. (The number of truly independent components is much smaller, because there any many relations among the R^ℓ_{ijk}.) When coordinates are changed as before, the transformation law for R is:

$$\tilde{R}^\alpha_{\beta\gamma\delta}(\tilde{x}) = \sum_{ijk\ell} R^i_{jk\ell}(x) \cdot \frac{\partial \tilde{x}^\alpha}{\partial x^i} \cdot \frac{\partial x^j}{\partial \tilde{x}^\beta} \cdot \frac{\partial x^k}{\partial \tilde{x}^\gamma} \cdot \frac{\partial x^\ell}{\partial \tilde{x}^\delta} \,. \tag{10.2}$$

So R (like any tensor), has the property that *if $R \equiv 0$ in some coordinate system, then $R \equiv 0$ in every coordinate system.* In other words, the vanishing of R is an *invariant property.*

Moreover, R can be computed by explicit formulas in terms of g. Precisely, one should first find the inverse g^{-1} of the matrix $g = (g_{ij})$. Using g^{ij} to denote its entries, so that $\sum_j g^{ij} g_{jk} = \delta^i_k$ (where $\delta^i_k = 0$ if $k \neq i$, $\delta^i_i = 1$), it turns out that

$$R^i_{jk\ell} = \frac{\partial \Gamma^i_{j\ell}}{\partial x^k} - \frac{\partial \Gamma^i_{jk}}{\partial x^\ell} + \sum_s \left(\Gamma^i_{sk} \Gamma^s_{j\ell} - \Gamma^i_{s\ell} \Gamma^s_{jk} \right), \tag{10.3}$$

where the Γ^k_{ij} are the so-called *Christoffel symbols*:

$$\Gamma^i_{jk} = \frac{1}{2} \sum_\ell g^{i\ell} \left(\frac{\partial g_{\ell k}}{\partial x^j} + \frac{\partial g_{\ell j}}{\partial x^k} - \frac{\partial g_{jk}}{\partial x^\ell} \right). \tag{10.4}$$

These formulas are of course familiar to those readers acquainted with differential geometry, and probably appear impenetrable and unmotivated to everybody else. But it should be clear from them that *if a metric is flat then R vanishes,* because for a flat metric all the Γ^k_{ij} vanish. Moreover, since the vanishing of R is an intrinsic property, we can test a metric g for flatness in *any* coordinate system, by computing R and seeing whether or not it vanishes. Actually, things are even nicer, because it turns out that the vanishing of R is in fact a sufficient as well a necessary condition for flatness.

10.9 Invariants and Klein's Erlangen program

In 1872, F. Klein formulated his *Erlangen program*, whose main idea is that each geometry is characterized by the *group of transformations* which preserve the relevant properties of the geometry. Associated to such a group there is usually a collection of *invariants* and a class of *covariant objects.* For example, for Euclidean space the transformation group is the orthogonal group, an obvious and important invariant is the distance between two points, and the components of a vector are covariant quantities. For the study of general differentiable manifolds, the natural "group" is really a *pseudogroup*, namely, the set of diffeomorphisms, and differential geometry has produced a huge zoo of covariant objects: tensors of various kinds (an in particular vector fields and differential forms), jet spaces, connections, etc.

For control systems, many of the properties of interest (e.g. the structure of minimum time trajectories, feedback stabilizability, feedback linearizability) are invariant under this general pseudogroup, so one should try to characterize them in a coordinate-free way, using invariants. A *complete set of invariants* for a class of properties is a set of invariants that complete determines these properties. For example, if we are interested in the solvability of a system of linear equations $Ax = b$, a complete set of invariants is provided by a pair of integers, namely, the rank of A and the rank of the "augmented matrix" $[A; b]$. $\diamondsuit$

Chapter 11

Beyond the maximum principle, geometrically

As we have shown, the work on necessary conditions for optimality began 300 years ago, reached new heights with the maximum principle, and is still continuing. But, valuable as they are, the results of this particular line of research are in a sense preliminary. Necessary conditions for optimality are tools for studying the properties of optimal curves, but in the end it is these properties that one really wants to understand. An undue emphasis on general necessary conditions for optimality for their own sake can lead to sterile near-tautologies, amounting to the assertion that "a necessary condition for a trajectory to be optimal is that it be optimal," skillfully packaged to convey an impression of depth and novelty. To avoid falling into this trap, the search for general conditions has to be coupled with the study of challenging concrete problems. These can be general (e.g. "what regularity properties, if any, are true of optimal trajectories for all possible smooth problems?") or can involve the complete understanding of the solution of a specific problem. We will give examples of both kinds of questions below.

A large body of work exists on the properties of optimal curves for the classical calculus of variations. But, within the much broader context of optimal control theory, little is known, and very exciting problems remain open. For example, the answers to some of the most basic questions about the general regularity properties of optimal trajectories[1] are not known.

[1]There are, of course, many other hard and important questions not touched upon in this book. For example, for finite-dimensional deterministic optimal control, there are problems having to do with optimal synthesis, the structure of reachable sets, controllability, computational methods, conditions involving state space constraints, etc. And there is an even larger area involving infinite-dimensional and stochastic systems.

Part of the problem here is to find the right language for talking about these properties. More precisely, one would like to do as in other mathematical theories where properties of the solutions of a problem are sought, and identify *invariants* of a problem that can be computed in terms of the data and provide the desired information about the solutions.

11.1 Covariance and invariance in control: Lie brackets

By now, the reader may be asking what our discussion of covariance and invariance in Chapter 10 had to do with optimal control. The answer is that it has a lot to do with optimal control and, more generally, with *control*, optimal or not.

To see this, let us recall that the main aspect that makes optimal control stand out and be different from the calculus of variations is the presence of a nontrivial dynamics, given by Equation (3.10). In the special but very important case of an *autonomous system*, i.e., when f does not depend on t, the dynamics (3.10) is really a family $\mathbf{f} = \{f(\cdot, u)\}_{u \in U}$ of vector fields, parametrized by $u \in U$. (From now on, we write $f_u(q)$ rather than $f(q, u)$, to emphasize that f_u should be regarded as a vector field.)

Vector fields are, of course, covariant[2] objects, given with respect to each coordinate system by a set of component functions that transform according to a specific law when one changes coordinates. Differential geometry tells us that, if we want to understand the properties of a family $\mathbf{f}$ of vector fields, *we should operate in a "coordinate-democratic" spirit, following the ideas of Lagrange and Riemann.* This means that we should look for other *covariant quantities* that are related to $\mathbf{f}$ in a natural way.

It turns out that, if f and g are two vector fields, then there is a natural covariant object derived from them,[3] namely, the *Lie bracket* $[f, g]$. This is defined in coordinates by the formula

$$[f, g] = \frac{\partial g}{\partial x}.f - \frac{\partial f}{\partial x}.g\,, \tag{11.1}$$

where we are writing f and g as columns of functions, and $\frac{\partial g}{\partial x}$, $\frac{\partial f}{\partial x}$ are the Jacobian matrices of f and g, which are square matrices of functions.

[2] In the general sense of covariance defined earlier. When one draws a distinction between "covariant" and "contravariant" objects, as is usually done in differential geometry, vector fields are of course contravariant, and covectors such as the "adjoint vector" of the maximum principle are covariant.

[3] Besides the trivial ones, such as the linear combinations of f and g.

COMMENT NO. 34. Here we must insert a box with the following simple example showing how the Lie bracket shows up. If you have m dollars in your bank account (that is, you have $m(t)$ dollars at time t, earning interest continuously at a rate ρ), and in addition at time t you spend $s(t)$ dollars per unit time (where $s(t)$ can be negative, in which case you are earning money), then m evolves according to

$$\dot{m}(t) = \rho m(t) - s(t)\,.$$

The function $t \mapsto s(t)$ is the control. The vector fields $m \mapsto f(m) = \rho m$ and $m \mapsto \rho m - 1$ describe, respectively, what happpens when you are spending nothing and when you are spending 1 dollar per unit of time. *These two behaviors do not commute,* because if you spend your money initially then this money will not earn interest, but if if you wait and spend it later then it will earn interest until you spend it. Mathematically, this is reflected in the fact that $[f, g] \neq 0$.

COMMENT NO. 35. At this point we should also include Vel Jurdjevic's famous pedagogical example: the operations of jumping into a swimming pool and filling the pool with water do not commute.

Then $[f, g]$ is a vector field, i.e., a tensor of exactly the same type as f and g. This suggests that $[f, g]$ should have an intrinsic, coordinate-free characterization, and indeed it does, since

$$[f, g](q) = \lim_{t \to 0} \frac{\delta(t) - q}{t^2}\,, \tag{11.2}$$

where

$$\delta(t) = q e^{tf} e^{tg} e^{-tf} e^{-tg}\,. \tag{11.3}$$

(Here $t \mapsto q e^{tX}$ denotes, for a vector field X, the integral curve of X that goes through q when $t = 0$.)

Exactly as in the case of Riemann's metric tensor g, a family $\{f_u\}$ of vector fields may have interesting invariant properties which are immediately detected in a particular coordinate system but are much harder to ascertain if the system is presented in different coordinates.

But certain things computable in terms of the f_u are independent of the choice of coordinates. For example, the fact that two vector fields f, g commute (i.e., $[f, g] = 0$), is intrinsic. Moreover, this property has a very

direct control-theoretic meaning: $[f, g] \equiv 0$ if and only if, for all control functions $u(\cdot)$, $v(\cdot)$, and all $\bar{q}$, the solution of

$$\dot{q}(t) = u(t)f(q(t)) + v(t)g(q(t))\,, \quad q(0) = \bar{q}\,, \tag{11.4}$$

is always given, for small t, by

$$q(t) = \bar{q}e^{U(t)f}e^{V(t)g}\,, \tag{11.5}$$

where $U(t) = \int_0^t u(s)ds$, and $V(t) = \int_0^t v(s)ds$. In other words, the effect of applying controls $t \mapsto u(t)$ and $t \mapsto v(t)$ during a time interval $[0, T]$ is completely determined by two numbers, namely, the total amounts $U(T)$ and $V(T)$ of u-control and v-control that have been applied. It does not matter whether all the u-control was applied first, or the v-control was applied first, or both u and v were evenly distributed on the interval. By contrast, if $[f, g] \neq 0$, then more than two numbers are needed to determine the final effect of applying two controls. For example, if f and g are the vector fields in the x, y plane $\mathbb{R}^3$ with components $(1, 0, 0)$ and $(0, 1, x)$, respectively, so that $[f, g]$ has components $(0, 0, 1)$, then a simple calculation shows that, to know $q(t)$ given $q(0)$, we need three independent numbers, namely, $U(t)$, $V(t)$, and the iterated integral $W(t) = \int_0^t U(s)v(s)ds$.

It then turns out that

(INV) *for real-analytic autonomous control systems $\dot{q} = f_u(q)$, the Lie bracket relations among the vector fields f_u at a point $\bar{q}$ form a complete set of invariants for the properties of the time-optimal trajectories through $\bar{q}$.*

The meaning of "complete set of invariants" has already been explained. A precise definition of "Lie bracket relations" would require more algebra than we care to develop here. Let us just show the main idea for the case of two vector fields f, g, without giving a complete definition. A *Lie bracket relation* between f and g at a point q is a formal linear combination of brackets of two indeterminates F, G, such that, when one plugs in f and g for F, G, and evaluates at q, the result is 0. For example, the expression $F + [F, G] + 3[F, [G, F]]$ is a Lie bracket relation between f and g at q iff $f(q) + [f, g](q) + 3[f, [g, f]](q) = 0$.

Remark 11.1.1 Actually, a stronger result is true, but we need more terminology to state it. For a system $\{f_u\}_{u\in U}$ of real-analytic vector fields, let L be the Lie algebra of vector fields generated by the f_u. Then *the Lie bracket relations at a point q form a complete set of invariants for all*

properties of the system that are preserved by local diffeomorphisms of the maximal integral manifold S of L through q.

What this means, in more pedestrian terms, is that the Lie bracket relations "see" everything that happens on the integral manifold but do not see what happens outside. For example, the property that S is a neighborhood of q is *not* determined by the Lie bracket relations, because this property has to do with how S sits in relation to the whole space, and is not about something that happens within S itself. But the structure of the time-optimal trajectories through q *is* determined by the Lie bracket relations at q, because all the trajectories through q stay in S.

Clearly, (INV) is a special case of this. Another important special case arises when the system has the "accessibility property" at q, that is, S is a neighborhood of q. In that case *all* local properties near q that are diffeomorphism-invariant—for example, asymptotic stabilizability by state feedback, local equivalence to a linear system, local feedback linearizability—are determined by the Lie bracket relations.

Naturally, properties that are not diffeomorphism-invariant—for example, convexity of reachable sets—need *not* be determined by the Lie bracket relations. ♢

11.2 Control, Lie brackets, zoology and medieval bestiaries

For the reasons described in the preceding remark, Lie brackets play a crucial role in optimal control theory in many ways and, more generally, they are the key objects in the theory of nonlinear finite-dimensional control. They play a central role in nonlinear controllability, feedback linearization, and system equivalence. Just to give a simple illustration having nothing to with optimality, let us consider a control theory problem totally analogous to that of the characterization of flatness for a metric. In control theory, "linear control systems"—i.e., systems of the form $\dot{q} = Aq + Bu$, evolving in $\mathbb{R}^n$, with u taking values in $\mathbb{R}^m$, and A, B matrices of the appropriate sizes—play a crucial role, comparable in importance to that of the flat metrics within the class of general Riemannian metrics. So it is important to know when a system that looks nonlinear because it is presented in "curvilinear coordinates," is in fact linear, in the sense that one can make it linear by changing coordinates. The answer turns out to be a criterion involving Lie brackets, analogous to the vanishing of Riemann's tensor R.

Precisely, for a system $\dot{q} = f(q) + u_1 g_1(q) + \ldots + u_m g_m(q)$, $q \in \mathbb{R}^n$, with C^∞ vector fields f, $g_1, \ldots, g_m$, and a point $\bar{q}$ in $\mathbb{R}^n$, a necessary condition for the system to be equivalent near $\bar{q}$ to a linear system is that all the iterated brackets of members of the set $\{f, g_1, \ldots, g_m\}$ involving two g_i's should vanish near $\bar{q}$. Under an extra technical condition ("strong accessibility," that is, the requirement that the Lie brackets $\mathrm{ad}_f^k(g_i)(\bar{q})$, $k \geq 0$, $i = 1, \ldots, m$ span the whole space) this condition is sufficient as well.

Using such a criterion, one can see, for example, that a system such as

$$\dot{x}_1 = x_2 + 2x_2x_3\,, \qquad \dot{x}_2 = x_3 - x_1x_2 + x_2^3\,, \tag{11.6}$$
$$\dot{x}_3 = u + x_2^2 + x_1x_3 - x_1^2x_2 + 2x_1x_2^3\,,$$

which superficially looks very nonlinear, is in fact a linear system in disguise, whereas a system such as

$$\dot{x}_1 = u_1\,,\ \ \dot{x}_2 = u_2\,,\ \ \dot{x}_3 = u_1x_2 - u_2x_1\,, \tag{11.7}$$

whose nonlinearity appears to be much milder, is in fact highly nonlinear.

So ***Lie brackets help us organize our knowledge of nonlinear systems in terms of their deep structural properties.*** This way of classifying systems is as superior to the naïve approach based on a superficial counting of the number of nonlinearities as the animal taxonomy of modern zoology is to the classifications found in medieval bestiaries. Exactly as a dolphin is structurally closer to a zebra than to a shark—in spite of the very visible ways, such as shape and habitat, in which a dolphin resembles a shark more than a zebra—System (11.6) is much closer to a linear system than System (11.7). Lie brackets enable us to discover this, and then one can find abundant confirmation by looking at the properties of both systems. (For example, (11.6) is globally stabilizable by a smooth state feedback, but (11.7) is not even locally stabilizable near any point.)

Control with Lie brackets is to control without Lie brackets like the taxonomy of modern zoology is to that of a medieval bestiary.

11.3 Using Lie brackets to understand optimal trajectories

For optimal control problems, the effort to understand the properties of optimal trajectories in terms of reasonable invariants has begun only recently, and most of the work remains to be done.

Just to give the flavor of some of these results, we concentrate on one problem that has attracted a lot of attention in recent years, namely, that of minimum time control. Suppose, for example, that we are looking at minimum time control of systems in $\mathbb{R}^n$ of the form $\dot{x} = f(x) + ug(x)$, with $|u| \leq 1$, where f and g are smooth vector fields. Here the control space U is the interval $[-1, 1]$. The Hamiltonian is $H = \varphi + u\psi - p_0$, where $\varphi = \langle p, f(x)\rangle$ and $\psi = \langle p, g(x)\rangle$. Then u will equal 1 when $\psi > 0$, and -1 when $\psi < 0$. A naïve genericity argument may suggest the wrong conclusion that $\psi = 0$ is a "rare" event, which will happen at most at isolated points, so that in fact the optimal trajectories will be "bang-bang," i.e., such that u is piecewise constant with values 1 and -1 and finitely many jumps. This, however, is quite wrong.

Example 11.3.1 Let $Q = \mathbb{R}^2$, and assume that $q = (x, y) \in Q$ evolves according to $\dot{x} = 1 - y^2$, $\dot{y} = u$, $|u| \leq 1$, so f, g are the vector fields $(x, y) \mapsto (1 - y^2, 0)$, $(x, y) \mapsto (0, 1)$, respectively. (Equivalently, $f = (1-y^2)\frac{\partial}{\partial x}$ and $g = \frac{\partial}{\partial y}$.) Suppose we want a trajectory that goes from $(0, 0)$ to $(1, 0)$ in minimum time. It is obvious that the solution is given by $u(t) \equiv 0$, $y(t) \equiv 0$, $x(t) \equiv t$, $0 \leq t \leq 1$, and the optimal time is 1. So the optimal control is "singular"—i.e., takes values in the open interval $]-1, 1[$ —and therefore drastically fails to be bang-bang. And one can easily verify that this example is "stable under perturbations" in any reasonable sense of the word. So it is not true that optimal controls are always, or even generically, bang-bang. ♢

In view of our previous remarks, it is to be expected that the *Lie brackets* of the vector fields f and g will play a crucial role in determining the properties of a solution. Not surprisingly, a close analysis of the problem, applying (MP), shows that this is indeed so.

For example, the "singular" control of Example 11.3.1 is closely related to a property of the Lie brackets of the vector fields f and g. Precisely, *the singular trajectory exists and is contained in the x axis because $[f, g]$ vanishes on the x axis.*

Since the computation of Example 11.3.1, as presented, does not explicitly display any Lie brackets, we will now make the Lie brackets stand out by looking, more generally, at systems of the form $\dot{x} = f(x) + ug(x)$, with $|u| \leq 1$, without specializing to a particular pair of vector fields f, g.

If one computes the derivative of ψ along an extremal γ, the result is ρ evaluated along γ, where $\rho = \langle p, [f, g]\rangle$. So, if n happens to be 2, and g and $[f, g]$ are linearly independent at each point, then ψ and $\dot{\psi}$ cannot vanish

simultaneously, for otherwise p would be orthogonal to $g(x)$ and $[f,g](x)$, so p would vanish, which is a contradiction. (Recall that (MP') says that for variable time problems H must vanish. Since $H = \langle p, f(x)+u\,g(x)\rangle - p_0$, it is clear that $p=0$ implies $p_0=0$, contradicting (NT).) Now, if a function and its derivative do not vanish simultaneously, it follows that the zeros of the function are isolated. In our case, this implies that the optimal controls must be bang-bang. So we have proved a simple theorem on the structure of optimal trajectories: *for a minimum time problem in $\mathbb{R}^2$ arising from a control system $\dot{x} = f(x) + ug(x)$, with $|u| \leq 1$, if g and $[f,g]$ are linearly independent at each point, then all optimal trajectories are bang-bang.* The reader can verify that in Example 11.3.1 the Lie bracket $[f,g]$ vanishes along the x axis, so the trajectory of that example is contained in the "singular set" of points where g and $[f,g]$ fail to be linearly independent, in perfect agreement with our theorem.

11.4 Implementing Lagrange's covariance idea in optimal control

Our elementary arguments about Lie brackets and minimum time control provide a good illustration of how Lagrange's principle of "coordinate democracy" applies in the optimal control setting, and why a truly intrinsic formulation of the maximum principle, on manifolds, is needed.

To understand optimal trajectories, one needs to study certain functions, such as the "switching function" ψ. We can think of ψ as a "generalized component" of the (co)vector p, in the sense that the ordinary components p_i of p are the inner products of p with the coordinate vector fields $\frac{\partial}{\partial q^i}$, and ψ is the product of p with the vector field g. The adjoint equation gives us an explicit formula for the functions $\dot{p}_i$. A truly "democratic" version of the maximum principle should treat all generalized components equally, that is, should give us in one swoop the derivative of *every* function $\langle p, X(q)\rangle$, for *every* vector field X.

With such a version one could, in each case, work with the generalized components of p that are most natural for the problem under study, such as, for example, the "switching function" $\langle p, g(q)\rangle$ for a system $\dot{q} = f(q)+ug(q)$. (Another example will be given in our differential geometric analysis of the brachistochrone in Chapter 13.) This would be, we believe, true in spirit to Lagrange's idea that it is good to have a method that "***expresses the equations of every problem in the most simple form relative to***

each set of variables and ... enables us to see beforehand which variables one should use in order to facilitate the integration as much as possible."

An intrinsic version, treating all "generalized components" (sometimes known as "momentum functions") equally, can be formulated in a number of ways, e.g. using Poisson brackets. For a general version of the maximum principle on manifolds, using generalized components as above (and also for an alternative version involving connections along curves), we refer the reader to [118]. The crucial point of the version given in [118] is—considering, for simplicity, the case of the minimum time problem for a control system of the form $\dot{q} = f(q, u)$—that the adjoint equation is rephrased as the following statement (the "intrinsic adjoint equation"):

(IAE) *for every smooth vector field X, the derivative with respect to t of the function $t \mapsto \langle p(t), X(q_*(t))\rangle$ equals $\langle p(t), [f_u, X](q_*(t))\rangle$, where we define $f_u(q) = f(q, u)$*

The point of using (IAE) is that it treats the inner product of with X—i.e., "the X component of p"—equally for all vector fields X, thus enabling us to follow Lagrange's prescription (cf. Page 189) that one should express "the equations of every problem in the most simple form relative to each set of variables and ... enables us to see beforehand which variables one should use in order to facilitate the integration as much as possible," by working in each case with the functions of the form $\langle p, X\rangle$ that are most suited to the problem.

11.5 Why the language of Lie brackets is natural for optimal control

The preceding discussion, based on rather trivial examples, is meant to illustrate what we have in mind when we talk about "understanding the qualitative properties of trajectories" and why we say that the appropriate language for discussing this question and formulating conditions is that of Lie brackets of vector fields.

The reader should think of the following analogy. When one tries to solve a simple 2×2 system of linear equations such as $2x + 3y = 5$, $6x - 2y = 4$, one may fail to notice that the number $2.(-2) - 3.6$ is important. If, however, one "does it with letters" (as a Calculus student would say), i.e., sets out to solve a ***general*** system $a_{11}x + a_{12}y = b_1$, $a_{21}x + a_{22}y = b_2$, then

one discovers immediately that the number $a_{11}a_{22} - a_{12}a_{21}$ determines the properties of the solutions, so it becomes natural to gives this number a name such as "determinant," and to look for a generalization of the concept to higher dimensions. Similarly, when one applies the maximum principle to study a particular problem such as the one of Example 11.3.1, one is in fact computing a Lie bracket, although this fact may go unnoticed. But when one "does it with letters," e.g. by working with a general problem of the form $\dot{x} = f(x) + ug(x)$, leaving f and g unspecified, then the Lie bracket $[f, g]$ stands out immediately as the object that "determines" what happens, exactly as the expression $a_{11}a_{22} - a_{12}a_{21}$ did in the linear algebra example.

It is a general fact that, when we apply the maximum principle to minimum time problems, we get conclusions of the form "if such and such thing is true of the Lie brackets of the vector fields involved, then the optimal trajectories have such and such properties." This can be ascertained in all kinds of examples, provided that one makes sure that the computations are always done "with letters." But the occurrence of the brackets is due to a profound *a priori* reason, namely, Statement (INV) of the previous section.

In recent years, ever since the advent of "differential geometric control theory," much has been done to determine, in a systematic way, how structural properties of a system, embodied in their Lie bracket relations, relate to properties of the optimal trajectories. For example, it has been known for a long time that for linear minimum time optimal control the "bang-bang property" holds. From a general nonlinear perspective, one can show that the bang-bang property holds when certain Lie brackets B_i can be expressed as linear combinations of some other brackets. Linear systems just happen to be those for which the B_i vanish. This provides an explanation of the linear bang-bang theorem from the nonlinear perspective, as well as a generalization.

Chapter 12

Some open problems

We supplement our introduction to what may lie beyond the maximum principle by challenging the reader with three open problems. Our challenges are not like the one issued by Johann Bernoulli, since we do not know the answers to our questions.

Open Problem 1. Consider a minimum time problem corresponding to a controlled dynamics of the form $\dot{q} = f(q, u)$, where (i) u takes values in a set U which is a closed cube, or a ball, or, more generally, a compact subanalytic subset of $\mathbb{R}^m$, (ii) q takes values in an open subset Q of $\mathbb{R}^n$, (iii) f is real-analytic as a function of q and u. Suppose a trajectory-control pair $\gamma_* = (q_*, u_*)$ (with u measurable and U-valued) of this system, defined on the interval $[0, T]$, is a solution of the minimum time problem with initial state $\bar{q}$ and terminal state $\hat{q}$. Suppose this problem has a unique solution (except for time translates of γ_*). *Does it follow that*

(1) u is real-analytic on the complement of a closed subset C of $[0, T]$ of measure zero,

or, perhaps, even more strongly, that

(2) u is real-analytic on the complement of a countable closed subset C of $[0, T]$?

The requirement that the minimum time problem have a unique solution is added to exclude trivial cases where every control is optimal. For example, for the system $\dot{x} = 1$, $\dot{y} = u$, $|u| \leq 1$, evolving in $\mathbb{R}^2$, it is clear that any trajectory going from (x_1, y_1) to (x_2, y_2) must do so in time $x_2 - x_1$, so all trajectories are optimal, and in particular there is no limit to the pathology of an optimal control.

The problem, as stated, is open even for simple systems of the form of the form $\dot{q} = f(q) + ug(q)$, $|u| \leq 1$, in $\mathbb{R}^3$. For these simple systems, the answer in $\mathbb{R}^2$ is known: (2) follows, and in fact, even more strongly, C can actually be taken to be finite. Examples in $\mathbb{R}^3$ show that the answer there is not as simple as for $\mathbb{R}^2$, since there are problems for which the unique optimal control has infinitely many points of discontinuity.

It is known in general that the following weaker conclusion does follow:

(0) *u is real-analytic on an open dense subset of $[0, T]$.*

This result depends crucially on real analyticity, and is not valid for C^∞ systems. In fact, if $u_* : [0, 1] \mapsto [-1, 1]$ is a completely arbitrary measurable function, then one can always find a pair (f, g) of C^∞ vector fields in $\mathbb{R}^3$ and two points $\bar{q}, \hat{q} \in \mathbb{R}^3$, such that u_* is the unique solution of the problem of steering $\bar{q}$ to $\hat{q}$ in minimum time by a trajectory of $\dot{q} = f(q) + ug(q)$, $|u| \leq 1$. So for C^∞ systems the only general answer to the question whether there are any nontrivial regularity properties shared by all uniquely optimal controls is that there are no such properties. For the real-analytic case, the answer is "yes," since (0) is an example of such a property. The question is whether (0) can be strengthened to (1), or perhaps even (2), or whether there are counterexamples.

For more details on this problem, the reader may consult [112, 113]. ♢

Open Problem 2. Let $f_1, \ldots, f_m$ be C^∞ vector fields on an open subset Q of $\mathbb{R}^n$. Let Λ be the Lie algebra of vector fields generated by $f_1, \ldots, f_m$. Suppose $\{g(q) : g \in \Lambda(q)\} = \mathbb{R}^n$ for every $q \in Q$. Consider the minimum time problem for the control system $\dot{q} = u_1 f_1(q) + \ldots + u_m f_m(q)$, with $u = (u_1, \ldots, u_m) \in U$, where U is the closed unit ball in $\mathbb{R}^m$. Assume $f_1(q), \ldots, f_m(q)$ are linearly independent at each point.

Are all minimum time trajectories smooth?

This is the problem of the smoothness of sub-Riemannian minimizers. Several papers with incorrect answers have been published. By applying the maximum principle, one can easily show that the "normal extremals" are smooth, but the regularity properties of abnormal extremals are not well understood. Examples of optimal abnormal extremals have been given in the literature, but they are all smooth. Nonsmooth abnormal extremals are easy to produce, but none so far have been proved to be optimal.

For more details on this problem and a long list of references, the reader is referred to [85]. ♢

Finally, for those readers who find the preceding two questions too general and abstract, we close with a very concrete problem.

Open Problem 3. Consider the minimum time problem for the system in $\mathbb{R}^4$ whose law of evolution is $\dot{x} = \cos\alpha$, $\dot{y} = \sin\alpha$, $\dot{\alpha} = \beta$, $\dot{\beta} = u$, $|u| \leq 1$. *What is the structure of the optimal trajectories? In particular: do all optimal controls for this problem have at most countably many points of nonanalyticity?*

The above system is a "unicycle" whose angular acceleration is being controlled. A simple computation using the maximum principle yields three "basic" families of optimal trajectories, namely, P, N and Z. The P and N trajectories, corresponding to $u = 1$ and $u = -1$, respectively, are clotoids (also known as "Cornu spirals"). The Z trajectories correspond to $u = 0$ and only occur when $\beta = 0$, in which case the unicycle is moving along a straight line in the (x, y) plane. A more careful analysis of the conclusions arising from the maximum principle shows that there cannot be a junction from a P or an N trajectory to a Z trajectory. On the other hand, the possibility of "infinite chattering" (i.e., an alternation of infinitely many P and N pieces, on intervals $[t_j, t_{j+1}]$, with $t_j < t_{j+1}$ and $\lim_{j\to\infty} t_j = \bar{t} < \infty$) cannot be excluded, and it is natural to conjecture that such an infinite chattering occurs.

In fact, if the fixed endpoints $\bar{q} = (\bar{x}, \bar{y}, \bar{\alpha}, \bar{\beta})$, $\hat{q} = (\hat{x}, \hat{y}, \hat{\alpha}, \hat{\beta})$ are such that $(\bar{x}, \bar{y})$ and $(\hat{x}, \hat{y})$ are very far, then one would expect that the optimal trajectory connecting them will involve rectilinear motion most of the time. However, rectilinear motion alone cannot drive $\bar{q}$ to $\hat{q}$ unless the angles $\bar{\alpha}$, $\hat{\alpha}$ are right—that is, $\bar{\alpha} = \hat{\alpha}$ and the line $t \mapsto (\bar{x} + t\cos\bar{\alpha}, \bar{y} + t\sin\bar{\alpha})$ passes through $(\hat{x}, \hat{y})$—and $\bar{\beta} = \hat{\beta} = 0$.

So in general one expects that the optimal trajectory will involve an initial maneuver to position the vehicle in the appropriate line, followed by a long stretch of rectilinear motion, followed by a terminal maneuver to get the vehicle to the desired target position. The initial and terminal maneuvers should involve concatenations of clotoids, but cannot consist of *finite* concatenations, for a finite set has a first and a last element, and then there would be a forbidden junction between a clotoid and a line. The next simplest guess is that they involve infinite chattering as described above, but this has not been proved, and even more pathological possibilities, such as an uncountable set of control switchings, could in principle occur.

For a general discussion of infinite chattering, we refer the reader to the book [136]. Some partial results on the unicycle with angular acceleration

control are given in [122]. ♢

Chapter 13

The brachistochrone, 300 years ago and today

We now look at the brachistochrone problem from various perspectives, starting with Johann's Bernoulli's analysis, [1] and then proceeding on to the calculus of variations and optimal control solutions, and the differential-geometric approach, showing in each case how the later method yields better information than its predecessors.

13.1 Johann Bernoulli's challenge

The June 1696 challenge took the form of an "Invitation to all Mathematicians to solve a new problem":

[1] Jakob's method was quite different fromm Johann's, and at first sight seemed clumsier, but in the long run it has turned out to be more akin to the mainstream ideas of the calculus of variations, Hamilton-Jacobi theory, and dynamic programming, and is therefore widely considered to be of great historical importance in the development of optimal control, as shown in our earlier quotation from Stillwell. It will not, however, be discussed here, due to lack of space. Goldstine's book [59] give an excellent account.

> *If in a vertical plane two points A and B are given, then it is required to specify the orbit AMB of the movable point M, along which it, starting from A, and under the influence of its own weight, arrives at B in the shortest possible time. So that those who are keen of such matters will be tempted to solve this problem, is it good to know that it is not, as it may seem, purely speculative and without practical use. Rather it even appears, and this may be hard to believe, that it is very useful also for other branches of science than mechanics. In order to avoid a hasty conclusion, it should be remarked that the straight line is certainly the line of shortest distance between A and B, but it is not the one which is traveled in the shortest time. However, the curve AMB—which I shall disclose if by the end of this year nobody else has found it—is very well known among geometers.*

COMMENT NO. 36. IN THE EARLIER VERSION WE HAD "DIVULGE." I LIKE "DISCLOSE" BETTER. IT'S CERTAINLY BETTER ENGLISH.

Later, Johann Bernoulli changed his original plan to disclose the solution by the end of 1696 and, following a suggestion of Leibniz, extended the deadline until Easter 1697.

COMMENT NO. 37. PUT PICTURE OF PAGE WITH JOHANN'S LATIN TEXT AROUND HERE

Responses came from the best mathematical minds of the time. Besides Johann's own solution, there was one by Leibniz, who had communicated it to Johann in a letter dated June 16, 1696; another one by Johann's elder brother Jakob; one by Tschirnhaus; one by l'Hôpital and, finally, one by Newton, which was presented to the Royal Society on February 24, 1697 and published, anonymously and without proof, in the *Philosophical Transactions*. (The identity of the anonymous author was clear to Johann Bernoulli, since he thought that you can know *ex ungue leonem*, i.e., that you can know the lion from its claws.) The May 1697 issue of *Acta Eruditorum* contains articles with Johann's and Jakob's solutions, contributions by Tschirnhaus[2] and l'Hôpital[3], a short note by Leibniz, remarking that he would not reproduce his own solution, since it was similar to that of the Bernoullis, and Newton's anonymous paper, reprinted from the *Philosophical Transactions*.

[2]See Scholium III.
[3]See Scholium IV.

SCHOLIUM IX. The cycloid's double role as the brachistochrone and the tautochrone

As is apparent from Johann Bernoulli's announcement, he was under the impression that the problem was new. However, Leibniz knew better: Galileo in his book on the *Two New Sciences* in 1638 formulated the brachistochrone problem and even suggested that the solution was a circle. (Galileo had showed—correctly—that an arc of a circle always did better than a straight line—except, of course, when B lies on the same vertical line as A—but he then wrongly concluded from this that the circle is optimal.

Johann Bernoulli considered the fact that Galileo had been wrong on two counts, in thinking that the catenary was a parabola, and that the brachistochrone was a circle, as definitive evidence of the superiority of differential calculus (or the *Nova Methodus* as he called this new set of techniques, following Leibniz).

He was thrilled by his discovery that the brachistochrone was a cycloid. This curve had been introduced by Galileo who gave it its name: *related to the circle*. Huygens had discovered that the the cycloid has a remarkable property: it is the only curve with the property that a body falling under its own weight is guided by this curve in such a way that it will oscillate with a period that is independent of the initial point where the body is released.[4] Therefore Huygens called this curve, the cycloid, the *tautochrone* from $\tau\alpha\grave{\upsilon}\,\tau\acute{o}\varsigma$: equal, and $\chi\rho\acute{o}\nu o\varsigma$: time).

Johann Bernoulli was amazed and somewhat puzzled, it seems, by the fact that the same curve had these two remarkable properties related to the time traveled on it by a body falling under its own weight:

[4]Contrary to what Galileo thought, the circle has this property only approximately: the period of oscillation of a pendulum is a function of its amplitude.

> Before I end I must voice once more the admiration that I feel for the unexpected identity of Huygens' tautochrone and my brachistochrone. I consider it especially remarkable that that this coincidence can take place only under the hypothesis of Galilei [HJS&JCW: that is, "that the velocities of falling bodies are to each other as the square roots of the altitudes traversed"], so the we even obtain from this a proof of its correctness. Nature alwasy tend to act in the simplest way, and so here lets one curve serve two different functions, while under any other hypothesis we would need two curves, one for tautochrone oscillations, the other for the most rapid fall. If, for example, the velocities were as the altitudes, then both curves would be algebraic, the one a circle, the other one a straight line.

13.2 Why was the brachistochrone so important in 1696-97?

Perhaps the crucial fact about Johann Bernoulli's question is that we are told, very clearly, that ***the solution is a curve that everybody knows, and yet there appears to be no way to guess that the solutions is a cycloid by means of some intuitive geometric argument.***

This situation should be compared with that arising in another famous curve minimization problem, the ***isoperimetric problem***. In the isoperimetric problem, we are asked to find the simple closed curve that encloses the largest possible area among all curves of a given length, and the answer also turns out to be "a curve that everybody knows," namely, a circle. But in this case it is fairly easy to guess directly that the answer is a circle, even though it is a delicate matter to give a rigorous proof. For example, it suffices to observe that the question is rotationally symmetric, so the answer ought to be a rotationally symmetric curve, and the only such curve is the circle[5]. This argument is rather convincing, and makes it quite obvious that

[5]This argument is not a rigorous proof for at least two reasons. First, it assumes existence of solutions. The second and most serious flaw in the argument is that it requires uniqueness of the solution, up to translations. Indeed, it might conceivably happen that the solution is not unique, and in that case the only conclusion one can draw from rotational symmetry is that every rotation will map a solution to a solution, which may be different. Consider, for example, the problem of *minimaizing*—rather than maximizing—enclosed area. If we insist on working in the space of simple closed curves, then the problem has no solution. If we work in the slightly larger space of closed curves that are not necessarily simple, then the problem has many solutions, since any curve that encloses zero area is a solution. Rotations map solutions to solutions, so the family of all solutions has rotational symmetry, but the individual solution curves are not rotationally symmetric.

the solution must be a circle.

Therefore, if Johann Bernoulli had challenged "the best mathematical minds" of his time

COMMENT NO. 38. CHECK TO SEE IF THESE WERE JB'S EXACT WORDS

to solve the isoperimetric problem, it would not have made sense for him to say, as he did in his brachistochrone challenge, that "the curve—which I shall disclose if by the end of this year nobody else has found it—is very well known among geometers." Everybody would have guessed immediately that the mysterious curve was the circle.

In the case of the brachistochrone, ***there is no way to guess***[6]. The problem was posed and solved at a time when the Calculus was being invented, and cannot be solved without using Calculus. So it provides an impressive illustration of the power of the Calculus to solve problems that cannot be solved in any other way, and to discover truths about the world that can be stated and understood in "well known" terms[7], but can only be arrived at by using this new mathematical tool.

