The Brachistochrone Problem and Modern Control Theory

Héctor J. Sussmann[1]
Department of Mathematics
Rutgers University
Hill center, Busch Campus
Piscataway, NJ 08854, USA
sussmann@math.rutgers.edu
http://www.math.rutgers.edu/~sussmann

Jan C. Willems
Department of Mathematics
University of Groningen
P.O. Box 800, 9700 AV Groningen
The Netherlands
J.C.Willems@math.rug.nl
http://www.math.rug.nl/~willems

Dedicated to Velimir Jurdjevic on his 60th birthday

1. Introduction

The purpose of this paper is to show that *modern control theory*, both in the form of the "classical" ideas developed in the 1950s and 1960s, and in that of later, more recent methods such as the "nonsmooth," "very nonsmooth" and "differential-geometric" approaches, provides the best and mathematically most natural setting to do justice to Johann Bernoulli's famous 1696 "brachistochrone problem." (For the classical theory, especially the smooth version of the Pontryagin Maximum Principle, see, e.g., Pontryagin *et al.* [**31**], Lee and Markus [**26**], Berkovitz [**3**]; for the nonsmooth approach, and the version of the Maximum Principle for locally Lipschitz vector fields, cf., e.g., Clarke [**12, 13**], Clarke *et al.* [**14**]; for very nonsmooth versions of the Maximum Principle, see Sussmann [**35, 36, 39, 40, 41, 42, 43, 44**]; for the differential-geometric approach, cf., e.g., Jurdjevic [**24**], Isidori [**19**], Nijmeijer and van der Schaft [**30**], Jakubczyk and Respondek [**20**], Sussmann [**37, 40**].)

We will make our case in favor of modern control theory in two main ways. *First*, we will look at four approaches to the brachistochrone problem, presenting them in chronological order, and comparing them. *Second*, we will look at several "variations on the theme of the brachistochrone," that is, at several problems closely related to the one of Johann Bernoulli.

The first of our two lines of inquiry will be pursued in sections 3, 4, 5, and 6, devoted, respectively, to

[1]The writing of this paper began while the first author was Johann Bernoulli Visiting Professor at the University of Groningen, in May, June and July 1999, and continued while he was back at Rutgers University, partially supported by the National Science Foundation under NSF Grant DMS 98-03411. It was concluded while he was Weston Visiting Professor at the Weizmann Institute of Science in Rehovot, Israel, from October 2000 to January 2001. He is grateful to the University of Groningen, the Weizmann Institute, and NSF, for their help and support.

1. Johann Bernoulli's own solution based on an analogy with geometrical optics,
2. the solution based on the classical calculus of variations,
3. the optimal control method,

and, finally,

4. the differential-geometric approach.

We will show that *each of the three transitions from one of the methods in the above list to the next one leads to real progress for the brachistochrone problem, by making it possible to derive stronger conclusions about it than those that were obtainable by means of previously existing techniques.*

Specifically, we will point out that

1.a the differential equation derived by Johann Bernoulli has spurious solutions (i.e., solutions other than the cycloids), as was first noticed by Taylor in 1715,

but

1.b the application of the Euler-Lagrange condition of the calculus of variations eliminates these solutions and leaves only the cycloids,

and that

2.a with the calculus of variations approach, the existence of solutions is a delicate problem, due to the non-coercivity of the Lagrangian,

but

2.b the optimal control control method renders this question trivial, reducing the proof of existence to a straightforward application of the Ascoli-Arzelà theorem,

and, moreover,

3.a the usual calculus of variations formulations require that one postulate that the optimal curve is the graph of a function $x \mapsto y(x)$,

but

3.b with the optimal control control method this postulate becomes a provable conclusion.

These observations show that the first two of the three transitions brought improvements to the analysis of Johann Bernoulli's problem. But the core of our argument is §6, where we look at the third transition and the advent of differential-geometric control theory. We will argue that,

on the one hand

4.a previous methods require that we assume extra knowledge of physics (in the form of the law of conservation of energy) in order even to be able to translate Johann's problem into mathematics,

and

5.a with those methods an anomalous situation appears to arise when one looks for a state space formulation of the problem and asks for its "Hamilton principal function" V (that is, the value function regarded as a function of both the initial and terminal states), since V ought to be a function of an even number ν of variables, but for the brachistochrone problem the best choice of ν seems to be *five*;

but on the other hand

4.b in the differential-geometric framework one does not need to bring in conservation of energy in order to *formulate* the problem; instead, one writes the equations as a control system in $\mathbb{R}^4$, and then computes the accessibility distribution, which turns out to be three-dimensional at each point; this shows that a nontrivial conserved quantity exists locally, and direct computation shows that it exists globally and is equal to the energy,

and

5.b the anomaly described in 5.a disappears, and the principal function V turns out to be defined on $\mathbb{R}^4 \times \mathbb{R}^4$ (although $V(q, q') < +\infty$ if and only if q and q' belong to the same leaf of the foliation of $\mathbb{R}^4$ by the level submanifolds of the energy function); the method provides a full explanation for the apparent anomaly, by showing that the full optimal synthesis requires impulse controls that lead to instantaneous jumps.