13.3 Formulation of the brachistochrone problem

We now present the formulation of the brachistochrone problem in modern mathematical language, as it is usually presented in modern books. (We will make a case below for a different formulation.) Choose x and y axes in the plane with the y axis pointing downwards. Use (a_1, a_2) and (b_1, b_2) to denote, respectively the coordinates of the end points A and B. Fix a number $E \in \mathbb{R}$. A path $f : [0, T] \to \mathbb{R}^2$, defined on an interval $[0, T]$, and having components $f_1(t), f_2(t)$, is said to be a *feasible trajectory* (or *feasible path*) if

(i) $f(0) = A$, $f(T) = B$, and f is Lipschitz continuous,

[6]Even now, 300 years later, we are not aware of any simple geometric argument that might lead to the cycloid without using Calculus.

[7]This is an important point. Compare the brachistochrone, for example, with Newton's 1685 work on the shape of a body with minimal drag. In this case, the solution also requires the Calculus, but once we have solved the problem all we find is a curve defined by some strange formula. This curve was not "well know to geometers," and would not have had any particular significance to mathematicians. who would only retain from going through Newton's solution the fact that the solution is "some curve." In the case of the brachistochrone, one does not the Calculus to explain what the solution is, since cycloids were already well understood curves, but one does need it to discover that the solution is a cycloid.

(ii) $\frac{1}{2}(|\dot{f}_1(t)|^2 + |\dot{f}_2(t)|^2) = E + g f_2(t)$ for almost all $t \in [0, T]$,

where g is the gravitational constant.

Condition (i) states that the path f must start at A and end at B. The number $E + ga_2$ is the initial kinetic energy of the body. (Often, only the case $E + ga_2 = 0$ is considered, corresponding to the body being dropped at time 0 with zero initial speed.) Condition (ii) reflects conservation of energy: at each instant t, the kinetic energy of the body must equal its initial kinetic energy $E + ga_2$ plus the potential energy $g(f_2(t) - a_2)$ due to its loss of height. (The law that a body which has fallen from a height h has kinetic energy proportional to h was due to Galileo, and was well known in Bernoulli's time.)

A feasible path $f : [0, T] \to \mathbb{R}^2$ is said to be *optimal* if there exists no feasible path $\tilde{f} : [0, \tilde{T}] \to \mathbb{R}^2$ for which $\tilde{T} < T$. A *brachistochrone* is a curve in $\mathbb{R}^2$ traversed by an optimal feasible path, i.e., a subset B of $\mathbb{R}^2$ of the form

$$B = \{(x, y) \in \mathbb{R}^2 : \text{ there exists } t \in [0, T^*], \text{ such that } (x, y) = f^*(t)\},$$

where $f^* : [0, T^*] \mapsto \mathbb{R}^2$ is an optimal feasible path.

Obviously, a feasible path must be entirely contained in the closed half-plane

$$H_+(\eta) = \{(x, y) : y \geq -\eta\}, \quad \text{where} \quad \eta = \frac{E}{g}, \tag{13.1}$$

since, for $f_2(t)$ to be $< -\eta$, the kinetic energy would have to be negative, which is of course impossible.

The following two facts are true, though neither one is completely obvious:

(1) an optimal feasible path exists for any $A, B \in H_+(\eta)$,

(2) the path of (1) is unique.

One fact which is truly obvious is that *the solution cannot always be a straight line*, a possibility that Johann Bernoulli rightly warns against. For example, consider the extreme case when $E = 0$ and $a_1 = a_2 = b_2 = 0 \neq b_1$. We will show that in this situation there is a feasible path that goes from A to B in finite time. Since the straight-line segment S from A to B is horizontal, (ii) implies that the speed of motion along S vanishes, and then it will follow that S cannot be an optimal path, because the motion along S takes infinite

time. To construct a feasible path from A to B we follow a circle C with center on the x axis. The circle can be parametrized by x, taking values in the interval $[0, b_1]$. The speed v of motion along C will be proportional to $\sqrt{y}$, and hence bounded away from 0 as long as x is bounded away from 0 and b_1. Near A, we can also use y as a parameter, and then $y \downarrow 0$ as $x \downarrow 0$, and v behaves like $\frac{dy}{dt}$ and also like $\sqrt{y}$, so $\frac{dt}{dy} \sim \frac{1}{\sqrt{y}}$, showing that t stays finite as x and y approach 0, since the function $y \mapsto \frac{1}{\sqrt{y}}$ is integrable near $y = 0$. A similar argument shows that t stays finite near B. So C joins A to B in finite time. ♢

13.4 Johann Bernoulli's solution

It turns out that the brachistochrone is a cycloid. When $E = 0$, it is the curve described by a point P in a circle that rolls without slipping on on the x axis, in such a way that P passes through A and then through B, without hitting the x axis in between. (It is easy to see that this defines the cycloid uniquely.)

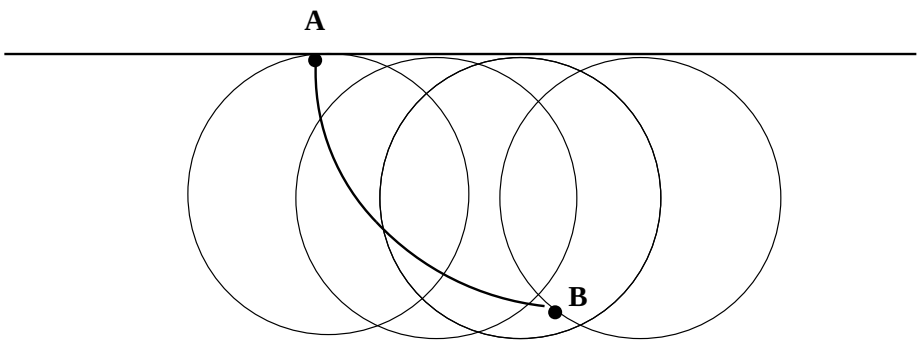


FIGURE 8. The cycloid

Johann Bernoulli's ingenious derivation of the brachistochrone has been the subject of numerous accounts, but since this event plays a crucial role in our own tale, we will briefly tell the story again.

Johann Bernoulli based his derivation on Fermat's minimum time principle. If we imagine for a moment that instead of dealing with the motion of a

moving body we are dealing with a light ray, Condition (ii) above gives us a formula for the "speed of light" c as a function of position: $c = \sqrt{2E + 2gy}$. From now on, let us change coordinates so that $E = 0$. (This amounts to moving the x axis upward to a height η. Then η is now equal to 0, so our feasible paths live in the half plane $H_+ \stackrel{\text{def}}{=} H_+(0)$.) Also, let us rescale—or, if the reader so prefers, "change our choice of physical units"—so that $2g = 1$. Then our problem is exactly equivalent to that of determining the light rays—i.e., the minimum-time paths—in a plane medium where the speed of light c varies continuously as a function of position according to the formula $c = \sqrt{y}$. It is at least intuitively clear that, if we discretize our problem by dividing the half-plane H_+ into horizontal strips $S_k = \{(x, y) : y_k \leq y \leq y_{k+1}\}$ of height δ, for $k = 0, 1, \ldots$, where $y_k = k\delta$, and treating c in each strip S_k as a constant c_k (by, say, setting $c_k = \sqrt{y_{k+1}}$), then the light rays for the discretized problem should approach those for the original problem as $\delta \downarrow 0$.

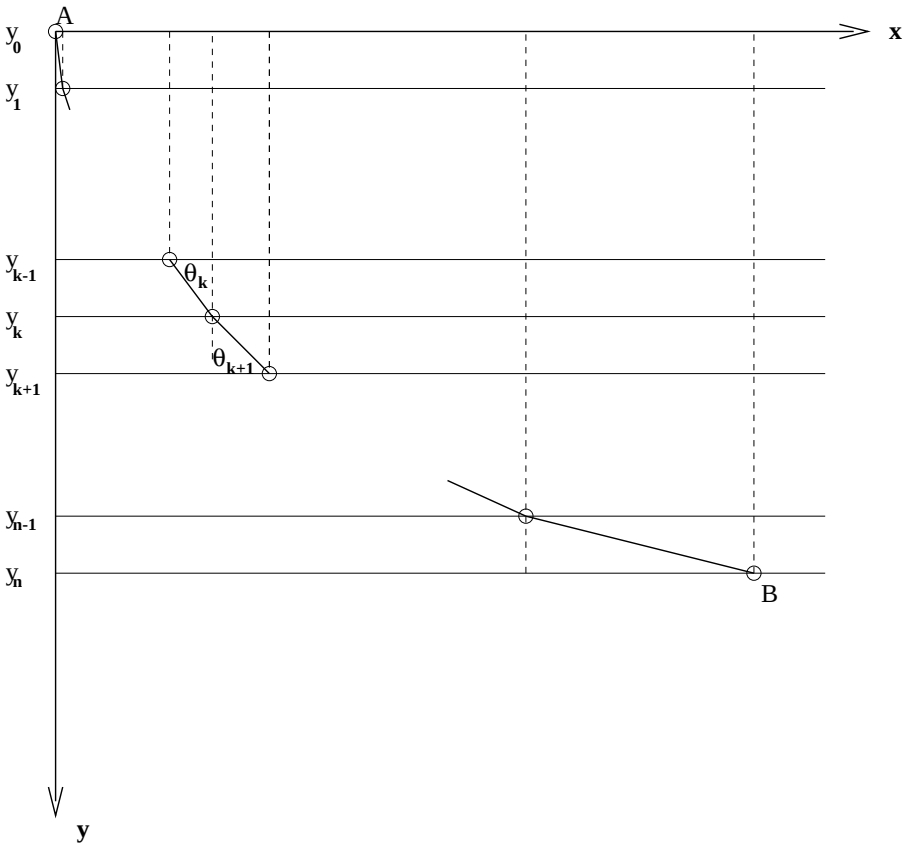


FIGURE 9. Bernouli"s discretization

The light rays of the discretized problem can be studied using Snell' law. Clearly, the paths will be straight-line segments within each individual strip, and all that needs to be done is to determine how these rays bend as they cross the boundary between two strips. The answer is provided by the laws of optics as developed by Snell, Descartes, Fermat, Leibniz and Huygens.

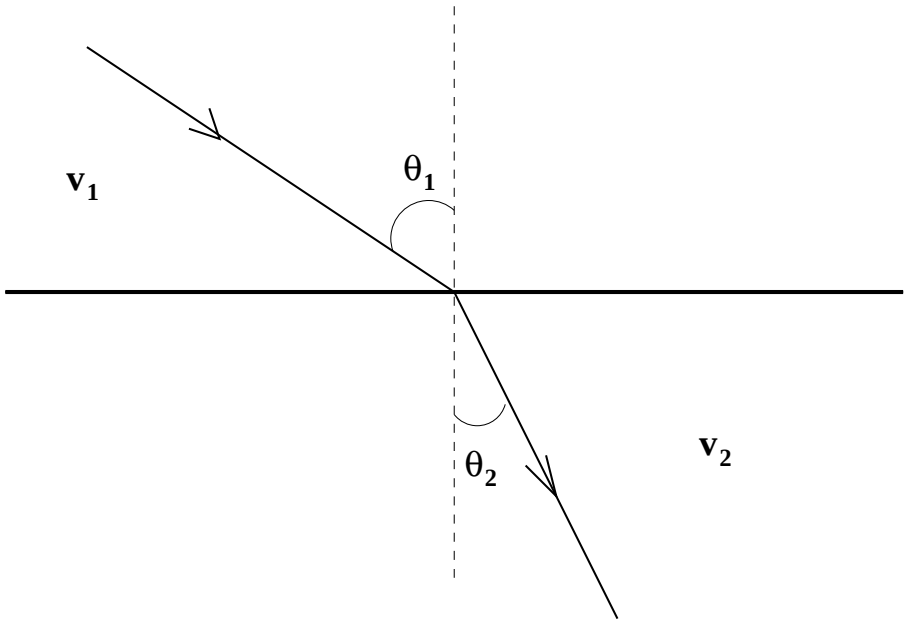


FIGURE 10. Snell's law

Snell had observed that, if two media—in the plane—are separated by a line L, and a light ray is refracted at the boundary between them, then the ratio of the sines of the incidence angles is constant. Fermat subsequently showed that this is precisely what happens when light is assumed to follow a minimum-time path. The law relating the incidence angles to the velocities of propagation is due to Leibniz and Huygens, and implies the law of Snell. It says that

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2} \qquad \text{or, equivalently,} \qquad \frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2}\,,$$

where θ_i is the angle of the line joining P_i to P' with the line perpendicular to L.

Johann Bernoulli used this law to conclude that the quantity $\frac{\sin\theta_k}{\sqrt{y_{k+1}}}$ will be a constant, since in each strip S_k the speed of our light ray is $\sqrt{y_{k+1}}$. Passing to the limit as $\delta \downarrow 0$, we conclude that the sine of the angle θ between the tangent to the brachistochrone and the vertical axis must be proportional to $\sqrt{y}$. Since $\sin\theta = \frac{dx}{\sqrt{dx^2+dy^2}}$, we find that $\frac{dx^2}{dx^2+dy^2} = Ky$, where K is a constant. Then $\frac{dx^2+dy^2}{dx^2} = \frac{1}{Ky}$, i.e., $1 + y'(x)^2 = \frac{C}{y}$, where $C = \frac{1}{K}$. So the curve described by expressing the y-coordinate of the brachistochrone as a function of its x-coordinate will satisfy the differential equation

$$y'(x) = \sqrt{\frac{C - y(x)}{y(x)}}\,, \tag{13.2}$$

with C a constant. The curves parametrized by

$$x(\varphi) = x_0 + \frac{C}{2}(\varphi - \sin\varphi)\,, \qquad y(\varphi) = \frac{C}{2}(1 - \cos\varphi)\,, \qquad 0 \le \varphi \le 2\pi\,, \tag{13.3}$$

satisfy this differential equation. It is easily seen that these equations specify the cycloid generated by a point P on a circle of diameter C that rolls without slipping on the horizontal axis, in such a way that P is at $(x_0, 0)$ when $\varphi = 0$. Moreover, it is also easy to check that

(*) given two points A and B in H_+ there is exactly one curve in this family passing through A and B.

(The parameter count for (*) is obviously right: (13.3) describes a 2-parameter family of curves; requiring that the curve pass through A fixes one of the parameters, and then asking that it pass through B fixes the other. Proving (*) requires a bit more work, but is still quite easy.)

13.5 Remarks on Johann Bernoulli's argument

The argument that we have presented is the one of Johann Bernoulli, and Equation (13.2) is the one that he wrote in his paper, followed by the statement "from which I conclude that the *Brachistochrone* is the ordinary *Cycloid*." (Johann Bernoulli actually wrote $dy = dx\sqrt{\frac{x}{a-x}}$, but he was using x for the vertical coordinate and y for the horizontal one. Cf. [109], p. 394.)

In present day notation, it is customary to adopt the convention that the symbol $\sqrt{r}$ stands for the *nonnegative* square root of r, but it is obvious that Johann Bernoulli did not have this in mind. What he meant was, clearly, what we would write as

$$y'(x) = \pm\sqrt{\frac{C - y(x)}{y(x)}}\,, \tag{13.4}$$

or, equivalently,

$$y(x)(1 + y'(x)^2) = \text{constant}\,. \tag{13.5}$$

In particular, the solution curves should be allowed to have a negative slope, contrary to what a "modern" interpretation of (13.2) might suggest. But y' should stay continuous, so that a switching from a + to a − solution of (13.4) is not permitted.

Even with the more accurate rewriting (13.5), the differential equation derived by Johann Bernoulli also has spurious solutions, not given by (13.3). Indeed, for any $\bar{y} > 0$, the constant function $y(x) = \bar{y}$ is a solution, corresponding to $C{=}\bar{y}$. More generally, one can take an ordinary cycloid given by (13.3), follow it up to $\varphi{=}\pi$—so that $dy/dx{=}0$—then follow the constant solution $y(x){=}C$ for an arbitrary time T, and then continue with a cycloid given by (13.3). Such paths are, indeed, compatible with Huygens' law of refraction.

Remark 13.5.1 The existence of the spurious solutions was pointed out by Brook Taylor (1685-1731) in 1715.

Taylor is known to mathematicians as the discoverer of "Taylor's expansion," the inventor of integration by parts, and the creator of the calculus of finite differences.

Like many other mathematicians of his time, he was involved in a priority dispute with Johann Bernoulli, involving Taylor's 1708 solution of the problem of the center of oscillation, that went unpublished until 1714. ♢

13.6 The calculus of variations approach

The spurious solutions, and all the other problems, such as the apparent arbitrariness of the requirement that y' be continuous, can be eliminated in a number of ways. For example, one can prove directly that the spurious trajectories are not optimal. Or one can use, as an alternative to Johann

Bernoulli's method, the calculus of variations approach, based on writing the Euler-Lagrange equation (4.3).

It is easy to see that the brachistochrone problem can be put in the "standard" form (3.5), provided we ***postulate***[8] that it suffices to consider curves in the x, y plane that are graphs of functions $y = y(x)$ defined on the interval $[a_1, b_1]$. Then Constraint (ii) can be written (taking $E=0$ and $2g=1$ as before) as $dx^2+dy^2=y\,dt^2$, which gives $dt = \frac{\sqrt{dx^2+dy^2}}{\sqrt{y}} = L(y, \dot{y})dx$, where $L(y, u) = \frac{\sqrt{1+u^2}}{\sqrt{y}}$, and we are using x rather than t for the time variable, and writing $\dot{y}$ for dy/dx. So Johann Bernoulli's problem becomes that of minimizing the integral $\int_{a_1}^{b_1} L(y(x), \dot{y}(x))dx$ subject to $y(a_1) = a_2$ and $y(b_1) = b_2$.

This gives the equation

$$1 + y'(x)^2 + 2y(x)y''(x) = 0\,, \tag{13.6}$$

which is *stronger* than (13.5), since (13.5) is equivalent to $y' + y'^3 + 2yy'y'' = 0$, i.e., to $y'(1 + y'^2 + 2yy'') = 0$, whose solutions are those of (13.6) plus the spurious solutions found earlier.

It is easy to verify that the solutions of the Euler-Lagrange equation (13.6) are *exactly* the curves given by (13.3), without any extra spurious solutions, showing that, for the brachistochrone problem, ***the Euler-Lagrange method gives better results than Johann Bernoulli's approach.***

We will soon see that ***optimal control is even better.***

13.7 The brachistochrone as an optimal control problem

We can formulate Johann Bernoulli's question as an optimal control problem, in which the motion takes place in the x, y plane, and the dynamics is given by

$$\dot{x} = u\sqrt{|y|}\,, \quad \dot{y} = v\sqrt{|y|}\,, \tag{13.7}$$

where the control is a 2-dimensional vector (u, v) taking values in the set $U = \{(u, v) : u^2+v^2=1\}$ (cf. Footnote 8).

The Hamiltonian $H(x, y, u, v, p, q, p_0, t)$ is then given (using $\alpha = \operatorname{sgn} y$) by the formula $H = (p_1u + p_2v)\sqrt{\alpha y} - p_0$, and the application of (NCO)

[8] With optimal control, this "postulate" becomes a *provable conclusion*, as shown below.

gives the conditions

$$u = \frac{p_1}{|p|}\,, \quad v = \frac{p_2}{|p|}\,, \tag{13.8}$$

where $|p| = \sqrt{p_1^2 + p_2^2}$, as well as the differential equations

$$\dot{p_1} = 0\,, \quad \dot{p_2} = -\alpha\,\frac{p_1 u + p_2 v}{2\sqrt{\alpha y}} = -\frac{\alpha|p|}{2\sqrt{\alpha y}}\,. \tag{13.9}$$

(Notice that $|p(t)| \neq 0$. Indeed, (MP') tells us that $H = 0$. So $|p| = 0$ would imply $p_0 = 0$, contradicting (NT).)

If the constant p_1 vanishes, then $\dot{x} \equiv 0$, so we get a vertical line. Otherwise, $\dot{x}$ is always $\neq 0$, showing that we can use x to parametrize our solution. Since

$$y'(x) = \frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{v}{u} = \frac{p_2}{p_1}\,, \tag{13.10}$$

we have

$$1 + y'(x)^2 = \frac{|p|^2}{p_1^2} \tag{13.11}$$

and

$$y''(x) = \frac{1}{p_1} \cdot \frac{dp_2}{dx} = \frac{\dot{p}_2}{p_1 \dot{x}}\,. \tag{13.12}$$

But (13.7) and (13.8) imply that

$$\dot{x} = \frac{p_1\sqrt{\alpha y}}{|p|}\,, \tag{13.13}$$

and then (13.9) and (13.12) yield

$$y''(x) = -\frac{|p|^2}{2yp_1^2}\,. \tag{13.14}$$

So $2yy'' = -\frac{|p|^2}{p_1^2} = -(1 + y'^2)$, and then

$$1 + y'^2 + 2yy'' = 0\,, \tag{13.15}$$

which is exactly Equation (13.6). As we showed before, this leads to the cycloids, with no "spurious solutions." Notice that this argument does not involve any discretization or any use of Snell' law of refraction.

Notice also that in our control argument *we have not postulated that the solution curves could be represented as graphs of functions $y(x)$. We have*

proved it! (Remember that in the calculus of variations case this was an extra assumption.)

This is one example showing that, for the brachistochrone problem, ***the optimal control method gives better results than the classical calculus of variations.***

13.8 Optimality proofs and the existence problem

For a second example showing the superiority of the optimal control method over the classical calculus of variations, for the specific case of the brachistochrone problem, we turn to the question of the rigorous proof of the optimality of Bernoulli's cycloids. Clearly, no argument based only on necessary conditions for optimality will ever prove that a trajectory is optimal, because it could happen, for example, that there are no optimal trajectories, in which case the statement that "every optimal trajectory is a cycloid" would be vacuously true. If we really want to prove the optimality of Bernoulli's cycloids, an extra step is needed.

In addition, even the classical calculus of variations and optimal control proofs that every optimal trajectory is a cycloid are of limited validity: they only work for trajectories that are entirely contained in the open half-plane $\overset{o}{H}_+ = \{(x, y) : y > 0\}$.

To give a truly rigorous proof of the optimality of the cycloids, valid not only when A and B are both in $\overset{o}{H}_+$ but also when they are in the closed half-plane $H_+ = \{x, y) : y \geq 0\}$, one needs a more elaborate argument.

One possibility (cf. Bliss [14]) is to use Hamilton-Jacobi theory. Another, much simpler approach, is to prove first the *existence* of an optimal trajectory. Indeed, once this is known for all $A, B \in H_+$, then it is clear that any optimal trajectory $q_* : [0, T] \mapsto H_+$ going from A to B must be such that its restriction $\xi \lceil [\alpha, \beta]$ to every subinterval $[\alpha, \beta]$ of $[0, T]$ is also optimal for the problem $P(\alpha, \beta)$ with endpoints $q_*(\alpha)$, $q_*(\beta)$. (This is a rather obvious point, known in optimal control theory as the "principle of optimality," and noticed in the case of the brachistochrone by Jakob Bernoulli.) If one assumes that $q_*(\alpha)$ and $q_*(\beta)$ belong to $\overset{o}{H}_+$ for $0 < \alpha < \beta < T$, then the necessary conditions for optimality give a unique candidate for the solution of $P(\alpha, \beta)$, namely, the unique curve in the family given by (13.3) that goes through $q_*(\alpha)$ and $q_*(\beta)$. So ξ is such that $\xi \lceil [\alpha, \beta]$ is given by (13.3) whenever $0 < \alpha < \beta < T$, and this easily implies that ξ itself is a cycloid given

by (13.3). To complete the argument, one has to exclude the possibility of a solution of the problem with endpoints A, B which intersects the x axis at some other point. This can be done by a simple qualitative argument that we omit. (The key lemma is this: suppose L is a horizontal line in the upper half-plane H_+. Let P, Q be distinct points of L. Then the straight-line segment joining P to Q is faster than any trajectory from P to Q that lies entirely below L.)

So everything hinges upon the existence result. It turns out that this is not at all trivial if the brachistochrone problem is reformulated in terms of the classical calculus of variations, because the Lagrangian $L(y,u) = \frac{\sqrt{1+u^2}}{\sqrt{y}}$ has a singularity at $y = 0$.[9] (Existence theorems for classical calculus of variations problems that include the brachistochrone can be found in Cesari [36], §14.3 and §14.4. But they are not at all simple.)

In the control theory setting, however, the existence question is easily settled by a completely trivial application of Ascoli's theorem: given $A, B \in H_+$, let T be the infimum of the times of all trajectories going from A to B; let $\{\xi_j\}$ be a sequence of trajectories from A to B, defined on intervals $[0, T_j]$, such that $T_j \downarrow T$. Using Ascoli's theorem, extract a subsequence that converges uniformly to a curve $\xi : [0,T] \mapsto H_+$. In principle, ξ is just a trajectory of the "convexified" system, in which the evolution equations are those of (13.7) but the control constraint is $u^2 + v^2 \leq 1$ rather than $u^2 + v^2 = 1$. However, it is easy to see that ξ is a true trajectory of the nonconvexified system because, if the inequality $u(t)^2 + v(t)^2 \leq 1$ was strict on a set of positive measure, then it would be possible to find an even faster trajectory from A to B by just reparametrizing ξ. So we see, once again, that *in the control setting a question that is hard from the point of view of the calculus of variations can become easy or even trivial.*

The key point in the above argument is that there is at least one trajectory going from A to B in finite time, as can easily be verified. The argument would not work if the speed of light was y rather than $\sqrt{y}$, for in that case there is no path joining A and B in finite time, unless both A and B belong to $\overset{o}{H}_+$ or are equal.

[9] We thank F. H. Clarke for bringing this point to our attention.

13.9 Hestenes' "variation of the brachistochrone problem"

In 1956, M. Hestenes wrote (in [66], p. 64):

> ... it is interesting to speculate on what type of brachistochrone problem the Bernoulli brothers might have formulated had they lived in modern times. An interesting variation of the brachistochrone problem can be described as follows: Consider an idealized rocket ship moving in a vertical plane. The ship is to be considered as a particle acted upon by a gravitational force g and a thrust of constant magnitude F and variable inclination φ ... There are no other forces acting on this ship. The pilot steers the ship by controlling the inclination φ, that is, by controlling the direction of the thrust. The problem is to find a path of least time under suitable initial and terminal conditions. Analytically, one seeks to find among all functions
>
> $$x = x(t) \qquad y = y(t) \qquad \varphi = \varphi(t) \qquad 0 \leq t \leq T$$
>
> subject to the constraints (with unit mass)
>
> $$\ddot{x} = F\cos\varphi \qquad \ddot{y} = F\sin\varphi - g$$
>
> and with x, y, $\dot{x}$, $\dot{y}$, prescribed at $t = 0$ and $t = T$, one which minimizes the time of flight T.
> It can be shown that the path of least time has the following interesting property: there is a line L with direction fixed in space, and with position fixed relative to the ship, such that if a particle M is moving along L with constant speed, the ship will traverse a path of least time by keeping the thrust aimed at the particle M.

Hestenes later on proceeds to solve the problem. Let us present a solution based on optimal control. The dynamical equations can be rewritten as

$$\dot{x} = z \qquad \dot{z} = F\cos\varphi \qquad \dot{y} = w \qquad \dot{w} = F\sin\varphi - g\,.$$

If $[0,T] \ni t \mapsto (x_*(t), z_*(t), y_*(t), w_*(t)) \in \mathbb{R}^2$ is an optimal trajectory, corresponding to a control function $[0,T] \ni t \mapsto \varphi_*(t) \in \mathbb{R}$, then the maximum principle yields a nonnegative constant π_0 and and a vector-valued function

$$[0,T] \ni t \mapsto \pi(t) = (\pi_x(t), \pi_z(t), \pi_y(t), \pi_w(t))\,.$$

The Hamiltonian is given by

$$H(x, z, y, w, \varphi, p_x, p_z, p_y, p_w, p_0) = p_x\, z + p_z\, F\cos\varphi + p_y\, w + p_w\,(F\sin\varphi - g) - p_0\,.$$

The adjoint equation becomes

$$\dot\pi_x = 0 \qquad \dot\pi_z = -\pi_x \qquad \dot\pi_y = 0 \qquad \dot\pi_w = -\pi_y\,.$$

Then, if we write $\omega(t) = (\pi_z(t), \pi_w(t))$, we have

$$\omega(t) = \omega_0 + t\omega_1\,,$$

where ω_0 and ω_1 are constant two-dimensional vectors.

The Hamiltonian maximization condition then says that the vector

$$\mathbf{u}(t) = (\cos\varphi(t), \sin\varphi(t))$$

is given by

$$\mathbf{u}(t) = \frac{\omega(t)}{\|\omega(t)\|}\,. \tag{13.16}$$

The Hamiltonian must vanish along our trajectory. Therefore, $\pi(t)$ can never vanish, for if $\pi(t) = 0$ it would follow that $\pi_0 = 0$, contradicting the nontriviality condition. Hence $\omega(t)$ cannot vanish identically. So ω is a linear function which is not identically equal to zero. Therefore there is at most one value of t such that $\omega(t) = 0$. It follows that the function $\mathbf{u}$ of (13.16) is well defined except possibly at one point $\bar t$. Moreover, if such a $\bar t$ exists, then $\omega(t) = (t - \bar t)\omega_1$, so

$$\mathbf{u}(t) = \begin{cases} -\omega_1 & \text{if} \quad t < \bar t \\ \omega_1 & \text{if} \quad t > \bar t. \end{cases}$$

The "line L" of Hestenes' "interesting property" is the line L_t with direction ω_1, with a point λ_t having a "position fixed relative to the ship" given by

$$\lambda_t = (x_*(t), y_*(t)) + \omega_0\,.$$

So in fact L_t is moving along with the ship. If a point q_t on L_t moves with constant velocity ω_1, starting at λ_0 at time 0, then the point's position at time t is $\lambda_t + t\omega_1$. Therefore the vector from the ship's position to q_t is precisely $\omega(t)$. We have shown that the direction of the thrust is $\omega(t)$, from which Hestenes' "geometric property" follows.

13.10 The reflected brachistochrone problem

So far, our analysis of the brachistochrone problem was carried out in a half-plane, by looking only for optimal trajectories that are entirely above the x axis, as in Johann Bernoulli's original problem.

Let us pursue Hestenes' idea that "it is interesting to speculate on what type of brachistochrone problem the Bernoulli brothers might have formulated had they lived in modern times." An obvious question for the brothers to ask would have been that of finding out what happens if we work in the whole plane. This is in fact a much more natural mathematical setting for the brachistochrone problem. We shall refer to this new question as the *reflected brachistochrone problem.* (It is because we were thinking all along of this extension that we wrote $\sqrt{|y|}$ rather that $\sqrt{y}$ in (13.7)).

Solving this more general problem amounts to finding the light rays when the medium is *the whole plane*, and the speed of light is $\sqrt{|y|}$. Notice that this problem is "completely controllable," in the sense that any two points A, B of $\mathbb{R}^2$, even if they lie on opposite sides of the x axis, can be joined by a feasible path. The right-hand side of (13.7) vanishes along the x axis, but this does not prevent the existence of feasible paths crossing the x-axis, because the function $\sqrt{|y|}$ is not Lipschitz near the x axis. (If the function was Lipschitz, then by the usual uniqueness theorem of ordinary differential equations, every solution going through a point on the x axis would have to be a constant curve.) However the same non-Lipschitz feature that makes the system controllable also renders the maximum principle inapplicable, in its classical and nonsmooth versions, including the Lojasiewicz version, since all these require a Lipschitz reference vector field.

Suppose we want to find the light rays for a medium which is the whole plane, with "speed of light" equal to $\sqrt{|y|}$. Can the necessary conditions help us find the solution? Suppose, for example, that we want to find an optimal trajectory from A to B, where A lies in the upper half-plane and B is in the lower half-plane. Then one can show, first of all, that an optimal trajectory ξ exists, using Ascoli's theorem as before. Next, using the usual necessary conditions for optimality, e.g. the Euler-Lagrange equation or the classical version of the maximum principle, one shows that any portion of an optimal curve which is entirely contained in the closed upper half plane or in the closed lower half plane is a cycloid given by (13.3), or a reflection or such a cycloid with respect to the x axis. Next, one sees that ξ cannot traverse the x axis more than once. (This requires an elementary qualitative lemma that we leave as exercise.) So we know that ξ consists of a cycloid

Γ_1 going from A to a point X in the x axis, followed by a second cycloid Γ_2 going from X to B. It remains to find X.

It turns out that ***the version of [119] applies***, since this result does not require Lipschitz continuity—or even continuity—of the right-hand side, and works as long as the reference trajectory arises from a semidifferentiable flow. The result is that ***the cycloids*** Γ_1***,*** Γ_2 ***must be such that the quantity***

$$C = |y(x)|(1 + y'(x)^2)\,. \tag{13.17}$$

which is constant along each Γ_i***, actually has the same value for both curves***. This extra condition determines X. We refer the reader to [119] for the details.

COMMENT NO. 39. JAN: Actually, I am not sure that "the details" are done in my Kobe CDC paper. (I'd have to check. I cannot do it right now.) And I am not sure that, even if the details turn out to be done there, it is a good idea to "refer the reader." Maybe we should put in the full computation. What do you think? If you are in favor, I volunteer to do it. I probably have the whole thing on line somewhere, so it should not be too much work, if I can find it.

The "extra condition" has a simple geometric meaning. Recall that the cycloids of Johann Bernoulli's problem are generated by a circle rolling along the x axis. The extra condition then says that

the rolling circles that generate the upper and lower cycloids must have equall radii.

COMMENT NO. 40. JAN: wouldn't it be nice to put a figure here of the two cycloids, displaying the similarity of this situation with that of Snell's law, with the segments replaced by cycloids?

13.11 A modest claim

If we are going to "speculate" as Hestenes did, it is natural to ask not only what questions the Bernoulli brothers would have asked if they had lived in modern times, but also what kinds of answers they would have liked.

This is of course a hard question to answer, and anything we say will have to be highly subjective. Granted this, we are firmly convinced that ***the***

Bernoulli brothers would have liked the "interesting property" of the solution of Hestenes' problem (which is almost certainly the reason why Hestenes chose to emphasize this property) and also that they would have found elegant the geometric condition of our solution of the reflected brachistochrone problem. We believe this should be obvious to all mathematical readers but, since *de gustibus non est disputandum*, we shall make no attempt to give reasons for this purely aesthetic assertion.

COMMENT NO. 41. JAN: I think this is a very important point, and the case we are making seems extremely convincing to me, but I am of course far from a neutral objective observer. Please think it over. If you agree that is very convincing, then maybe you can find a better way to say what I am trying to say, or you can add some more stuff on this?

13.12 Hamilton's principal function for the brachistochrone problem, and the differential-geometric connection

The mathematical formulation of the brachistochrone problem, as presented so far, is not completely natural, because it takes it for granted that we know that energy is conserved. More precisely, "energy conservation" is "assumed" in at least two conceptually different ways. First, ***the fact of energy conservation is a prerequisite for even stating the problem***. This is clearly not intellectually satisfactory, because it would be much better to state the problem first, and then use energy conservation to solve it (or, even better, to let the mathematical analysis discover energy conservation for us). *Second*, ***the problem of "finding the minimum time curve from A to B" does not make sense for general A and B, unless the value of the energy E is specified***. The problem usually studied in textbooks is another one, namely, that of "finding the minimum-time curve from A to B subject to an extra constraint $E = \bar{E}$, for a given $\bar{E}$." If this constraint is not given, then we really have, for a given A, B, a one-parameter family of problems, one for each value of $\bar{E}$. (Alternatively, we may want to look at the *unconstrained* problem, in which we are also minimizing over all values of $\bar{E}$. But this is clearly not what Johann Bernoulli or anyone else who studied the problem had in mind, and in addition is easy to solve: the infimum of the times is zero, and there is no minimizer.)

13.13 What is Hamilton's principal function for the brachistochrone?

Another way to see why there is a problem is to ask ***what is the state space of the brachistochrone problem, and what is its "principal function" in Hamilton's sense?*** Or, using the precise definitions of the appendix, ***what is, exactly, the controlled dynamical system with cost that corresponds to the brachistochrone problem?*** Let $\mathcal{C} = (S, U, f, \mathcal{A}, J)$ be this system. It is clear that J should be time. But what, exactly, should we choose for S? Can we take $S = \mathbb{R}^2$? This would mean that one can specify arbitrary points A and B and define $V_{\mathcal{C}}(A, B)$ unambiguously. But we know that this is not so. There is a well defined value $V^E(A, B)$ for each value of E, but these values are in general different, and the infimum of all of them is zero.

A common way to get around this difficulty is by stipulating that we are only interested in dropping the particle with zero initial kinetic energy. (One can even interpret Johann Bernoulli's statement of the problem, quoted in Chapter 13, as suggesting this possibility.) This, however, will not do, because it would have the following unnatural consequence: *if we pick two points CD in an optimal arc Γ going from A to B, and use $\Gamma(C, D)$ to denote the piece of Γ from C to D, then $\Gamma(C, D)$ will not in general be an optimal arc from C to D.* (The reason should by now be obvious: the kinetic energy of the particle that started falling with zero velocity at A will not in general have zero velocity at C.) In modern terminology, *the problem with zero initial kinetic energy does not satisfy the principle of optimality.*[10]

Perhaps the state space is bigger than the plane? In other words, it might be the case that one can specify the position and something else at the starting time and a similar object at the terminal time. This will not do either. We cannot, for example, specify the initial and terminal positions and velocity vectors, or even just the positions and the lengths of the velocities. What can be specified is A, B, and the energy—or something yielding equivalent information, e.g. the length of the initial velocity vector—and from this there is no way to derive a state space S such that V is defined on S.

[10] Jakob Bernoulli's solution of the brachistochrone problem makes heavy use of the fact that a portion of a minimizer is itself a minimizer. So Jakob cannot possibly have been thinking that one should only consider the case of zero initial kinetic energy.

COMMENT NO. 42. WE NEED ONE MORE SUBSECTION HERE. Topic: the Hestenes problem and the brachistochrone are closely related. In both cases we are moving in $\mathbb{R}^2$ and controlling acceleration. In the BP we control acceleration by requiring that acceleration is that of gravity plus something perpendicular to the velocity (D'Alembert's virtual force principle). In Hestenes acceleration=gravity plus thrust. So in both cases the problem should have $\mathbb{R}^4$ as its state space. Then how come the BP is formulated in $\mathbb{R}^2$? It should not be!!! And how come for the BP we specify initial and terminal positions, but for Hestenes we specify initial and terminal positions AND VELOCITIES? Isn't that an anomaly? Then we should give a first hint of the answer that is coming later: the BP is also a 4-dimensional problem, really, but it is anomalous in two ways: first, because there is a nontrivial conserved quantity—the energy—so the state space is foliated by 3-dimensional manifolds such that every trajectory is contained in one of these manifolds. Second, because for the BP the controls are unbounded, even within a single leaf existence of an optimal control joining any two points does not follow. These two anomalies conspire to make the BP look two-dimensional and to conceal its true 4-dimensional nature. It should be obvious from the above that I think I know what I want to say, but if you want to give try your hand and write something, go ahead.

COMMENT NO. 43. Maybe we should make a complete study of the general acceleration control problem in $\mathbb{R}^2$ of which both Hestenes and BP are particular cases? Compute the Lie algebra in general and show how it degenerates for the BP? What do you say?

COMMENT NO. 44. One the previous GNUs are taken care of, I'd like to go back to the beginning of the book and put in a few remarks on how the BP is a "degenerate problem" and how this fact has obscured its true meaning and how it fits within a general theory.

13.14 Lie brackets, integrals of motion, and the differential-geometric perspective

A much better way to pose the problem is to write down the equations of motion that truly correspond to Johann Bernoulli's problem as he stated it in June 1696. One should then let the mathematical analysis lead us to

the discovery of energy conservation and the simplification resulting from it. This will lead us naturally to a four-dimensional problem whose analysis is best carried out from the differential-geometric perspective, by means of Lie-algebraic calculations.

The true equations of motion for the brachistochrone are

$$\dot{x} = v, \quad \dot{y} = w\,, \quad \dot{v} = uw\,, \quad \dot{w} = -uv - g\,. \tag{13.18}$$

Here x and y are the horizontal and vertical coordinates of our moving point, and v, w are the components of the velocity vector. The requirement that the point is freely falling "under the influence of its own weight" means that the force effectively acting on the point should be equal to a vector proportional to $[0, -g]$ plus a "virtual force" that does no work, i.e., is perpendicular to the velocity. Equation (13.18) captures these requirements, by introducing a virtual force vector of the form $[uw, -uv]$, where u is an arbitrary "control," taking values in $\mathbb{R}$.

Using $q = [x,\, y,\, v,\, w]^\dagger$, we get a familiar equation, namely,

$$\dot{q} = F(q) + uG(q)\,, \tag{13.19}$$

where $F = [v, w, 0, -g]^\dagger$, and $G = [0, 0, w, -v]^\dagger$.

This is a *4-dimensional system*, with state space $\mathbb{R}^4$.

13.15 The diamond Lie algebra

Having expressed our equations of motion in the form (13.19), the first thing that one should do, from the differential-geometric perspective is to compute the iterated Lie brackets of F and G, and determine the ***Lie algebra sturcture of our problem.*** It turns out that the four vector fields F, G, $[F, G]$ and $[F, [F, G]]$ are linearly independent over $\mathbb{R}$ (that is, there exists no nontrivial linear combination of them with constant coefficients that vanishes identically), and in addition the relations

$$[G, [F, G]] = F\,, \quad [F, [F, [F, G]]] = [G, [F, [F, G]]] = 0 \tag{13.20}$$

hold. It then follows easily that, if L denotes the Lie algebra of vector fields generated by F and G, then L is four-dimensional. As an abstract Lie algebra, L is isomorphic to the *diamond Lie algebra* G_4, a solvable Lie algebra that plays an important role in representation theory (cf. [11], p. 59). This Lie algebra has a basis H, P, Q, E, with the commutation relations

$[H,P] = Q$, $[H,Q] = -P$, $[P,Q] = E$, $[H,E] = [P,E] = [Q,E] = 0$. (To construct an isomorphism Φ from G_4 to L, define $\Phi(P) = [F,G]$, $\Phi(Q) = F$, $\Phi(H) = G$, $\Phi(E) = -[F,[F,G]]$. A concrete realization of G_4 as an algebra of differential operators—acting of smooth functions on the real line—can be defined by "quantizing the harmonic oscillator," i.e., by taking $P = \frac{d}{dx}$, $Q =$ multiplication by x, $E =$ identity, $H = \frac{1}{2}(-P^2+Q^2) = \frac{1}{2}\left(-\frac{d^2}{dx^2}+x^2\right)$, so H is the Hermite operator.

13.16 Energy conservation

Although F, G, $[F,G]$ and $[F,[F,G]]$ are linearly independent over $\mathbb{R}$, a simple computation shows that

$$G = \psi_1 F + \psi_2 [F,G] + \psi_3 [F,[F,G]] \quad \text{and} \quad [G,[F,G]] = F\,, \tag{13.21}$$

where the ψ_i are smooth functions of q, given by

$$\psi_1 = \frac{v}{g}\,, \quad \psi_2 = -\frac{w}{g}\,, \quad \psi_3 = -\frac{v^2+w^2}{2g^2}\,. \tag{13.22}$$

From these facts one can easily show that every iterated Lie bracket of F and G is a linear combination of F, $[F,G]$, and $[F,[F,G]]$ with smooth coefficients. Since $F(q)$, $[F,G](q)$, and $[F,[F,G]](q)$ are linearly independent at each point $q \in \mathbb{R}^4$, we can conclude that the *distribution* Λ generated by L is 3-*dimensional* at each point. (By definition, $\Lambda(q) \stackrel{\text{def}}{=} \{X(q) : X \in L\}$.) This means—using, for example, Frobenius' theorem on the existence of integral manifolds—that, at least locally, there is a nontrivial integral of motion, i.e., a function with nonzero gradient which is constant along all integral curves of F and G, and then also along all solutions of (13.19). This integral of motion can be easily computed and turns out to be the energy E, given by

$$E(x,y,v,w) = \frac{v^2+w^2}{2} + gy\,. \tag{13.23}$$

Notice that E is a smooth function on $\mathbb{R}^4$, with nowhere vanishing gradient. Therefore the level hypersurfaces $S_{\bar{E}}$ of E—i.e., the sets $E = \bar{E}$, for $\bar{E} \in \mathbb{R}$—are smooth 3-dimensional manifolds. (The set $S_{\bar{E}}$ is given by the equation $y = \frac{2c-v^2+w^2}{2g}$, which shows that each $S_{\bar{E}}$ is the graph of a smooth function $(x,v,w) \mapsto y$.)

13.17 Naïve application of the maximum principle

The fact that our 4-dimensional state space $\mathbb{R}^4$ is foliated by 3-dimensional "leaves" S_E has profound implications for the optimization problem. Indeed, *if we try to find the optimal trajectories by applying the maximum principle, we get* ***nothing***. This is a consequence of more general result, namely, that,

(#) *whenever a control system satisfies a nontrivial "holonomic constraint" such as* $E =$ constant*, then the maximum principle is uninformative, because* ***every trajectory is an extremal***.

(The word "nontrivial" here means that ∇E never vanishes.) This explains why in our case the maximum principle, naïvely applied, yields nothing.

13.18 Application of the maximum principle on manifolds

Using the integral of motion E, we can regard $\mathbb{R}^4$ as foliated by the 3-dimensional level hypersurfaces of E, and conclude that every trajectory is contained in one of the leaves $S_{\bar{E}}$. If $\bar{E} \in \mathbb{R}$, then the maximum principle on manifolds (cf. Chapter 11 and [118]) can be applied to the problem restricted to $S_{\bar{E}}$, and the conclusion turns out to be exactly the same as that for the unrestricted problem, except that now the nontriviality condition says that $p_*(t)$ cannot be orthogonal to $S_{\bar{E}}$ at $q_*(t)$. (In differential-geometric terms, $p_*(t)$ is really a *covector* on $S_{\bar{E}}$ at $q_*(t)$, and has to be nontrivial as such. Equivalently, if one insists on regarding $p_*(t)$ as a vector in $\mathbb{R}^4$, then it should not be orthogonal to the tangent space to $S_{\bar{E}}$ at $q_*(t)$.) Naturally, one could also avoid using the maximum principle on manifolds by using the fact that $S_{\bar{E}}$ is identified with $\mathbb{R}^3$ in obvious way, since it is the graph of a smooth function $y(x, v, w)$. This would amount to eliminating y from the equations by substituting for it its expression in terms of the other variables, for each fixed $\bar{E}$.

The Hamiltonian is $H(q, u, p, p_0) = \varphi(p, q)\rangle + u\psi(q, p) - p_0$, where the functions φ, ψ are defined by $\varphi(p, q) = \langle p, F(q)\rangle$ and $\psi(q, p) = \langle p, G(q)\rangle$. Since u is completely unrestricted, H cannot have a maximum as a function of u unless $\psi(q, p) = 0$. So ψ must vanish along every lifted extremal. Differentiation of $\psi = 0$ yields $\rho = 0$, where $\rho(q, p) = \langle p, [F, G](q)\rangle$. A further differentiation—together with the fact that $[G, [F, G]] = F$—yields

$\theta + u\varphi = 0$, where $\theta(q,p) = \langle p, [F,[F,G]](q)\rangle$. This gives $u = -\varphi^{-1}\theta$, i.e.,

$$u = -\frac{\langle p, [F,[F,G]](q)\rangle}{\langle p, F(q)\rangle}\,. \tag{13.24}$$

This, together with the equalities $\psi = \rho = 0$, determine u as a function of q, i.e., provide an optimal control in the form of a feedback law $u = u(q)$. In fact, (13.21) implies that $\psi = \psi_1\varphi + \psi_2\rho + \psi_3\theta$ along our trajectory. Since $\psi = \rho = 0$, and $\theta = -u\varphi$, we get (using the fact that $\varphi \neq 0$, because p is nontrivial as a linear functional on the span of F, $[F,G]$ and $[F,[F,G]]$, so that $(\varphi,\rho,\theta) \neq (0,0,0)$, while on the other hand $\psi = \rho = 0$, so that if $\varphi = 0$ implies $\theta = 0$) $u = \psi_3^{-1}\psi_1$, i.e.,

$$u = \frac{2vg}{v^2 + w^2}\,. \tag{13.25}$$

It is then clear that u is a smooth function on the complement of the set

$$\Sigma = \{(x,y,v,w) : v^2 + w^2 = 0\}\,. \tag{13.26}$$

The "singular" set Σ corresponds, naturally, to the points where the kinetic energy vanishes, which as we know from our other analyses is where some singular behavior occurs.

If we define

$$Z(q) = F(q) - \psi_1(q)G(q)\,, \tag{13.27}$$

we see that *Z is a smooth vector field on $\mathbb{R}^4\backslash\Sigma$ whose trajectories are the extremals in the sense of the maximum principle.*

The reader is urged to carry out the remaining computations explicitly as an exercise, and verify that the final result is once again Johann Bernoulli's family of cycloids.

Remark 13.18.1 The preceding calculations provide another illustration of the implementation of ***Lagrange's idea of choosing the best possible coordinates for a particular problem***. Notice how everything is made much simpler by the fact that, using the intrinsic version of the adjoint equation, we can do all our computations directly for the switching function $\langle p, G\rangle$ and other functions of the same type, without ever writing the components of p. ♢

13.19 The complete synthesis and Hamilton's principal function for the 4-dimensional brachistochrone problem

We haven't yet completely answered all our earlier questions and determined Hamilton's principal function V.

It is clear that $V(A, a, B, b) = +\infty$ unless both (A, a) and (B, b) lie in the same energy hypersurface, because if the energies are different then the points cannot be joined by a trajectory, and the infimum of the empty set is $+\infty$.

On the other hand, if (A, a), (B, b) belong to the same level hypersurface $S_{\bar{E}}$, then we expect that $V(A, a, B, b)$ will be finite. Moreover, although it would be a "rash conclusion" to say that there must be a minimizing curve—cf. our earlier quotation from Synge's paper, in Chapter 4—it is natural to expect that such a curve will exist, and if it does not then we ought to be able to explain why.

The classical analysis carried out earlier seems to suggest that, given A, B and $\bar{E}$, there will be only one optimal arc joining them. So, if we give A and B but do not fix the energy, then there should be a 1-parameter family of such arcs (one arc for each value of E). Yet we now seem to need a 3-parameter family, because we want to accomodate all possible pairs (a, b) (i.e., four parameters) as long as they give rise to the same energy (one constraint, and $4 - 1 = 3$). Equivalently, ***given $\bar{E}$, we ought to be able to choose the directions θ_a, θ_b, of the vectors a and b arbitrarily.*** (Their magnitudes are determined by the fact that the energy at A and B must equal $\bar{E}$.) But ***given A, B and $\bar{E}$ there is a unique cycloid from A to B corresponding to the energy $\bar{E}$***, and in particular the directions $\bar{\theta}_a$, $\bar{\theta}_b$ of this cycloid at A and B are completely determined by A, B and $\bar{E}$.

In other words, ***we ought to be able to choose freely the two angles θ_a, θ_b, but it appears that we cannot do so.***

The solution of this apparent paradox is as follows. Suppose we fix an energy level $\bar{E}$. Given any A, the set $C(A, \bar{E})$ of those a's such that $(A, a) \in S_{\bar{E}}$ is either empty or a circle, except for the degenerate case when $(A, 0) \in S_{\bar{E}}$, in which case $C(A, \bar{E})$ is a point. If $C(A, \bar{E})$ is a circle, then through each point of $\{A\} \times C(A, \bar{E})$ there passes an integral curve of Z. The set of all points lying on such curves is 2-dimensional, and intersects the 1-dimensional set $\{B\} \times C(B, \bar{E})$ at just one point. This means that there is a unique pair $(\bar{a}, \bar{b})$ such that $(A, \bar{a})$ and $(B, \bar{b})$ lie on the same integral

curve of Z. The portion of that curve joining $(A, \bar{a})$ to $(B, \bar{b})$ is, of course, our famous cycloid.

So, then, what is the optimal trajectory joining (A, a) to (B, b) for general a, b? What is the value of $V(A, a, B, b)$? The answer is quite simple: the "optimal trajectory" is obtained by applying an ***impulse control*** (cf. Ch. 3) to rotate a to $\bar{a}$ instantaneously, then following the cycloid from $(A, \bar{a})$ to $(B, \bar{b})$, and then applying another impulse to rotate $\bar{b}$ to b. (These are not true trajectories, of course, but it is easy to translate the above into a statement about a minimizing sequence of true trajectories.) The value $V(A, a, B, b)$ is therefore equal to $V(A, \bar{a}, B, \bar{b})$.

13.20 Deriving Snell' law from the maximum principle

We conclude our discussion by going back to an even earlier episode in our story, namely, Fermat's minimum time principle and Snell' law of refraction. We have already explained how this law played a crucial role in Johann Bernoulli's solution of the brachistochrone problem, which then gave rise to the calculus of variations and optimal control. It would therefore be nice to end our discussion by "closing the loop" and showing that the results attained during this 300-year old endeavor do apply, in particular, to their venerable distant ancestor from geometrical optics.

So let us consider a medium such as the one we described when discussing Snell' law, in which the speed of light is constant on each of the two sides H_-, H_+ of a hyperplane H in $\mathbb{R}^n$, but the constant values c_-, c_+, are different. Let us ask the obvious question, namely, whether Snell' law be derived from the maximum principle.

It is easy to see that there is no way to formulate this problem as a minimum time problem with a controlled dynamics $\dot{q} = f(q, u, t)$ in which f is continuous with respect to q, since $||f||$ has to jump as we cross H.

In our previous discussion in Chapters 8 and 9 we have explained how one can get rid of requirements that f and L be differentiable, and even continuous, with respect to u. But here the situation is different, since we want to be able to work with an f that is discontinuous with respect to q.

Once again, it turns out that our recent version of (MP) given in [119] applies to this problem as well. The reason is that in [119] all that is needed is that the reference trajectory q_* arise from a flow $\{\Phi_{t',t}\}$ consisting of semidifferentiable maps, and there is no need for this flow to be generated

by from a continuous vector field. One can formulate our refraction problem in terms of the dynamical law $\dot{q} = f(q, u, t)$, where the control u is now a pair (u_-, u_+) of unit vectors, and $f(q, u, t) = c_- u_-$ for $q \in H_-$, $f(q, u, t) = c_+ u_+$ for $q \in H_+$. If $\bar{q} \in H_-$, $\hat{q} \in H_+$, it is easily seen, by applying the classical version of (MP), that an optimal trajectory from $\bar{q}$ to $\hat{q}$ has to consist of a segment going from $\bar{q}$ to some point P in H, followed by another segment from P to $\hat{q}$. The question is to find P. It is easy to verify that, if $q_* : [a, b] \mapsto \mathbb{R}^n$ is a trajectory of this type, then q_* is in fact a trajectory of a constant control $u = (u_-, u_+)$, such that both u_- and u_+ point to the same side of H. Although the vector field $q \mapsto f(q, u)$ is discontinuous, one can easily see that the corresponding flow maps $\Phi_{t,a}$ are continuously differentiable (and *a fortiori* semidifferentiable) at $\bar{q}$, for all t except the one such that $\Phi_{t,a}(\bar{q}) \in H$. This suffices to make the maximum principle of [119] applicable, and derive Snell' law.

Chapter 14

Conclusion

We have shown that the results discovered in the 300 years that followed Johann Bernoulli's challenge yield new and better insights into his "splendid problem." First, the Euler-Lagrange equations eliminate the spurious solutions. Second, classical optimal control theory makes it possible to settle easily the existence question, and to avoid the *Ansatz* that the optimal arcs are graphs of functions $y = y(x)$. Third, modern differential-geometric optimal control—which is heavily indebted to Lagrange's idea of coordinate invariance—allows the problem to be formulated and understood in its proper 4-dimensional setting, the only one that makes Hamilton's principal function meaningful and avoids the apparently paradoxes pertaining to the correct choice of endpoint conditions. Finally, modern methods, using nonsmooth techniques, allow us to deal with the very natural "reflected brachistochrone" problem, and to honor the ancestors of Bernoulli's own work—Fermat's principle and Snell' law—by including them in a single, unified formulation of the necessary conditions.

Would Johann Bernoulli and the other five contributors to the May 1697 *Acta Eruditorum* have have appreciated these new ways of looking at Johann's "splendid problem? Would Lagrange, Hamilton and Weierstrass have liked them? We hope we have presented strong evidence that they would have, and leave the final verdict up to the reader.

Appendix: Some formal definitions.

COMMENT NO. 45. This appendix is very weak. It's got to be completely rewritten and improved.

A.1. Controlled dynamical systems

A *controlled dynamical law* (CDL) is a triple $\mathcal{L} = (S, U, f)$ where S—the "state space" of $\mathcal{L}$—is a subset of some Euclidean space $\mathbb{R}^n$, U—the "control space"—is a set, and f is a mapping from $S \times U$ to $\mathbb{R}^n \cup \{\#\}$. (Here $\#$ is a fixed object, subject only to the requirement that $\# \notin \mathbb{R}^n$. It stands for "undefined," so that $f(x,u) = \#$ is just a way of saying, for a particular (x,u), that f is not defined for that (x,u). This device allows us to consider CDL's with a partially defined map f.) An *open-loop control* for a CDL $\mathcal{L} = (S, U, f)$ is a map η from a subinterval I of $\mathbb{R}$ to U. A *trajectory-control pair* (TCP) of $\mathcal{L}$ is a pair (ξ, η) such that η is an open-loop control and ξ is a locally absolutely continuous map from the domain $D(\eta)$ of η to S, with the property that $\dot{\xi}(t) = f(\xi(t), \eta(t))$ for almost all $t \in D(\eta)$. We use $TCP(\mathcal{L})$ to denote the set of all TCP's of $\mathcal{L}$. (More precisely, we identify two TCP's $\gamma_1 = (\xi_1, \eta_1)$, $\gamma_2 = (\xi_2, \eta_2)$ if $D(\eta_1) = D(\eta_2)$, $\xi_1(t) = \xi_2(t)$ for all $t \in D(\eta_1)$, and $\eta_1(t) = \eta_2(t)$ for almost all $t \in D(\eta_1)$. So $TCP(\mathcal{L})$ really is a set of equivalence classes of TCP's.)

It is clear how to define the *restrictions* $\eta \lceil I'$, $\gamma \lceil I'$ of an open-loop control η and of a TCP $\gamma = (\xi, \eta)$ to a subinterval I' of $D(\eta)$. Also, one can define the *concatenation* $\eta_1 * \eta_2$ of two controls η_1, η_2, provided that their domains I_1, I_2 are such that $I_1 \cup I_2$ is an interval, $I_1 \cap I_2$ is a null set (i.e., $\mathrm{card}(I_1 \cap I_2) \leq 1$) and I_1 lies to the left of I_2. If $\gamma_1 = (\xi_1, \eta_1)$, $\gamma_2 = (\xi_2, \eta_2)$ are two TCP's with domains I_1, I_2 such that $\eta_1 * \eta_2$ is defined, then the concatenation $\gamma_1 * \gamma_2$ is defined—in an obvious way—if $\xi_1(t) = \xi_2(t)$ for $t \in I_1 \cap I_2$, and the curve ξ whose graph is the union of the graphs of of ξ_1 and ξ_2 is locally absolutely

continuous on $I_1 \cup I_2$. (For example, if $I_i = [a_i, b_i]$ are compact intervals, then $\gamma_1 * \gamma_2$ is defined iff $b_1 = a_2$ and $\xi_1(b_1) = \xi_2(a_2)$.)

A *controlled dynamical system* (CDS) is a 4-tuple $\Sigma = (S, U, f, \mathcal{A})$ such that (S, U, f) is a CDL and $\mathcal{A}$—the set of "admissible TCP's of Σ"—is a subset of $TCP(M, U, f)$ which is closed under restrictions (that is, if $\gamma \in \mathcal{A}$ and $\gamma' = \gamma\lceil I'$ for some subinterval I' of the domain of γ, then $\gamma' \in \mathcal{A}$). Often, $\mathcal{A}$ is defined by singling out a "class of admissible controls." That is, we specify a class $\mathcal{U}$ of open-loop controls, and then take $\mathcal{A} = \mathcal{A}_{\mathcal{U}}$, where $\mathcal{A}_{\mathcal{U}}$ is the set of all TCP's (ξ, η) of (S, U, f) such that $\eta \in \mathcal{U}$.

A.2. Cost functionals, optimal control problems, and the value

A *cost functional* for a CDS $(S, U, f, \mathcal{A})$ is a function $J : \mathcal{A} \mapsto \mathbb{R} \cup \{+\infty\}$. The functional J is *additive* if $J(\gamma_1 * \gamma_2) = J(\gamma_1) + J(\gamma_2)$ whenever γ_1, γ_2 belong to $\mathcal{A}$ and $\gamma_1 * \gamma_2$ is defined and belongs to $\mathcal{A}$.

A *controlled dynamical system with cost* (CDSWC) is a five-tuple

$$\mathcal{C} = (S, U, f, \mathcal{A}, J)$$

such that $(S, U, f, \mathcal{A})$ is a CDS and J is a cost functional for $(S, U, f, \mathcal{A})$.

To a CDSWC $\mathcal{C} = (S, U, f, \mathcal{A}, J)$ we can associate *optimal control problems* in various ways. Given any subset $\mathcal{B}$ of $\mathcal{A}$ we can consider "the problem $P_{\mathcal{C}}(\mathcal{B})$ of minimizing $J(\gamma)$ for $\gamma \in \mathcal{B}$." The *value* $V_{\mathcal{C}}(\mathcal{B})$ of this problem is the extended real number

$$V_{\mathcal{C}}(\mathcal{B}) = \inf\{J(\gamma) : \gamma \in \mathcal{B}\}\,. \tag{A.1}$$

Then $V_{\mathcal{C}}(\mathcal{B}) = +\infty$ iff $J(\gamma) = +\infty$ for all $\gamma \in \mathcal{B}$. (In particular, $V_{\mathcal{C}}(\mathcal{B}) = +\infty$ when $\mathcal{B}$ is empty.) It is also possible for $V_{\mathcal{C}}(\mathcal{B})$ to be $-\infty$, which will happen, of course, iff J is not bounded below on $\mathcal{B}$. A $\gamma \in \mathcal{B}$ will be called a *solution* of $P_{\mathcal{C}}(\mathcal{B})$ if $J(\gamma) = V_{\mathcal{C}}(\mathcal{B})$.

A.3. Generalized solutions

Naturally, the infimum defining $V_{\mathcal{C}}(\mathcal{B})$ need not be attained as a minimum, so a solution of $P_{\mathcal{C}}(\mathcal{B})$ may fail to exist. IN that case, it is often convenient to enlarge the set of TCP's of a system by introducing new, "generalized trajectory-control pairs," so that in the enlarged space the problem has a true solution.

Example 14.0.1 For a trivial example of this situation, consider the problem of reaching $0 \in \mathbb{R}$ in "minimum time" from an initial point $x \in \mathbb{R}$, when the dynamics is $\dot{x} = u$, $u \in U$, with $U = \mathbb{R}$. It is clear that when $x \neq 0$ the value of this problem is zero, but this value is not attained as a true minimum by any trajectory-control pair arising from an "ordinary" control, where "ordinary" means, for example, Lebesgue integrable. On the other hand, if we allow the "controls" to contain impulses, then we can take, for $x \in \mathbb{R}$, $\eta_x = -x.\delta$, where "δ" is Dirac's Delta function, and the control η_x obviously steers x to 0 in time 0.

We can define $\hat{\mathcal{B}}$ to be the set of all sequences $\Gamma = \{\gamma_j\}_{j=1}^{\infty}$ of members of $\mathcal{B}$ for which the limit $\hat{J}(\Gamma) = \lim_{j\to\infty}(J(\gamma))$ exists. We can then think of the members of $\hat{\mathcal{B}}$ as "generalized TCP's," and of the functional $\hat{J} : \hat{\mathcal{B}} \to \mathbb{R} \cup \{-\infty, +\infty\}$ as the extension of J from $\mathcal{B}$ to $\hat{\mathcal{B}}$. (There is a an obvious canonical embedding of $\mathcal{B}$ in $\hat{\mathcal{B}}$, obtained by regarding each $\gamma \in \mathcal{B}$ as a constant sequence.) It is clear that if $\mathcal{B} \neq \emptyset$ then $\hat{J}$ attains a minimum on $\hat{\mathcal{B}}$ at some $\Gamma \in \hat{\mathcal{B}}$, and $\hat{J}(\Gamma) = V_{\mathcal{C}}(\mathcal{B})$. In order words, if $\mathcal{B} \neq \emptyset$ then one can always construct a *minimizing sequence* Γ, i.e., a sequence $\{\gamma_j\}$ in $\mathcal{B}$ whose costs $J(\gamma_j)$ converge to $V_{\mathcal{C}}(\mathcal{B})$. If one regards such a Γ as a solution of $P_{\mathcal{C}}(\mathcal{B})$ in a "generalized sense," or as a solution of the "extended problem" in which we seek to minimize $\hat{J}$ on $\hat{\mathcal{B}}$ rather than minimizing J on $\mathcal{B}$, then the original problem $P_{\mathcal{C}}(\mathcal{B})$ always has a generalized solution, and the entended problem always has a solution, as long as $\mathcal{B} \neq \emptyset$. In this abstract setting, the concept of a generalized solution is rather trivial and uninteresting, but it turns out that in many cases one can realize generalized solutions as much more concrete objects, e. g. by showing that minimizing sequences converges in some sense, such as that of Schwartz distributions or relaxed trajectories.

Remark 14.0.2 In the previous example the true solution of the minimization problem was a distribution in the sense of L. Schwartz's theory. The reader should not extrapolate from this example and imagine that for general optimal control problems the appropriate class of generalized controls is that of Schwartz distributions. It turns out that one needs a rather different class of generalized objects. This is discussed in detail in papers by A. Bressan, F. Rampazzo, and their coworkers, cf. [23, 26, 27, 91, 92, 99]. ♢

A.4. The value function (i.e., Hamilton's principal function)

A particularly important class of optimal control problems is that of *fixed endpoint problems*. Given $x, x' \in S$, we can consider the set $\mathcal{B}(x, x')$ of all TCP's $\gamma = (\xi, \eta) \in \mathcal{A}$ such that γ goes from x to x'. (Precisely, $\gamma \in \mathcal{B}(x, x')$ iff $\gamma \in \mathcal{A}$, $D(\eta)$ is a compact interval $[a, b]$, $\xi(a) = x$, and $\xi(b) = x'$.) We will use $P_{\mathcal{C}}(x, x')$—rather than $P_{\mathcal{C}}(\mathcal{B}(x, x'))$—to denote this problem, and will refer to it by means of self-explanatory expressions such as "the problem of driving x to x' with minimum cost $J(\gamma)$, subject to the dynamical constraint $\dot{x} = f(x, u)$ and the requirement that $\gamma \in \mathcal{A}$." The value of $P_{\mathcal{C}}(x, x')$ will be written $V_{\mathcal{C}}(x, x')$, and $V_{\mathcal{C}} : S \times S \mapsto (\mathbb{R} \cup \{-\infty, +\infty\}) \times (\mathbb{R} \cup \{-\infty, +\infty\})$ is the *value function* of the CDSWCF $(S, U, f, \mathcal{A}, J)$.

The value function, as defined here, is essentially the same as ***Hamilton's principal (or characteristic) function***, except that the modern definition

- avoids the issue of the existence of a minimizing arc, by taking $V_{\mathcal{C}}(x, x')$ to be the *infimum* of the costs of all arcs from x to x',
- is formulated directly for control problems rather than for the more restricted calculus of variations,

and

- allows the values of $V_{\mathcal{C}}(x, x')$ to be $-\infty$ and $+\infty$, the latter case arising when there are no trajectories joining x to x'[1].

We emphasize that a "controlled dynamical system with cost" (CDSWC) $\mathcal{C}$ is not at all the same as an "optimal control problem." A CDSWC gives rise to ***many*** optimal control problems, such as for example the fixed-endpoint problems $P_{\mathcal{C}}(x, x')$ described before, as well as other problems obtained by making different choices of $\mathcal{B}$. (For example, one could choose $\mathcal{B}$ to be the set of all TCP's in $\mathcal{A}$ that start at some given subset A and end at some other subset B of S, obtaining the problem of "joining A to B with minimum cost by a TCP γ of $\dot{x} = f(x, u)$, $u \in U$, $\gamma \in \mathcal{A}$.") An optimal control problem has a *value*, which is an extended real number, and if the set of trajectories is $\neq \emptyset$ then the problem may have a *solution*—i.e., a minimizer—or at least a minimizing sequence or generalized solution, which *realizes* this value. A CDSWC $\mathcal{C}$ has a *value function* $V_{\mathcal{C}}$. The object that realizes the value is

[1]This situation is, of course, impossible in the classical calculus of variations, but it can easily occur in control problems.

something called an "optimal synthesis," by which it is meant—roughly—a family $\{\gamma_{x,x'}\}_{(x,x')\in C}$ in $\mathcal{A}$ that selects, for each $(x,x') \in C$ (where $C = \{(x,x') \in S \times S : \mathcal{B}(x,x') \neq \emptyset\}$, an—ordinary or generalized—solution $\gamma_{x,x'}$ of the problem $P_{\mathcal{C}}(x,x')$. (This is only a first approximation to the concept of a synthesis. In the control literature one looks for syntheses having extra regularity properties, i.e., in selecting solutions $\gamma_{x,x'}$ that depend "nicely"—e.g. continuously, or smoothly, or piecewise smoothly—on x and x'.)

An important type of cost functionals are the *Lagrangian costs*, obtained when J is of the form

$$J(\xi,\eta) = \int_{D(\eta)} L(\xi(t),\eta(t))dt\,, \tag{A.2}$$

for some function $L : S \times U \mapsto \mathbb{R} \cup \{+\infty\}$ having the property that L is "admissible for $\mathcal{A}$," i.e., that for every $(\xi,\eta) \in \mathcal{A}$ the integrand of (A.2) is measurable and such that $\int_{D(\eta)} \max(0, -L(\xi(t),\eta(t)))dt < +\infty$.

Lagrangian cost functionals are always additive. Within the Lagrangian costs, an important example is *time*, i.e., the functional J obtained by taking $L \equiv 1$.

If the cost functional J is additive, and the class $\mathcal{A}$ is closed under concatenations, then the value function $V_{\mathcal{C}}$ satisfies

a. the ***dynamic programming inequality***

$$V_{\mathcal{C}}(x,x'') \leq V_{\mathcal{C}}(x,x') + V_{\mathcal{C}}(x',x'') \;\text{ for all }\; x,x',x'' \in S\,, \tag{A.3}$$

b. the ***principle of optimality***, which says that *any part of an optimal trajectory is optimal.* (That is, if $\gamma = (\xi,\eta) \in \mathcal{A}$ is a solution of the problem $P_{\mathcal{C}}(x,x')$, $D(\eta) = [a,b]$, $a \leq \alpha \leq \beta \leq b$, $y = \xi(\alpha)$ and $y' = \xi(\beta)$, then the restriction $\gamma\lceil[\alpha,\beta]$ of γ to the interval $[\alpha,\beta]$ is a solution of $P_{\mathcal{C}}(y,y')$.)

A CDSWC $\mathcal{C} = (S,U,f,\mathcal{A},J)$ is of the *classical calculus of variations type* if S is an open subset of $\mathbb{R}^n$, $U = \mathbb{R}^n$, $f(x,u) = u$, and J is Lagrangian. In that case, the CDL (S,U,f) is rather trivial, so nothing interesting really really happens until one brings in the cost J, i.e., the Lagrangian L.

We illustrate the above definitions with a simple example. Consider the formal expression $\dot{x} = u$, where we want to think of x and u as real-valued variables, x being the "state" and u the "control." Let us take S, the state space, to be $\mathbb{R}$. If we specify in addition a set U of values of the control u, then we get a CDL (S,U,f), by taking f to be given by $f(x,u) = u$. If, for

example, we take U to be the interval $[-1, 1]$, then we can obtain different CDS's—often referred to as "the system $\dot{x} = u$, with state $x \in \mathbb{R}$ and control constraint $|u| \leq 1$"—by making different choices of $\mathcal{A}$. One standard way to do that would be to choose a class $\mathcal{U}$ of open-loop controls, and let $\mathcal{A} = \mathcal{A}_{\mathcal{U}}$. One could, for example, take $\mathcal{U}$ to be the class of all continuous U-valued maps defined on subintervals of $\mathbb{R}$, or the set of all maps of class C^k for some k, or the class of all polynomial maps. In all these cases, the resulting class $\mathcal{A}_{\mathcal{U}}$ will not be closed under concatenations. One could also take $\mathcal{U}$ to be the class of all measurable U-valued maps, or that of all piecewise constant ones (or piecewise continuous, or piecewise C^1). For these classes, $\mathcal{A}_{\mathcal{U}}$ is be closed under concatenations. By restricting $\mathcal{U}$ further to consist of controls η whose domain $D(\eta)$ is a compact interval, one can associate a CDSWC to each of the above CDS's, by taking J to be time. The resulting optimal control problems are the *minimum time problems* associated to the CDS's $(\mathbb{R}, [-1, 1], \mathcal{A}_{\mathcal{U}})$. The value function for all these problems is given by $V_{\mathcal{C}}(x, x') = |x - x'|$.

One could do the same thing with $U = \mathbb{R}$, in which case we would be considering "the CDS $\dot{x} = u$ with $x \in \mathbb{R}$ and unrestricted $u \in \mathbb{R}$," with the proviso, of course, that to specify a CDS we have to be precise about the choice of the class $\mathcal{A}$. If, for example, we choose $\mathcal{U}$ to be the set of Lebesgue integrable controls—that is, $\mathcal{U} = \cup_{-\infty<a\leq b<+\infty} L^1(\mathbb{R}, \mathbb{R})$—and take $\mathcal{A} = \mathcal{A}_{\mathcal{U}}$, then we get the system $\dot{x} = u$, with $x \in \mathbb{R}$, $u \in \mathbb{R}$, and $u(\cdot) \in L^1$. If we take the cost functional to be time, as before, we get a CDSWC, the minimum-time problem for the system $\dot{x} = u$ with unrestricted $u \in \mathbb{R}$ and $u(\cdot)$ integrable. The value function for this system is of course given by $V_{\mathcal{C}}(x, y) = 0$, since any two points can be connected in arbitrarily small time by using very large controls. On the other hand, if $x \neq y$ then a minimizer does not exist but, as explained in Example 14.0.1, one can arrange for the problem to have a true solution by considering a suitable extension $\hat{\mathcal{U}}$ of $\mathcal{U}$ to a space of "generalized controls," which in this case can be taken to be any space which is large enough to contain all multiples of the Delta function. (For example, we could take $\hat{\mathcal{U}}$ to be the set of all distributional derivatives of Borel measures on $\mathbb{R}$.)

A.5. Systems on manifolds, control bundles, partially defined vector fields

The above definitions can be extended in various ways. For example, one can allow S to be a subset of a manifold M other than $\mathbb{R}^n$. In that case,

the condition that f be a map from $S \times U$ to $\mathbb{R}^n \cup \{\#\}$ is replaced by the requirement that f be a—possibly partially defined—map on $S \times U$ such that $f(x, u)$ is a tangent vector to M at x whenever $x \in S$, $u \in U$, and $f(x, u)$ is defined.

One can also allow U to be a "control bundle" rather than a "control set." That is, U can be taken to be a set $U(x)$ that depends on the point $x \in S$. In this case, the concept of an "open-loop control" ceases to be meaningful, but that of a TCP makes perfect sense, exactly as in the case of a state-independent control space.

A special case of the control bundle situation is that of *calculus of variations problems on manifolds*, i.e., CDSWC's where the state space is a manifold M, the control bundle U is the tangent bundle of M (that is, $U(x)$ is the tangent space at x for each x), the dynamics f is given by $f(x, u) = u$, and the cost functional is Lagrangian. Another important example of CDSWC's is obtained by taking the state space M to be a Riemannian manifold, $U(x)$ to be the closed unit ball in the tangent space to M at x, and $f(x, u) = u$. Taking the cost functional to be time, we see that the minimizers are the arcs of minimum length, and the value function is the Riemannian distance. This example can be generalized in an obvious way to case of a *subriemannian manifold*, i.e., a smooth manifold M endowed with a smooth subbundle E of the tangent bundle, and with a Riemannian metric defined on E. As in the Riemannian case, one can take $U(x)$ to be the closed unit ball in $E(x)$, and define $f(x, u) = u$.

In addition, it is perfectly possible to allow f—and the Lagrangians L—to depend on time.

Finally, one can try to define analogues of all these concepts for state spaces S more general than subsets of finite-dimensional manifolds. For example, one can take S to a be a Banach space—often a function space—and be led to the very rich and active field of "distributed parameter control." Or one can consider many other types of "controlled systems," such as discrete-time or stochastic systems.

Bibliography

[1] Agrachev, A. A. and R. V. Gamkrelidze, *A second-order optimality principle for a time-optimal problem.* Math. Sbornik **100**, 142, 1976. [GNU: UPDATE THIS REFERENCE, IF NEEDED. PAGE NOS. MISSING.]

[2] Agrachev, A. A. and R. V. Gamkrelidze, *The exponential representation of flows and the chronological calculus.* Math. Sbornik **109**, 149, 1978. [GNU: UPDATE THIS REFERENCE, IF NEEDED. PAGE NOS. MISSING.]

[3] Agrachev, A. A. and R. V. Gamkrelidze, *The Morse index and the Maslov index for smooth control systems.* Doklady Akad. Nauk USSR **287**, 1986. [GNU: UPDATE THIS REFERENCE, IF NEEDED. PAGE NOS. MISSING.]

[4] Anglin, W. S., *Mathematics: A Concise History and Philosophy.* Springer-Verlag, New York, 1994.

[5] Ball, J. M. and V. J. Mizel, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation.* Arch. Rational Mech. Anal. **63**, 1985, pp. 273–294.

[6] Bell, E. T., *Men of Mathematics.* Simon and Schuster, New York, 1937.

[7] Bell, E. T., *The development of Mathematics.* McGraw-Hill, New York, 1940.

[8] Bell, D. J. and D. H. Jacobson, *Singular Optimal Control Problems.* Academic Press, New York, 1973.

[9] Beltrami, E., *Saggio di interpetazione della geometria non-euclidea.* Giornale di Matematiche **VI**, 1868, pp. 284-312 (in Itialian). Eng. transl. in Stillwell, [108], pp. 7–34.

[10] Berkovitz, L. D., *Optimal Control Theory.* Springer-Verlag, New York, 1974.

[11] Berna, P., N. Conze, M. Duflo, M. L'evy-Nahas, M. Rais, P. Renouard, M. Vergne, *Représentations des groupes résolubles.* Dunod, Paris, 1972.