The discussion in §6 will also show that the differential-geometric approach brings into the problem interesting connections with other mathematical ideas. To begin with, the analysis leads naturally to a minimum time optimal control problem with a four-dimensional state space and an unbounded scalar control entering linearly into the dynamics. Moreover, this problem is quite degenerate, because its accessibility distribution happens to be three-dimensional, as pointed out above. Since the problem is affine in the control but not linear, complete controllability on each leaf is in principle problematic. The accessibility Lie algebra turns out to be the "diamond Lie algebra," a well known four-dimensional solvable Lie algebra that occurs, for example, in the study of the quantized harmonic oscillator. To establish controllability on each leaf one needs additional arguments, and it turns out that the issue can be settled by computing the Jurdjevic-Kupka Lie saturate (cf. Bonnard *et al.* [**8, 9**], Jurdjevic-Kupka [**21, 22**], Jurdjevic [**24**]) of the family of system vector fields. The minimum time problem then makes sense in principle for every pair of points belonging to the same leaf. But, since the control is unbounded, it is not at all obvious that an optimal control joining A to B exists for every pair (A, B) of points belonging to a given leaf L. An application of the Maximum Principle yields a vector field Z tangent to the leaves such that all optimal arcs are integral curves of Z. Since, in general, two arbitrarily chosen points A, B of a leaf L do not lie on the same integral curve of Z, it follows that for general pairs $(A, B) \in L \times L$ an optimal control joining A to B does not exist. On the other hand, when a control system is completely controllable and a suitable properness condition is satisfied[2], then most mathematicians would agree that optimal controls ought to exist, and if they do not then this can only happen because we have not made a good choice of our space $\mathcal{U}$ of admissible controls, so $\mathcal{U}$ has to be enlarged by adding suitably defined "generalized controls." (For a trivial example, consider the problem of moving from A to B in minimum time with the dynamics $\dot{x} = u$, u arbitrary, so in particular no bound is prescribed for the velocity u. Clearly, the "minimum time," is 0, and to attain it one needs to allow instantaneous jumps, which means

[2]See Definition 6.9.1 for a precise statement of the properness condition, and Proposition 6.9.2 for the proof that the condition is satisfied for the brachistochrone problem.

that the class of ordinary controls has to be enlarged to allow for Delta functions. We will show in §6 that something like this happens for the brachistochrone.) In our case, the true complete optimal synthesis of the problem turns out to require impulse controls, a fact known to be common in problems with "cheap" unbounded controls, and studied in other (linear quadratic) contexts by Jurdjevic and Kupka, cf. Jurdjevic-Kupka [**23**], Kupka [**25**].

Our second group of arguments in favor of modern control is presented in §7, and consists of "five modern variations on the theme of the brachistochrone." Following a suggestion made by M. Hestenes in 1956, we "speculate on what type of brachistochrone problem the Bernoulli brothers might have formulated had they lived in modern times" (Hestenes [**18**], p. 64), and we extend Hestenes' idea by speculating not only on "what type of ... problem the ... brothers might have formulated" but also on *what type of answer they might have liked for such problems.*

We deal with five questions. The first one is Hestenes' own "variation of the brachistochrone problem," which turns out to be a problem solvable by optimal control but not by classical calculus of variations techniques.

The second "variation" is the "reflected brachistochrone," a very natural extension of Johann's problem, in which the state space is the whole plane rather than a half plane. This problem is no longer solvable by means of traditional smooth or nonsmooth optimal control methods—because it is a minimum time optimal control problem with a non-Lipschitzian right-hand side—but turns out to lend itself to the "very nonsmooth" approach, which yields a solution with a simple geometric interpretation that, in our view, the Bernoulli brothers would have liked.

The third "variation" is the *brachistochrone problem with friction*, which corresponds, mathematically, to "unfolding" the original Johann Bernoulli question by embedding it in a one-parameter family of problems, the parameter being, of course, the friction coefficient ρ. These problems turns out to be nondegenerate, in the sense that, for example, their accessibility distribution has rank four at each point. Consequently, their time-optimal controls behave qualitatively like the time-optimal controls for generic control-affine systems in $\mathbb{R}^4$ with scalar unbounded controls. In particular, through every point there pass a one-parameter family of optimal curves, since the optimal control is given in feedback form as a function of the state and of an additional scalar parameter that stays constant along each trajectory. The analysis clarifies the way in which the original problem of Johann Bernoulli is degenerate: the accessibility Lie algebra is eight-dimensional for $\rho \neq 0$ but four-dimensional for $\rho = 0$; the accessibility distribution has rank four everywhere if $\rho \neq 0$ but has rank three everywhere if $\rho = 0$; the optimal "feedback" control is a function of the state and an extra parameter, but this parameter only occurs in a term that is multiplied by ρ, so that when $\rho = 0$ the optimal control is a function of the state only.

The fourth "variation" deals with the derivation of Snell's law of refraction from general necessary conditions for an optimum. Since Snell's law plays a crucial role in Johann's solution of the brachistochrone problem, we speculate that a natural question for him to have asked "if he had lived in modern times" is whether the general theories about minimizing curves that were created in the three centuries after Johann's work actually cover the Snell law case. Our answer is negative *if only the classical calculus of variations and pre-1990 smooth and nonsmooth optimal* *control are considered*, but becomes positive if we add to the list our recent very nonsmooth methods

To conclude the paper, we take the liberty of including a fifth "variation" that we find amusing but does not really have much to do with control theory. In this "variation"—which is a digression on a topic of great importance and related to geometry, to the brachistochrone, and to Johann Bernoulli—we "speculate" that Johann Bernoulli would have found it very interesting to know[3] that with his work on the brachistochrone he had been extremely close to solving one of the most important open problems in Geometry, namely, that of the independence of Euclid's fifth postulate from the other ones. (The problem had intrigued mathematicians since ancient times—cf., e.g., Bonola [**10**]—and was very much in the forefront of research in geometry in Johann's times, as shown by the publication of Saccheri's book [**32**] in 1733. A definitive solution was found by Beltrami in 1868, cf. Beltrami [**2**] and Stillwell [**33**].)

REMARK 1.1. An earlier, related version of the argument presented here appeared in [**38**]. That paper also dealt with the evolution from the brachistochrone to the calculus of variations and optimal control, but focused on a different set of issues, especially on the search for the correct Hamiltonian formulation of the necessary condition for an optimum and how opportunities were missed before the "control Hamiltonian" was finally discovered in the 1950s. ♢

2. Johann Bernoulli and the brachistochrone problem

In the June 1696 issue of the journal *Acta Eruditorum*, Johann Bernoulli—a Swiss-born professor of mathematics at the University of Groningen, in the Netherlands[4]—challenged mathematicians to solve the "problema novum" of finding the curve that—following Leibniz' suggestion—he named the "brachistochrone," from the Greek $\beta\rho\acute{\alpha}\chi\iota\sigma\tau o\varsigma$: shortest, and $\chi\rho\acute{o}\nu o\varsigma$: time[5]. The journal's May 1697 issue contains articles on this problem by six of the most renowned mathematicians of the time: Johann Bernoulli, Johann's elder brother Jakob, Leibniz, Newton, the marquis de l'Hôpital, and Tschirnhaus.

Leibniz called Johann Bernoulli's question a "splendid" problem. The fact that he and five other thinkers of such unparalleled distinction chose to become actively engaged in the search for its solution shows that at the time of its formulation the problem was perceived by the leaders of the mathematical community as a major challenge, lying at the very frontiers of their research.

It would not have been easy at that point to predict the future impact on science of Johann Bernoulli's challenge. Yet, the six famous authors of the 1697 papers must have had their own reasons to sense that an exceptionally promising line of inquiry was being opened, and the developments of the 300 years that followed have lent ample support to their intuition.

[3]Although he probably would not have thought of asking the question.

[4]The Bernoulli family was originally from Antwerp in Flanders, where they lived until 1583. Like many other Flemish Protestants, they fled Flanders to escape religious oppression by the Spanish rulers. They spent some time in Frankfurt, and finally settled in Basel, Switzerland, early in the seventeenth century.

[5]So the brachistochrone is the solution of a *minimum time* problem, i.e., of a genuine *optimal control* problem in the modern sense of the term!

Today, it is clear that the events of 1696-97 represented a critical turning point in the history of mathematics, and it is widely agreed that with them an important new field, later to be called the "calculus of variations," was born.

Remark 2.1. The words "calculus of variations" were first used by Euler in a 1760 paper, cf. [**15**]. Euler had received, in 1755, a letter from a 19-year-old mathematician from Turin, called Ludovico de la Grange Tournier, where a new method was proposed for the study of variational problems, by using what we would now call "variations" (cf. Goldstine [**16**], p. 110.). This was quite different from the approach Euler had been using until then, and presented in his 1744 book entitled *Methodus inveniendi maximi minimeve proprietate gaudentes sive solutio problematis isoperimetrici latissimo sensu accepti*, that is "A method for discovering curved lines having a maximum or minimum property or the solution of the isoperimetric problem taken in its widest sense." Euler's earlier method was based on a time-discretization followed by a passage to the limit[6], as the size h of the time intervals goes to 0. Euler was so enthusiastic about the new idea of the unknown youngster from Turin that he dropped his own method, adopted that of Lagrange instead, and renamed the subject the *calculus of variations*. In the summary to his first paper using variations, Euler says "Even though the author of this [Euler] had meditated a long time and had revealed to friends his desire yet the glory of first discovery was reserved to the very penetrating geometer of Turin LA GRANGE, who having used analysis alone, has clearly attained the same solution which the author had deduced by geometrical considerations." ♢

The newborn subject was destined to become a major area of mathematics. Moreover, its implications would soon extend to physics, whose very foundations would be shown in the 18th century—by Euler, Maupertuis, Lagrange, and others—to depend on principles formulated in the language of the calculus of variations.

So it is not surprising that the story of the brachistochrone is told in many books, and that the importance of this curve is stressed in most histories of mathematics. The brachistochrone has received honors not commonly bestowed on mathematical curves. For example, it is depicted in a stained-glass window of the Academy building (the main venue of the university) in Groningen.

Moreover, *the brachistochrone is probably the only mathematical curve that has been deemed worthy of a monument in its honor.* In 1994, on the occasion of the 375th anniversary of its foundation, the University of Groningen decided to honor in various ways the most famous former members of its faculty. In the case of Johann Bernoulli, who had been a professor from 1695 to 1705, a decision was made to honor him by erecting a monument to the brachistochrone, one of the most significant discoveries—and one of which he was particularly proud—that he made during his Groningen period. The unveiling of the brachistochrone monument, located in the Zernike complex of the University, took place in 1996, coinciding with the 300th anniversary of the publication of Johann's challenge in *Acta Eruditorum*.

2.1. Johann Bernoulli's challenge. The June 1696 challenge took the form of an "Invitation to all Mathematicians to solve a new problem":

[6]Euler himself talked about "letting h be infinitesimal."

> *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.*

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

Responses came from the best mathematical minds of the time. In addition to 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 Jakob; one by Tschirnhaus; one by l'Hôpital and, finally, one by Newton. Newton's solution 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 the solutions by Johann [**5**] and Jakob [**6**], as well as contributions by Tschirnhaus and l'Hôpital, 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*.

2.2. 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 a remarkable fact about the cycloid: *it is the only curve with the property that, when a body oscillates by falling under its own weight while being guided by this curve, then the period of the oscillations is independent of the* *initial point where the body is released.*[7] Therefore Huygens called this curve, the cycloid, the *tautochrone* (from $\tau\alpha\acute{\upsilon}\tau\acute{o}\varsigma$: 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:

> *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 that we even obtain from this a proof of its correctness. Nature always tends 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.*

2.3. Why was the brachistochrone so important in 1696-97? Perhaps the crucial fact about Johann Bernoulli's question is that he is telling us very clearly that *the solution is a curve that everybody knows*, and yet *there appears to be no way to guess which curve it is by means of some intuitive geometric argument.*

This situation should be compared with that arising in another famous and much older question about curve minimization, the *isoperimetric problem.* In this problem, we are asked to find the simple closed rectifiable curve that encloses the largest possible area among all simple closed rectifiable curves of a given length, and the answer also turns out to be a curve that is certainly "very well known among geometers," 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. Although this argument does not amount to a completely rigorous proof[8], it is undoubtedly rather convincing, and makes it quite obvious that the solution must be a circle.

Therefore, if Johann Bernoulli had challenged "the best mathematical minds" of his time 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," since everybo,y would have guessed immediately that the mysterious curve was the circle.

[7]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.

[8]The argument assumes the existence of a solution, as well as its uniqueness 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 S to a solution S', which may be different from S.