[12] Bernoulli, Johann, *Curvatura radii in diaphanis nonuniformibus ... [The curvature of a ray in nonuniform media ...].* Acta Eruditorum, May 1697, pp. 206–211. (*Opera omnia* **I**, pp 187–193

[13] Bernoulli, Jakob, *Solutio problematum fraternorum ... una cum propositione reciproca aliorum [Solution of problems of my brother ... together with the proposition of others in turn].* Acta Eruditorum, May 1697, pp. 211–217. (*Opera* **I**, pp 768–778.

[14] Bliss, G. A., *Calculus of Variations.* Open Court Publishing Co., Chicago, 1925.

[15] Bolza, O., *Über der 'Anormalen Fall' beim Lagrangeschen und Mayerschen Problem mit gemischten Bedingungen und variablen Endpunkten.* Math. Ann. **LXXIV**, 1913, pp. 430–446.

[16] Bonola, R., *Non-Euclidean Geometry.* (Italian original: *La Geometria non-Euclidea,* The Open Court Publishing Co., 1912.) Eng. transl. by H. S. Carslaw, Dover, New York, 1955.

[17] Borelli, G. A., *Euclides restitutus.* Pisa, Italy, 1658.

[18] Bressan, A., *On differential relations with lower continuous right-hand side. An existence theorem.* J. Diff. Equations **37**, 1980, pp. 89–97.

[19] Bressan, A., *Solutions of lower semicontinuous differential inclusions on closed sets.* Rend. Sem. Mat. Univ. Padova **69**, 1983, pp. 99–107.

[20] Bressan, A., *A high-order test for optimality of bang-bang controls.* S.I.A.M. J. Control Opt. **23**, 1985, pp. 38–48.

[21] Bressan, A., *Directionally continuous selections and differential inclusions.* Funk. Ekvac. **31**, 1988, pp. 459–470.

[22] Bressan, A. and G. Colombo, *Extensions and selections of maps with decomposable values.* Studia Math. **90**, 1988, pp. 69–86.

[23] Bressan, A. and F. Rampazzo, *On differential systems with vector-valued impulsive controls.* Boll. Unione Matematica Italiana B(7) **3**, 1988, pp. 641–656.

[24] Bressan, A., *On the qualitative theory of lower semicontinuous differential inclusions.* J. Diff. Equations **77**, 1989, pp. 379–391.

[25] Bressan, A., *Upper and lower semicontinuous differential inclusions: a unified approach.* In *Nonlinear Controllability and Optimal Control*, H. J. Sussmann Ed., Marcel Dekker, New York, 1990, pp. 21–31.

[26] Bressan, A. and F. Rampazzo, "Impulsive control systems with commutative vector fields," *J. Optim. Theory Appl.* **71** (1991), pp.67-83.

[27] Bressan, A. and F. Rampazzo, "Impulsive control systems without commutativity assumptions, *J. Optim. Theory Appl.* **81** (1994), pp. 435-457.

[28] Brockett, R. W., "System theory on group manifolds and coset spaces," *S.I.A.M. J. Control* 10: 265-284 (1972).

[29] Buttazzo, G., V. Ferone, and B. Kawohl, *Minimum problems over sets of concave functions and related questions.* Math. Nachr. **173**, 1995, pp. 71–89.

[30] Cajori, F., *A History of Mathematics.* Chelsea Pub. Co., New York, 1980.

[31] Canudas de Wit, C. A., *Adaptive control for partially know systems: theory and applications.* Elsevier, Amsterdam-New York, 1988.

[32] Carathéodory, C., *Calculus of Variations and Partial Differential Equations of the First Order. Part II: Calculus of Variations.* Eng. transl. by R. B. Dean, 2nd edition, Holden Day, Inc., San Francisco, 1967. (Originally published as *Variationsrechnung und Partielle Differentialgleichungen erster Ordnung.* B. G. Teubner, Berlin, 1935.

[33] Casti, J. L., *Alternate realities: mathematical models of nature and man.* Wiley-Interscience, New York, 1989.

[34] Cataldi, P. A., *Operetta delle linee rette equidistanti et non equidistanti.* Bologna, Italy, 1603.

[35] Cataldi, P. A., *Agguinta all'operetta delle linee rette equidistanti et non equidistanti.* Bologna, Italy, 1604.

[36] Cesari, L., *Optimization—Theory and Applications.* Springer-Verlag, New York, 1983.

[37] Clarke, F. H., *The maximum principle under minimal hypotheses.* S.I.A.M. J. Control Optim. **14**,n 1976, pp. 1078–1091.

[38] Clarke, F. H., *Optimization and Nonsmooth Analysis.* Wiley Interscience, New York, 1983.

[39] Clarke, F. H. and R. B. Vinter, "The relationship between the maximum principle and dynamic programming," *S.I.A.M. J. Control Optim.* **25** (1987), no. 5, pp. 1291-1311.

[40] Clarke, F. H., *Methods of Dynamic and Nonsmooth Optimization.* CBMS-NSF Regional Conference Series in Applied Mathematics **57**. Society for Industrial and Applied Mathematics (S.I.A.M.), Philadelphia, PA, 1989.

[41] Clarke, F. H., Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory.* Grad. Texts in Math. **178**, Springer-Verlag, Berlin, 1998.

[42] Conan Doyle, A., *Silver Blaze.* In *The Complete Sherlock Holmes,* Doubleday, New York, 1930.

[43] Coolidge, J. L., *A History of Geometrical Methods.* Dover , New York, 1963. (Original edition; Oxford Univ. Press, 1940.)

[44] Descartes, R., *Dioptrique*, 1637.

[45] Dubins, L. E., *On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents.* Amer. J. Math. **79**, 1957, pp. 497-516.

[46] Dugundji, J., *Topology.* Allyn and Bacon, Inc., Boston, 1966.

[47] Dyson, F. J., *Missed opportunities.* Bull. Amer. Math. Soc. **78**, 1972, pp. 635-652.

[48] Elworthy, K. D., W. N. Everitt and E. B. Lee Eds., *Differential equations, dynamical systems, and control science : a Festschrift in honor of Lawrence Markus.* Marcel Dekker, New York, 1994.

[49] *Encyclopædia Britannica; Macropædia*. Vol. 17. Chicago, 1980.

[50] Euler, L., *Methodus inveniendi ... accepti.* Lausanne, Geneva, 1744; *Opera omnia*, ser. I, vol. 25, 1952.

[51] Fermat, P., *Letters to Martin Cureau de la Chambre.* In *Oeuvres, II*, pp. 355–359 (1657) and pp. 457–463 (1662).

[52] Fleming, W. H., and H. Mete Soner, *Controlled Markov Processes and Viscosity Solutions.* New York, Springer-Verlag, 1993.

[53] Fuller, A. T., *Study of an optimum nonlinear system.* J. Electronics Control **15**, 1963, pp. 63–71.

[54] Gabasov, V., and F. M. Kirillova, *High-order necessary conditions for optimality.* S.I.A.M. J. Control **10**, 1972, pp. 127–168.

[55] Gandz, S., *Die Harpedonaptai oder Seilspanner.* Quellen und Studien Gesch. der Math. **B 1**, 1930, pp. 255-277.

[56] Giaquinta, M., and S. Hildebrandt, *Calculus of Variations.* Springer-Verlag, Berlin-New York, 1996 (2 vols.).

[57] Goh, B. S., *The second variation for the Bolza problem.* S.I.A.M. J. Control **4**, 1966, pp. 309–325.

[58] Goldstein, H., *Classical Mechanics.* Addison-Wesley, 1980.

[59] Goldstine, H. H., *A History of the Calculus of Variations from the 17th to the 19th Century.* Springer-Verlag, New York, 1981.

[60] Guillen, M., *Five equations that changed the world: the power and poetry of mathematics.* New York, Hyperion, 1995.

[61] Heath, T. L., *A Manual of Greek Mathematics.* Clarendon Press, Oxford, 1931.

[62] Heeres, J. E., *The Discovery of Tasmania: Extracts from the Journal of Abel Janszoon Tasman.* Government Printer, Hobart, Tasmania, 1985.

[63] Hermann, R., *On the accessibility problem in control theory.* In International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz Eds., Academic Press, New York, 1963.

[64] Hermes, H., *Lie algebras of vector fields and local approximation of attainable sets.* S.I.A.M. J. Control and Optimization **16**, 1974, pp. 715–727.

[65] Hestenes, M. R., *A general problem in the calculus of variations with applications to paths of least time.* Rand Corporation RM-100, ASTIA Document No. AD-112382, 1950.

[66] Hestenes, M. R., *Elements of the calculus of variations.* In *Modern Mathematics for the Engineer,* E. F. Beckenbach Ed., McGraw-Hill, New York, 1956.

[67] Hestenes, M. R., *Calculus of Variations and Optimal Control Theory.* Wiley, 1966.

[68] Hughes, R., *The Fatal Shore.* Vintage Books, 1988; republished by Harvill Press, 1996.

[69] Huygens, C., *Traité de la lumière* (Leiden, 1690). In *Oeuvres complètes XIX*, 1737.

[70] Jurdjevic, V., and H. J. Sussmann, *Control systems on Lie groups.* J. Diff. Equations **12**, 1972, pp. 313–329.

[71] Jurdjevic, V. *Geometric Control Theory.* Cambridge U. Press, New York, 1997.

[72] Kazarinoff, N. D., *Geometric Inequalities.* New Mathematical Library No. 4, The Mathematical Association of America, Washington, DC, 1961.

[73] Kelley, H. J., R. E. Kopp and H. G. Moyer, *Singular extremals.* In *Topics in Optimization*, G. Leitman ed., Academic Press, New York, 1967.

[74] Kline, M., *Mathematics, the Loss of Certainty.* Oxford University Press, New York, 1980.

[75] Knobloch, H. W., *High Order Necessary Conditions in Optimal Control.* Springer-Verlag, Berlin, 1975.

[76] Krener, A. J., *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems.* S.I.A.M. J. Control **12**, 1974, pp. 43–52.

[77] Krener, A. J., *The higher order maximum principle and its application to singular extremals.* S.I.A.M. J. Control Opt. **15**, 1977, pp. 256–293.

[78] Kupka, I., *The ubiquity of Fuller's phenomenon.* In *Nonlinear Controllability and Optimal Control,* H. J. Sussmann Ed., Marcel Dekker, New York 1990, pp. 313–350.

[79] Lanczos, C., *The Variational Principles of Mechanics.* Univ. of Toronto Press, 1949.

[80] Landau, Y. D., *Adaptive control: the model reference approach.* Marcel Dekker, New York, 1979.

[81] Lavrentiev, M., *Sur quelques problèmes du calcul des variations.* Ann. Mat. Pura Appl. **4**, 1926, pp. 7–28.

[82] Lee, E. B., and L. Markus, *Foundations of Optimal Control Theory.* Wiley, New York, 1967.

[83] Leibniz, G. W., *Unicum opticae, catoptricae et dioptricae principium.* Acta Eruditorum **1**, 1683, pp. 186-190.

[84] Leibniz, G. W., *Nova methodus pro maximis et minimis, itemque tangentibus, qua nec fractas nec irrationales quantitates moratur, et singulari pro illis calculus genus.* Acta Eruditorum **3**, 1684, pp. 467-473. Engl. transl. in Struik, [109], pp. 272-280.

[85] Liu, W. S., and H. J. Sussmann, *Shortest paths for sub-riemannian metrics on rank-2 distributions.* Memoirs of the American Math. Society **564**, Vol. 118, Nov. 1995.

[86] Lobry, C., *Controlabilité des systèmes non-linéaires.* S.I.A.M. J. Control **8**, 1970, pp. 573–605.

[87] Lochak, G., *La Géométrisation de la Physique.* Flammarion, Paris, 1994.

[88] Marchal, C., *Chattering arcs and chattering controls.* J. Optim Theory Appl. **11**, 1973, pp. 441–468.

[89] Markov, A. A., *Some examples of the solution of a special kind of problem in greatest and least quantities,* (in Russian). Soobshch. Karkovsk. Mat. Obshch. **1**, 1887, pp. 250–276.

[90] McShane, E. J., *On multipliers for Lagrange problems.* Amer. J. Math. **LXI**, 1939, pp. 809–819.

[91] Motta, M. and F. Rampazzo, *Dynamic programming for nonlinear systems driven by ordinary and impulsive controls.* S.I.A.M. J. Control Optim. **34**, 1996, pp. 199–225.

[92] Motta, M. and F. Rampazzo, *Nonlinear systems with unbounded controls and state constraints: a problem of proper extension.* NoDEA – Nonlinear Differential Equations and Applications **3,2**, 1996, pp. 191–216. [GNU: CHECK THIS REFERENCE, IF NEEDED]

[93] Pars, L. A., *An Introduction to the Calculus of Variations.* Heinemann, London, 1962.

[94] Pears, I., *An Instance of the Fingerpost.* Vintage, London, 1998.

[95] Perec, G., *La Disparition.* Éditions Denoël, Paris, 1969.

[96] Perec, G., *A Void*, English translation of [95], by G. Adair, Harper-Collins, London, 1994.

[97] Poincaré, H., *Theory of Fuchsian groups.* Acta Math. **I**, 1882, pp. 1–62.

[98] Pontryagin, L. S., V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes.* Wiley, New York, 1962.

[99] Rampazzo, F. and C. Sartori, *The minimum time function with unbounded controls.* to appear in J. of Math. Est. Systems Control. [GNU: UPDATE THIS REFERENCE, IF NEEDED]

[100] Reeds, J. A., and L. A. Shepp, *Optimal paths for a car that goes both forwards and backwards.* Pacific J. Math. **145**, 1990, pp. 367–393.

[101] Saccheri, G., *Logica demonstrativa.* Turin, Italy, 1697.

[102] Saccheri, G., *Euclides ab omni naevo vindicatus: sive conatus geometricus quo stabiliuntur prima ipsa universae Geometriae Principia.* Milan, Italy, 1733.

[103] Sastry, S. S., *Adaptive control: stability, convergence and robustness.* Prentice Hall, Englewood Cliffs NJ, 1989.

[104] Seidenberg, A., *Peg and chord in ancient Greek geometry.* Scripta Mathematica **XXIV**, 1959, pp. 107–122.

[105] Seidenberg, A., *The ritual origin of geometry.* Arch. Hist. Exact Sciences **1**, 1962, pp. 488–527.

[106] Seidenberg, A., *The origin of mathematics.* Arch. Hist. Exact Sciences **18**, 1977, pp. 301–342.

[107] Stillwell, J., *Mathematics and its History.* Springer-Verlag, New York, 1989.

[108] Stillwell, J., *Sources of Hyperbolic Geometry.* History of Mathematics, Vol. 10, Amer. Math. Society, Providence, RI, 1996.

[109] Struik, D. J., Ed., *A Source Book in Mathematics, 1200-1800.* Harvard U. Press, Cambridge, Mass. 1969.

[110] Subbotina, N. N., *The maximum principle and the superdifferential of the value function.* Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform. **18**, 1989, no. 3, pp. 151–160.

[111] Sussmann, H. J., and V. Jurdjevic, *Controllability of nonlinear systems.* J. Diff. Equations **12**, 1972, pp. 95–116.

[112] Sussmann, H. J., *Trajectory regularity and real analyticity: some recent results.* In *Proc. 25th I.E.E.E. Conference on Decision and Control, Athens, Greece, 1986,* I.E.E.E. publications, New York, 1986, pp. 3306–3312.

[113] Sussmann, H. J., *A weak regularity theorem for real analytic optimal control problems.* Re_vis_ta Matemática Iberoamericana **2**, No. 3, 1986, pp. 307–317.

[114] Sussmann, H. J., *A strong version of the Lojasiewicz maximum principle.* In *Optimal Control of Differential Equations,* N. H. Pavel Ed., Marcel Dekker, New York, 1994, pp. 293–309.

[115] Sussmann, H. J., *A strong version of the Maximum Principle under weak hypotheses.* In *Proc. 33rd I.E.E.E, Conference on Decision and Control, Orlando, FL, 1994*, I.E.E.E. publications, New York, 1994, pp. 1950-1956.

[116] Sussmann, H. J., and G. Q. Tang, *Shortest paths for the Reeds-Shepp car: a worked out example of the use of geometric techniques in nonlinear optimal control.* Rutgers Center for Systems and Control (Sycon) Report 91-10, September 1991. To appear in S.I.A.M. J. Control.

[117] Sussmann, H. J., *Shortest 3-dimensional paths with a prescribed curvature bound.* In *Proc. 34th I.E.E.E. Conference on Decision and Control, New Orleans, La, Dec. 1995*, I.E.E.E. publications, New York, 1995, pp. 3306–3312.

[118] Sussmann, H. J., *An introduction to the coordinate-free maximum principle.* In *Geometry of Feedback and Optimal Control,* B. Jakubczyk and W. Respondek Eds., Marcel Dekker, New York, 1997, pp. 463–557.

[119] Sussmann, H. J., *A strong maximum principle for systems of differential inclusions.* In *Proc. 35th I.E.E.E. Conference on Decision and Control, Kobe, Japan, Dec. 1996,* I.E.E.E. publications, 1996, pp. 1809–1814.

[120] Sussmann, H. J., *Some recent results on the maximum principle of optimal control theory.* In *Systems and Control in the Twenty-First Century,* C. I. Byrnes, B. N. Datta, D. S. Gilliam and C. F. Martin Eds., Birkhäuser, Boston, 1997, pp. 351–372.

[121] Sussmann, H. J., and J. C. Willems, *300 years of optimal control: from the brachystochrone to the maximum principle.* I.E.E.E. Control Systems Magazine, June 1997.[GNU: PAGE NUMBERS]

[122] Sussmann, H. J. *The Markov-Dubins problem with angular acceleration control.* In *Proc. 36th I.E.E.E. Conference on Decision and Control (CDC), San Diego, CA, Dec. 1997.* I.E.E.E. Publications, New York, 1997, pp. 2639-2643.

[123] Sussmann, H. J., *Multidifferential calculus: chain rule, open mapping and transversal intersection theorems.* In *Optimal Control: Theory, Algorithms, and Applications,* W. W. Hager and P. M. Pardalos, Editors, Kluwer Academic Publishers, 1998, pp. 436–487.

[124] Sussmann, H. J., *Geometry and optimal control.* In *Mathematical Control Theory*, J. Baillieul and J. C. Willems, Eds., Springer-Verlag, New York, 1998, pp. 140–198.

[125] Sussmann, H. J., *Transversality conditions and a strong maximum principle for systems of differential inclusions.* In Proc. 37th I.E.E.E. Conference on Decision and Control, Tampa, FL, Dec. 1998. I.E.E.E. publications, 1998, pp. 1–6.

[126] Sussmann, H. J., *Generalized differentials, flows, variations, and the maximum principle.* To appear in the book of proceedings of the 1999 AMS Short Course on Nonlinear Control, held in San Antonio, TX, in January 1999, during the joint winter meeting of the AMS and the MAA. The book will be published in the Symposia on Pure and Applied Mathematics series.

[127] Synge, J. L., *The life and early work of Sir William Rowan Hamilton.* Scripta Mathematica **X**, 1944, pp. 13–25.

[128] Taylor, H. G., *The Discovery of Tasmania.* Cat & Fiddle Press, Hobart, Tasmania, 1973.

[129] Tulczyjew, W. M., *A Symplectic Formulation of Relativistic Particle Dynamics.* Acta Physica Polonica, Vol. **B**8, 1977, p. 39. GNU NEED COMPLETE PAGE REF

[130] Tulczyjew, W. M., *Control of Static Mechanical Systems.* In *Proceedings of "Dynamical Systems and Microphysics".* CISM, Udine, Italy, 1984, p. 359. GNU NEED COMPLETE PAGE REF

[131] van der Waerden, B. L., *Geometry and Algebra in Ancient Civilizations.* Springer-Verlag, New York, 1983.

[132] Vitale, G., *Euclide restituto over gli antichi elementi geometrici ristaurati, e facilitati. Libri XV.* Rome, Italy, 1680.

[133] Wallis, J., *De Postulato Quinto; et Definitione Quinta; Lib. 6 Euclidis; discertatio geometrica.* Opera Math. t. II, Oxford, 1693, pp. 669–678.

[134] Yourgrau, W., and S. Mandelstam, *Variational Principles in Dynamics and Quantum Theory.* W. B. Saunders & Co., Philadelphia, 1968.

[135] Weierstrass, K., *Vorlesungen über Variationsrechnung; Mathematische Werke, VII,* (in German). Leipzig, Akadem. Verlagsges, 1927. Reprinted by Johnson Reprint Corp., New York, 1967.

[136] Zelikin, M. I., and V. F. Borisov, *Theory of Chattering Control, with Applications to Astronautics, Robotics, Economics and Engineering.* Birkhäuser, Boston, 1994.

COMMENT NO. 46. THE MAP IN PAGE 16 COMES FROM THE WEB SITE

http://pacific.vita.org/pacific/dutch/tasman.htm

IS THERE A COPYRIGHT PROBLEM?

COMMENT NO. 47. THE PICTURE IN PAGE 11 COMES FROM THE WEB SITE

http://www.norfacad.pvt.k12.va.us/project/bernoull/bern.htm

IS THERE A COPYRIGHT PROBLEM?

COMMENT NO. 48. Re: Page 44. Maybe we should put in some stuff about transversality conditions, after all!

COMMENT NO. 49. Re: Page 44. Say something about dynamic programming here.

COMMENT NO. 50. Re: Footnote 17 in page 45. Do we really do that in the appendix?

COMMENT NO. 51. Re: Remark 3.9.4 in page 49: must put more stuff on the Dubins-Reeds-Shepp problem.

COMMENT NO. 52. Re: Page 44. Check that we do indeed show in §3.10 what we say we do.

COMMENT NO. 53. Re: Page 67 Put explanation of why (II) is equivalent to (I).

COMMENT NO. 54. Re: Page 68. Put elementary proof of Z1 and Z2

COMMENT NO. 55. The Euler footnote 7 in page 72 has to be expanded, especially adding stuff about Euler after 1727, and in particular on Euler's friendship with Daniel B., and how it affected the relation with Johann B. (I am thinking of Euler passing on to Johann the manuscript he got from Daniel on the hydrodinamics law, which was then plagiarized by Johann. **Maybe this footnote should be changed into a Scholium?**

COMMENT NO. 56. Re: Page 72. We need something biographical on Lagrange. It should include the moving story of his great love for a much younger woman when he was old. (I think her name was Renée).

COMMENT NO. 57. Re: Page 74 We need some biographical stuff on Legendre.

COMMENT NO. 58. Re: page 74. Need biographical sketch for Hamilton

COMMENT NO. 59. Re: page 74. We should clarify the following issue: was Hamilton the first person to write "Hamilton's equations"? I have heard somewhere (I wish I could remember where) that it was Cauchy. Notice that in the text I talk about the "classical Hamiltonian" but I avoid saying "the Hamiltonian that Hamilton wrote" for this reason. We need some research here.

COMMENT NO. 60. Re: Page 84 Need biographical stuff about Weierstrass

COMMENT NO. 61. Re: Page 90. Need biographical stuff about Caratheorody

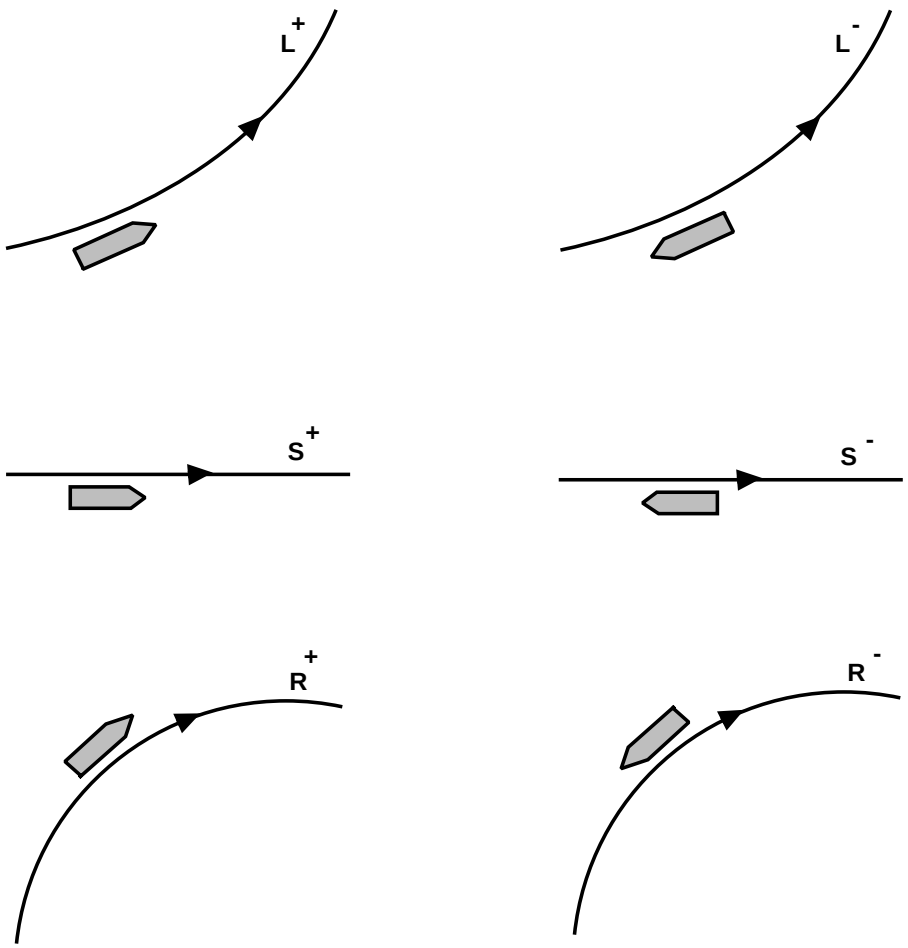


FIGURE 11. Snell's law

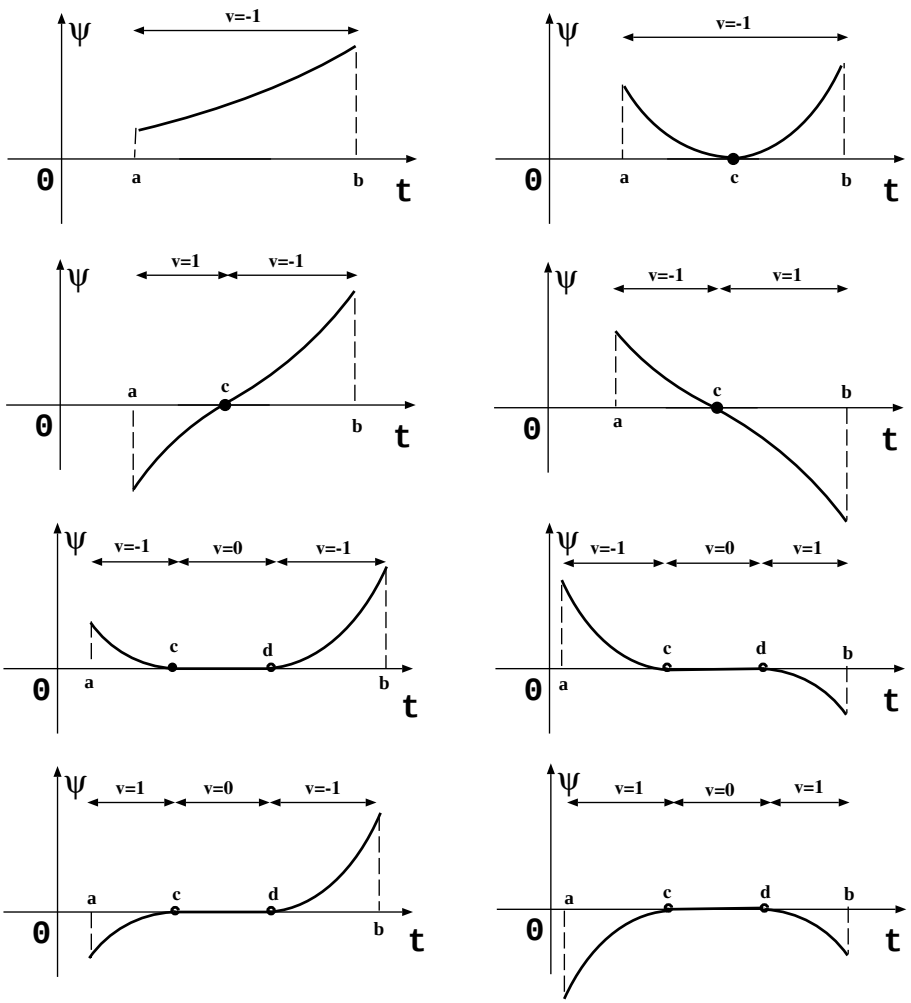


FIGURE 12. Snell's law

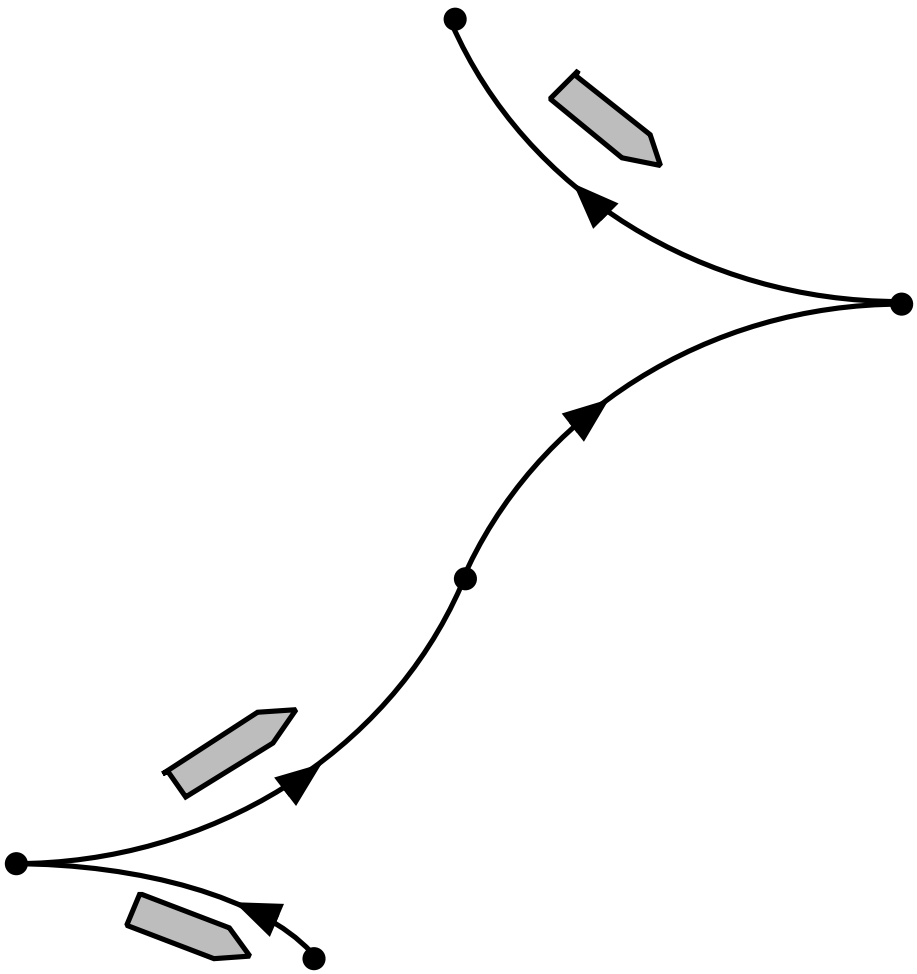


FIGURE 13. Snell's law

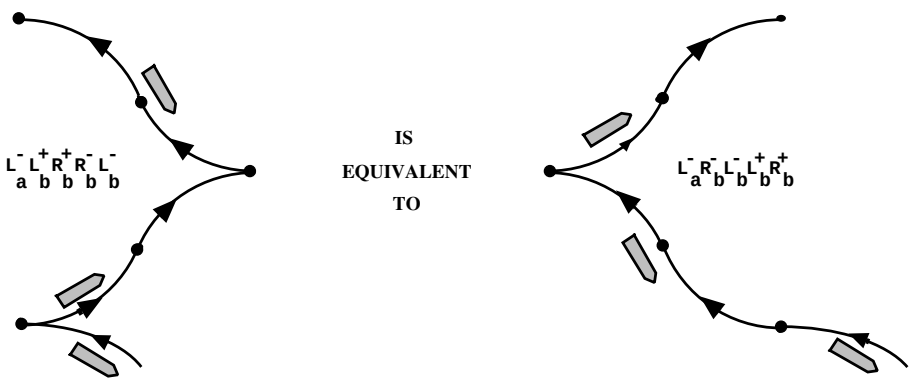
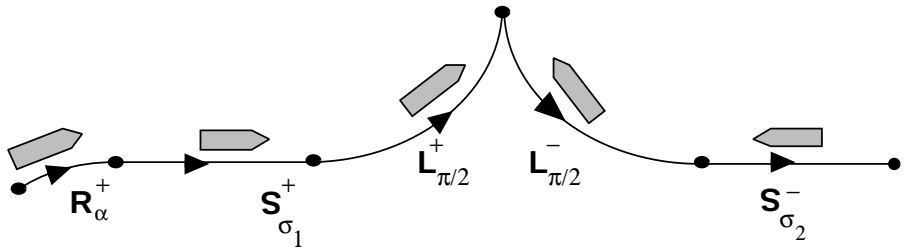
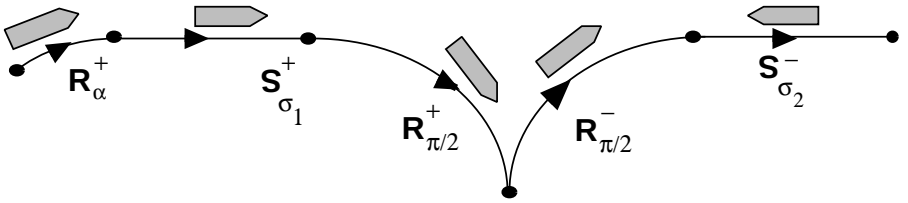


FIGURE 14. Snell's law



AND



ARE EQUIVALENT TO

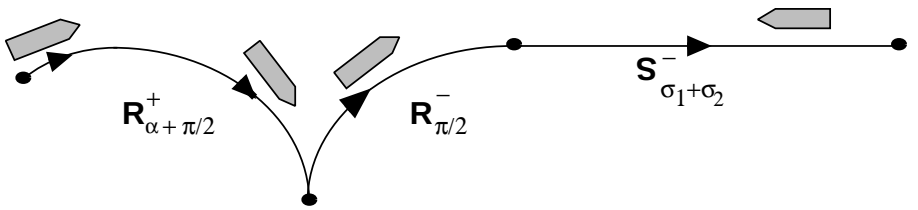


FIGURE 15. Snell's law

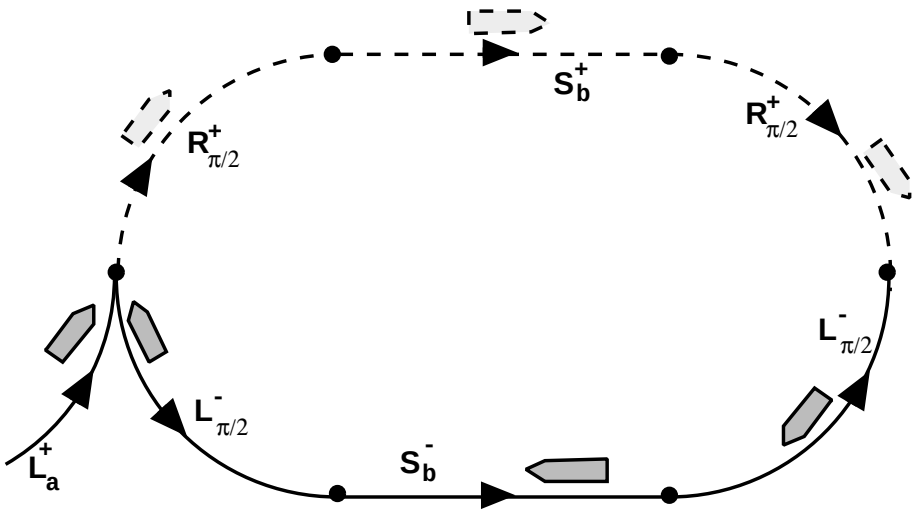


FIGURE 16. Snell's law

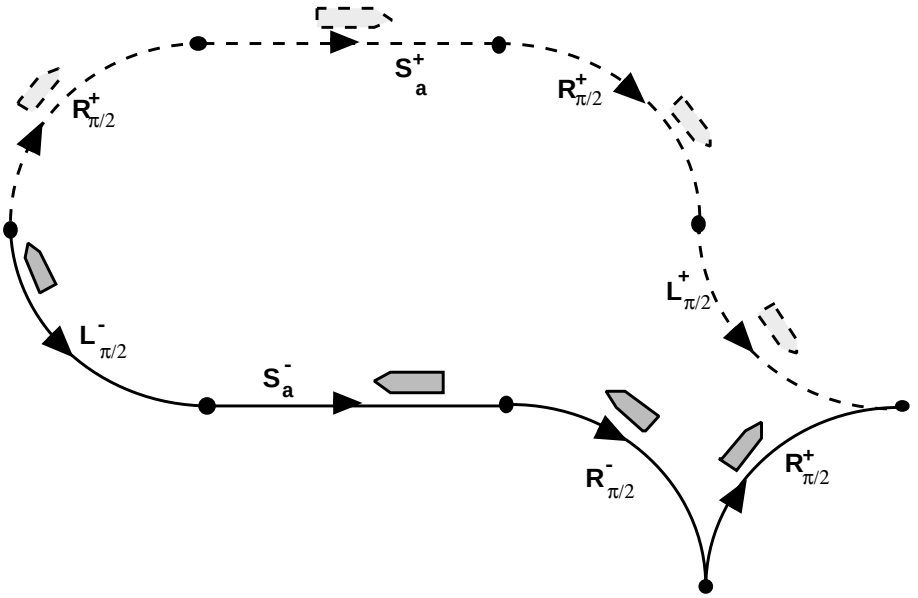


FIGURE 17. Snell's law

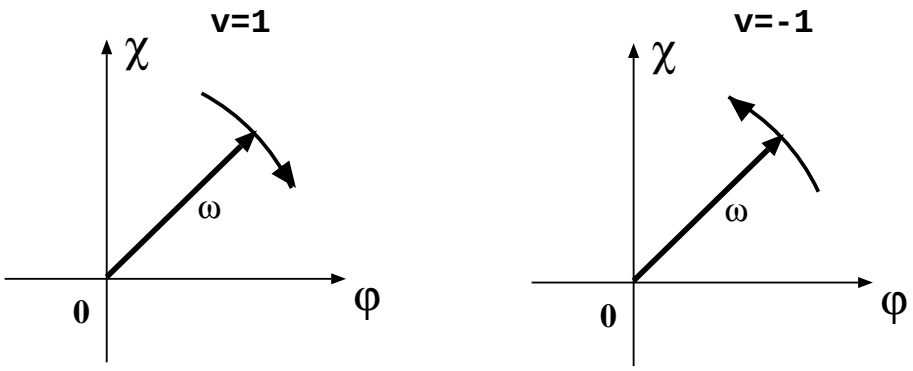


FIGURE 18. Snell's law

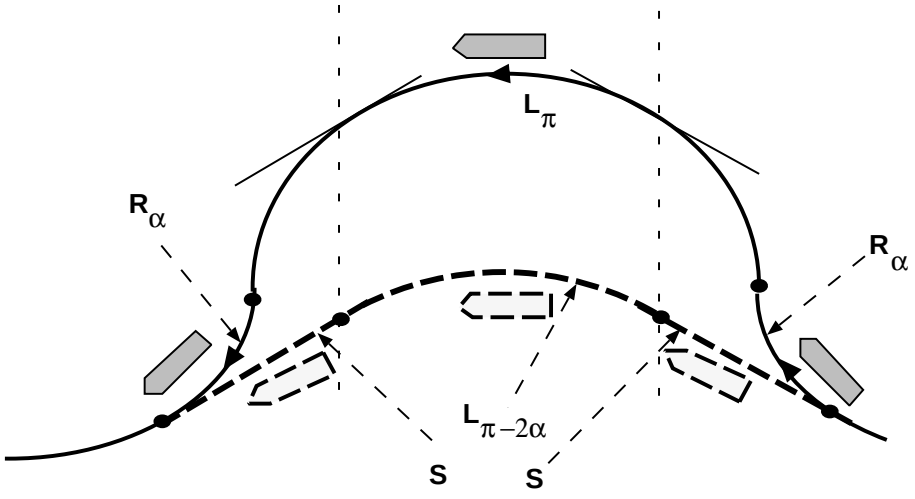
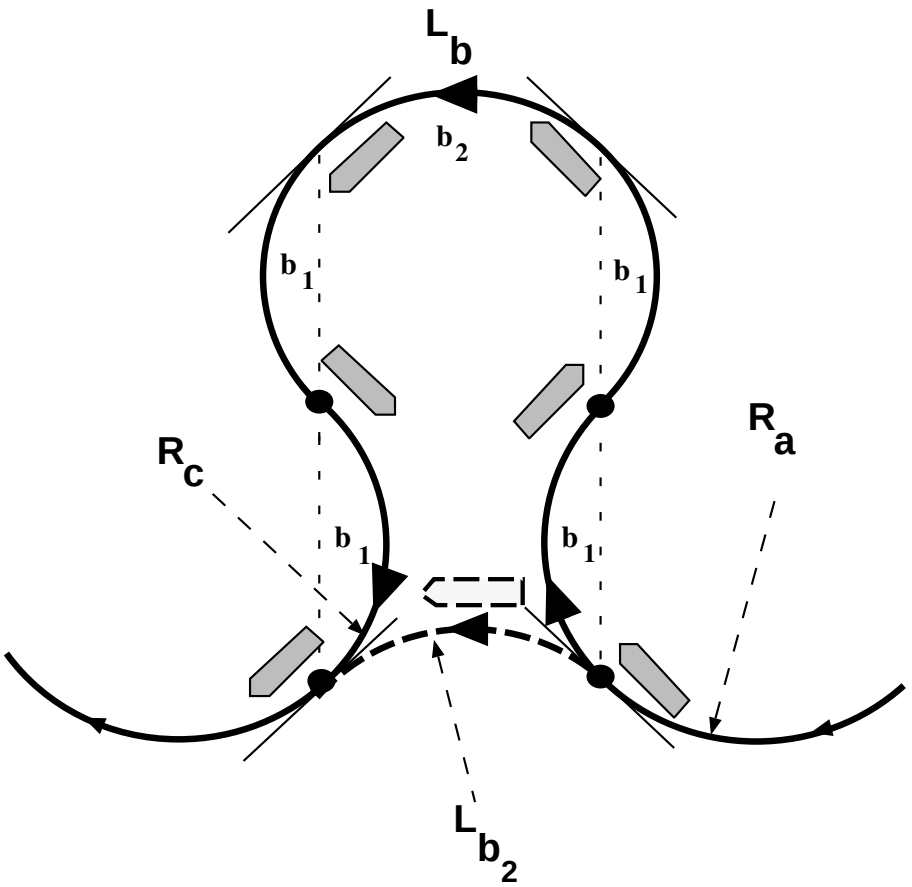