In the case of the brachistochrone, *there is no way to guess*[9]. 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[10], but can only be arrived at by using this new mathematical tool.

3. The standard formulation and Johann Bernoulli's solution

We now discuss the "standard formulation" of the brachistochrone problem in modern mathematical language, as it is usually found in modern books, and Johann Bernoulli's ingenious solution based on an analogy with geometrical optics.

In our view, the standard formulation is questionable. We will explain our objections in §6 below, where we will make a case for a different approach, based on differential-geometric control. But first it is necessary to present the classical point of view in detail.

The key fact is that the "movable point M" is supposed to move "under the influence of its own weight," which presumably means "under the influence of its own weight *and nothing else other than whatever force is needed to keep M on the given curve.*" This means, in particular, that *energy is conserved.*

3.1. Formulation of the brachistochrone problem. We choose x and y axes in the plane with the y axis pointing downwards, use (a, α) and (b, β) to denote, respectively the coordinates of the endpoints A and B of our curve, and fix a number $E \in \mathbb{R}$, thought of as the total energy of the particle. The condition characterizing the feasible paths

$$[0, T] \ni t \mapsto \xi(t) = (x(t), y(t)) \in \mathbb{R}^2 \tag{3.1}$$

will then be—using m, g to denote, respectively, the particle's mass and the gravitational acceleration—that the sum of the kinetic energy $\frac{m}{2}(\dot{x}(t)^2 + \dot{y}(t)^2)$ and the potential energy $-gmy(t)$ is equal to E. We can certainly choose units so that $m = 1$. The condition then becomes

$$\frac{1}{2}(\dot{x}(t)^2 + \dot{y}(t)^2) - gy(t) = E\,.$$

So we formally define an (A, B, E)-*feasible trajectory* (or (A, B, E)-*feasible path*) to be a map (3.1), defined on some interval $[0, T]$, such that

(i) $\xi(0) = A$, $\xi(T) = B$,
(ii) ξ is absolutely continuous,
(iii) $\frac{1}{2}(\dot{x}(t)^2 + \dot{y}(t)^2) = E + gy(t)$ for almost all $t \in [0, T]$.

---

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

[10] 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 known 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 need 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.

Notice that Condition (iii) implies in particular that the derivatives $\dot{x}$ and $\dot{y}$ are bounded, so a feasible path is in fact Lipschitz continuous.

Condition (i) states that the path ξ must start at A and end at B. The number $E + g\alpha$ is the initial kinetic energy of the body. (Often, only the case $E + g\alpha = 0$ is considered, corresponding to the body being dropped at time 0 with zero initial speed.) Condition (ii) is of course a modern technicality that would have made no sense to Johann Bernoulli, who was far from imagining that there could exist curves that cannot be recovered by integrating their derivative. Condition (iii) reflects conservation of energy. (The law that the kinetic energy of a body which has fallen from a height h is increased by an amount proportional to h was due to Galileo, and was well known in Bernoulli's time.)

An (A, B, E)-feasible path $\xi : [0, T] \to \mathbb{R}^2$ is said to be *optimal* if there exists no (A, B, E)-feasible path $\tilde{\xi} : [0, \tilde{T}] \to \mathbb{R}^2$ for which $\tilde{T} < T$. An (A, B, E)-*brachistochrone* is either (a) an optimal (A, B, E)-feasible path, or (b) the geometric locus in $\mathbb{R}^2$ of such a path, that is, a subset $\mathcal{B}$ of $\mathbb{R}^2$ such that

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

for some optimal (A, B, E)-feasible path $\xi^* : [0, T^*] \mapsto \mathbb{R}^2$. A *brachistochrone* is an object which is an (A, B, E)-brachistochrone for some A, B, E.

Obviously, a feasible path[11] (3.1) must be entirely contained in the closed half-plane

$$H_+(\eta) = \{(x, y) : y \geq -\eta\}, \quad \text{where} \quad \eta = \frac{E}{g}, \tag{3.2}$$

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

It is clear that *the solution cannot always be a straight line*, as Johann Bernoulli rightly points out in his "Invitation." For example, consider the extreme case when $E = 0$ and $a = \alpha = \beta = 0 \neq b$. (That is, when the endpoints of the curve are at the same height but do not coincide, and the initial kinetic energy vanishes.) 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, (iii) 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 $[0, b]$. 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. ♢

3.2. 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

[11] From now on, as long as A, B and E are fixed, we just talk about "feasible paths" and "brachistochrones," without explicitly specifying A, B and E.

through B, without hitting the x axis in between. (It is easy to see that this defines the cycloid uniquely.)

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 (iii) above gives us a formula for the "speed of light" $c(x, y)$ as a function of the position (x, y):

$$c(x, y) = \sqrt{2E + 2gy}\,.$$

We now 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

$$c(x, y) = \sqrt{y}\,.$$

It is then reasonable to expect that, if we use a suitable discretization of our problem, then the optimal paths of the discretized problem will approach those of the original problem as the discretization parameter δ goes to 0. Johann Bernoulli did this as follows. Let us define

$$y_k^\delta = k\delta \quad \text{for} \quad \delta > 0 \text{ and } k = 0, 1, \ldots ,$$

and then divide the half-plane H_+ into horizontal strips

$$S_k^\delta \stackrel{\text{def}}{=} \{(x, y) \in \mathbb{R}^2 : y_k^\delta \le y \le y_{k+1}^\delta\} \quad \text{for} \quad k = 0, 1, \ldots ,$$

of height δ. Johann's discretization is then obtained by replacing c by a constant c_k^δ in each strip S_k^δ. The constants c_k^δ are defined by

$$c_k^\delta = \sqrt{y_{k+1}^\delta} \quad \text{for} \quad \text{each} \quad k\,.$$

He then computed the light rays for the original problem by taking the limit of the light rays for the discretized problem as $\delta \downarrow 0$.

The light rays of the discretized problem can be studied using Snell's 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.

Snell had observed that, if the speeds of light on the two sides H_1, H_2 of a line L in the plane are different constants v_1, v_2, and a light ray R traverses H_1 and then enters H_2 after being refracted at L, then the ratio $\frac{\sin\theta_1}{\sin\theta_2}$ of the sines of the incidence and refraction angles is a constant, independent of θ_1. (By definition, θ_i is, for $i = 1, 2$, the angle formed by the part R_i of R that lies in H_i with the line perpendicular to L. It is clear that the R_i are straight lines.) 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

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2} \quad \text{or, equivalently,} \quad \frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2}\,, \tag{3.3}$$

Johann Bernoulli used (3.3) to study the light rays of the discretized problem. If R is such a ray, and we let R_k be the part of R that lies in the strip S_k^δ, and use θ_k^δ to denote the angle of R_k with a vertical line, then (3.3) enabled him to conclude that the quantity

$$\nu = \frac{\sin\theta_k^\delta}{\sqrt{y_{k+1}^\delta}}$$

is a constant, since in each strip S_k^δ the speed of our light ray is $\sqrt{y_{k+1}^\delta}$.

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

$$\frac{dx^2}{dx^2+dy^2} = \nu y\,,$$

where ν is a constant. Then

$$\frac{dx^2+dy^2}{dx^2} = \frac{1}{\nu y}\,,$$

i. e.,

$$1 + y'(x)^2 = \frac{C}{y}\,, \ \text{ where } \ C = \frac{1}{\nu}\,.$$

So the curve described by expressing the y-coordinate of the brachistochrone as a function of its x-coordinate will satisfy the differential equation

$$y'(x) = \sqrt{\frac{C-y(x)}{y(x)}}\,, \tag{3.4}$$

for some constant C.

The parametrized curves $\varphi \mapsto (x(\varphi), y(\varphi))$ given by

$$x(\varphi) = x_0 + \frac{C}{2}(\varphi - \sin\varphi)\,, \qquad y(\varphi) = \frac{C}{2}(1-\cos\varphi)\,, \qquad 0 \le \varphi \le 2\pi\,, \tag{3.5}$$

satisfy (3.4). 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 passing through A and B and belonging to the family of curves (3.5).*

(The parameter count for (*) is obviously right: (3.5) 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.)

4. Spurious solutions and the calculus of variations approach

4.1. Remarks on Johann Bernoulli's argument. The argument that we have presented is the one of Johann Bernoulli, and Equation (3.4) is the one that he wrote in his paper, followed by the statement "from which I conclude that the *Brachistochrone* is the ordinary *Cycloid*." (He 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. [**34**], 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 today as

$$y'(x) = \pm\sqrt{\frac{C-y(x)}{y(x)}}\,, \tag{4.1}$$

or, equivalently,

$$y(x)(1+y'(x)^2) = \text{constant}\,. \tag{4.2}$$

In particular, the solution curves should be allowed to have a negative slope, contrary to what a "modern" interpretation of (3.4) might suggest. But y' should stay continuous, so that a switching from a $+$ to a $-$ solution of (4.1) is not permitted.

Even with the more accurate rewriting (4.2), the differential equation derived by Johann Bernoulli also has *spurious solutions*, not given by (3.5). 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 (3.5), 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 (3.5). Such paths are, indeed, compatible with Huygens' law of refraction.

Remark 4.1.1. 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.) ♢

4.2. 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. This equation gives a necessary condition for a function y_* to minimize the integral

$$J = \int_a^b L(y(x), y'(x), x)\,dx$$

within the class $\mathcal{Y}$ of all functions

$$[a,b] \ni x \mapsto y(x) \in \mathbb{R}$$

that satisfy some appropriate technical conditions (for example, the requirement that $y(\cdot)$ is Lipschitz continuous) and are such that $y(a)$ and $y(b)$ have given values α, β.

Here, the "Lagrangian"

$$\Omega \times \mathbb{R} \times [a,b] \ni (y,u,x) \mapsto L(y,u,x) \in \mathbb{R}$$

is a given function, and Ω is an open interval in $\mathbb{R}$. Under suitable technical conditions on L, y_*, and the set $\mathcal{Y}$, if y_* is a minimizer then the condition

$$\frac{d}{dx}\left(\frac{\partial L}{\partial u}\Big(y_*(x), y_*'(x), x\Big)\right) = \frac{\partial L}{\partial y}\Big(y_*(x), y_*'(x), x\Big) \tag{4.3}$$

must be satisfied for almost all x.

It is easy to see that the above result can be applied to the brachistochrone problem, provided that we *postulate*[12] that it suffices to consider curves in the x,y plane that are graphs of functions $x \mapsto y(x)$ defined on the interval $[a,b]$. Then Constraint (iii) can be written (taking $E = 0$ and $2g = 1$ as before) as

$$dx^2 + dy^2 = y\,dt^2\,,$$

which gives

$$dt = \frac{\sqrt{dx^2+dy^2}}{\sqrt{y}} = L(y,y')dx\,,$$

where

$$L(y,u) = \frac{\sqrt{1+u^2}}{\sqrt{y}}\,. \tag{4.4}$$

So Johann Bernoulli's problem becomes that of minimizing the integral

$$J = \int_a^b L(y(x), y'(x))dx\,,$$

subject to $y(a) = \alpha$ and $y(b) = \beta$, where L is given by (4.4).

This gives the equation

$$1 + y'(x)^2 + 2y(x)y''(x) = 0\,, \tag{4.5}$$

which is *stronger* than (4.2), since (4.2) is equivalent to $y' + {y'}^3 + 2yy'y'' = 0$, i.e., to $y'(1 + {y'}^2 + 2yy'') = 0$, whose solutions are those of (4.5) plus the spurious solutions found earlier.

It is easy to verify that the solutions of the Euler-Lagrange equation (4.5) are *exactly* the curves given by (3.5), without any extra spurious solutions, showing that, for the brachistochrone problem, *the Euler-Lagrange method gives better results than Johann Bernoulli's approach.*

We now show that *optimal control is even better.*

---

[12]With optimal control, this "postulate" becomes a *provable conclusion*, as will be shown in §5 below.