FIGURE 19. Snell's law



$$
\begin{aligned}
2\,b_1 + b_2 &= b \\
b_1 + b_2 &= \pi \\
b_1 &= b - \pi
\end{aligned}
$$

FIGURE 20. Snell's law

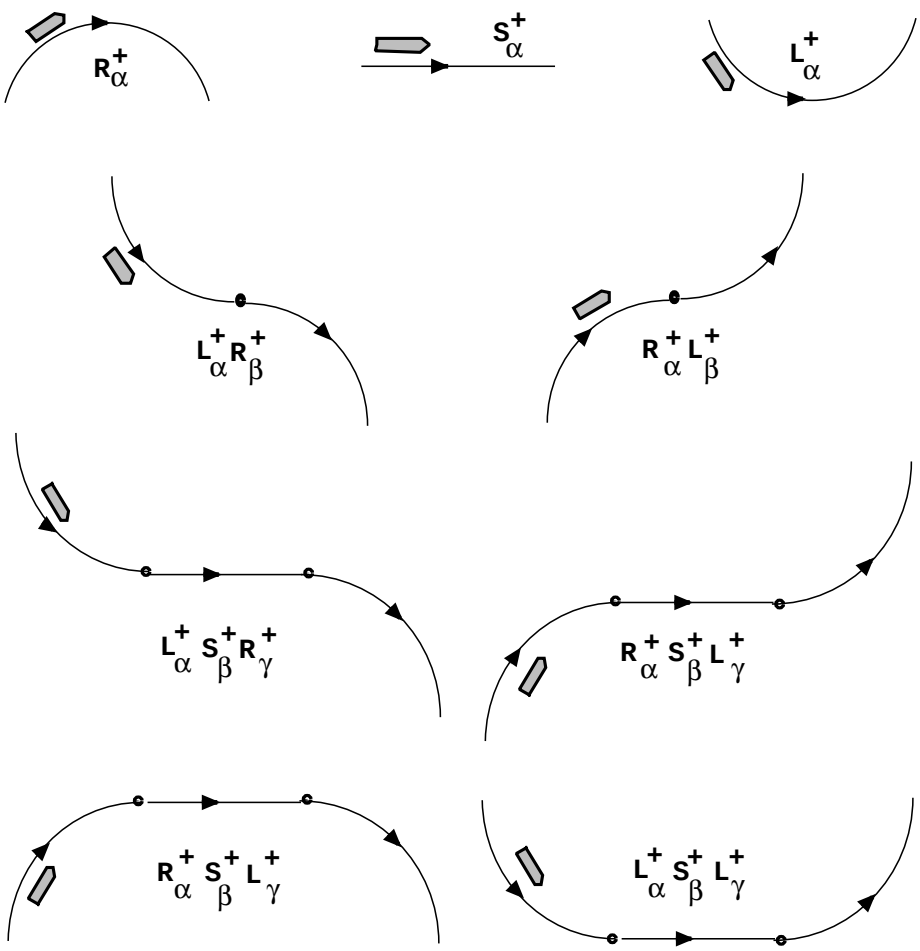


FIGURE 21. Snell's law

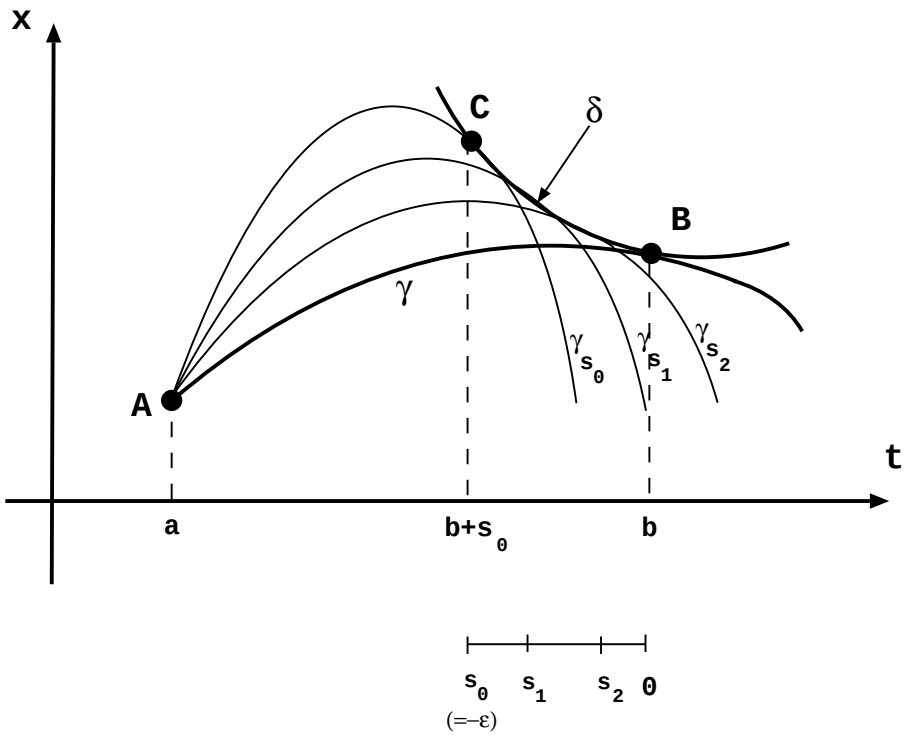


FIGURE 22. Snell's law

14.1 Example II: the isoperimetric problem

.

14.1.1 Problem statement

The standard statement of the isoperimetric problem is as follows:

(IP) *Of all simple, closed, rectifiable curves of a given length L, find those that enclose the largest possible area.*

To make this mathematically precise we would have, to begin with, to give precise definitions of all the terms involved. We then have to give a formula for the area enclosed by a simple closed curve, and solve the maximization problem. Some of these issues—those of defining "simple," "closed," "rectifiable," and "length"— are easy to take care of, so we will do it here.

Here we will show how to use the necessary conditions for an optimum to derive the classical result that the solution of the isoperimetric problem exists and is a circle. If this is so, the radius R of this circle is determined by the equation $2\pi R = L$, and then the enclosed area is πR^2, i.e., $\frac{L^2}{4\pi}$.

The solution presented here is not completely rigorous, because

1. We will assume that the "area enclosed" by a simple, closed, rectifiable curve ξ is the line integral

$$A(\xi) \stackrel{\text{def}}{=} \frac{1}{2}\int_a^b (x(t)\dot{y}(t) - y(t)\dot{x}(t))\,dt\,, \tag{A.4}$$

   if $\xi : [a, b] \mapsto \mathbb{R}^2$ and we write $\xi(t) = (x(t), y(t))$.

2. We will prove that the solution is a circle under the assumption that a related maximization problem has a solution.

The precise justification of these two assumptions will be given later, in §14.2.

14.1.2 Problem reformulations

From now on, L is a fixed positive real number. We let Γ_L^1 be the set of all curves $[0, 1] \ni t \mapsto \gamma(t) = (x(t), y(t)) \in \mathbb{R}^2$ such that

1. γ is absolutely continuous;

2. $\gamma(1) = \gamma(0)$;

3. γ has length L, that is,

$$\int_0^1 \|\dot{\gamma}(t)\| = \int_0^1 \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\,dt = L\,;$$

4. γ is simple (that is, the restriction of γ to the half-open interval $[0, 1[$ is one-to-one);

5. γ is oriented counterclockwise.

The precise meaning of the words "oriented counterclockwise" is as follows: if $\gamma : [0,1] \mapsto \mathbb{R}^2$ is continuous, simple, and such that $\gamma(0) = \gamma(1)$, then the Jordan Curve Theorem (cf. Theorem 14.2.23 below) implies that the complement of the set $|\gamma| = \{\gamma(t) : t \in [0,1]\}$ has two connected components, one bounded and one unbounded; let $\mathrm{Ins}(\gamma)$ denote the bounded connected component of $\mathbb{R}^2 \backslash |\gamma|$; then the winding number $\nu(\xi, p)$ is either equal to 1 for all $p \in \mathrm{Ins}(\gamma)$ or equal to -1 for all $p \in \mathrm{Ins}(\gamma)$, by Theorem 14.2.24. (The definitions of "connected component" and "winding number" are given later, in §14.2.11 and §14.2.12.)

If $\gamma \in \Gamma^1_L$, then $\mathrm{Ins}(\gamma)$ is the "region enclosed by γ." We will admit that

(A1) *If a curve* $[0,1] \ni t \mapsto \gamma(t) = (x(t), y(t)) \in \mathbb{R}^2$ *belongs to* Γ^1_L*, then the area of the region* $\mathrm{Ins}(\gamma)$ *is equal to* $A(\gamma)$*, where*

$$A(\gamma) \stackrel{\mathrm{def}}{=} \frac{1}{2} \int_0^1 (x(t)\dot{y}(t) - y(t)\dot{x}(t))\, dt\,.$$

(This is "Green's formula," proved below, cf. Theorem 14.3.4.)

To any curve $\gamma \in \Gamma^1_L$ we can associate the curve $\xi_\gamma : [0, L] \mapsto \mathbb{R}^2$ obtained from γ by *reparametrization by arc-length*. Precisely, write $\gamma(t) = (x(t), y(t))$ for $t \in [0,1]$, and define a function $s_\gamma : [0,1] \mapsto [0, L]$ (the "arc-length along γ") by letting

$$s_\gamma(t) = \int_0^t \sqrt{\dot{x}(u)^2 + \dot{y}(u)^2}\, du\,.$$

Then s_γ is a monotonically nondecreasing continuous function from $[0,1]$ onto $[0, L]$. Furthermore, if $0 \le t_1 < t_2 \le 1$ and $s_\gamma(t_1) = s_\gamma(t_2)$, then $\int_{t_1}^{t_2} \sqrt{\dot{x}(u)^2 + \dot{y}(u)^2}\, du = s_\gamma(t_2) - s_\gamma(t_1) = 0$, so $\dot{x}(u) \equiv \dot{y}(u) \equiv 0$ for almost all $t_1 \le u \le t_2$, and then $\gamma(s_\gamma(t_2)) = \gamma(s_\gamma(t_1))$. Therefore there exists a unique map $\xi_\gamma : [0, L] \mapsto \mathbb{R}^2$ such that $\xi_\gamma(s) = \gamma(t)$ whenever $s = s_\gamma(t)$. If $0 \le s_1 < s_2 \le L$, then we can pick $t_1, t_2 \in [0,1]$ such that $s_1 = s_\gamma(t_1)$ and $s_2 = s_\gamma(t_2)$. Then $t_1 < t_2$, and

$$\begin{aligned} \|\xi_\gamma(s_2) - \xi_\gamma(s_1)\| &= \|\gamma(t_2) - \gamma(t_1)\| \\ &= \|\int_{t_1}^{t_2} \dot{\gamma}(t)\, dt\| \\ &\le \int_{t_1}^{t_2} \|\dot{\gamma}(t)\|\, dt \\ &= s_\gamma(t_2) - s_\gamma(t_1) = s_2 - s_1\,. \end{aligned}$$

Therefore $\|\xi_\gamma(s_2) - \xi_\gamma(s_1)\| \leq |s_2 - s_1|$ whenever $s_1, s_2 \in [0, L]$. So ξ_γ *is Lipschitz, with constant* 1. It follows that ξ_γ is absolutely continuous and $\left\|\frac{d}{ds}(\xi_\gamma(s))\right\| \leq 1$ for almost all $s \in [0, L]$.

Formally, the identity $\xi_\gamma \circ s_\gamma = \gamma$ implies that

$$\frac{d}{dt}(\gamma(t)) = \frac{d}{ds}(\xi_\gamma(s))\Big|_{s=s_\gamma(t)} \cdot \frac{d}{dt}(s_\gamma(t))\,,$$

and then

$$\left\|\frac{d}{dt}(\gamma(t))\right\| = \left\|\frac{d}{ds}(\xi_\gamma(s))\Big|_{s=s_\gamma(t)}\right\| \cdot \frac{d}{dt}(s_\gamma(t))\,,$$

so that, making the change of variables $s = s_\gamma(t)$, we find

$$\begin{aligned}\int_0^L \left\|\frac{d}{ds}(\xi_\gamma(s))\right\| ds &= \int_0^1 \left\|\frac{d}{ds}(\xi_\gamma(s))\Big|_{s=s_\gamma(t)}\right\| \cdot \frac{ds_\gamma}{dt}(t)\, dt \\ &= \int_0^1 \left\|\frac{d}{dt}(\gamma(t))\right\| dt \\ &= L\,.\end{aligned}$$

Therefore

$$\left\|\frac{d}{ds}(\xi_\gamma(s))\right\| = 1 \text{ for almost all } s \in [0, L]\,. \tag{A.5}$$

The above argument is of course nonrigorous. For a rigorous proof of (A.30), see Proposition 14.2.7 below.

Clearly,

So our problem can be restated as follows:

(IP') Find a curve $\gamma_* \in \Gamma_L$ that maximizes the functional $\Gamma_L \ni \gamma \mapsto A(\gamma) \in \mathbb{R}$.

We associate to (IP') another problem involving a larger class of curves. We let $\tilde{\Gamma}_L$ be the set of all curves $\gamma : [0, L] \mapsto \mathbb{R}^2$ such that

1. γ is absolutely continuous,
2. $\|\dot{\gamma}(t)\| \leq 1$ for almost all $t \in [0, L]$,
3. $\gamma(L) = \gamma(0) = (0, 0)$,

and consider the problem

(IP") Find a curve γ_* belonging to $\tilde{\Gamma}_L$ that maximizes the functional $\tilde{\Gamma}_L \ni \gamma \mapsto A(\gamma) \in \mathbb{R}$.

It is cleat that $\Gamma_L \subseteq \tilde{\Gamma}_L$. Therefore

Proposition 14.1.1 *If (IP") has a solution γ_* such that $\gamma_* \in \Gamma_L$, then γ_* is a solution of (IP').* ♢

14.1.3 Finding the solutions of (IP")

To find the solutions of (IP"), we apply the necessary conditions for optimality given in the conjecture. Suppose that

$$[0, L] \ni t \mapsto \gamma_*(t) = (x(t), y(t)) \in \mathbb{R}^2$$

is a solution of (IP"). It then follows from the conjecture on necessary conditions for an optimum that there exists an absolutely continuous "momentum" function

$$[0, L] \ni t \mapsto \mu(t) = (m_x(t), m_y(t)) \in \mathbb{R}^2$$

such that the Hamilton equations

$$\dot{m}_x(t) = -\frac{\partial H}{\partial x}(\gamma_*(t), \mu(t), \dot{\gamma}(t)) , \tag{A.6}$$

$$\dot{m}_y(t) = -\frac{\partial H}{\partial y}(\gamma_*(t), \mu(t), \dot{\gamma}(t)) , \tag{A.7}$$

and the Hamiltonian maximization condition

$$H(\gamma_*(t), \mu(t), \dot{\gamma}_*(t)) = \max \left\{ H(\gamma_*(t), \mu(t), w) : w \in \mathbb{R}^2, \|w\| \leq 1 \right\} \tag{A.8}$$

hold for almost all $t \in [0, L]$. Here H is the Hamiltonian, that is, the function

$$\mathbb{R}^6 \ni (x, y, p_x, p_y, u, v) \mapsto H(x, y, p_x, p_y, u, v) \stackrel{\text{def}}{=} p_x u + p_y v + xv - yu . \tag{A.9}$$

In view of (A.9), we can rewrite (A.6), (A.7), as

$$\dot{m}_x(t) = -\dot{y}(t) , \tag{A.10}$$

$$\dot{m}_y(t) = \dot{x}(t) . \tag{A.11}$$

Define

$$\alpha(t) = m_x(t) - y(t) , \qquad \beta(t) = m_y(t) + x(t) .$$

Then

$$\dot{\alpha}(t) = -2\dot{y}(t)\,, \tag{A.12}$$
$$\dot{\beta}(t) = 2\dot{x}(t)\,, \tag{A.13}$$

and

$$H(\gamma_*(t), \mu(t), u, v) = \alpha(t)u + \beta(t)v\ . \tag{A.14}$$

Therefore, if we define

$$\nu(t) = \sqrt{\alpha(t)^2 + \beta(t)^2}\,, \tag{A.15}$$

(A.8) says that

$$\left.\begin{aligned} \dot{x}(t) &= \tfrac{\alpha(t)}{\nu(t)}\,, \\ \dot{y}(t) &= \tfrac{\beta(t)}{\nu(t)}\,, \end{aligned}\right\} \quad \text{if} \quad \nu(t) \neq 0\,. \tag{A.16}$$

If $\nu(t) = 0$ then (A.8) allows the vector $\dot{\gamma}_*(t)$ to be an arbitrary member of $\{w \in \mathbb{R}^2 : \|w\| \leq 1\}$. We can combine both cases into the following restatement of (A.16) which is also valid when $\nu(t) = 0$:

$$\nu(t)\dot{x}(t) = \alpha(t) \qquad \text{and} \qquad \nu(t)\dot{y}(t) = \beta(t)\,. \tag{A.17}$$

A simple calculation then shows that

$$\frac{d}{dt}(\nu^2) = \frac{d}{dt}(\alpha^2 + \beta^2) = 2\alpha\dot{\alpha} + 2\beta\dot{\beta} = 2\nu\dot{x}\cdot(-2\dot{y}) + 2\nu\dot{y}\cdot 2\dot{x} = 0\,.$$

Therefore *the function ν is constant.* Let $\bar{\nu}$ be the constant value of $\nu(t)$. If $\bar{\nu} = 0$ then $\alpha(t) \equiv \beta(t) \equiv 0$, so $\dot{\alpha}(t) \equiv \dot{\beta}(t) \equiv 0$, and then (A.12) and (A.13) imply that $\dot{x}(t) \equiv \dot{y}(t) \equiv 0$, so γ_* is a constant curve and $A(\gamma_*) = 0$. It is clear that there do exist curves $\gamma \in \tilde{\Gamma}_L$ for which $A(\gamma) > 0$. Therefore the fact that $A(\gamma_*) = 0$ contradicts the assumption that γ_* is a solution of (IP"). So we can exclude the possibility that $\bar{\nu} = 0$.

From now on we assume that $\bar{\nu} > 0$. Then (A.16) implies that $\dot{\gamma}$ is continuous, and

$$\|\dot{\gamma}(t)\| = 1 \qquad \text{for all } t \in [0, L]\,. \tag{A.18}$$

Write

$$\xi(t) = \dot{x}(t)\,, \qquad \eta(t) = \dot{y}(t)\,.$$

Then, using (A.12), (A.13), and (A.16) we can conclude that

$$\dot{\xi} \equiv -\frac{2\eta}{\bar{\nu}} \quad \text{and} \quad \dot{\eta} \equiv \frac{2\xi}{\bar{\nu}}\,.$$

Therefore

$$\ddot{\xi} + \frac{4\xi}{\bar{\nu}^2} \equiv \ddot{\eta} + \frac{4\eta}{\bar{\nu}^2} \equiv 0\,.$$

This implies that

$$\eta(t) = c_1 \cos\Big(\frac{2t}{\bar{\nu}}\Big) + c_2 \sin\Big(\frac{2t}{\bar{\nu}}\Big)$$

for some constants c_1, c_2. If we choose $t_0 \in \mathbb{R}$ and $r \in [0, \infty[$ such that

$$c_1 = r\cos\Big(\frac{2t_0}{\bar{\nu}}\Big)\,, \qquad c_2 = r\sin\Big(\frac{2t_0}{\bar{\nu}}\Big)\,,$$

we find that

$$\eta(t) = r\cos\Big(\frac{2(t-t_0)}{\bar{\nu}}\Big)\,.$$

Then

$$\dot{\xi}(t) = \frac{\bar{\nu}\eta(t)}{2} = -r\sin\Big(\frac{2(t-t_0)}{\bar{\nu}}\Big)\,.$$

Since $\|\dot{\gamma}(t)\| = 1$, we have $r = 1$, and we have shown that

$$\xi(t) = -\sin\Big(\frac{2(t-t_0)}{\bar{\nu}}\Big) \qquad \text{and} \qquad \eta(t) = \cos\Big(\frac{2(t-t_0)}{\bar{\nu}}\Big)\,. \tag{A.19}$$

The fact that $\gamma(L) = \gamma(0)$ implies that $\bar{\nu} = \frac{L}{\pi k}$ for some nonzero integer k. Therefore

$$\xi(t) = -\sin\Big(\frac{2\pi k(t-t_0)}{L}\Big) \qquad \text{and} \qquad \eta(t) = \cos\Big(\frac{2\pi k(t-t_0)}{L}\Big)\,. \tag{A.20}$$

We then get, by integration,

$$\begin{aligned} x(t) &= \bar{x} + \frac{L}{2\pi k}\cos\Big(\frac{2\pi k(t-t_0)}{L}\Big)\,, \\ y(t) &= \bar{y} + \frac{L}{2\pi k}\sin\Big(\frac{2\pi k(t-t_0)}{L}\Big)\,, \end{aligned}$$

where

$$\begin{aligned} \bar{x} &= \frac{-L}{2\pi k}\cos\Big(\frac{2\pi k t_0}{L}\Big)\,, \\ \bar{y} &= \frac{L}{2\pi k}\sin\Big(\frac{2\pi k t_0}{L}\Big)\,. \end{aligned}$$

The above formulas define a family $\Big\{\gamma^{t_0,k}\Big\}_{t_0\in\mathbb{R},k\in\mathbb{Z}}$ of curves belonging to $\tilde{\Gamma}_L$, depending on the real parameter t_0 and the nonzero integer parameter k.

If we let $\bar{\gamma}^{t_0,k}$ be the point with coordinates $\frac{-L}{2\pi k}\cos\left(\frac{2\pi k t_0}{L}\right)$, $\frac{L}{2\pi k}\sin\left(\frac{2\pi k t_0}{L}\right)$, then the curves $\gamma^{t_0,k}$ satisfy

$$\left\|\gamma^{t_0,k}(t) - \bar{\gamma}^{t_0,k}\right\| \equiv \frac{L}{2\pi k}\,.$$

Therefore *the locus* $|\gamma^{t_0,k}|$ *of* $\gamma^{t_0,k}$ *is a circle of radius* $\frac{L}{2\pi k}$. It is easy to verify that *the map* $[0,L] \ni t \mapsto \gamma^{t_0,k}(t) \in \mathbb{R}^2$ *is parametrized by arc-length and covers the circle* $|\gamma^{t_0,k}|$ $|k|$ *times.* Moreover, *the orientation derived from this map is counterclockwise if* $k > 0$ *and clockwise if* $k < 0$.

Let us now compute the numbers $A(\gamma^{t_0,k})$. We have

$$\begin{aligned} A(\gamma^{t_0,k}) &= \int_0^L (x(t)(\dot{y}(t) - y(t)\dot{x}(t))\,dt \\ &= \int_0^L \Big(\bar{x} + \frac{L}{2\pi k}\cos\Big(\frac{2\pi k(t-t_0)}{L}\Big)\Big)\cos\Big(\frac{2\pi k(t-t_0)}{L}\Big)\,dt \\ &\quad - \int_0^L \Big(\bar{y} + \frac{L}{2\pi k}\sin\Big(\frac{2\pi k(t-t_0)}{L}\Big)\Big)\Big(-\sin\Big(\frac{2\pi k(t-t_0)}{L}\Big)\Big)\,dt \\ &= \frac{L^2}{2\pi k}\,. \end{aligned}$$

The numbers $A(\gamma^{t_0,k})$ are therefore negative for $k < 0$ (corresponding to the fact that for such values of k the curve $\gamma^{t_0,k}$ is oriented clockwise) and positive for $k > 0$. Furthermore, the largest of these values is $\frac{L^2}{2\pi}$, which is attained for $k = 1$.

In addition, when $k = 1$ the curves $\gamma^{t_0,k}$ belong to Γ_L, because they are simple, parametrized by arc-length, and oriented counterclockwise. It follows that *the curves* $\gamma^{t_0,1}$ are solutions of (IP') if they are solutions of (IP"). Moreover, if a solution of (IP") exists, then it has to be one of the $\gamma^{t_0,1}$, because we have shown that the $\gamma^{t_0,k}$ are the only members of $\tilde{\Gamma}$ that satisfy the necessary conditions for an optimum. In that case, then all the $\gamma^{t_0,1}$ are solutions, because the functional A has the same value for all of them.

Summarizing, we have shown

Proposition 14.1.2 *Assume that (IP") has a solution. Then the solutions of (IP") are the circles* $\gamma^{t_0,1}$, *for* $t_0 \in \mathbb{R}$. *Moreover, these circles are also the solutions of the isoperimetric problem (IP'). In particular, the area enclosed by the area-maximizing circles is* $\frac{L^2}{4\pi}$. ♢

14.2 Rigorous solution of the isoperimetric problem

14.2.1 Existence of area-maximizers

To conclude our complete solution of (IP'), we have to prove that (IP") has a solution. The proof is quite simple. The set $\tilde{\Gamma}_L$ is a subset of $C^0([0,L],\mathbb{R}^2)$ of continuous functions $\gamma : [0,L] \mapsto \mathbb{R}^2$, which is a Banach space in a well-known waym with the sup norm. We first prove the following result.

Proposition 14.2.1 *$\tilde{\Gamma}_L$ is a compact subset of $C^0([0,L],\mathbb{R}^2)$.*

PROOF. Let $\{\gamma_j\}_{j=1}^{\infty}$ be a sequence of members of $\tilde{\Gamma}_L$. We want to extract a subsequence $\{\gamma_{j(k)}\}_{k=1}^{\infty}$ that converges uniformly to a limit $\gamma \in \tilde{\Gamma}_L$.

Since every member γ of $\tilde{\Gamma}_L$ is absolutely continuous and satisfies $\|\dot{\gamma}(t)\| \leq 1$ for a.e. t, and $\gamma(0) = (0,0)$, it is clear that the sequence $\{\gamma_j\}_{j=1}^{\infty}$ is uniformly bounded and equicontinuous. So there exists a subsequence $\{\gamma_{j(k)}\}_{k=1}^{\infty}$ that converges uniformly to a limit γ, and our conclusion will follow if we prove that $\gamma \in \tilde{\Gamma}_L$. Since $\gamma_j(0) = \gamma_j(L) = (0,0)$ for every j, it is clear that $\gamma(0) = \gamma(L) = (0,0)$. Since γ_j is absolutely continuous and $\|\dot{\gamma}_j(t)\| \leq 1$ for almost all j, it follows that

$$\|\gamma_j(t) - \gamma_j(s)\| = \| \int_s^t \dot{\gamma}_j(r)\, dr\| \leq |t-s|$$

for all j and all $t, s \in [0,L]$. Therefore

$$\|\gamma(t) - \gamma(s)\| \leq |t-s| \quad \text{for all} \quad t,s \in [0,L]\,.$$

This says that γ is Lispchitz with constant 1, so γ is absolutely continuous and $\|\dot{\gamma}(t)\| \leq 1$ for a.e. t. Therefore $\gamma \in \tilde{\Gamma}_L$. ♢

We now prove the continuity of A.

Proposition 14.2.2 *The functional A is continuous on $\tilde{\Gamma}_L$.*

PROOF. Let $\{\gamma_j\}_{j=1}^{\infty}$ be a sequence of members of $\tilde{\Gamma}_L$ that converges uniformly to a limit $\gamma_\infty \in \tilde{\Gamma}_L$. We want to prove that

$$\lim_{j\to\infty} \Big(A(\gamma_j) - A(\gamma_\infty)\Big) = 0\,. \tag{A.21}$$

Write

$$\gamma_j(t) = (x_j(t), y_j(t)) \quad \text{for} \quad j \in \{1, 2, \ldots\} \cup \{\infty\}\,,\ t \in [0, L]\,.$$

Then

$$A(\gamma_j) - A(\gamma_\infty) = I_j^1 - I_j^2\,,$$

where

$$\begin{aligned} I_j^1 &= \int_0^L \Big(x_j(t)\dot{y}_j(t) - x_\infty(t)\dot{y}_\infty(t)\Big)\,dt\,, \\ I_j^2 &= \int_0^L \Big(y_j(t)\dot{x}_j(t) - y_\infty(t)\dot{x}_\infty(t)\Big)\,dt\,. \end{aligned}$$

We must therefore show that

$$\lim_{j\to\infty} I_j^1 = \lim_{j\to\infty} I_j^2 = 0\,.$$

Let us prove that $\lim_{j\to\infty} I_j^1 = 0$. (The proof that $\lim_{j\to\infty} I_j^2 = 0$ is similar.) Write

$$I_j^1 = \hat{I}_j + \tilde{I}_j\,,$$

where

$$\begin{aligned} \hat{I}_j &= \int_0^L \Big(x_j(t) - x_\infty(t)\Big)\dot{y}_j\,dt\,, \\ \tilde{I}_j &= \int_0^L x_\infty(t)\Big(\dot{y}_j(t) - \dot{y}_\infty(t)\Big)\,dt\,. \end{aligned}$$

We will show that

$$\lim_{j\to\infty} \hat{I}_j = \lim_{j\to\infty} \tilde{I}_j = 0\,.$$

The fact that $\lim_{j\to\infty} \hat{I}_j = 0$ is trivial, because $x_j \to x_\infty$ uniformly, and the functions $\dot{y}_j$ are uniformly bounded, since $|\dot{y}_j(t)| \le \|\dot{\gamma}_j(t)\| \le 1$.

To prove that $\lim_{j\to\infty} \tilde{I}_j = 0$, we pick a real number ε such that $\varepsilon > 0$, and find J such that $|\tilde{I}_j| \le \varepsilon$ whenever $j \ge J$. Let h be a function of class C^1 on $[0, L]$ such that

$$|x_\infty(t) - h(t)| \le \frac{\varepsilon}{4L} \quad \text{whenever} \quad t \in [0, L]\,.$$

(The existence of such an h follows, for example, from the Stone-Weierstrass theorem, since x_∞ is continuous.) Then

$$\tilde{I}_j = \int_0^L \Big(x_\infty(t) - h(t)\Big)\Big(\dot{y}_j(t) - \dot{y}_\infty(t)\Big)\,dt + \int_0^L h(t)\Big(\dot{y}_j(t) - \dot{y}_\infty(t)\Big)\,dt\,.$$

The absolute value of the first integral is bounded by $\frac{\varepsilon}{2}$, because $|x_\infty(t) - h(t)| \leq \frac{\varepsilon}{4L}$ and $|\dot{y}_j(t) - \dot{y}_\infty(t)| \leq 2$. To estimate the second integral, we first rewrite it using integrationby parts:

$$\begin{aligned}\int_0^L h(t)\Big(\dot{y}_j(t) - \dot{y}_\infty(t)\Big)\,dt &= h(L)\Big(y_j(L) - y_\infty(L)\Big) - h(0)\Big(y_j(0) - y_\infty(0)\Big) \\ &\quad - \int_0^L \dot{h}(t)\Big(y_j(t) - y_\infty(t)\Big)\,dt \\ &= -\int_0^L \dot{h}(t)\Big(y_j(t) - y_\infty(t)\Big)\,dt\,,\end{aligned}$$

using the fact that $y_j(0) = y_\infty(0) = y_j(L) = y_\infty(L) = 0$. Since $y_j \to y_\infty$ uniformly, and $\dot{h}$ is a fixed bounded function, the limit of $\int_0^L h(t)\Big(\dot{y}_j(t) - \dot{y}_\infty(t)\Big)\,dt$ as $j \to \infty$ is zero. Therefore we can find J such that

$$\Big|\int_0^L h(t)\Big(\dot{y}_j(t) - \dot{y}_\infty(t)\Big)\,dt\Big| \leq \frac{\varepsilon}{2} \quad \text{whenever} \quad j \geq J\,.$$

It then follows that

$$|\tilde{I}_j| \leq \varepsilon \quad \text{whenever} \quad j \geq J\,,$$

completing our proof. ♢

It then follows from Propositions 3 and 4 that

Proposition 14.2.3 *The maximization problem (IP") has a solution.* ♢

14.2.2 The final result

By combining Propositions 14.1.2 and 14.2.3, we get

Theorem 14.2.4 *The solutions of the modified isoperimetric problem (IP") are the circles $\gamma^{t_0,1}$, for $t_0 \in \mathbb{R}$. Moreover, these circles are also the solutions of the original isoperimetric isoperimetric problem (IP'). In particular, the area enclosed by the area-maximizing circles is $\frac{L^2}{4\pi}$.* ♢

14.2.3 Polygonal loops

If n is a positive integer, a *plane polygonal loop with n sides* (or *plane n-polygonal loop*) is a loop ξ in $\mathbb{R}^2$ such that there exist times $t_1, t_2, \ldots, t_n$ for

which $0 = t_1 < t_2 < \ldots < t_n < 1$ having the property that for $j = 1, \ldots, n$ the map $[t_j, t_{j+1}] \ni t \mapsto \xi(e^{2\pi i t}) \in \mathbb{C}$ is linear. A plane n-polygonal loop is completeTly determined by the times $t_1, t_2, \ldots, t_n$ and the n points $A_1, \ldots, A_n$ defined by letting $A_j = \xi(e^{2\pi i t_j})$ for $j = 1, \ldots, n$. Furthermore, changing the choices of $t_1, t_2, \ldots, t_n$ results in a reparametrization of ξ, which does not change the carrier $|\xi|$ nor the winding numbers $\nu(\xi, p)$ for $p \notin |\xi|$. We can therefore talk unambiguously about the carrier and the winding numbers of the "polygonal with vertices $A_1, \ldots, A_n$."

A *plane triangle* is a plane 3-polygonal loop.

The *vertices* of a triangle

A triangle ξ is *nondegenerate* if $|\xi|$ is not entirely contained in a straight line. It is easy to see that a triangle ξ with vertices A_1, A_2, A_3 is nondegenerate if and only if the vectors $A_2 - A_1$ and $A_3 - A_1$ are linearly independent. If ξ is a nondegenerate triangle, we say that ξ *oriented counterclockwise* if the determinant of the 2 by 2 suqare matrix with columns $A_2 - A_1$ and $A_3 - A_1$ is positive.

14.2.4 Rectifiable curves

A *partition* of a compact interval $[a, b]$ is a finite sequence $\mathbf{t} = (t_0, t_1, \ldots, t_N)$ such that $a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$.

If $\xi : [a, b] \mapsto \mathbb{R}^n$ is a curve, and $\mathbf{t} = (t_0, t_1, \ldots, t_N)$ is a partition of $[a, b]$, then the $\mathbf{t}$-*approximate length* of ξ is the number $\mathrm{Le}(\xi, \mathbf{t})$ defined by

$$\mathrm{Le}(\xi, \mathbf{t}) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{N} \|\xi(t_j) - \xi(t_{j-1}\| \, .$$

The *length* $\mathrm{Le}(\xi)$ of ξ is the supremum of the numbers $\mathrm{Le}(\xi, \mathbf{t})$ over all the partitions $\mathbf{t}$ of $[a, b]$. The curve ξ is said to be *rectifiable* if $\mathrm{Le}(\xi)$ is finite.

14.2.5 Absolutely continuous curves

Proposition 14.2.5 *Let $\gamma : [a, b] \mapsto \mathbb{R}^n$ be an absolutely continuous map. Then γ is rectifiable and $\mathrm{Le}(\gamma) = \int_a^b \|\dot{\gamma}(t)\| \, dt$.*

PROOF. Let $\mathbf{t} = (t_0, \ldots, t_N)$ be a partition of the interval $[a, b]$. Then $\gamma(t_j) - \gamma(t_{j-1}) = \int_{t_{j-1}}^{t_j} \dot{\gamma}(t) \, dt$, so $\|\gamma(t_j) - \gamma(t_{j-1})\| \leq \int_{t_{j-1}}^{t_j} \|\dot{\gamma}(t)\| \, dt$. There-

fore

$$\mathrm{Le}(\gamma, \mathbf{t}) = \sum_{j=1}^{N} \|\gamma(t_j) - \gamma(t_{j-1})\| \le \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} \|\dot\gamma(t)\|\, dt = \int_a^b \|\dot\gamma(t)\|\, dt\,.$$

Hence

$$\mathrm{Le}(\gamma) \le \int_a^b \|\dot\gamma(t)\|\, dt\,, \tag{A.22}$$

and in particular γ is rectifiable. To complete our proof, we have to show that $\mathrm{Le}(\gamma) \ge \int_a^b \|\dot\gamma(t)\|\, dt$.