5. The optimal control approach

Having shown that the Euler-Lagrange equation gives better results than Johann Bernoulli's method, we now look at another important step in the evolution of the theory of minimizing curves, namely, optimal control theory. We show that optimal control provides even better results in two ways:

1. With optimal control it is no longer necessary to postulate that the optimal curve is the graph of a function $x \mapsto y(x)$.
2. In the optimal control setting, the problem of the existence of an optimum becomes trivial.

5.1. 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 dynamical behavior is given by

$$\dot{x} = u\sqrt{|y|}\,, \quad \dot{y} = v\sqrt{|y|}\,. \tag{5.1}$$

Here the control is a 2-dimensional vector (u, v) taking values in the set

$$U = \{(u, v) : u^2 + v^2 = 1\}\,. \tag{5.2}$$

(Actually, for the "true" brachistochrone problem the motion is restricted to the upper half plane, so we could have written $\sqrt{y}$ rather than $\sqrt{|y|}$. We choose to use the more general expression $\sqrt{|y|}$ because later, in the study of the reflected brachistochrone, we will want to work in the whole plane.)

Let us use p, q, p_0 to denote, respectively, the momentum variables conjugate to x and y, and the abnormal multiplier. Then the control theory Hamiltonian $H(x, y, u, v, p, q, p_0, t)$ is given (using $\sigma = \operatorname{sgn} y$) by the formula

$$H(x, y, u, v, p, q, p_0, t) = (pu + qv)\sqrt{\sigma y} - p_0\,.$$

The Pontryagin Maximum Principle then tells us that if a curve

$$[0, T] \ni t \mapsto \xi(t) = (x(t), y(t))$$

is optimal then there exist absolutely continuous functions $t \mapsto p(t)$ and $t \mapsto q(t)$, and a nonnegative constant p_0, such that, if we let $u = \dot{x}$, $v = \dot{y}$, and write $\vec{p}$ to denote the momentum vector (p, q), and $\|\vec{p}\|$ to denote its Euclidean norm $\sqrt{p^2 + q^2}$, then the Hamiltonian maximization conditions

$$u(t) = \frac{p(t)}{\|\vec{p}(t)\|}\,, \quad v(t) = \frac{q(t)}{\|\vec{p}(t)\|}\,, \tag{5.3}$$

as well as the "adjoint system" of differential equations

$$\dot{p}(t) = 0\,, \quad \dot{q}(t) = -\sigma\,\frac{p(t)u(t) + q(t)v(t)}{2\sqrt{\sigma y(t)}} = -\frac{\sigma\|\vec{p}\|}{2\sqrt{\sigma y(t)}}\,, \tag{5.4}$$

are satisfied for almost all t. (Notice that $\|\vec{p}(t)\| \neq 0$. Indeed, the Maximum Principle also gives the condition that $H = 0$. So $\|\vec{p}\| = 0$ would imply $p_0 = 0$, contradicting the nontriviality of the triple $(p(t), q(t), p_0)$.)

If the constant p 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

$$y'(x) = \frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{v}{u} = \frac{q}{p}\,, \tag{5.5}$$

we have

$$1 + y'(x)^2 = \frac{\|\vec{p}\|^2}{p^2} \tag{5.6}$$

and

$$y''(x) = \frac{1}{p} \cdot \frac{dq}{dx} = \frac{\dot{q}}{p\dot{x}}\,. \tag{5.7}$$

But (5.1) and (5.3) imply that

$$\dot{x} = \frac{p\sqrt{\sigma y}}{\|\vec{p}\|}\,, \tag{5.8}$$

and then (5.4) and (5.7) yield

$$y''(x) = -\frac{\|\vec{p}\|^2}{2yp^2}\,. \tag{5.9}$$

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

$$1 + {y'}^2 + 2yy'' = 0\,, \tag{5.10}$$

which is exactly Equation (4.5). As before, this leads to the cycloids, with no "spurious solutions." Notice that this argument does not involve any discretization or any use of Snell's 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!* This should be contrasted with the analysis based on the calculus of variations, where this fact had to be postulated.)

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

5.2. Optimality proofs and the existence problem. For a second example showing the advantages 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 solely 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.

To give a rigorous proof of the optimality of the cycloids, valid for all choices of A and B in the closed half-plane $H_+ = \{x, y) : y \geq 0\}$, one can proceed in two ways.

One possibility (cf. Bliss [**7**]) is to use Hamilton-Jacobi theory. Another, much simpler, approach, is to prove first the *existence* of an optimal trajectory. Indeed, once this existence result is known for all $A, B \in H_+$, then it is clear that any optimal trajectory $\xi_* : [0, T] \mapsto H_+$ 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 $\xi_*(\alpha)$, $\xi_*(\beta)$. (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 $\xi_*(\alpha)$ and $\xi_*(\beta)$ are interior points of 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 (3.5) that goes through $\xi_*(\alpha)$ and $\xi_*(\beta)$. So ξ_* is such that $\xi_*\lceil[\alpha,\beta]$ is given by (3.5) whenever $0<\alpha<\beta<T$, and this easily implies that ξ_* itself is a cycloid given by (3.5).

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, using the following lemma.

LEMMA 5.2.1. *Suppose L is a horizontal line contained in the interior of the upper half-plane H_+. Let P, Q be distinct points of L. Then the straight-line segment joining P to Q is strictly faster than any trajectory ζ from P to Q such that ζ lies entirely in the lower closed half-plane determined by L and at least one point of ζ lies strictly below L.*

PROOF. Let T_ξ, T_ζ be the corresponding times. Let $\bar{y}$ be the y coordinate of the points of L, so $\bar{y}>0$. Write $P=(p,\bar{y})$, $Q=(q,\bar{y})$. Let $\xi:[0,T]\mapsto H_+$ be the horizontal segment from P to Q. Then

$$T_\xi=\frac{|q-p|}{\sqrt{\bar{y}}}\,.$$

Now let $[a,b]\ni t\mapsto\zeta(t)=(x(t),y(t))\in H_+$ be a trajectory from P to Q as in our statement. Then

$$\dot{x}(t)\le\|\dot{\zeta}(t)\|=\sqrt{y(t)}\le\sqrt{\bar{y}}\,,$$

and the inequality is strict for almost all t in some interval I of positive measure. So

$$|q-p|=|x(b)-x(a)|=\Big|\int_a^b\dot{x}(t)\,dt\Big|<(b-a)\sqrt{\bar{y}}=T_\zeta\sqrt{\bar{y}}\,.$$

Therefore

$$T_\zeta>\frac{|q-p|}{\sqrt{\bar{y}}}=T_\xi\,,$$

and our proof is complete. ♢ □

The following is an immediate consequence of the lemma.