Fix a positive number ε. Then pick positive numbers β and λ such that $2\beta + 2\lambda(b-a) < \varepsilon$, and choose α such that $\alpha > 0$ and $\int_E \|\dot\gamma\| < \beta$ whenever E is a measurable subset of $[a,b]$ such that $\mathrm{meas}(E) < \alpha$. Using Lusin's theorem, select a compact subset K of $[a,b]$ such that $\mathrm{meas}([a,b]\backslash K) < \alpha$ and the restriction of $\dot\gamma$ to K is continuous. Let δ be such that $\delta > 0$ and $\|\dot\gamma(t_2) - \dot\gamma(t_1)\| < \lambda$ whenever $t_1, t_2 \in K$ and $|t_2 - t_1| < \delta$. Let $E = [a,b]\backslash K$.

Now pick a partition $\mathbf{t} = (t_0, \ldots, t_N)$ of $[a,b]$ such that $t_j - t_{j-1} < \delta$ for $j = 1, \ldots, N$. Then

$$\int_a^b \|\dot\gamma(t)\|\, dt - \sum_{j=1}^{N} \|\gamma(t_j) - \gamma(t_{j-1})\| = \sum_{j=1}^{N} I_j\,,$$

where $I_j = I_j^1 - I_j^2$, $I_j^1 = \int_{t_{j-1}}^{t_j} \|\dot\gamma(t)\|\, dt$, $I_j^2 = \|U_j\|$, and $U_j = \int_{t_{j-1}}^{t_j} \dot\gamma(t)\, dt$. Write $\mu_j = \mathrm{meas}(K \cap [t_{j-1}, t_j])$.

Fix j, and assume that $K \cap [t_{j-1}, t_j] \ne \emptyset$. Then

$$\int_{K\cap[t_{j-1},t_j]} \dot\gamma(t)\, dt = U_j - \int_{[t_{j-1},t_j]\cap E} \dot\gamma(t)\, dt\,,$$

so

$$\Big\| \int_{K\cap[t_{j-1},t_j]} \dot\gamma(t)\, dt \Big\| \le \|U_j\| + \int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\, dt\,.$$

Pick $\hat t \in K \cap [t_{j-1}, t_j] \ne \emptyset$. Then

$$\mu_j \dot\gamma(\hat t) = \int_{K\cap[t_{j-1},t_j]} \dot\gamma(t)\, dt + \int_{K\cap[t_{j-1},t_j]} (\dot\gamma(\hat t) - \dot\gamma(t))\, dt\,,$$

from which it follows that

$$\begin{aligned} \mu_j \|\dot\gamma(\hat t)\| &\le \| \int_{K\cap[t_{j-1},t_j]} \dot\gamma(t)\, dt\| + \int_{K\cap[t_{j-1},t_j]} \|\dot\gamma(\hat t) - \dot\gamma(t)\|\, dt \\ &\le \|U_j\| + \int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\, dt + \lambda\mu_j\,. \end{aligned}$$

On the other hand, if $t \in K \cap [t_{j-1}, t_j]$, then

$$\|\dot\gamma(t)\| = \|\dot\gamma(t) - \dot\gamma(\hat t) + \dot\gamma(\hat t)\| \leq \|\dot\gamma(t) - \dot\gamma(\hat t)\| + \|\dot\gamma(\hat t)\| \leq \lambda + \|\dot\gamma(\hat t)\| \,.$$

Therefore

$$\int_{K\cap[t_{j-1},t_j]} \|\dot\gamma(t)\|\,dt \leq \lambda\mu_j + \mu_j\|\dot\gamma(\bar t)\| \leq \|U_j\| + \int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt + 2\lambda\mu_j \,.$$

If we add $\int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt$ to both sides, we get

$$\begin{aligned} I_j^1 &= \int_{t_{j-1}}^{t_j} \|\dot\gamma(t)\|\,dt \\ &\leq \lambda\mu_j + \mu_j\|\dot\gamma(\hat t)\| \\ &\leq \|U_j\| + 2\int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt + 2\lambda\mu_j \,. \end{aligned} \tag{A.23}$$

This inequality was proved assuming that $K \cap [t_{j-1}, t_j] \neq \emptyset$, but is obviously true as well if the set $K \cap [t_{j-1}, t_j]$ is empty, because in that case $\int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt = I_j^1$. Therefore (A.23) holds for all j. Hence

$$I_j = I_j^1 - \|U_j\| \leq 2\int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt + 2\lambda\mu_j \,.$$

If we add over all j, we find

$$\sum_{j=1}^N I_j \leq 2\int_E \|\dot\gamma\| + 2\lambda(b-a), \leq 2\beta + 2\lambda(b-a) < \varepsilon \,.$$

Therefore $\int_a^b \|\dot\gamma(t)\|\,dt \leq \varepsilon + \sum_{j=1}^N \ \|\gamma(t_j) - \gamma(t_{j-1})\| \leq \varepsilon + \mathrm{Le}(\gamma)$. Since ε is arbitrary, we have shown that

$$\int_a^b \|\dot\gamma(t)\|\,dt \leq \mathrm{Le}(\gamma) \,. \tag{A.24}$$

In view of (A.22) and (A.24), our conclusion is proved. ♢

14.2.6 Parametrization by arc length

If $\xi : [a, b] \mapsto \mathbb{R}^n$ is a curve, then we can define the *arc-length function* s_ξ of ξ to be the function $s_\xi : [a, b] \mapsto \mathbb{R} \cup \{+\infty\}$ given by

$$s_\xi(t) = \mathrm{Le}(\xi_t) \,,$$

where, for $t \in [a, b]$, we write ξ_t to denote the restriction of ξ to $[a, t]$, i.e., the curve $[a, t] \ni \tau \mapsto \xi(\tau) \in \mathbb{R}^n$.

Proposition 14.2.6 *Suppose that $a, b \in \mathbb{R}$, $a \le b$, and $\xi : [a,b] \mapsto \mathbb{R}^n$ is a rectifiable curve, and $L = \mathrm{Le}(\xi)$. Then the arc length s_ξ is a continuous monotonically nondecreasing function from $[a,b]$ to $\mathbb{R}$ that satisfies $s_\xi(a) = 0$ and $s_\xi(b) = L$.*

PROOF. The facts that $s_\xi(a) = 0$ and $s_\xi(b) = L$ follow trivially from the definition of length. If $a \le t < t' \le b$, and $\mathbf{t} = (t_0, t_1, \ldots, t_N)$ is a partition of $[a,t]$, then $\mathbf{t}' = (t_0, t_1, \ldots, t_N, t')$ is a partition of $[a,t']$, and $\mathrm{Le}(\xi_{t'}, \mathbf{t}') = \mathrm{Le}(\xi_t, \mathbf{t}) + \|\xi(t') - \xi(t)\|$, so $\mathrm{Le}(\xi_t, \mathbf{t}) \le \mathrm{Le}(\xi_{t'}, \mathbf{t}') \le s_\xi(t')$. Taking the supremum over all the partitions $\mathbf{t}$ of $[a,t]$, we find $s_\xi(t) \le s_\xi(t')$. This proves the monotonicity of s_ξ, from which it follows that the values of s_ξ are finite, since $s_\xi(b) < \infty$.

To prove that s_ξ is continuous, we fix a $t \in [a,b]$ and let

$$\lambda_- = \lim_{\tau < t, \tau \uparrow t} s_\xi(\tau)\,, \qquad \lambda_+ = \lim_{\tau > t, \tau \downarrow t} s_\xi(\tau)\,.$$

(If $t = a$ we let $\lambda_- = 0$. If $t = b$ we let $\lambda_+ = L$.)

We will prove that $\lambda_+ = \lambda_-$. Let $\alpha = \lambda_+ - \lambda_-$. Fix a positive number ε, and then choose a positive δ such that $\|\xi(\tau) - \xi(t)\| \le \varepsilon$ whenever $\tau \in [a,b]$ and $|\tau - t| \le \delta$. Then pick $t' \in [a,b]$ such that $t \le t'$ and $s_\xi(t') \le \lambda_+ + \varepsilon$. Choose a partition $\mathbf{t}' = (t_0, t_1, \ldots, t_N)$ of the interval $[a,t']$ such that $\mathrm{Le}(\xi_{t'}, \mathbf{t}') \ge s_\xi(t') - \varepsilon$. Without loss of generality, we may assume that t itself is one of the partition points, and $t_j - t_{j-1} < \delta$ for all j, for otherwise we could modify $\mathbf{t}'$ by adding new partition points, making sure that t is one of them, and the inequality $\mathrm{Le}(\xi_{t'}, \mathbf{t}') \ge s_\xi(t') - \varepsilon$ would still hold. Let $j \in \{0, \ldots, N\}$ be such that $t = t_j$.

Suppose first that $0 < j < N$. We claim that, if $\hat{\mathbf{t}} = (t_0, \ldots, t_{j+1})$, then

$$\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) \ge s_\xi(t_{j+1}) - \varepsilon \ge \lambda_+ - \varepsilon\,. \tag{A.25}$$

Indeed, the second inequality of (A.25) follows from the definition of λ_+, the monotonicity of s_ξ, and the fact that $t_{j+1} > t$. If $\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) < s_\xi(t_{j+1}) - \varepsilon$, then there would exist a number ρ and a partition $\hat{\mathbf{u}} = (u_0, \ldots, u_M)$ of $[a, t_{j+1}]$ such that

$$\rho > \varepsilon\,, \quad \mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) < s_\xi(t_{j+1}) - \rho\,, \quad \text{and} \quad \mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{u}}) \ge s_\xi(t_{j+1}) - (\rho - \varepsilon)\,.$$

Then $\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{u}}) > \varepsilon + \mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}})$. If we let $\mathbf{u}' = (u_0, \ldots, u_M, t_{j+2}, \ldots, t_N)$, then $\mathbf{u}'$ is a partition of $[a,t']$ and $\mathrm{Le}(\xi_{t'}, \mathbf{u}') > \varepsilon + \mathrm{Le}(\xi_{t'}, \mathbf{t}') \ge s_\xi(t')$. The inequality $\mathrm{Le}(\xi_{t'}, \mathbf{u}') > s_\xi(t')$ clearly contradicts the definition of $s_\xi(t')$, so the first inequality of (A.25) is proved.

On the other hand, if $\tilde{\mathbf{t}} = (t_0, \ldots, t_{j-1})$, then

$$\mathrm{Le}(\xi_{t_{j-1}}, \tilde{\mathbf{t}}) \leq s_\xi(t_{j-1}) \leq \lambda_- \,. \tag{A.26}$$

It follows from (A.25) and (A.26) that

$$\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) - \mathrm{Le}(\xi_{t_{j-1}}, \tilde{\mathbf{t}}) \geq \lambda_+ - \varepsilon - \lambda_- = \alpha - \varepsilon \,. \tag{A.27}$$

But

$$\begin{aligned} \mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) - \mathrm{Le}(\xi_{t_{j-1}}, \tilde{\mathbf{t}}) &= \|\xi(t_{j+1} - \xi(t_j)\| + \|\xi(t_j - \xi(t_{j-1})\| \\ &= \|\xi(t_{j+1} - \xi(t)\| + \|\xi(t) - \xi(t_{j-1})\| \\ &\leq 2\varepsilon \,, \end{aligned}$$

and then (A.27) implies that $\alpha \leq 3\varepsilon$.

The inequality $\alpha \leq 3\varepsilon$ also follows easily if $j = 0$ or $j = N$. Since ε was arbitrary, α must vanish, and the continuity of s_ξ at t is proved. $\diamondsuit$

To any rectifiable curve $\xi : [a, b] \mapsto \mathbb{R}^n$ of length L we can associate the curve $\gamma_\xi : [0, L] \mapsto \mathbb{R}^n$ obtained from ξ by *reparametrization by arc-length*. The precise definition of γ_ξ is as follows. We know that the arc length s_ξ is a monotonically nondecreasing continuous function from $[a, b]$ onto $[0, L]$. Furthermore, s_ξ satisfies the inequality

$$s_\xi(t) + \|\xi(t') - \xi(t)\| \leq s_\xi(t') \quad \text{whenever} \quad a \leq t < t' \leq b \,. \tag{A.28}$$

(Proof: the inequality is trivial if $t = t'$; assume that $t < t'$; if $\mathbf{t} = (t_0, \ldots, t_N)$ is a partition of $[a, t]$, then $\mathbf{t}' = (t_0, \ldots, t_N, t')$ is a partition of $[a, t']$, and $\mathrm{Le}(\xi_t, \mathbf{t}) + \|\xi(t') - \xi(t)\| = \mathrm{Le}(\xi_{t'}, \mathbf{t}') \leq s_\xi(t')$. Taking the supremum over $\mathbf{t}$, we find that (A.28) holds.) It follows from (A.28) that if $a \leq t < t' \leq b$ and $s_\xi(t) = s_\xi(t')$ then $\xi(t) = \xi(t')$. Therefore there exists a unique map $\gamma_\xi : [0, L] \mapsto \mathbb{R}^n$ having the property that $\gamma_\xi(s) = \xi(t)$ whenever $s = s_\xi(t)$. If $0 \leq s < s' \leq L$, then we can pick $t, t' \in [a, b]$ such that $s = s_\xi(t)$ and $s' = s_\xi(t')$. Then $t < t'$, and (A.28) implies that $\|\gamma_\xi(s') - \gamma_\xi(s)\| \leq s' - s$.

Therefore

$$\|\gamma_\xi(s') - \gamma_\xi(s)\| \leq |s' - s| \quad \text{whenever} \quad s, s' \in [0, L] \,. \tag{A.29}$$

So γ_ξ *is Lipschitz with constant* 1. It follows that γ_ξ is absolutely continuous and $\left\| \frac{d}{ds}(\gamma_\xi(s)) \right\| \leq 1$ for almost all $s \in [0, L]$.

Formally, the identity $\xi_\gamma \circ s_\gamma = \gamma$ implies that

$$\frac{d}{dt}(\gamma(t)) = \frac{d}{ds}(\xi_\gamma(s))\Big|_{s=s_\gamma(t)} \cdot \frac{d}{dt}(s_\gamma(t))\,,$$

and then

$$\Big\|\frac{d}{dt}(\gamma(t))\Big\| = \Big\|\frac{d}{ds}(\xi_\gamma(s))\Big|_{s=s_\gamma(t)}\Big\| \cdot \frac{d}{dt}(s_\gamma(t))\,,$$

so that, making the change of variables $s = s_\gamma(t)$, we find

$$\begin{aligned}\int_0^L \Big\|\frac{d}{ds}(\xi_\gamma(s))\Big\|\,ds &= \int_0^1 \Big\|\frac{d}{ds}(\xi_\gamma(s))\Big|_{s=s_\gamma(t)}\Big\| \cdot \frac{ds_\gamma}{dt}(t)\,dt \\ &= \int_0^1 \Big\|\frac{d}{dt}(\gamma(t))\Big\|\,dt \\ &= L\,.\end{aligned}$$

Therefore

$$\Big\|\frac{d}{ds}(\xi_\gamma(s))\Big\| = 1 \text{ for almost all } s \in [0,L]\,. \tag{A.30}$$

The above argument is of course nonrigorous. For a rigorous proof of (A.30), see Proposition 14.2.7 below.

Proposition 14.2.7 *Let $\xi : [a,b] \mapsto \mathbb{R}^n$ be a rectifiable map of length L, and let γ_ξ be the reparametrization of ξ by arc-length. Then γ_ξ is Lipschitz and $\|\dot\gamma_\xi(s)\| = 1$ for almost all $s \in [0,L]$.*

PROOF. Inequality (A.29), which was rigorously proved in page 284, shows that γ_ξ is Lipschitz with constant 1. Hence $\|\dot\gamma_\xi(s)\| \le 1$ for almost all $s \in [0,L]$.

Let us show that $\mathrm{Le}(\gamma_\xi) = L$. Let $\mathbf{t} = (t_0,\ldots,t_N)$ be a partition of $[a,b]$. Let $\mathbf{s} = (s_0,\ldots,s_N)$, where $s_j = s_\xi(t_j)$. Then $\mathbf{s}$ is a partition of $[0,L]$, except for the fact that the s_j need not be all different. Let $\hat{\mathbf{s}}$ be the partition obtained by eliminating from $\mathbf{s}$ every s_j that is equal to some previous s_k. It is then clear that

$$\mathrm{Le}(\gamma_\xi) \ge \mathrm{Le}(\gamma_\xi,\hat{\mathbf{s}}) = \sum_{j=1}^N \|\gamma_\xi(s_j) - \gamma_\xi(s_{j-1})\| = \sum_{j=1}^N \|\xi(t_j) - \xi(s_{j-1})\|\,.$$

Hence $\mathrm{Le}(\gamma_\xi) \ge \mathrm{Le}(\xi,\mathbf{t})$ for an arbitrary partition $\mathbf{t}$ of $[a,b]$. Therefore $\mathrm{Le}(\gamma_\xi) \ge \mathrm{Le}(\xi) = L$. On the other hand, Proposition (14.2.5) tells us that $\mathrm{Le}(\gamma_\xi) = \int_0^L \|\dot\gamma_\xi(s)\|\,ds$, so $\mathrm{Le}(\gamma_\xi) \le L$, because $\|\dot\gamma_\xi(s)\| \le 1$ for almost all $s \in [0,L]$. Therefore $\mathrm{Le}(\gamma_\xi) = L$. Since $\int_0^L \|\dot\gamma_\xi(s)\|\,ds = L$, and $\|\dot\gamma_\xi(s)\| \le 1$ for almost all $s \in [0,L]$, it must be the case that $\|\dot\gamma_\xi(s)\| = 1$ for almost all $s \in [0,L]$, and our proof is complete. ♢

14.2.7 Changing variables

Suppose $a < b$, $c < d$, $f : [c, d] \mapsto \mathbb{R}^n$ is an integrable vector-valued function, and $\sigma : [a, b] \mapsto [c, d]$ is an absolutely continuous, monotonically nondecreasing function such that $\sigma(a) = c$ and $\sigma(b) = d$. Can we "change variables" in the integral $\int_c^d f(s)\, ds$ by writing $s = \sigma(t)$ and $ds = \dot{\sigma}(t)\, dt$, and conclude that the formula

$$\int_c^d f(s)\, ds = \int_a^b \dot{\sigma}(t) \cdot f(\sigma(t))\, dt \tag{A.31}$$

holds?

Notice, first of all, that if f is an "integrable function" in the usual sense (that is, an equivalence class of "true functions" from $[c, d]$ to $\mathbb{R}^n$ modulo equality almost everywhere, such that every member of the class is Lebesgue-integrable), then $f \circ \sigma$ is not a well defined "integrable function" on $[a, b]$, because if two functions f_1, f_2 on $[c, d]$ coincide on the complement of a null set E, and $g_1 = f_1 \circ \sigma$, $g_2 = f_2 \circ \sigma$, then all we can conclude is that g_1 and g_2 coincide on the complement of $\sigma^{-1}(E)$, and $\sigma^{-1}(E)$ need not be a null set. It turns out, however that this is not a serious difficulty because, as we will show, if E is a null set then $\dot{\sigma}(t) = 0$ for almost all $t \in \sigma^{-1}(E)$, so the function $\dot{\sigma} \cdot (f \circ \sigma)$ is well defined modulo a.e. equality. Furthermore, $\dot{\sigma} \cdot (f \circ \sigma)$ is integrable, because $\dot{\sigma}$ is integrable and $f \circ \sigma$ is bounded. So it is at least plausible that (A.31) might be true. The following proposition says that this is indeed so. For the reasons explained above, we state it first for "true functions," where a *true function* on $[a, b]$ is a function on $[a, b]$, rather than an equivalence class modulo a.e. equality.

Proposition 14.2.8 *Let $n \in \mathbb{N}$, $a, b, c, d \in \mathbb{R}$ be such that $a < b$ and $c < d$. Let $f : [c, d] \mapsto \mathbb{R}^n$ be a Lebesgue integrable true function. Assume that $\sigma : [a, b] \mapsto \mathbb{R}$ is an absolutely continuous, monotonically nondecreasing map such that $\sigma(a) = c$ and $\sigma(b) = d$. Then the true function*

$$[a, b] \ni t \mapsto \dot{\sigma}(t) \cdot f(\sigma(t))$$

is integrable, and (A.31) holds.

PROOF. If I is a closed interval $[\gamma, \delta]$ (where $c \leq \gamma \leq \delta \leq b$), then $\sigma^{-1}(I)$ is the interval $J = [\alpha, \beta]$, where $\alpha = \min\{s \in [a, b] : \sigma(s) = \gamma\}$ and $\beta = \max\{s \in [a, b] : \sigma(s) = \delta\}$. Therefore

$$\text{meas}(I) = \delta - \gamma = \sigma(\beta) - \sigma(\alpha) = \int_\alpha^\beta \dot{\sigma}(t)\, dt = \int_{\sigma^{-1}(I)} \dot{\sigma}(t)\, dt\,.$$

A similar argument applies if I is half-open half-closed, or if I is open. (For example, if $I =]\gamma, \delta]$, then β is defined as above, but α and J must be taken to be $\max\{s \in [a,b] : \sigma(s) = \gamma\}$, and $]\alpha, \beta]$, respectively.) So $\text{meas}(I) = \int_{\sigma^{-1}(I)} \dot{\sigma}(t)\, dt$ for all intervals I such that $I \subseteq [c,d]$. Using the formula $\sigma^{-1}(E \cup F) = \sigma^{-1}(E) \cup \sigma^{-1}(F)$, it follows that $\text{meas}(E) = \int_{\sigma^{-1}(E)} \dot{\sigma}(t)\, dt$ for every subset of $[c,d]$ which is a finite union of intervals, since every such subset is in fact a finite union of pairwise disjoint intervals.

Let $\mathcal{E}$ be the class of all Borel measurable subsets E of $[c,d]$ such that $\text{meas}(E) = \int_{\sigma^{-1}(E)} \dot{\sigma}(t)\, dt$, and let $\mathcal{I}$ be the class of subsets of $[c,d]$ that are finite unions of intervals. Then $\mathcal{I} \subseteq \mathcal{E}$. Furthermore, it is easy to see that $\mathcal{E}$ is a monotone class of subsets of $[c,d]$, that is, (i) if $\{E_j\}_{j=1}^{\infty}$ is a sequence of members of $\mathcal{E}$ such that $E_j \subseteq E_{j+1}$ for every j, and $E = \bigcup_{j=1}^{\infty} E_j$, then $E \in \mathcal{E}$, and (ii) if $\{E_j\}_{j=1}^{\infty}$ is a sequence of members of $\mathcal{E}$ such that $E_{j+1} \subseteq E_j$ for every j, and $E = \bigcap_{j=1}^{\infty} E_j$, then $E \in \mathcal{E}$. (Indeed, if $E \in \mathcal{E}$ and $F = [c,d] \backslash E$, then

$$\begin{aligned} \int_{\sigma^{-1}(F)} \dot{\sigma} &= \int_a^b \dot{\sigma} - \int_{\sigma^{-1}(E)} \dot{\sigma} = \sigma(b) - \sigma(a) - \text{meas}(E) \\ &= d - c - \text{meas}(E) = \text{meas}([c,d]) - \text{meas}(E) = \text{meas}(F)\,, \end{aligned}$$

so $F \in \mathcal{E}$. This shows that $\mathcal{E}$ is closed under complementation, from which it follows that if we prove (i) then (ii) is true as well. To prove (i), we assume that $E = \bigcup_{j=1}^{\infty} E_j$, $E_j \in \mathcal{E}$, and $\{E_j\}_{j=1}^{\infty}$ is increasing. Then

$$\begin{aligned} \text{meas}(E) &= \lim_{j\to\infty} \text{meas}(E_j) = \lim_{j\to\infty} \int_{\sigma^{-1}(E_j)} \dot{\sigma} = \lim_{j\to\infty} \int_a^b \dot{\sigma} \cdot \chi_{\sigma^{-1}(E_j)} \\ &= \int_a^b \dot{\sigma} \cdot \chi_{\sigma^{-1}(E)} = \int_{\sigma^{-1}(E)} \dot{\sigma}\,, \end{aligned}$$

by the Lebesgue monotone convergence theorem. So $E \in \mathcal{E}$.)

It follows from the monotone convergence theorem (cf. Halmos, vaca, Theorem vaca) that $\mathcal{E}$ is the set of all Borel subsets of $[c,d]$. So we have proved that $\text{meas}(E) = \int_{\sigma^{-1}(E)} \dot{\sigma}(t)\, dt$ for all Borel subsets of $[c,d]$.

Now[2] suppose that E is a Lebesgue measurable subset of $[c,d]$. Then there exist Borel measurable subsets E_1, E_2 of $[c,d]$ such that $E_1 \subseteq E \subseteq E_2$ and $\text{meas}(E_2 \backslash E_1) = 0$. Let $F_1 = \sigma^{-1}(E_1)$, $F = \sigma^{-1}(E)$, $F_2 = \sigma^{-1}(E_2)$.

[2]If E is Borel measurable then $\sigma^{-1}(E)$ is also Borel measurable, because σ is continuous. But if E is Lebesgue measurable then it does not follow that $\sigma^{-1}(E)$ is Lebesgue measurable as well.

Then F_1 and F_2 are Borel measurable, $F_1 \subseteq F \subseteq F_2$, and $\int_{F_2 \backslash F_1} \dot{\sigma} = 0$. Since $\dot{\sigma}$ is nonnegative and measurable, there exists a Lebesgue null subset N of $F_2 \backslash F_1$ such that $\dot{\sigma} = 0$ on $(F_2 \backslash F_1) \backslash N$. It follows that the functions $\dot{\sigma} \cdot \chi_F$ and $\dot{\sigma} \cdot \chi_{F_2}$ coincide on the set $[a,b] \backslash N$. Since N is a null set, and $\dot{\sigma} \cdot \chi_{F_2}$ is integrable, we conclude that $\dot{\sigma} \cdot \chi_F$ is integrable as well, and

$$\int_a^b \dot{\sigma} \cdot \chi_F = \int_a^b \dot{\sigma} \cdot \chi_{F_2} = \int_{F_2} \dot{\sigma} = \text{meas}(E_2) = \text{meas}(E)\,.$$

Since $\chi_F = \chi_E \circ \sigma$, we have shown that $\dot{\sigma} \cdot (\chi_E \circ \sigma)$ is integrable and

$$\int_a^b \dot{\sigma} \cdot (\chi_E \circ \sigma) = \int_c^d \chi_E$$

whenever E is a Lebesgue measurable subset of $[c,d]$. In other words, we have proved that $\dot{\sigma} \cdot (f \circ \sigma)$ is integrable and (A.31) holds whenever f is the indicator function of a Lebesgue measurable subset of $[c,d]$.

It then follows easily that $\dot{\sigma} \cdot (f \circ \sigma)$ is integrable and (A.31) holds whenever f is a simple function (i.e., a finite linear combination of indicator functions of a Lebesgue measurable subsets of $[c,d]$). Then $\dot{\sigma} \cdot (f \circ \sigma)$ is integrable and (A.31) holds whenever f is a bounded Lebesgue measurable real-valued function on $[c,d]$, since every such function is a uniform limit of simple functions. If f is nonnegative and measurable, then $f = \lim f_j$, where $f_j(s) = f(s)$ when $f(s) \leq j$ and $f_j(s) = j$ when $f(s) > j$. Therefore $f \circ \sigma = \lim f_j \circ \sigma$, so $\dot{\sigma} \cdot f \circ \sigma$ is measurable and (A.31) holds. In particular, if f is integrable (i.e., if $\int_a^b f < \infty$) then $\dot{\sigma} \cdot f \circ \sigma$ is integrable as well. Now, if $f : [c,d] \mapsto \mathbb{R}$ is integrable, then we can write $f = f_+ - f_-$, where both f_+ and f_- are integrable and nonnegative, and conclude that $\dot{\sigma} \cdot (f \circ \sigma)$ is integrable and (A.31) holds. Finally, if f is vector-valued, then the desired conclusion follows by applying what we have shown to each of the components of f. $\diamondsuit$

Now suppose that

(A1) Q is an open subset of $\mathbb{R}^n$,

(A2) $Q \times \mathbb{R}^n \ni (q,u) \mapsto L(q,u) \in \mathbb{R}$ is a continuous function, thought of as a time-independent Lagrangian,

(A3) L is positively homogeneous of degree 1 with respect to the velocity, that is, $L(q,ru) = rL(q,u)$ whenever $q \in Q$, $u \in \mathbb{R}^n$ and $r \geq 0$,

(A4) $\xi : [c,d] \mapsto Q$ is an absolutely continuous curve in Q,

(A5) $\sigma : [a,b] \mapsto \mathbb{R}$ is absolutely continuous, monotonically nondecreasing, and such that $\sigma(a) = c$ and $\sigma(b) = d$,

(A6) $\gamma = \xi \circ \sigma$.

Proposition 14.2.9 *Assume that conditions (A1)$\cdots$(A6) above hold. Then γ is absolutely continuous,*

$$\frac{d\gamma}{dt}(t) = \dot{\sigma}(t) \cdot \frac{d\xi}{ds}(\sigma(t)) \quad \text{for almost all} \quad t \in [a,b]\,, \tag{A.32}$$

and

$$\int_c^d L\Big(\xi(s), \frac{d\xi}{ds}(s)\Big)\, ds = \int_a^b L\Big(\gamma(t), \frac{d\gamma}{dt}(t)\Big)\, dt\,. \tag{A.33}$$

PROOF. Let $f(s) = \frac{d\xi}{ds}(s)$, $g(t) = \dot{\sigma}(t) \cdot f(\sigma(t))$. Then Proposition 14.2.8 implies that g is integrable. Furthermore, we can pick an arbitrary $t \in [a,b]$ and apply Proposition 14.2.8 to the interval $[a,t]$, with σ replaced by its restriction to $[a,t]$, and conclude that

$$\gamma(t) - \gamma(a) = \xi(\sigma(t)) - \xi(c) = \int_c^{\sigma(t)} f(s)\, ds = \int_a^t g(\tau)\, d\tau\,.$$

Since g is integrable and t is arbitrary, γ is absolutely continuous and (A.32) holds.

Now let $h(s) = L\Big(\xi(s), \frac{d\xi}{ds}(s)\Big)$. Then Proposition 14.2.8 implies that

$$\begin{aligned}
\int_c^d h(s)\, ds &= \int_a^b \dot{\sigma}(t) \cdot h(\sigma(t)) \\
&= \int_a^b \dot{\sigma}(t) \cdot L\Big(\xi(\sigma(t)), \frac{d\xi}{ds}(\sigma(t))\Big)\, dt \\
&= \int_a^b L\Big(\gamma(t), \dot{\sigma}(t) \cdot \frac{d\xi}{ds}(\sigma(t))\Big)\, dt \\
&= \int_a^b L\Big(\gamma(t), \frac{d\gamma}{dt}(t)\Big)\, dt\,,
\end{aligned}$$

where in the last two steps we have used the homogeneity of L and (A.32).

Therefore (A.33) holds, and our proof is complete. $\diamondsuit$

14.2.8 The complex exponential

We will need to use the *exponential function* $\mathbb{C} \ni z \mapsto e^z \in \mathbb{C}$. By definition,

$$e^{x+iy} = e^x \cdot e^{iy} = e^x(\cos y + i \sin y)\,. \tag{A.34}$$

The following well-known properties of this function (whose proofs can be found in any elementary texbook of complex function theory) will be used:

(E1) $e^{z_1+z_2} = e^{z_1} \cdot e^{z_2}$ whenever $z_1, z_2 \in \mathbb{C}$,

(E2) $e^0 = 1$,

(E3) $e^z \neq 0$ whenever $z \in \mathbb{C}$,

(E4) if $z \in \mathbb{C}$, then $e^z = 1$ if and only if $z = 2\pi ik$ for some integer k,

(E5) the series $z - 1 - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} - \frac{(z-1)^4}{4} + \ldots$ is convergent for $z \in \mathbb{C}$ such that $|z-1| < 1$, and defines a continuous function

$$\{u \in \mathbb{C} : |u-1| < 1\} \ni z \mapsto \mathrm{Ln}(z) \in \mathbb{C}$$

such that $e^{\mathrm{Ln}(z)} = z$ whenever $|z-1| < 1$.

Since $e^{\mathrm{Ln}(z)} = z$ whenever $|z-1| < 1$, the function Ln is called "the standard branch of the natural logarithm."

If follows from the above properties that

(E6) if $z_0 \in \mathbb{C}$ and $z_0 \neq 0$, then there exists continuous function

$$\{u \in \mathbb{C} : |u - z_0| < |z_0|\} \ni z \mapsto \Phi(z) \in \mathbb{C}$$

such that $e^{\Phi(z)} = z$ whenever $|z - z_0| < |z_0|$.

(E7) if $[a,b] \ni t \mapsto \gamma(t) \in \mathbb{C}$ is an arc in the complex plane (i.e., a continuous map with values in $\mathbb{C}$) such that $\gamma(t) \neq 0$ for all $t \in [a,b]$, then there exists a continuous function $[a,b] \ni t \mapsto f(t) \in \mathbb{C}$ such that $e^{f(t)} = \gamma(t)$ for all $t \in [a,b]$.

Indeed, to prove (E6), we write $z_0 = x_0 + iy_0$, $r_0 = \sqrt{x_0^2 + y_0^2}$, $x_0 = r_0 \cdot \cos\theta_0$, $y_0 = r_0 \cdot \sin\theta_0$, and observe that $r_0 > 0$, so r_0 has a natural logarithm λ_0, and then $e^{\lambda_0} = r_0$. Therefore, if we write $s_0 = \lambda_0 + i\theta_0$, (A.34) tells us that

$e^{s_0} = z_0$. Then, if $|z - z_0| < |z_0|$, we have $\left|\frac{z}{z_0} - 1\right| < 1$, so $\mathrm{Ln}\left(\frac{z}{z_0}\right)$ is well defined. We then define $\Phi(z) = s_0 + \mathrm{Ln}\left(\frac{z}{z_0}\right)$, so that

$$e^{\Phi(z)} = e^{s_0} \cdot e^{\mathrm{Ln}\left(\frac{z}{z_0}\right)} = z_0 \cdot \frac{z}{z_0} = z\,.$$

To prove (E7), we let $\varepsilon = \min\{\,|\gamma(t)| : t \in [a,b]\,\}$, so that $\varepsilon > 0$, and use the fact that the map γ is uniformly continuous to obtain a positive δ such that $|\gamma(t_1) - \gamma(t_2)| < \varepsilon$ whenever $t_1, t_2 \in [a,b]$ and $|t_1 - t_2| < \delta$. We then find $t_0, t_1, \ldots, t_N$ such that $a = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N = b$ and $t_j - t_{j-1} < \delta$ if $j = 1, 2, \ldots, N$. We write $z_j = \gamma(t_j)$ and $D_j = \{z \in \mathbb{C} : |z - z_j| < |z_j|\}$ for $j = 1, \ldots, N$, and use (E6) to pick, for each $j \in \{1, \ldots, N\}$, a continuous function $\Phi_j : D_j \mapsto \mathbb{C}$ such that $e^{\Phi_j(z)} = z$ whenever $z \in D_j$. We then modify the Φ_j recursively, and define new functions $\Psi_j : D_j \mapsto \mathbb{C}$, as follows: we take $\Psi_1 = \Phi_1$; if $1 < j \leq N$, and $\Psi_1, \ldots, \Psi_{j-1}$ have been defined and are such that $e^{\Psi_\ell(z)} = z$ whenever $z \in D_\ell$, $0 \leq \ell < j$, then we observe that $z_j \in D_{j-1}$, because $|z_j - z_{j-1}| = |\gamma(t_j) - \gamma(t_{j-1})| < \varepsilon \leq |z_{j-1}|$, since $t_j - t_{j-1} < \delta$; therefore $e^{\Psi_{j-1}(z_j)} = z_j = e^{\Phi_j(z_j)}$, from which it follows that $\Phi_j(z_j) = \Psi_{j-1}(z_j) + 2\pi i k_j$ for some integer k_j; we then define $\Psi_j(z) = \Phi_j(z_j) - 2\pi i k_j$ for $z \in D_j$, so that $e^{\Psi_j(z)} = z$ whenever $z \in D_j$. Using the Ψ_j, we define $f(t)$, for $t \in [a,b]$, by letting $f(t) = \Psi_j(\gamma(t))$ whenever $t_{j-1} \leq t \leq t_j$, $j \in \{1, \ldots, N\}$. Then f is well defined. (Indeed, the only values of t for which the condition $t_{j-1} \leq t \leq t_j$ does not determine j uniquely are $t_1, \ldots, t_{N-1}$. For each of these t's, if $t = t_j$, $1 \leq j < N$, then $t \in [t_{j-1}, t_j]$ and $t \in [t_j, t_{j+1}]$, so the definition would yield both $f(t) = \Psi_j(\gamma(t)) = \Psi_j(z_j)$ and $f(t) = \Psi_{j+1}(\gamma(t)) = \Psi_{j+1}(z_j)$, but

$$\Psi_{j+1}(z_j) = \Phi_{j+1}(z_j) - 2\pi i k_{j+1} = \Big(\,\Psi_j(z_j) + 2\pi i k_j\,\Big) - 2\pi i k_{j+1} = \Psi_j(z_j)\,,$$

so the two values of $f(t)$ coincide.)

14.2.9 Loops

We think of $\mathbb{R}^2$ as the complex plane $\mathbb{C}$. We use $\mathbb{S}^1$ to denote the unit circle $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} = \{z \in \mathbb{C} : |z| = 1\}$.

For each positive real number L, we let $\kappa_{_L}$ be the map

$$\mathbb{R} \ni t \mapsto \kappa_{_L}(t) \stackrel{\mathrm{def}}{=} e^{\frac{2\pi i}{L}t} \in \mathbb{S}^1\,.$$

A *loop* (or *closed curve*) in $\mathbb{R}^2$ is a continuous map $\xi : \mathbb{S}^1 \mapsto \mathbb{R}^2$. Then $C^0(\mathbb{S}^1, \mathbb{R}^2)$—or $C^0(\mathbb{S}^1, \mathbb{C})$— is the set of all loops in $\mathbb{R}^2$. The *carrier*, or *locus*, of the loop ξ is the set $\{\xi(s) : s \in \mathbb{C}\}$. Clearly, $|\xi|$ is compact.

If $L > 1$, then $C^0(\mathbb{S}^1, \mathbb{R}^2)$ can be identified in an obvious way with $C^0_{per,L}(\mathbb{R}, \mathbb{R}^2)$, the set of all continuous maps $\zeta : \mathbb{R} \mapsto \mathbb{R}^2$ that are L-periodic, i.e., such that $\zeta(t + L) = \zeta(t)$ for all $t \in \mathbb{R}$. Naturally, the identification is made by assigning to each $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ the map $\zeta = \xi \circ \kappa_{_L} \in C^0_{per,L}(\mathbb{R}, \mathbb{R}^2)$.

A loop $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ is said to be *simple* if it is a one-to-one map, i.e., if $\xi(z_1) \neq \xi(z_2)$ whenever $z_1, z_2 \in \mathbb{S}^1$ and $z_1 \neq z_2$.

A loop $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ is said to be *rectifiable* if it is absolutely continuous, in the sense that for some positive L the function $\zeta = \xi \circ \kappa_{_L}$ is locally absolutely continuous (in which case it follows easily that $\xi \circ \kappa_L$ is locally absolutely continuous for every positive L).

14.2.10 Reparametrization of loops

A *reparametrization* of a loop $\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ is a loop of the form $\xi \circ \psi$, where $\psi : \mathbb{S}^1 \mapsto \mathbb{S}^1$ is an orientation-preserving homeomorphism. ("Orientation-preserving" means, for example, that if $\bar{t}$ is such that $\psi(e^{2\pi i \bar{t}}) = 1$, then the imaginary part of $\psi(e^{2\pi i(\bar{t}+\varepsilon)})$ is positive for all sufficiently small positive ε.)

A rectifiable loop $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$ is said to be *parametrized by constant times arc-length* (abbr. PCAL) if there exists a positive constant C such that $\|\dot{\zeta}(t)\| = C$ for all $t \in \mathbb{R}$, where $\zeta(t) \stackrel{\text{def}}{=} \xi(e^{2\pi i t})$.

To any rectifiable loop $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$ we can associate another loop $\gamma_\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ obtained from ξ by *reparametrization by constant times arc-length*. Precisely, γ_ξ is defined as follows. Let $\zeta(t) = \xi(e^{2\pi i t})$, for $t \in \mathbb{R}$. Write $\zeta(t) = x(t) + iy(t)$ for $t \in \mathbb{R}$, and define a function $s_\xi : \mathbb{R} \mapsto \mathbb{R}$ (the "arc-length along ζ") by letting

$$s_\xi(t) = \int_0^t \sqrt{\dot{x}(u)^2 + \dot{y}(u)^2}\, du\,.$$

Then s_ξ is a monotonically nondecreasing continuous function from $\mathbb{R}$ to $\mathbb{R}$. If $L = s_\xi(1)$ (so that L is the length of ξ) then s_ξ maps $[0, 1]$ onto $[0, L]$. Furthermore, if $0 \leq t_1 < t_2 \leq 1$ and $s_\xi(t_1) = s_\xi(t_2)$, then

$$\int_{t_1}^{t_2} \sqrt{\dot{x}(u)^2 + \dot{y}(u)^2}\, du = s_\xi(t_2) - s_\xi(t_1) = 0\,,$$

so $\dot{x}(u) \equiv \dot{y}(u) \equiv 0$ for almost all $t_1 \leq u \leq t_2$, and then $\xi(t_2) = \xi(t_1)$. Therefore there exists a unique map $\theta_\xi : [0, L] \mapsto \mathbb{C}$ such that $\theta_\xi(s) = \zeta(t)$

whenever $s = s_\xi(t)$, $t \in [0,1]$. In particular, this map satisfies the identities $\theta_\xi(L) = \zeta(1) = \zeta(0) = \theta_\xi(0)$, since $L = s_\xi(1)$. So we can define a continuous map $\gamma_\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ by letting $\gamma_\xi(e^{2\pi i t}) = \theta_\xi(Lt)$ for $0 \le t \le 1$.

If $0 \le s_1 < s_2 \le L$, then we can pick $t_1, t_2 \in [0,1]$ such that $s_1 = s_\xi(t_1)$ and $s_2 = s_\xi(t_2)$. Then $t_1 < t_2$, and

$$\begin{aligned} \|\theta_\xi(s_2) - \theta_\xi(s_1)\| &= \|\zeta(t_2) - \zeta(t_1)\| \\ &= \|\int_{t_1}^{t_2} \dot\zeta(t)\, dt\| \\ &\le \int_{t_1}^{t_2} \|\dot\zeta(t)\|\, dt \\ &= s_\xi(t_2) - s_\xi(t_1) = s_2 - s_1\,. \end{aligned}$$

Therefore $\|\theta_\xi(s_2) - \theta_\xi(s_1)\| \le |s_2 - s_1|$ whenever $s_1, s_2 \in [0, L]$. So θ_ξ *is Lipschitz with constant* 1. It follows that θ_ξ is absolutely continuous and $\left\| \frac{d}{ds}(\theta_\xi(s)) \right\| \le 1$ for almost all $s \in [0, L]$.

Formally, the identity $\theta_\xi \circ s_\xi = \zeta$ implies that

$$\frac{d}{dt}(\zeta(t)) = \frac{d}{ds}(\theta_\xi(s))\Big|_{s=s_\xi(t)} \cdot \frac{d}{dt}(s_\xi(t))\,,$$

and then

$$\|\dot\zeta(t)\| = \left\| \frac{d}{ds}(\theta_\xi(s))\Big|_{s=s_\xi(t)} \right\| \cdot \|\dot\zeta(t)\|\,,$$

so that

$$\left\| \frac{d}{ds}(\theta_\xi(s)) \right\| = 1 \text{ for almost all } s \in [0, L]\,. \tag{A.35}$$

The above argument is of course nonrigorous. For a rigorous proof of (A.35), see Proposition 14.2.7 below.

It then follows that

$$\left\| \frac{d}{dt}(\gamma_\xi(e^{2\pi i t}) \right\| = L \text{ for almost all } t \in [0,1]\,. \tag{A.36}$$

Therefore

Proposition 14.2.10 *If $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$ is a rectifiable loop of length L, and the loop γ_ξ is defined as above, then γ_ξ is parametrized by constant times arc-length (the constant being L), and $\xi = \gamma \circ \psi$, where ψ is the map*

$$\mathbb{S}^1 \ni e^{2\pi i t} \mapsto \psi(e^{2\pi i t}) \stackrel{\text{def}}{=} e^{2\pi i \lambda s_\xi(t)} \in \mathbb{S}^1\,,$$

and $\lambda = L^{-1}$.

14.2.11 Connected components

Vaca.

14.2.12 The winding number

If $\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ is a loop, and $p \in \mathbb{C}\backslash|\xi|$, then the *winding number* of ξ about p in the integer $\nu(\xi, p)$ defined as follows: pick any positive L, and use (E7) to choose a continuous function $f : \mathbb{R} \mapsto \mathbb{C}$ such that $e^{f(t)} = (\xi \circ \kappa_{_L})(t) - p$ for all $t \in \mathbb{R}$. Then (E4) implies that the number $\frac{f(t+L)-f(t)}{2\pi i}$ necessarily has to be an integer, because

$$e^{f(t+L)} = (\xi \circ \kappa_{_L})(t+L) - p = (\xi \circ \kappa_{_L})(t) - p = e^{f(t)} .$$

It is clear that this integer depends continuously on t. Furthermore, if $f_1, f_2 : \mathbb{R} \mapsto \mathbb{C}$ are any two continuous maps such that $e^{f_1(t)} = (\xi \circ \kappa_{_L})(t) - p$ and $e^{f_2(t)} = (\xi \circ \kappa_{_L})(t) - p$ for all $t \in \mathbb{R}$, then f_1 and f_2 must differ by a constant, because $e^{f_1(t)-f_2(t)} = 1$, so that $f_1(t) - f_2(t) = 2\pi i k(t)$, where $k(t)$ is an integer that depends continuously on t. So $\frac{f(t+L)-f(t)}{2\pi i}$ is in fact independent of t and f, and it is easy to see that it is also independent of L. We define $\nu(\xi, p) \stackrel{\text{def}}{=} \frac{f(t+L)-f(t)}{2\pi i}$ for any $t \in \mathbb{R}$, where L and f are as above.

It is clear that reparametrization preserves carriers and winding numbers. Precisely,

Proposition 14.2.11 *If η is a reparametrization of a loop $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$, then $|\eta| = |\xi|$ and $\nu(\eta, p) = \nu(\xi, p)$ for all $p \in \mathbb{C}\backslash|\xi|$.*

PROOF. The fact that $|\eta| = |\xi|$ is trivial. Let $f : \mathbb{R} \mapsto \mathbb{C}$ be continuous and such that $e^{f(t)} = \eta(e^{2\pi i t})$ for all t. Let $\psi : \mathbb{S}^1 \mapsto \mathbb{S}^1$ be an orientation-preserving homeomorphism such that $\eta = \xi \circ \psi$. Then there exist $a \in \mathbb{R}$ and a homeomorphism $\tilde{\psi} : [0, 1] \mapsto [a, a + 1]$ such that

$$e^{2\pi i \tilde{\psi}(t)} = \psi(e^{2\pi i t}) \tag{A.37}$$

for all $t \in [0, 1]$, $\tilde{\psi}(0) = a$ and $\tilde{\psi}(1) = a + 1$. Extend $\tilde{\psi}$ to the whole real line by defining $\tilde{\psi}(k + t) = k + \psi(t)$ for $t \in [0, 1]$ and $k \in \mathbb{Z}$. Let $g = f \circ \tilde{\psi}$. Then (A.37) holds for all $t \in \mathbb{R}$. Moreover,

$$e^{g(t)} = e^{f(\tilde{\psi}(t))} = \eta(e^{2\pi i \tilde{\psi}(t)}) = \eta\Big(\psi(e^{2\pi i t})\Big) = (\eta \circ \psi)(e^{2\pi i t}) = \xi(e^{2\pi i t}) .$$

Therefore $\nu(\xi, p) = \frac{g(1)-g(0)}{2\pi i} = \frac{f(\tilde{\psi}(1))-f(\tilde{\psi}(0))}{2\pi i} = \frac{f(a+1)-f(a)}{2\pi i} = \nu(\eta, p)$. $\diamondsuit$

If a loop ξ is rectifiable, then we can express the winding number as an integral.

Proposition 14.2.12 *Let ξ be a rectifiable loop in $\mathbb{C}$. Then*

$$\nu(\xi, p) = \frac{1}{2\pi i} \int_0^L \frac{\dot{\zeta}(s)}{\zeta(s) - p}\, ds \text{ if } p \in \mathbb{C}\backslash|\xi|\,,\ L > 0\,,\ \text{and } \zeta = \xi \circ \kappa_L. \tag{A.38}$$

PROOF. We can choose f by letting

$$f(t) = f_0 + \int_0^t \frac{\dot{\zeta}(s)}{\zeta(s) - p}\, ds\,, \tag{A.39}$$

where $f_0 \in \mathbb{C}$ is such that $e^{f_0} = \zeta(0) - p$. Indeed, if we define $g : \mathbb{R} \mapsto \mathbb{C}$ by letting $g(t) = (\zeta(t) - p).e^{-f(t)}$, we have

$$\begin{aligned} \dot{g}(t) &= -(\zeta(t) - p).\dot{f}(t).e^{-f(t)} + \dot{\zeta}(t).e^{-f(t)} \\ &= -(\zeta(t) - p).\frac{\dot{\zeta}(t)}{\gamma(t) - p}.e^{-f(t)} + \dot{\zeta}(t).e^{-f(t)} \\ &= -\dot{\zeta}(t).e^{-f(t)} + \dot{\zeta}(t).e^{-f(t)} \\ &= 0\,. \end{aligned}$$

Therefore $g(t) \equiv$ constant. Since $g(0) = 1$, we have $g(t) \equiv 1$ for all t, so $e^{f(t)} \equiv \gamma(t) - p$. Therefore $\nu(\xi, p) = \frac{1}{2\pi i}(f(L) - f(0))$, and then (A.39) implies that (A.38) holds. ♢

Proposition 14.2.13 *The winding number $\nu(\xi, p)$ depends continuously on the point p and the curve ξ as long as p and ξ vary in such a way that $p \in \mathbb{C}\backslash|\xi|$. More precisely, if $\{p_j\}_{j=1}^{\infty}$ is a sequence of points of $\mathbb{C}$ that converges to $p \in \mathbb{C}$, $\{\xi_j\}_{j=1}^{\infty}$ is a sequence of loops in $\mathbb{C}$ that converges uniformly to a loop ξ, and $p \notin \mathbb{C}\backslash|\xi|$, then there exists a positive integer j_* such that $p_j \notin \mathbb{C}\backslash|\xi_j|$ and $\nu(\xi_j, p_j) = \nu(\xi, p)$ whenever $j \geq j_*$*

PROOF. Let f be a continuous map from $\mathbb{R}$ to $\mathbb{C}$ having with the property that $e^{f(t)} = \xi(e^{2\pi i t}) - p \stackrel{\text{def}}{=} h(t)$ for all $t \in \mathbb{R}$. Let $h_j(t) = \xi_j(e^{2\pi i t}) - p_j$. Then $h_j \to h$ as $j \to \infty$, uniformly on compact sets, so $h_j \to h$ uniformly, since the h_j are 1-periodic. Since $h(t) \neq 0$ for all t, the functions $g_j \stackrel{\text{def}}{=} \frac{h_j}{h}$ converge uniformly to the constant 1 as $j \to \infty$. Let D be the open disc $\{z \in \mathbb{C} : \|z - 1\| < 1\}$. Let $\mathrm{Ln} : D \mapsto \mathbb{C}$ be the function defined in (E5). Fix

$j_* \in \mathbb{N}$ such that $g_j(t) \in D$ whenever $t \in \mathbb{R}$ and $j \geq j_*$. Then, if $j \geq j_*$, $h_j(t)$ must be $\neq 0$ for every t, so $p_j \notin |\xi_j|$. Define

$$f_j(t) = f(t) + \mathrm{Ln}\Big(g_j(t)\Big) \text{ for } t \in \mathbb{R}\,,\, j \geq j_*.$$

Then

$$e^{f_j(t)} = e^{f(t)}.e^{\mathrm{Ln}(g_j(t))} = h(t).g_j(t) = h_j(t) = \xi_j(e^{2\pi i t}) - p_j\,.$$

Therefore

$$\begin{aligned}
\nu(\xi_j, p_j) &= \frac{1}{2\pi i}(f_j(1) - f_j(0)) \\
&= \frac{1}{2\pi i}\Big(f(1) - f(0) + \mathrm{Ln}(g_j(1)) - \mathrm{Ln}(g_j(0))\Big) \\
&= \frac{1}{2\pi i}\Big(f(1) - f(0)\Big) \\
&= \nu(\xi, p)
\end{aligned}$$

whenever $j \geq j_*$, since $\mathrm{Ln}(g_j(1)) = \mathrm{Ln}(g_j(0))$, because the function g_j is 1-periodic.) ♢

14.2.13 The no-retraction theorem

Theorem 14.2.14 *Let $R > 0$, and write*

$$D(R) = \{z \in \mathbb{C} : |z| \leq 1\}\,, \qquad S(R) = \{z \in \mathbb{C} : |z| = 1\}\,.$$

Then there does not exist a continuous map $\rho : D(R) \mapsto S(R)$ such that $\rho(z) = z$ whenever $z \in S(R)$.

PROOF. Suppose ρ exists. For $0 \leq r \leq 1$, define a loop $\xi_r : \mathbb{S}^1 \mapsto \mathbb{C}$ by letting $\xi_r(e^{2\pi i t}) = \rho(re^{2\pi i t})$. Then each ξ_r satisfies $|\xi_r(e^{2\pi i t})| = 1$ for all t, so $0 \notin |\xi_r|$. Let $w(r) = \nu(\xi_r, 0)$. Therefore Proposition 14.2.13 implies that the integer-valued function $[0,1] \ni r \mapsto w(r) \in \mathbb{R}$ is continuous. Hence $w(r) \equiv$ constant, so $w(0) = w(1)$. Furthermore, ξ_0 is a constant map, from which it follows easily that $\nu(\xi_0, 0) = 0$. On the other hand, since $\xi_1(e^{2\pi i t}) = \rho(e^{2\pi i t}) = e^{2\pi i t}$, the function $f : \mathbb{R} \mapsto \mathbb{C}$ given by $f(t) = 2\pi i t$ satisfies $e^{f(t)} = \xi_1(e^{2\pi i t})$, so $\nu(\xi_1, 0) = \frac{f(1)-f(0)}{2\pi i} = 1$. So $w(1) \neq w(0)$, and we have reached a contradiction. ♢

14.2.14 The Jordan curve theorem

If K is a compact subset of $\mathbb{C}$, and $p \in \mathbb{C}\backslash K$, the *Borsuk map* $\beta_{p,K} : K \mapsto \mathbb{S}^1$ is defined by

$$\beta_{p,K}(z) = \frac{z-p}{|z-p|} \qquad \text{for } z \in K\,.$$

Proposition 14.2.15 *Let K be a compact subset of $\mathbb{C}$, let $p \in \mathbb{C}\backslash K$, and let U be the connected component of $\mathbb{C}\backslash K$ that contains p. Assume that U is bounded. Then the Borsuk map $\beta_{p,K}$ cannot be extended to a continuous map $\beta : K \cup U \mapsto \mathbb{S}^1$.*

PROOF. Assume β exists. Let R be positive and so large that $K \cup U$ is entirely contained in the disc $D(p,R) = \{z \in \mathbb{C} : |z-p| \le R\}$. Define a map $\tilde{\beta} : D(p,R) \mapsto \partial D(p,R)$ by letting

$$\tilde{\beta}(z) = \begin{cases} p + R \cdot \beta(z) & \text{if} \quad z \in K \cup U\,, \\ p + R \cdot \frac{z-p}{|z-p|} & \text{if} \quad z \in D(p,R)\backslash U\,. \end{cases}$$

Then $\tilde{\beta}$ is well defined, because $(D(p,R)\backslash U) \cap (K \cup U) \subseteq K$ and $\beta(z) = \frac{z-p}{|z-p|}$ if $z \in K \cup U$. Furthermore, $\tilde{\beta}$ is continuous, because both sets $D(p,R)\backslash U$ and $K \cup U$ are closed. So $\tilde{\beta}$ is a continuous map from $D(p,R)$ to $\partial D(p,R)$. Moreover, it is clear that $\tilde{\beta}(z) = z$ if $z \in \partial D(p,R)$. So $\tilde{\beta}$ is a continuous retraction from $D(p,R)$ to $\partial D(p,R)$. Since such a retraction does not exist, we have reached a contradiction. ♢

Proposition 14.2.16 *Let K be a compact subset of $\mathbb{C}$, and let Q be a subset of $\mathbb{C}\backslash K$ that contains at least one point of each bounded connected component of $\mathbb{C}\backslash K$. Then every continuous map $f : K \mapsto \mathbb{S}^1$ can be extended to a continuous map $F : \mathbb{C}\backslash Q \mapsto \mathbb{S}^1$.*

PROOF. For each positive number r such that $r < \sqrt{2}$, we define

$$S_r = \{s \in \mathbb{S}^1 : |s-1| \le r\}\,,$$

and let $\rho(r) = \frac{r}{2}\sqrt{4-r^2}$. Then the set S_r can be identified with the real interval $I_r \stackrel{\text{def}}{=} [-\rho(r), \rho(r)]$ by means of the map

$$I_r \ni u \mapsto \mu_r(u) \stackrel{\text{def}}{=} \sqrt{1-u^2} + ui \in \mathbb{S}^1\,. \tag{A.40}$$

Now fix α such that $0 < \alpha < \frac{\sqrt{2}}{3}$. Choose a positive number δ such that $|f(z) - f(z')| < \alpha$ whenever $z, z' \in K$ and $|z - z'| \leq 2\sqrt{2}\delta$. Let G^δ be the δ-grid of $\mathbb{C}$, i.e., the set of all rectangles

$$R^\delta_{n,m} = \{x + iy : \delta n \leq x \leq \delta(n+1), \delta m \leq x \leq \delta(m+1)\}\,,$$

for all $n, m \in \mathbb{Z}$. For each $R \in G^\delta$ we let $V(R)$, $E(R)$ be, respectively, the set of vertices of R and the set of edges of R.

Let G^δ_K be the set of those rectangles $R \in G^\delta$ such that R intersects K. Define

$$|G^\delta_K| \stackrel{\text{def}}{=} \bigcup\{R : R \in G^\delta\}\,, \qquad \tilde{K} \stackrel{\text{def}}{=} K \cup |G^\delta_K|\,.$$

We will first extend f to a continuous map $\tilde{f} : \tilde{K} \mapsto \mathbb{S}^1$. For this purpose, we pick a point $z_R \in K \cap R$ for each $R \in G^\delta_K$. Let

$$V_K = \bigcup\{V(R) : R \in G^\delta_K\}\,.$$

Let $K_1 = K \cup V_K$. We then extend f to a continuous map $f_1 : K_1 \mapsto \mathbb{S}^1$ by defining $f_1(z)$ as follows, if $z \in K_1$:

1. If $z \in K$, we define $f_1(z) = f(z)$.
2. If $z \in V_K \backslash K$, we take $f_1(z)$ to be $f(z_R)$, where R is such that $z \in R$ and $R \in G^\delta_K$. (For a given z, the rectangle R exists, but need not be unique, since z could belong to two, three or four rectangles in G^δ_K. If R is not unique, we choose one for each z, in an arbitrary fashion.)

With this definition, f_1 satisfies the inequality

$$|f_1(z) - f_1(z_R)| < \alpha \quad \text{whenever} \quad z \in R \cap K_1\,,\ R \in G^\delta_K\,. \tag{A.41}$$

Indeed, if $z \in R \cap K$, then $|z - z_R| \leq \sqrt{\delta}$, so $|f_1(z) - f_1(z_R)| < \alpha$. If $z \in V_K$ but $z \notin K$, then $f_1(z) = f(z_{R'})$ for some rectangle $R' \in G^\delta_K$ such that $R' \cap R \neq \emptyset$. Then $|z_{R'} - z_R| \leq 2\sqrt{2}\delta$, so $|f_1(z_{R'}) - f_1(z_R)| < \alpha$, and then $|f_1(z) - f_1(z_R)| < \alpha$.

Next, we let

$$E_K = \bigcup\{E(R) : R \in G^\delta_K\}\,, \qquad |E_K| = \bigcup\{e : e \in E_K\}\,.$$

Let $K_2 = K \cup |E_K| = K_1 \cup |E_K|$. We extend f_1 to a continuous map $f_2 : K_2 \mapsto \mathbb{S}^1$ by defining f_2 as follows:

1. We let $f_2(z) = f_1(z)$ if $z \in K_1$.

2. For each edge $e \in E_K$, we pick a point $z_e \in e \cap K_1$. (This is possible because $e \cap K_1 \neq \emptyset$, since the vertices of e belong to K_1.) Then the inequality $|f_1(z) - f_1(z_e)| < 2\alpha$ holds for all $z \in e \cap K_1$, because there exists $R \in G_K^\delta$ such that $e \subseteq R$, and then $|f_1(z) - f_1(z_R)| < \alpha$ for all $z \in e \cap K_1$. Then
$$e \cap K_1 \ni z \mapsto \frac{f_1(z)}{f_1(z_e)} \in \mathbb{S}^1$$
is a continuous map from $e \cap K_1$ to the set $S_{2\alpha}$ (thanks to the identity $\left|\frac{f_1(z)}{f_1(z_e)} - 1\right| = |f_1(z) - f_1(z_e)|$, which is true because $|f_1(z_e)| = 1$). Since $2\alpha < \sqrt{2}$, we know that $S_{2\alpha}$ can be identified with the real interval $I_{2\alpha}$ by means of the map $\mu_{2\alpha}$ defined in (A.40). Then the Tietze extension theorem implies that there exists a map $\varphi_e : e \mapsto S_{2\alpha}$ such that $\varphi_e(z) = \frac{f_1(z)}{f_1(z_e)}$ if $z \in e \cap K_1$. We can then define f_2 on e by letting $f_2(z) = \varphi_e(z) \cdot f_1(z_e)$ for $z \in e$.

We now claim that f_2 satisfies the inequality

$$|f_2(z) - f_2(z_R)| < 3\alpha \quad \text{whenever} \quad z \in R \cap K_2\,,\ R \in G_K^\delta\,. \tag{A.42}$$

Indeed, if $z \in R \cap K_1$, then $f_2(z) = f_1(z)$ and $f_2(z_R) = f_1(z_R)$, so (A.41) implies that $|f_2(z) - f_2(z_R)| < \alpha$. If $z \in R$ but $z \notin K_1$, then z belongs to and edge $e \in E(R)$. Then (A.41) implies that $|f_2(z_e) - f_2(z_R)| = |f_1(z_e) - f_1(z_R)| < \alpha$. On the other hand,

$$|f_2(z) - f_2(z_e)| = |f_2(z_e)| \cdot \left|\frac{f_2(z)}{f_2(z_e)} - 1\right| = |\varphi_e(z) - 1| < 2\alpha\,.$$

Therefore $|f_2(z) - f_2(z_R)| < 3\alpha$.

We now define the extension $\tilde{f} : \tilde{K} \mapsto \mathbb{S}^1$. For each rectangle $R \in G_K^\delta$, the map f_2 is defined on the compact set $K_2 \cap R$, which contains the boundary of R. Furthermore, (A.42) implies that

$$K_2 \cap R \ni z \mapsto \frac{f_2(z)}{f_2(z_R)} \in \mathbb{S}^1$$

is a continuous map to the set $S_{3\alpha}$. Since $3\alpha < \sqrt{2}$, the set $S_{3\alpha}$ can be identified with the real interval $I_{3\alpha}$ by means of the map $\mu_{3\alpha}$ of (A.40). Then the Tietze extension theorem implies that there exists a map $\psi_R : R^\delta \mapsto S_{3\alpha}$

such that $\psi_R(z) = \frac{f_2(z)}{f_2(z_R)}$ if $z \in K_2 \cap R$. We then define $\tilde{f}$ on R by letting $\tilde{f}(z) = \psi_R(z) \cdot f_2(z_R)$ for $z \in R$.

Let W be the set of all vertices of G^δ that do not belong to $\tilde{K}$. We let $\tilde{K}_1 = \tilde{K} \cup W$, and extend $\tilde{f}$ to a map $\tilde{f}_1 : \tilde{K}_1 \mapsto \mathbb{S}^1$ by choosing $\tilde{f}_1(z) \in \mathbb{S}^1$ in an arbitrary fashion if $z \in W$. We then let H be the set of all edges of G^δ that are not contained in $\tilde{K}$, and define $|H| = \bigcup\{h : h \in H\}$. We define $\tilde{K}_2 = \tilde{K}_1 \cup |H|$, and extend $\tilde{f}_1$ to a continuous map $\tilde{f}_2 : \tilde{K}_2 \mapsto \mathbb{S}^1$ as follows: for each $h \in H$, let v, w be the vertices of h, pick real numbers a, b such that $e^{2\pi i a} = v$ and $e^{2\pi i b} = w$, and define $\tilde{f}_2(tv + (1-t)w) = e^{2\pi i(ta+(1-t)b)}$ if $0 \leq t \leq 1$.

Next, we let $\mathcal{R} = G^\delta \backslash G^\delta_K$, so $\mathcal{R}$ is the set of all rectangles $R \in G^\delta$ that do not intersect K. For each $R \in \mathcal{R}$, use c_R to denote the center of R. The map $\tilde{f}_2$ is defined on the boundary ∂R. Pick, for each R, a continuous map $p_R : R \backslash \{c_R\} \mapsto \partial R$ such that $p_R(z) = z$ whenever $z \in \partial R$. We then let $\tilde{K}_3 = \tilde{K}_2 \cup \bigcup\{R \backslash \{c_R\} : R \in \mathcal{R}\}$ and extend $\tilde{f}_2$ to $\tilde{K}_3$ by letting

$$\tilde{f}_3(z) = \begin{cases} \tilde{f}_2(z) & \text{if} \quad z \in \tilde{K}_2\,, \\ \tilde{f}_2(p_R(z)) & \text{if} \quad z \in R \backslash \{c_R\} \text{ and } R \in \mathcal{R}\,. \end{cases}$$

We have now extended f to a continuous map $\tilde{f}_3 : \tilde{K}_3 \mapsto \mathbb{S}^1$, where $\tilde{K}_3$ is the whole complex plane $\mathbb{C}$, minus the discrete set $C = \{c_R : R \in \mathcal{R}\}$. Then $C \subseteq \mathbb{C} \backslash K$. If U is any bounded connected component of $\mathbb{C} \backslash K$, then $C \cap U$ is finite. Pick a point $q_U \in Q \cap U$, and define $C^0_U = (C \cap U) \cup \{q_U\}$.

The map $\tilde{f}_3$ is then continuous on $U \backslash C^0_U$. We show how to replace the restriction $\tilde{f}_{3,U}$ of $\tilde{f}_3$ to $U \backslash C^0_U$ by a map $\tilde{f}_{4,U}$ into $\mathbb{S}^1$ which is continuous on $U \backslash \{q_U\}$ and coincides with $\tilde{f}_{3,U}$ outside a compact subset of U that contains C^0_U. If $C^0_U = \{q_U\}$, then of course we can take $\tilde{f}_{4,U} = \tilde{f}_{3,U}$. Otherwise, we pick a point P of $C^0_U \backslash \{q_U\}$, and construct a polygonal curve $\xi : [a, b] \mapsto U$ such that $\xi(a) = P$, $\xi(b) = q_U$, the map ξ is one-to-one, and $\xi(t) \notin C^0_U$ whenever $a < t < b$. Let $(P_0, \ldots, P_k)$ be the ordered sequence of vertices of ξ, so that $P_0 = P$, $P_k = q_U$, and there exist $t_0, \ldots, t_k$ with the property that $a = t_0 < t_1 < \ldots < t_{k-1} < t_k = b$, $\xi(t_j) = P_j$ for $j = 0, \ldots, k$, and ξ is linear on each interval $[t_{j-1}, t_j]$.

Let $\sigma_j = \{\xi(t) : t_{j-1} \leq t \leq t_j\}$, so σ_j is the segment joining P_{j-1} to P_j. Then σ_0 does not contain any points of C^0_U other than P and q_U. Therefore the map $\tilde{f}_{3,U}$ is continuous on a neighborhood of σ_0, except possibly at the endpoints of σ_0. Let τ_0 be the compact neighborhood $\{z : \text{dist}(z, \sigma_0) \leq \theta\}$ of σ_0, where θ is positive and small enough, so that $\tau_0 \subseteq U$ and τ_0 does not contain any points of C^0_U other than P and q_U. Then $\tilde{f}_{3,U}$ is continuous

on τ_0, except possibly at P_0 and P_1. The set $\tau_0 \backslash \{P_1\}$ obviously admits a continuous retraction $\kappa_0 : \tau_0 \backslash \{P_1\} \mapsto \partial\tau_0$. Replace $\tilde{f}_{3,U}$ by $\tilde{f}_{3,U} \circ \kappa_0$ on $\tau_0 \backslash \{P_1\}$. Then the modified map $\tilde{f}^0_{3,U}$ is continuous wherever $\tilde{f}_{3,U}$ was, except possibly at P_1, and is also continuous at P_0. So the discontinuity at P_0 has been eliminated, and a possible discontinuity at P_1 has been introduced. Repeating this procedure, we replace the discontinuity at P_1 by a possible discontinuity at P_2, and continue in this way until, of all the points of ξ, only the terminal point q_U remains as a possible discontinuity of the new map. This means that the set C^0_U of points of possible discontinuity of $\tilde{f}_{3,U}$ has been reduced by eliminating one point other than q_U. We then repeat this procedure, until all points of discontinuity other than q_U have been eliminated.

By carrying out this elimination of discontinuities separately on each bounded connected component of $\mathbb{C}\backslash K$, we obtain a continuous extension $\hat{f} : \mathbb{C}\backslash(Q \cup D) \mapsto \mathbb{S}^1$ of f to the whole complex plane, minus $Q \cup D$, where D is a discrete subset which is entirely contained in the unbounded component Ω of $\mathbb{C}\backslash K$. The set D can then be completely eliminated, by first picking a positive L such that $K \subseteq \{z : |z| \leq L\}$, and then eliminating all the discontinuities in $\{z : |z| \leq L\}$ as was done for the bounded components. We then let κ be the retraction $z \mapsto L\frac{z}{|z|}$ from $\{z : |z| \geq L\}$ to $\{z : |z| = L\}$, and replace $\hat{f}$ by $\hat{f} \circ \kappa$ on $\{z : |z| \geq L\}$. The resulting map F now has no discontinuities whatsoever on Ω, so F is the desired extension. $\diamondsuit$

If K is a subset of $\mathbb{C}$, and f_0, f_1 are continuous maps from K to $\mathbb{S}^1$, a *homotopy* from f_0 to f_1 is a continuous map $h : K \times [0,1] \mapsto \mathbb{S}^1$ such that $h(z,0) = f_0(z)$ and $h(z,1) = f_1(z)$ whenever $z \in K$. We say that f_0 and f_1 are *homotopic* if there exists a homotopy from f_0 to f_1.

Proposition 14.2.17 *Let K be a compact subset of $\mathbb{C}$, and let C be a closed subset of $\mathbb{C}$ such that $K \subseteq C$. Let f_0, f_1 be continuous maps from K to $\mathbb{S}^1$, such that f_0 and f_1 are homotopic, and f_0 admits a continuous extension to a map $F_0 : C \mapsto \mathbb{S}^1$. Then f_1 admits a continuous extension to a map $F_1 : C \mapsto \mathbb{S}^1$.*

PROOF. Let $h : K \times [0,1] \mapsto \mathbb{S}^1$ be a continuous map that satisfies the conditions $h(z,0) = f_0(z)$ and $h(z,1) = f_1(z)$ whenever $z \in K$. The map h is uniformly continuous on $K \times [0,1] \mapsto \mathbb{S}^1$, so there exists a positive δ such that $|h(z,t) - h(z,s)| < 1$ whenever $z \in K$, $t, s \in [0,1]$, and $|t - s| \leq \delta$. Let N be a positive integer such that $N\delta > 1$. For $k = 0, \ldots, N$, let g_k be the

map

$$K \ni z \mapsto g_k(z) \stackrel{\text{def}}{=} h\Big(z, \frac{k}{N}\Big) \in \mathbb{S}^1 \, .$$

Then $g_0 = f_0$ and $g_N = f_1$. We prove by induction on k that g_k has a continuous extension to C. The fact that g_0 has a continuous extension follows because $g_0 = f_0$. Assuming that g_k has a continuous extension G_k, and $k < N$, use the inequality $|g_k(z) - g_{k+1}(z)| < 1$, which is valid for all $z \in K$, to conclude that $|\theta_k(z) - 1| < 1$ for all $z \in K$, where $\theta_k(z) \stackrel{\text{def}}{=} \frac{g_{k+1}(z)}{g_k(z)}$.

So θ_k is a continuous map from K to the set $S_1 = \{z \in \mathbb{S}^1 : |z - 1| \leq 1\}$. This set can be identified with the real interval $I_1 = [-\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2}]$, as explained in the proof of Proposition 14.2.16. Therefore the Tietze extension theorem yields an extension of θ_k to a continuous map $\Theta_k : C \mapsto \mathbb{S}^1$. Define $G_{k+1}(z) = \Theta_k(z) \cdot G_k(z)$ for $z \in C$. Then $G_{k+1}(z)$ is a continuous map from C to $\mathbb{S}^1$, and $G_{k+1}(z) = g_{k+1}(z)$ whenever $z \in K$.

In particular, g_N has a continuous extension. Since $g_N = f_1$, our proof is complete. ♢

Proposition 14.2.18 *Let K be a compact subset of $\mathbb{C}$. Then $\mathbb{C}\backslash K$ is connected if and only if every continuous map $f : K \mapsto \mathbb{S}^1$ is homotopic to a constant.*

PROOF. If $\mathbb{C}\backslash K$ is connected, and $f : K \mapsto \mathbb{S}^1$ is continuous, then Proposition 14.2.16 implies that f has a continuous extension $F : \mathbb{C} \mapsto \mathbb{S}^1$. If we define $h(z, t) = F((1 - t)z)$ for $z \in K$ and $t \in [0, 1]$, then h is clearly a homotopy from f to the constant map $K \ni z \mapsto F(0) \in \mathbb{S}^1$.

To prove the converse, suppose every continuous map $f : K \mapsto \mathbb{S}^1$ is homotopic to a constant. Since a constant map obviously admits a continuous extension to $\mathbb{C}$, Proposition 14.2.17 implies that every continuous map $f : K \mapsto \mathbb{S}^1$ admits a continuous extension to a map $F : \mathbb{C} \mapsto \mathbb{S}^1$. In particular, if U is a bounded connected component of $\mathbb{C}\backslash K$, and $p \in U$, then the Borsuk map $\beta_{p,K}$ would admit a continuous extension. But Proposition 14.2.15 says that $\beta_{p,K}$ does not have a continuous extension. Therefore $\mathbb{C}\backslash K$ has no bounded connected components. Since $\mathbb{C}\backslash K$ has only one unbounded connected component, we have established that $\mathbb{C}\backslash K$ is connected, as desired. ♢

Proposition 14.2.19 *Let K be a compact subset of $\mathbb{C}$. Assume that K is homeomorphic to the interval $[0, 1]$. Then $\mathbb{C}\backslash K$ is connected.*

PROOF. It is clear that every continuous map from $[0,1]$ to $\mathbb{S}^1$ is homotopic to a constant. The conclusion then follows from Proposition 14.2.18. $\diamondsuit$

Proposition 14.2.20 *Let K be a compact subset of $\mathbb{C}$. Assume that K is homeomorphic to $\mathbb{S}^1$. Then*

1. *$\mathbb{C}\backslash K$ is not connected;*
2. *if U is any connected component of $\mathbb{C}\backslash K$, then the closure of U is the set $U \cup K$.*

PROOF. The identity map $z \mapsto z$ from $\mathbb{S}^1$ to $\mathbb{S}^1$ is not homotopic to a constant. (For example, if it was homotopic to a constant then Proposition 14.2.17 would yield a continuous extension to the disc $D(0,1) = \{z \in \mathbb{C} : |z| \leq 1\}$. This extension would be a continuous retraction from the disc to its boundary, contradicting the no-retraction theorem.)

Therefore not every continuous map from $\mathbb{S}^1$ to $\mathbb{S}^1$ is homotopic to a constant. So not every continuous map from K to $\mathbb{S}^1$ is homotopic to a constant. Therefore Proposition 14.2.18 implies that $\mathbb{C}\backslash K$ is not connected.

Now let U be a connected component of $\mathbb{C}\backslash K$, and let $\bar{U}$ denote its closure. Then $U \subseteq \bar{U} \subseteq U \cup K$. To prove our second conclusion, we have to pick $z \in K$ and show that $z \in \bar{U}$. Let B be an open disc with center z, and let B' be a compact ball with center z such that $B' \subseteq B$. Let J be a nonempty subset of K which is contained in B and is such that $K' = K\backslash J$ is homeomorphic to $[0,1]$. Let V be another connected component of $\mathbb{C}\backslash K$. Pick points $p \in U$, $q \in V$. Then p and q belong to $\mathbb{C}\backslash K'$, which is connected by Proposition 14.2.19. So there exists a continuous path $\xi : [0,1] \mapsto \mathbb{C}\backslash K'$ such that $\xi(0) = p$ and $\xi(1) = q$. Then ξ cannot be contained in $\mathbb{C}\backslash K$, because it joins points in two different components of $\mathbb{C}\backslash K$. Therefore $\xi(t) \in J$ for some t, and then $\xi(t) \in B'$ for some t. Let $\bar{t}$ be the smallest $t \in [0,1]$ such that $\xi(t) \in B'$. If $\bar{t} = 0$, then $p \in B'$, so $B' \cap U \neq \emptyset$ and then $B \cap U \neq \emptyset$. If $\bar{t} > 0$, then the set $\Xi = \{\xi(t) : 0 \leq t < \bar{t}\}$ is connected and entirely contained in U (because Ξ does not intersect K). Since $\xi(\bar{t}) \in B'$, $\xi(t)$ must belong to B for $t < \bar{t}$, t sufficiently close to $\bar{t}$. Therefore $B \cap U \neq \emptyset$.