COROLLARY 5.2.2. *Let A, B be points of H_+ such that $A\ne B$, and assume that $\xi:[a,b]\mapsto H_+$ is time optimal. Then $\xi(t)$ is an interior point of H_+ whenever $a<t<b$.*

PROOF. Let t be such that $a<t<b$. Assume $\xi(t)$ belongs to the boundary ∂H_+ of H_+. Then it cannot be the case that $\xi(s)\in\partial H_+$ for every $s\in[a,t]$, because if this was the case then the speed of motion on the interval $[a,t]$ would vanish, so $\xi(a)=\xi(t)$, and then we could obtain a trajectory with the same endpoints as ξ and strictly faster than ξ by just replacing ξ by its restriction to $[t,b]$. A similar argument shows that it is not true that $\xi(s)\in\partial H_+$ for all $s\in[t,b]$. It then follows that there exist a positive number δ, and times s_1, s_2 such that $a\le s_1<t<s_2\le b$, having the property that $\xi(s_1)\in L$ and $\xi(s_2)\in L$, where L is the line $\{(x,y):y=\delta\}$. The set $S=\{s\in[a,b]:\xi(s)\notin L\}$ is a relatively open subset of the interval $[a,b]$. Let I be the connected component of S that contains t. Then I is an interval, and I is relatively open in $[a,b]$. Since $a\le s_1<t<s_2\le b$, $s_1\notin S$, and $s_2\notin S$, the interval I is in fact open in $\mathbb{R}$, so $I=]\tau_1,\tau_2[$ where $s_1\le\tau_1<t<\tau_2\le s_2$. Let ζ be the restriction of ξ to the compact interval $[\tau_1,\tau_2]$.

Then Lemma 5.2.1 implies that ζ is not optimal, contradicting the fact that ξ is optimal. Therefore $\xi(t) \notin \partial H_+$. $\diamondsuit$ $\square$

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[13], because

(a) the Lagrangian L given by $L(y,u) = \frac{\sqrt{1+u^2}}{\sqrt{y}}$ has a singularity at $y = 0$,

and, even more importantly,

(b L is not sufficiently coercive for the usual existence theorems to apply.

(The precise meaning of (b) is that, as a function of u, L just grows like a constant times $|u|$, whereas for the usual existence theorems to apply L would have to grow like a constant times $|u|^r$ for some r such that $r > 1$. More sophisticated existence theorems for classical calculus of variations problems that include the brachistochrone are stated in Cesari [**11**], §14.3 and §14.4, but these conditions are not at all simple, and in addition they depend on other results by Cesari that were to appear in a book that was never published.)

In the control theory setting, however, the existence question is easily settled by a completely trivial application of the Ascoli-Arzelà 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 $T_j \downarrow T$. Then the ξ_j are all contained in a fixed compact subset of H_+, as can easily be shown by applying Gronwall's inequality and using the fact that the right-hand side of (5.1) has less than linear growth as a function of x and y. Then (5.1) implies that the derivatives $\dot{\xi}_j$ are uniformly bounded as well. So we can use the Ascoli-Arzelà theorem to 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 (5.1) 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. (If one of the points A, B belongs to ∂H_+, and $A \neq B$, then the argument would not work if the speed of light was y rather than $\sqrt{y}$, for in that case there would be no path joining A and B in finite time.)

6. The differential-geometric connection

6.1. Is the standard formulation of the brachistochrone problem satisfactory? The mathematical formulation of the brachistochrone problem, as presented in §3.1, 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***

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

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 let the mathematical analysis discover energy conservation for us, after which we could use this new fact to find the solution. *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 each given pair (A, B) of endpoints, 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.)

6.2. Hamilton's principal function. At this point, we digress to discuss an important idea of Hamilton, namely, the "principal" or "characteristic" function. This idea was not truly appreciated for more than century, until optimal control theory emerged to provide the right setting that would do it justice. So we will only describe it here in broad outline, following the 1944 article [**45**] by J. L. Synge. For the specific case of the brachistochrone, we will show below that *the search for Hamilton's principal function leads directly to the discovery of the differential-geometric structure underlying the problem.*

Synge wrote that

> *To Hamilton, optics and dynamics were merely two aspects of the calculus of variations. He was not interested in experiments. To him, optics was the investigation of the mathematical properties of of curves giving stationary values to an integral of the type*
>
> $$V = \int v\Big(x, y, z, \frac{dx}{du}, \frac{dy}{du}, \frac{dz}{du}\Big)du\,. \tag{6.1}$$
>
> *Hamilton was, in fact, a great contributor—probably the gratest single contributor of all time—to the calculus of variations.*
> *Consider two points, A' with coordinates x', y', z' and A with coordinates x, y, z. Consider all possible curves connecting A' and A. To each curve corresponds a value of the above integral.* [Note by HJS&JCW: Hamilton was working with Lagrangians that are homogeneous of degree one with respect to the velocities. In that case, the choice of the interval of definition of the curves is immaterial, since every curve can be reparametrized to be defined on [0,1] without changing the value of the cost integral.] *Now compare these values. Is there one curve that gives a smaller value to the integral than all others? It would seem that there must be, but it would be a rash conclusion. Suffice it to say that "in general" there is a curve giving a smallest value to the integral. Then such a curve is a* **ray** *in optics. In the language of the calculus of variations, it is an extremal.*
> *Now here is Hamilton's great central idea.* **Regard the minimum (or stationary) value of the integral as a function of the six coordinates** x', y', z', x, y, z **of** A' **and** A**.** *This function is Hamilton's characteristic or principal function. There is nothing hard to understand about that.*

> *What is hard to follow is Hamilton's plan to develop all the properties of the extremals from the characteristic function. In fact, it is so hard that few mathematicians in the ensuing century have given serious consideration to this plan.*

With the rise of optimal control, "Hamilton's plan to develop all the properties of the extremals from the characteristic function" has become one of the best and most widely used tools in the modern analysis of optimal control problems, under the name of the "dynamic programming" approach. Synge's complaint that "few mathematicians in the ensuing century have given serious consideration to this plan" may have been true in 1944 but is certainly not true today.