We have shown that $B \cap U \neq \emptyset$ if B is an arbitrary open disc with center z. Therefore $z \in \bar{U}$, and our proof is complete. $\diamondsuit$

Proposition 14.2.21 *Let K be a compact subset of $\mathbb{C}$, and let $f : K \mapsto \mathbb{C}$ be a continuous map. Then f is homotopic to a constant map if and only if*

there exists a continuous function $\varphi : K \mapsto \mathbb{R}$ *such that* $f(z) = e^{2\pi i\varphi(z)}$ *for all* $z \in K$.

PROOF. If φ exists, then we can define $h(z,t) = e^{2\pi i(1-t)\varphi(z)}$ for $z \in K$ and $t \in [0,1]$, and it is clear that h is a homotopy from f to a constant map.

Conversely, suppose $q \in \mathbb{S}^1$ and $h : K \times [0,1] \mapsto \mathbb{S}^1$ be a continuous map that satisfies the conditions $h(z,0) = f(z)$ and $h(z,1) = q$ whenever $z \in K$. Fix α such that $0 < \alpha < \frac{1}{2}$. Then the map $[-\alpha,\alpha] \ni ue^{2\pi iu}$ is a bijection onto the compact subset $S = \{e^{2\pi iu} : u \in [-\alpha,\alpha]\}$ of $\mathbb{S}^1$. Fix β such that $0 < \beta$ and the set $V = \{z \in \mathbb{S}^1 : |z-1| \leq \beta\}$ is contained in S.

The map h is uniformly continuous on $K \times [0,1] \mapsto \mathbb{S}^1$, so there exists a positive δ such that $|h(z,t) - h(z,s)| < \beta$ whenever $z \in K$, $t, s \in [0,1]$, and $|t-s| \leq \delta$. Let N be a positive integer such that $N\delta > 1$. For $j = 1, \ldots, N$, let g_j be the map

$$K \ni z \mapsto g_j(z) \stackrel{\text{def}}{=} \frac{h\Big(z, \frac{j-1}{N}\Big)}{h\Big(z, \frac{j}{N}\Big)} \in \mathbb{S}^1 .$$

Let $G(z) \equiv q$. Then $f = g_1 \cdot g_2 \cdot \ldots \cdot g_N \cdot G$. Clearly, each g_k takes values in S, so there exists a continuous $\psi_j : K \mapsto \mathbb{R}$ such that $g_j(z) = e^{2\pi i\psi_j(z)}$ for all $z \in K$. Let $c \in \mathbb{R}$ be such that $q = e^{2\pi ic}$. Define $\varphi = \psi_1 + \cdots + \psi_N + c$. Then $f(z) = e^{2\pi i\varphi(z)}$ for all $z \in K$. ♢

Proposition 14.2.22 *Let* K *be a compact subset of* $\mathbb{C}$. *Assume that* K *is the union of two compact sets* K_1, K_2, *such that* $K_1 \cap K_2$ *is connected. Let* p_1, p_2 *be points of* $\mathbb{C}\backslash K$ *that belong to the same connected component of* $\mathbb{C}\backslash K_1$ *and to the same connected component of* $\mathbb{C}\backslash K_2$. *Then* p_1, p_2 *belong to the same connected component of* $\mathbb{C}\backslash K$.

PROOF. Since p_1 and p_2 belong to the same connected component of $\mathbb{C}\backslash K_1$, the Borsuk maps $\beta_{p_1,K_1}, \beta_{p_2,K_1} : K_1 \mapsto \mathbb{S}^1$ are homotopic. Therefore, if we let

$$f_1(z) \stackrel{\text{def}}{=} \frac{\beta_{p_2,K_1}(z)}{\beta_{p_1,K_1}(z)} \qquad \text{for} \quad z \in K_1 ,$$

it follows that the map f_1 is homotopic to a constant. By Proposition 14.2.21, this implies that there exists a continuous function $\varphi_1 : K_1 \mapsto \mathbb{R}$ such that $f_1(z) = e^{2\pi i\varphi_1(z)}$ whenever $z \in K_1$. A similar argument shows that there exists a continuous function $\varphi_2 : K_2 \mapsto \mathbb{R}$ such that $f_2(z) = e^{2\pi i\varphi_2(z)}$

whenever $z \in K_2$, where

$$f_2(z) \stackrel{\text{def}}{=} \frac{\beta_{p_2,K_2}(z)}{\beta_{p_1,K_2}(z)} \qquad \text{for} \quad z \in K_2\,.$$

The function $\psi = \varphi_2 - \varphi_1 : K_1 \cap K_2 \mapsto \mathbb{R}$ is continuous and integer-valued. Since $K_1 \cap K_2$ is connected, ψ is in fact constant. After subtracting this constant from φ_2, we may assume that $\varphi_2 = \varphi_1$ on $K_1 \cap K_2$. We can the define $\varphi : K \mapsto \mathbb{R}$ by letting $\varphi(z) = \varphi_j(z)$ whenever $z \in K_j$, $j = 1, 2$. Then, if we let

$$f(z) \stackrel{\text{def}}{=} \frac{\beta_{p_2,K}(z)}{\beta_{p_1,K}(z)} \qquad \text{for} \quad z \in K\,,$$

we have $f(z) = e^{2\pi i\varphi(z)}$. Therefore f is homotopic to a constant, and then the maps $\beta_{p_1,K}$ and $\beta_{p_2,K}$ are homotopic.

Clearly, $\beta_{p_2,K}$ has a continuous extension to a map $\mu : \mathbb{C}\backslash\{p_2\}$. Since $\beta_{p_1,K}$ is homotopic to $\beta_{p_2,K}$, Proposition 14.2.17 tells us that $\beta_{p_1,K}$ also has a continuous extension to a map $\nu : \mathbb{C}\backslash\{p_2\}$.

If the connected components U_1, U_2 of p_1, p_2 in $\mathbb{C}\backslash K$ were different, p_2 would not belong to U_1, and then ν could be restricted to $K \cup U_1$, yielding a continuous extension of $\beta_{p_1,K}$ to $K \cup U_1$. On the other hand, Proposition 14.2.15 tells us that such an extension does not exist. Therefore $U_1 = U_2$, completing our proof. ♢

Theorem 14.2.23 *(The Jordan curve theorem.) Let K be a compact subset of $\mathbb{C}$. Assume that K is homeomorphic to $\mathbb{S}^1$. Then*

1. *$\mathbb{C}\backslash K$ has exactly two connected components;*
2. *if U is any connected component of $\mathbb{C}\backslash K$, then* $\text{Closure}(U) = U \cup K$.

PROOF. The second conclusion was already proved in Proposition 14.2.20. Furthermore, we also know from Proposition 14.2.20 that $\mathbb{C}\backslash K$ has at least two connected components. Therefore, it suffices to show that $\mathbb{C}\backslash K$ cannot have two different bounded connected components.

Let $\xi : \mathbb{S}^1 \mapsto K$ be a homeomorphism, and let $\gamma(t) = \xi(e^{2\pi it})$ for $0 \le t \le 1$, so γ is a continuous map from $[0, 1]$ onto K such that $\gamma(0) = \gamma(1)$ and γ is one-to-one on $[0, 1[$.

Suppose U_1, U_2 are different bounded connected components of $\mathbb{C}\backslash K$. Let U_0 be the unbounded component of $\mathbb{C}\backslash K$. Fix points p_0, p_1, p_2 in U_0, U_1, U_2, respectively. Fix eight times t_1, t_2, ..., t_8 having the property

that $0 < t_1 < t_2 < t_3 < t_4 < t_5 < t_6 < t_7 < t_8 < 1$. Write $J_\ell = [t_{2\ell-1}, t_{2\ell}]$ and $I_\ell =]t_{2\ell-1}, t_{2\ell}[$ for $\ell = 1, 2, 3, 4$. For each $\ell \in \{1, 2, 3, 4\}$, define a compact set K_ℓ by letting $K_\ell = K \backslash \gamma(I_\ell)$. Then each K_ℓ is homeomorphic to a compact interval, so the open sets $\Omega_\ell = \mathbb{C} \backslash K_\ell$ are connected by Proposition 14.2.19.

Since p_0 and p_1 belong to Ω_ℓ for each ℓ, there exist continuous one-to-one maps $\eta_\ell : [0, 1] \mapsto \Omega_\ell$ such that $\eta_\ell(0) = p_0$ and $\eta_\ell(1) = p_1$. (We are only interested in the curves η_1 and η_3.) Since p_0 and p_1 belong to different connected components of $\mathbb{C} \backslash K$, the curves η_ℓ must intersect K. Since η_ℓ does not intersect K_ℓ, it must intersect $\gamma(J_\ell)$. If we let s_ℓ be the smallest $s \in [0, 1]$ such that $\eta_\ell(s) \in \gamma(J_\ell)$, then $\eta_\ell(s_\ell) \in \gamma(J_\ell)$, and $\eta_\ell(s) \in U_0$ for $0 \leq s < s_\ell$. Define $\zeta_\ell : [0, 1] \mapsto \mathbb{C}$ by restricting η_ℓ to $[0, s_\ell]$ and reparametrizing time so that the interval $[0, s_\ell]$ becomes $[0, 1]$. Then ζ_ℓ is a continuous one-to-one map from $[0, 1]$ to $\mathbb{C}$, such that $\zeta_\ell(0) = p_0$, $\zeta_\ell(1) \in \gamma(J_\ell)$, and $\zeta_\ell(t) \in U_0$ for $0 \leq t < 1$.

The sets $\zeta_3([0, 1])$ and $\gamma(J_1)$ are compact and do not intersect, because $\zeta_3([0, 1]) \cap K = \{\zeta_3(1)\}$, and $\zeta_3(1) \notin \gamma(J_1)$. Therefore $\zeta_1(t) \notin \zeta_3([0, 1])$ if t is close enough to 1. Let $\bar{t}_1$ be the largest $t \in [0, 1]$ such that $\zeta_1(t) \in \zeta_3([0, 1])$. Let $\bar{t}_3$ be the unique $t \in [0, 1]$ such that $\zeta_3(t) = \zeta_1(\bar{t}_1)$. If $\bar{t}_1 > 0$, replace p_0 by $\zeta_1(\bar{t}_1)$, and then replace ζ_1 and ζ_3 by their restrictions to $[t_1, 1]$ and $[t_3, 1]$, reparametrized so that the intervals $[t_1, 1]$, $[t_3, 1]$ become $[0, 1]$.

We then have two continuous maps $\zeta_1, \zeta_3 : [0, 1] \mapsto \mathbb{C}$ such that

1. ζ_1 and ζ_3 are one-to-one,
2. $\zeta_1(0) = \zeta_3(0) = p_0$,
3. $\zeta_1(t) \in U_0$ and $\zeta_3(t) \in U_0$ whenever $0 \leq t < 1$,
4. $\zeta_1(1) \in \gamma(J_1)$ and $\zeta_3(1) \in \gamma(J_3)$,
5. $\zeta_1(t) \neq \zeta_3(s)$ whenever $0 \leq t \leq 1$ $0 \leq s \leq 1$, and $t + s > 0$.

Let τ_1, τ_3 be the unique points in J_1, J_3 such that $\gamma(\tau_1) = \zeta_1(1)$ and $\gamma(\tau_3) = \zeta_3(1)$. We construct two simple closed curves Λ_1, Λ_2, as follows. The curve Λ_1 is the concatenation of ζ_1, followed by the restriction of γ to the interval $[\tau_1, \tau_3]$, followed by ζ_3 run backwards. The curve Λ_3 is defined similarly, except for the fact that, to go from $\gamma(\tau_1)$ to $\gamma(\tau_3)$ we move along K in the other direction (that is, we follow the map $s \mapsto \xi(e^{2\pi i(\tau_1 - s)})$ from $s = 0$ until $s = 1 + \tau_1 - \tau_3$, at which point $\xi(e^{2\pi i(\tau_1 - s)})$ becomes $\gamma(\tau_3)$).

We now show that p_1 and p_2 can be connected by a continuous path ω_1 in $\mathbb{C} \backslash |\Lambda_1|$, where Λ_1 is, of course, the carrier of Λ_1. To see this, it suffices to

construct continuous paths $\theta_j : [0,1] \mapsto \mathbb{C}$, for $j = 1, 2$, such that $\theta_1(0) = p_1$, $\theta_2(0) = p_2$, $\theta_1(t) \in U_1$ for $0 \le t < 1$, $\theta_2(t) \in U_2$ for $0 \le t < 1$, $\theta_1(1) \in \gamma(J_4)$, and $\theta_2(1) \in \gamma(J_4)$. We then define ω_1 by concatenating θ_1, a portion of γ going from $\theta_1(1)$ to $\theta_2(1)$ along $\gamma(J_4)$, and θ_2 run backwards. The resulting path ω_1 is entirely contained in $U_1 \cup U_2 \cup \gamma(J_4)$, so ω_1 does not intersect $|\Lambda_1|$.

A similar argument shows that there exists a continuous path ω_3 connecting p_1 and p_2 in $\mathbb{C}\backslash|\Lambda_3|$. Since $|\Lambda_1| \cap |\Lambda_3| = |\zeta_1| \cup |\zeta_2|$, which is homeomorphic to a compact interval, Proposition 14.2.22 tells us that there exists a continuous path ω connecting p_1 and p_2 in $\mathbb{C}\backslash(|\Lambda_1| \cup |\Lambda_3|)$. Since $K \subseteq |\Lambda_1| \cup |\Lambda_3|)$, ω connects p_1 and p_2 in $\mathbb{C}\backslash K$, contradicting the assumption that U_1 and U_2 are different connected components of $\mathbb{C}\backslash K$. $\diamondsuit$

14.2.15 The winding numbers of a simple loop

The *Jordan curve theorem* says that if ξ is a simple loop then $\mathbb{R}^2\backslash|\xi|$ has exactly two connected components. In other words, $\mathbb{R}^2\backslash|\xi|$ is the union of two disjoint connected open subsets Ω_1, Ω_2 of $\mathbb{R}^2$. If $R \in \mathbb{R}$ is such that $R > 0$ and $\|\xi(s)\| \le R$ for all $s \in \mathbb{S}^1$, then the set $S = \{q \in \mathbb{R}^2 : \|q\| > R\}$ is open, connected, and disjoint from $|\xi|$. Therefore S is contained in one of the components Ω_1, Ω_2. The other component must then be contained in $\{q \in \mathbb{R}^2 : \|q\| \le R\}$. So one of the components is unbounded and the other one is bounded. We use $\mathrm{Ins}(\xi)$, $\mathrm{Out}(\xi)$ (the "inside" and the "outside" of ξ) to denote, respectively, the bounded and the unbounded connected component of $\mathbb{R}^2\backslash|\xi|$. We refer to $\mathrm{Ins}(\xi)$ as *the region enclosed by the simple loop* ξ. We use $A(\xi)$ to denote the area of $\mathrm{Ins}(\xi)$.

Theorem 14.2.24 *Let $\xi : \mathbb{S}^1 \mapsto \mathbb{R}^2 = \mathbb{C}$ be a simple loop. Then the funcion $\mathbb{C}\backslash|\xi| \ni p \mapsto \nu(\xi, p) \in \mathbb{Z}$ is constant on $\mathrm{Out}(\xi)$ and on $\mathrm{Ins}(\xi)$, and has the value 0 on $\mathrm{Out}(\xi)$ and the value 1 or -1 on $\mathrm{Ins}(\xi)$.*

PROOF. Proposition 14.2.13 tells us that the winding number $\nu(\xi, p)$ depends continuously on p as long as p varies on $\mathbb{C}\backslash|\xi|$. Since this number is always an integer, the function $\mathbb{C}\backslash|\xi| \ni p \mapsto \nu(\xi, p) \in \mathbb{Z}$ must be constant on $\mathrm{Out}(\xi)$ and on $\mathrm{Ins}(\xi)$.

Fix a positive $R \in \mathbb{R}$ such that $|\xi(e^{2\pi i t})| \le R$ for all $t \in R$, and a $p_* \in \mathbb{C}$ such that $|p| > R$. Then $\left|\frac{\xi(e^{2\pi i t})}{p_*}\right| < 1$ for every $t \in \mathbb{R}$, so we can define a continuous function $\sigma : \mathbb{R} \mapsto \mathbb{C}$ by letting

$$\sigma(t) = \mathrm{Ln}\Big(1 - \frac{\xi(e^{2\pi i t})}{p_*}\Big) \quad \text{for} \quad t \in \mathbb{R}\,.$$

Let $f_0 \in \mathbb{C}$ be such that $e^{f_0} = -p_*$. Define $f(t) = f_0 + \sigma(t)$. Then f is continuous and

$$e^{f(t)} = e^{\bar{f}}.e^{\sigma(t)} = (-p_*).\Big(1 - \frac{\xi(e^{2\pi it})}{p_*}\Big) = \xi(e^{2\pi it}) - p_*\,.$$

Therefore $\nu(\xi, p_*) = \frac{1}{2\pi i}(f(1) - f(0)) = 0$, since f is clearly 1-periodic. So $\nu(\xi, p_*) = 0$, and then $\nu(\xi, p) = 0$ for all $p \in \mathrm{Out}(\xi)$.

Now choose a point $\hat{p} \in \mathrm{Ins}(\xi)$. Fix a point $\hat{q}$ of $|\xi|$ closest to $\hat{p}$. Let $\hat{r} = |\hat{p} - \hat{q}|$. Then the open disc $\hat{D} = \{z : |z - \hat{p}| < \hat{r}\}$ does not intersect $|\xi|$ and is entirely contained in $\mathrm{Ins}(\xi)$. After a translation and a rotation of the coordinate axes of the plane, we may assume that $\hat{p} = 0$ and $\hat{q} = \hat{r}$. Fix a $\hat{t} \in \mathbb{R}$ such that, $\xi(e^{2\pi i\hat{t}}) = \hat{r}$. Fix a number $\hat{\varepsilon}$ such that $0 < \hat{\varepsilon} < \frac{1}{2}$, and two continuous functions $\rho : [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}] \mapsto]0, \infty[$, $\varphi : [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}] \mapsto \mathbb{R}$, such that $\xi(e^{2\pi it}) = \rho(t)e^{2\pi i\varphi(t)}$ whenever $t \in [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}]$, and $\varphi(\hat{t}) = 0$. Then $\rho(t) \geq \hat{r}$ for all $\in [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}]$, and $\rho(\hat{t}) = \hat{r}$.

Then, if $0 < \varepsilon \leq \hat{\varepsilon}$, it cannot be the case that $\varphi(t)$ vanishes for all $t \in [\hat{t} - \varepsilon, \hat{t} + \varepsilon]$. (Indeed, suppose that $\varphi(t) = 0$ whenever $t \in [\hat{t} - \varepsilon, \hat{t} + \varepsilon]$. Then $\xi(e^{2\pi it}) = \rho(t) \in \mathbb{R}$ for $t \in [\hat{t} - \varepsilon, \hat{t} + \varepsilon]$. Therefore the function $[\hat{t} - \varepsilon, \hat{t} + \varepsilon] \ni t \mapsto \xi(e^{2\pi it})$ would be real-valued and would have a local minimum at $\hat{t}$, contradicting the fact that ξ is one-to-one on some neighborhood of $\hat{t}$.) It follows that there exists a sequence $\{\tau_j\}_{j=1}^{\infty}$ in $\mathbb{R}$ such that $\tau_j \to \hat{t}$ as $j \to \infty$ and $\varphi(\tau_j) \neq 0$ for all j. If we let L_j be the closed half-line $\{re^{2\pi i\varphi(\tau_j)} : r \geq 0\}$, then $L_j \cap |\xi|$ is a compact set. Furthermore, $L_j \cap |\xi|$ is nonempty, because $\xi(\tau_j) = \rho(\tau_j)e^{2\pi i\varphi(\tau_j)} \in L_j \cap |\xi|$. So $L_j \cap |\xi|$ has a point q_j closest to 0. Write $q_j = \xi(e^{2\pi it_j})$, $t_j \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}]$. Also, let $q_j = r_j e^{2\pi i\varphi(t_j)}$, so $\hat{r} \leq r_j \leq \rho(\tau_j)$.

We claim that $\lim_{j\to\infty} t_j = \hat{t}$. Indeed, if this was not so there would exist a subsequence $\{t_{j(\ell)}\}_{\ell=1}^{\infty}$ of $\{t_j\}_{j=1}^{\infty}$ that converges to a $\tilde{t} \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}] \backslash \{\hat{t}\}$. Let $\tilde{q} = \xi(e^{2\pi i\tilde{t}})$. Then $\tilde{q} = \lim_{\ell\to\infty} q_{j(\ell)}$. Since $e^{2\pi i\varphi(\tau_j)} \to 1$ as $j \to \infty$, and $e^{2\pi i\varphi(t_j)} = e^{2\pi i\varphi(\tau_j)}$ (because $q_j \in L_j$), we conclude that $e^{2\pi i\varphi(t_j)} \to 1$ as $j \to \infty$, so $\tilde{q}$ is a positive real number $\tilde{r}$, and $\tilde{r} = \lim_{\ell\to\infty} r_{j(\ell)}$. Since $\hat{r} \leq r_{j(\ell)} \leq \rho(\tau_{j(\ell)})$, we have $\tilde{r} = \hat{r}$, and then $\tilde{q} = \hat{q}$. Since $|\tilde{t} - \bar{t}| \leq \frac{1}{2}$, the fact that $\xi(e^{2\pi i\tilde{t}}) = \tilde{q} = \hat{q} = \xi(e^{2\pi i\bar{t}})$ implies $\tilde{t} = \bar{t}$, which is a contradiction, proving our claim.

By passing to a subsequence we may assume that $|t_j - \hat{t}| < \hat{\varepsilon}$ for all j, and either $t_j < \hat{t}$ for all j or $t_j > \hat{t}$ for all j. Then $q_j = \rho(t_j)e^{2\pi i\varphi(t_j)}$—so $r_j = \rho(t_j)$ and $\varphi(t_j) = \varphi(\tau_j)$—and $re^{2\pi i\varphi(t_j)} \notin |\xi|$ whenever $0 \leq r < r_j$. By passing to a subsequence, we assume that either $\varphi(t_j) < 0$ for all j or $\varphi(t_j) > 0$ for all j.

Under our conditions, one of the following possibilities occurs:

I. $t_j < \hat{t}$ and $\varphi(t_j) < 0$ for all j,

II. $t_j < \hat{t}$ and $\varphi(t_j) > 0$ for all j,

III. $t_j > \hat{t}$ and $\varphi(t_j) < 0$ for all j,

IV. $t_j > \hat{t}$ and $\varphi(t_j) > 0$ for all j.

We will discuss Case I, and prove that $\nu(\xi, 0) = 1$. (The other three cases are similar. In Cases II and III one can prove that $\nu(\xi, 0) = -1$, and in Case IV that $\nu(\xi, 0) = 1$.)

Write $\varphi_j = -\varphi(t_j)$, so $\varphi_j > 0$. Fix a positive real number δ such that $\delta < \hat{r} = \text{dist}(0, |\xi|)$. Let $\eta_j : [0, 3] \mapsto \mathbb{C}$ be the curve given by

$$\eta_j(u) = \begin{cases} \Big((1-u)r_j + u\delta\Big)e^{-2\pi i\varphi_j} & \text{if} \quad 0 \le u \le 1\,, \\ \delta e^{-2\pi i((2-u)\varphi_j + 1 - u)} & \text{if} \quad 1 \le u \le 2\,, \\ (u-2)\hat{r} + (3-u)\delta & \text{if} \quad 2 \le u \le 3\,. \end{cases}$$

Then, as u goes from 0 to 3, $\eta_j(u)$ starts at the point $\xi(e^{2\pi i t_j}) = r_j e^{-2\pi i\varphi_j}$, moves along the straight-line segment M_j^1 from $r_j e^{-2\pi i\varphi_j}$ to $\delta e^{-2\pi i\varphi_j}$, then follows an arc M_j^2 of length $\delta.(2\pi - \varphi(t_j))$ of the circle $\{z : |z| = \delta\}$, clockwiset from $\delta e^{-2\pi i\varphi_j}$ to δ, and finally moves along the straight-line segment M_3 from δ to $\hat{r}$, ending at $\hat{r} = \xi(e^{2\pi i\hat{t}})$. It is clear that η_j is a one-to-one map. Furthermore, $\eta_j(u) \notin |\xi|$ whenever $0 < u < 3$.

Define $\hat{\zeta}_j : [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}] \mapsto \mathbb{C}$ by letting

$$\hat{\zeta}_j(t) = \begin{cases} \xi(e^{2\pi i t}) & \text{if} \quad t \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}] \backslash [t_j, \hat{t}]\,, \\ \eta_j\Big(\frac{3(t-t_j)}{\hat{t}-t_j}\Big) & \text{if} \quad t \in [t_j, \hat{t}]\,. \end{cases}$$

Then $\hat{\zeta}_j$ is continuous, and $\hat{\zeta}_j(\hat{t} + \frac{1}{2}) = \hat{\zeta}_j(\hat{t} + \frac{1}{2})$, so we can extend $\hat{\zeta}_j$ to a 1-periodic map ζ_j from $\mathbb{R}$ to $\mathbb{C}$. Write $\zeta_j(t) = \omega_j(e^{2\pi i t})$ for $t \in \mathbb{R}$. Then ω_j is a simple loop. Furthermore, $|\omega_j| \subseteq \{z : |z| \le R\}$, so $p_* \in \text{Out}(\omega_j)$, and then $\nu(\omega_j, p_*) = 0$.

Let $I_j = \{\xi(e^{2\pi i t}) : t_j < t < \hat{t}\}$. Let $K_j = |\xi| \backslash I_j$. Let $U_j = \mathbb{C} \backslash K_j$. Then K_j is a compact subset of $\mathbb{C}$ which is homeomorphic to the interval $[0, 1]$. Therefore K_j does not disconnect U_j. Since p_* and 0 obviously belong to U_j, there exists a continuous path $\gamma_j : [0, 1] \mapsto U_j$ such that $\gamma_j(0) = p_*$ and $\hat{\gamma}_j(1) = 0$. This path cannot be contained in $\mathbb{C} \backslash |\xi|$, because p_* and

0 belong to different connected components of $\mathbb{C}\backslash|\xi|$. Therefore the set $G_j = \{s \in [0,1] : \hat{\gamma}_j(s) \in |\xi|\}$ is nonempty. Since G_j is obviously compact, it has a smallest member s_j. Then $\gamma_j(s_j) \in |\xi|$ and $\gamma_j(s) \notin |\xi|$ whenever $0 \leq s < s_j$. Since γ_j is contained in U_j, $\gamma_j(s_j)$ cannot belong to K_j, so $\gamma_j(s_j) \in I_j$. Therefore $\gamma_j(s_j) = \xi(e^{2\pi i v_j})$ for some $v_j \in I_j$. Pick an $\ell(j)$ such that $t_j < t_{\ell(j)} < \hat{t}$. Define a curve $\mu_j : [0,3] \mapsto \mathbb{C}$ by letting

$$\mu_j(u) = \begin{cases} \gamma_j(s_j u) & \text{if} \quad 0 \leq u \in\leq 1\,, \\ \xi(e^{2\pi i((2-u)v_j + (u-1)t_{\ell(j)})}) & \text{if} \quad 1 \leq u \leq 2\,, \\ (3-u) r_j e^{2\pi i \varphi(t_{\ell(j)})} & \text{if} \quad 2 \leq u \leq 3\,. \end{cases}$$

Then μ_j is continuous, $\mu_j(0) = p_*$, and $\mu_j(3) = 0$. Moreover, μ_j never goes through $|\omega_j|$. Therefore $\nu(\omega_j, 0) = \nu(\omega_j, p_*) = 0$.

We now relate $\nu(\omega_j, 0)$ to $\nu(\xi, 0)$. Let $f : \mathbb{R} \mapsto \mathbb{C}$ be a continuous map such that $e^{f(t)} = \xi(e^{2\pi i t})$ for all $t \in \mathbb{R}$. Define $g_j : [0,3] \mapsto \mathbb{C}$ by

$$g_j(u) = \begin{cases} \ln\Big((1-u)r_j + u\delta\Big) - 2\pi i \varphi_j & \text{if} \quad 0 \leq u \leq 1\,, \\ \ln \delta - 2\pi i((2-u)\varphi_j + 1 - u) & \text{if} \quad 1 \leq u \leq 2\,, \\ \ln\Big((u-2)\hat{r} + (3-u)\delta\Big) + 2\pi i & \text{if} \quad 2 \leq u \leq 3\,. \end{cases}$$

Then g_j is continuous and $e^{g_j(u)} = \eta_j(u)$ for $0 \leq u \leq 3$. Define

$$h_j(t) = \begin{cases} f(t) & \text{if} \quad \hat{t} - \frac{1}{2} \leq t \leq t_j, \,, \\ g_j\Big(\frac{3(t-t_j)}{\hat{t}-t_j}\Big) + f(t_j) - g_j(0) & \text{if} \quad t_j \leq t \leq \hat{t}\,, \\ f(t) + f(t_j) - g_j(0) + g_j(3) - f(\hat{t}) & \text{if} \quad \hat{t} \leq t \leq \hat{t} + \frac{1}{2}\,. \end{cases}$$

Then h_j is continuous and $e^{h_j(t)} = \zeta_j(t)$ whenever $t \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}]$. Therefore $h_j(\hat{t} + \frac{1}{2}) - h_j(\hat{t} - \frac{1}{2}) = 2\pi i\, \nu(\omega_j, 0) = 0$. Then

$$f\Big(\hat{t}+\tfrac{1}{2}\Big) + f(t_j) - g_j(0) + g_j(3) - f(\hat{t}) = h_j\Big(\hat{t}+\tfrac{1}{2}\Big) = h_j\Big(\hat{t}-\tfrac{1}{2}\Big) = f\Big(\hat{t}-\tfrac{1}{2}\Big)\,,$$

so that

$$f\Big(\hat{t} + \tfrac{1}{2}\Big) - f\Big(\hat{t} - \tfrac{1}{2}\Big) = f(t_j) - g_j(0) + g_j(3) - f(\hat{t})\,.$$

It follows from the definition of g_j that

$$g_j(0) = \ln r_j - 2\pi i \varphi_j \qquad \text{and} \qquad g_j(3) = \ln \hat{r} + 2\pi i\,.$$

Since $\lim_{j\to\infty} r_j = \hat{r}$ (because $\hat{r} \leq r_j \leq \rho(\tau_j)$, $\tau_j \to \hat{t}$, and $\rho(\hat{t}) = \hat{r}$), and $\lim_{j\to\infty} \varphi_j = \lim_{j\to\infty}(-\varphi(\tau_j)) = -\varphi(\hat{t}) = 0$, we can conclude that $\lim_{j\to\infty}(g_j(3) - g_j(0)) = 2\pi i$. Therefore

$$f\Big(\hat{t}+\tfrac{1}{2}\Big)-f\Big(\hat{t}-\tfrac{1}{2}\Big)=\lim_{j\to\infty}\Big(f(t_j)-g_j(0)+g_j(3)-f(\hat{t})\Big)=2\pi i\,,$$

since $\lim_{j\to\infty} t_j=\hat{t}=2\pi i$. Hence

$$\nu(\xi,0)=\frac{1}{2\pi i}\Big(\,f\Big(\hat{t}+\frac{1}{2}\Big)-f\Big(\hat{t}-\frac{1}{2}\Big)\,\Big)=1\,.$$

Therefore $\nu(\xi,z)=1$ for every $z\in\mathrm{Ins}(\xi)$, and our proof is complete. $\diamondsuit$

14.3 Green's formula

If $\xi\in C^0(\mathbb{S}^1,\mathbb{C})$ is a loop in $\mathbb{C}$), the *winding number function* of ξ is the function

$$\mathbb{C}\backslash|\xi|\ni z\mapsto\nu_\xi(z)\stackrel{\mathrm{def}}{=}\nu(\xi,z)\in\mathbb{Z}\,.$$

We want to prove that ν_ξ is a well defined member of $L^1(\mathbb{R}^2,\mathbb{R})$ if ξ is rectifiable. For this purpose, we have to know first of all that the set $|\xi|$ has measure zero, so the fact that ν_ξ is not well defined on $|\xi$ does not matter, and ν_ξ is well defined as a measurable function from $\mathbb{R}^2$ to $\mathbb{R}$ up to equality almost everywhere.

Proposition 14.3.1 *Let $\xi\in C^0(\mathbb{S}^1,\mathbb{C})$ be a rectifiable loop. Then the 2-dimensional Lebesgue measure of $|\xi|$ is equal to zero.*

PROOF. In view of Proposition vaca, we may assume that ξ is Lipschitz. Let C be a Lipschitz constant for the map $\mathbb{R}\ni t\mapsto\eta(t)=\xi(e^{2\pi it})$. For each $n,k\in\mathbb{N}$ such that $1\le k\le n$, let $D_{k,n}$ be the disc with center $\eta\Big(\frac{2k-1}{2n}\Big)$ and radius $\frac{C}{2n}$. Then $|\xi|\subseteq\bigcup_{k=1}^n D_{k,n}$, so

$$\mathrm{meas}(|\xi|)\subseteq\sum_{k=1}^n\mathrm{meas}(D_{k,n})=n\pi\frac{C^2}{4n^2}=\frac{\pi C^2}{4n}\,.$$

Since n is arbitrary, $\mathrm{meas}(|\xi|)=0$, as stated. $\diamondsuit$

Proposition 14.3.2 *Let $C>0$, and let $\xi\in C^0(\mathbb{S}^1,\mathbb{C})$ be a Lipschitz loop with Lipschitz constant C. Then*

1. *The winding number function ν_ξ is Lebesgue integrable.*
2. *There exists a sequence $\{\xi_j\}_{j=1}^\infty$ of polygonal loops in $\mathbb{C}$ with Lipschitz constant C such that, as $j\to\infty$,*

2.a. $\xi_j \to \xi$ *uniformly,*

2.b. the functions ν_{ξ_j} *converge to* ν_ξ *in* $L^1(\mathbb{R}^2, \mathbb{R})$.

PROOF. Let $\eta(t) = \xi(e^{2\pi i t})$, so $\eta : \mathbb{R} \mapsto \mathbb{C}$ is a 1-periodic function. For $j = 1, 2, \ldots$, we define $\eta_j : \mathbb{R} \mapsto \mathbb{C}$ by letting

$$\eta_j(t) = (k - 2^j t)\xi\Big(\frac{k-1}{2^j}\Big) + (2^j t + 1 - k)\xi\Big(\frac{k}{2^j}\Big) \quad \text{if} \quad k \in \mathbb{Z}\,,\ \frac{k-1}{2^j} \le t \le \frac{k}{2^j}\,.$$

Then each η_j is 1-periodic, and linear on each interval $I_{j,k} \stackrel{\text{def}}{=} \Big[\frac{k-1}{2^j}, \frac{k}{2^j}\Big]$. If we define $\xi_j : \mathbb{S}^1 \mapsto \mathbb{C}$ by letting $\xi_j(e^{2\pi i t}) = \eta_j(t)$, then ξ_j is a polygonal loop with 2^j sides which is Lipschitz with constant C. Furthermore, the ξ_j converge uniformly to ξ as $j \to \infty$.

For $k = 0, 1, \ldots, 2^j$, let $A_k^j = \xi\Big(\frac{k}{2^j}\Big)$, so $A_j^0, \ldots, A_j^{2^j-1}$ are the vertices of ξ_j, $A_j^{2^j} = A_j^0$, and the segments $\sigma_j^k = \overline{A_j^{k-1} A_j^k}$, for $k = 1, \ldots, 2^j$, are the sides of ξ_j. Then ξ_{j+1} is obtained from ξ_j by replacing each side σ_j^k by the concatenation of the two segments σ_{j+1}^{2k-1} and σ_{j+1}^{2k}. That means that the polygonal ξ_{j+1} is obtained from ξ_j by adjoining to it the 2^j triangles $T_j^k = A_j^{k-1} A_{j+1}^{2k-1} A_j^k$, for $k = 1, \ldots, 2^j$. Therefore

$$\nu_{\xi_{j+1}} = \nu_{\xi_j} + \sum_{k=1}^{2^j} \nu_{T_j^k}\,.$$

For given j, k, The function $\nu_{T_j^k}$ has the value 1 or -1 inside T_j^k, and 0 outside. Therefore, $\nu_{T_j^k}$ is integrable, and its L^1 norm is equal to the area inside T_j^k, which is bounded by $C^2 2^{-2j-2}$. It follows that

$$\Big\|\nu_{\xi_{j+1}} - \nu_{\xi_j}\Big\|_{L^1} \le C^2 2^{-j-2}\,.$$

This shows that the sequence $\{\nu_{\xi_j}\}_{j=1}^\infty$ converges in $L^1(\mathbb{R}^2, \mathbb{R})$ to a limit $\hat{\nu}$.

On the other hand, if $z \notin |\xi|$, then $z \notin |\xi_j|$ if j is large enough, and $\lim_{j\to\infty} \nu_{\xi_j}(z) = \nu_\xi(z)$. In view of Proposition 14.3.1, $\nu_{\xi_j}(z) \to \nu_\xi(z)$ for almost all $z \in \mathbb{C}$. Therefore $\nu_\xi = \hat{\nu}$, and our conclusion is proved. ♢

Remark 14.3.3 The results of Propositions 14.3.1 and 14.3.2 are also true, with almost identical proof, if the curve ξ is Hölder continuous with exponent α, as long as $\alpha > \frac{1}{2}$. Indeed, to prove Proposition 14.3.1 in this case it suffices

to let $D_{k,n}$ be the disc with center $\eta\left(\frac{2k-1}{2n}\right)$ and radius $\left(\frac{C}{2n}\right)^{\alpha}$. To prove Proposition 14.3.2, we use the same ξ_j s, and observe that the area inside T_j^k is bounded by $C^2 2^{-2\alpha j-2}$. Then

$$\left\|\nu_{\xi_{j+1}} - \nu_{\xi_j}\right\|_{L^1} \leq C^2 2^{-(2\alpha-1)j-2}\,,$$

from which it still follows that $\{\nu_{\xi_j}\}_{j=1}^{\infty}$ converges in $L^1(\mathbb{R}^2, \mathbb{R})$. ♢

Theorem 14.3.4 *(Green's formula) Let $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$ be a rectifiable loop. Then*

$$A(\xi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \nu_\xi(x+iy)\,dx\,dy\,. \tag{A.43}$$

PROOF. Both sides of (A.43) are invariant under time reparametrization, so we may replace ξ by its arc-length reparametrization, and assume that ξ is a Lipschitz loop with Lipschitz constant 1.

By Proposition 14.3.2, there exists a sequence $\{\xi_j\}_{j=1}^{\infty}$ of polygonal loops that are Lipschitz with constant 1, converge uniformly to ξ, and are such that the functions ν_{ξ_j} converge to ν_ξ in L^1. Then Proposition vaca implies that $A(\xi_j) \to A(\xi)$. Therefore the validity of (A.43) for all rectifiable loops ξ follows if (A.43) is true for all polygonal loops.

To prove (A.43) for polygonal loops, it suffices to observe that each polygonal loop can be obtained from a triangle by successive adjunctions of triangles, and both sides of (A.43) are additive with respect to adjunction. Hence it suffices to prove (A.43) for triangles. Finally, the proof of (A.43) for triangles is an elementary direct calculation. ♢

Corollary 14.3.5 *Let $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$ be a simple rectifiable loop. Let $\omega = 1$ if ξ is oriented cuonterclockwise, and $\omega = -1$ if ξ is oriented clockwise. Then the area of the region* $\mathrm{Ins}(\xi)$ *is equal to $\omega A(\xi)$.*