6.3. What is Hamilton's principal function for the brachistochrone? Let us look at the brachistochrone problem from the modern dynamic programming point of view. The natural questions to ask are

1. *what is the state space for the brachistochrone problem,*

and

2. *what is the problem's "principal function" in Hamilton's sense?*

More precisely, we would like to know

3. *what, exactly, is the controlled dynamical system*

$$\dot{q} = f(q, u)\,, \quad q \in Q\,, \quad u \in U \tag{6.2}$$

   that corresponds to the brachistochrone problem? (and, in particular, *what is the system's state space Q?*)
4. *what is the cost functional J?*

and

5. *what is the value function V, regarded as a function*

$$Q \times Q \ni (q_i, q_t) \mapsto V(q_i, q_t) \in \mathbb{R} \cup \{-\infty, +\infty\}$$

   of the initial and terminal points, as in Hamilton's definition?

REMARK 6.3.1. Naturally, Hamilton's definition and Synge's version of it have to be modified to adjust them to contemporary standards of mathematical precision. Synge's text quoted above seems to require the existence of an optimum for the principal function to be well defined. We take the point of view that *$V(q_i, q_t)$ is always well defined as an extended real number—i.e., a member of $\mathbb{R} \cup \{-\infty, +\infty\}$.* Precisely, $V(q_i, q_t)$ is the *infimum* of the costs of all the trajectories going from q_i to q_t. In particular, $V(q_i, q_t) = +\infty$ if and only if there exist no trajectories from q_i to q_t with cost $< +\infty$. ♢

We want to address these questions taking seriously the facts that *energy conservation need not be known for it to be possible to state the problem mathematically,* and that, consequently, *there should not be a different brachistochrone problem for each value E of the energy, but only one brachistochrone problem.*

It is clear that J should be time. But what exactly should we choose for Q, U, and f? Can we, for example, take $Q = \mathbb{R}^2$? This would mean that one can specify arbitrary points q_i and q_t in $\mathbb{R}^2$ and define $V(q_i, q_t)$ unambiguously. But we know that this is not so. Given q_i and q_t, the minimization problem as formulated in §3.1 only makes sense *for a fixed value E of the energy*, and *only if both q_i and q_t belong to the E-dependent closed half-plane $H_+\left(\frac{E}{g}\right)$*. Therefore there is a well defined value $V^E(q_i, q_t)$ for each value of E such that $q_i \in H_+\left(\frac{E}{g}\right)$ and $q_t \in H_+\left(\frac{E}{g}\right)$, 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 §2.1, as suggesting this possibility.) This, however, will not do, because it would have the following unnatural consequence: *if we pick two points q, q' in an optimal arc Γ going from q_i to q_t, and use $\Gamma(q, q')$ to denote the piece of Γ from q to q', then $\Gamma(q, q')$ will not in general be an optimal arc from q to q'* (because the particle that started falling with zero velocity at q_i will not in general have zero velocity at q). In modern terminology, *the problem with zero initial kinetic energy does not satisfy the principle of optimality.*[14]

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 q_i, q_t, 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.

The key problem is that *Hamilton's principal function must be a function of an even number of parameters,* since it is a function of the initial and terminal states, but *the natural number of variables for the value function of the brachistochrone problem seems to be* **five**, *namely, two coordinates for the initial position, two for the terminal position, and one for the energy.*

6.4. 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

$$\left\{\begin{array}{rcl} \dot{x} & = & v\,, \\ \dot{y} & = & w\,, \\ \dot{v} & = & -uw\,, \\ \dot{w} & = & uv - g\,. \end{array}\right. \tag{6.3}$$

Here x, y, v and w are the state variables, and u is the control. The values of u are allowed to be arbitrary real numbers.

[14]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.

The variables 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 (6.3) 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 "†" to denote "transpose," and writing $q=[x,\ y,\ v,\ w]^\dagger$, we get a familiar equation, namely,

$$\dot{q} = F(q) + uG(q)\,, \tag{6.4}$$

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

This is a *4-dimensional system*, with state space Q equal to $\mathbb{R}^4$, and control space U equal to $\mathbb{R}$.

6.5. The diamond Lie algebra. Having expressed our equations of motion in the form (6.4), 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 structure of our problem.*

Precisely, we use L to denote the "accessibility Lie algebra" of our system, that is, the Lie algebra of vector fields generated by F and G. Our first task will then be to compute L.

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

$$[G,[F,G]] = F\,, \tag{6.5}$$

$$[F,[F,[F,G]]] = [G,[F,[F,G]]] = 0 \tag{6.6}$$

hold. It then follows easily that 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. [**4**], p. 59). This Lie algebra has a basis H,P,Q,E, with the commutation relations

$$\begin{aligned} [H,P] &= Q\,,\\ [H,Q] &= -P\,,\\ [P,Q] &= E\,,\\ [H,E] &= [P,E]=[Q,E]=0\,. \end{aligned}$$

(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 on smooth functions on the real line—can be defined by "quantizing the harmonic oscillator," i.e., by taking

$$\begin{aligned} P &= \frac{d}{dx}\,, \\ Q &= \text{multiplication by } x\,, \\ E &= \text{identity}\,, \\ H &= \frac{1}{2}(-P^2+Q^2) = \frac{1}{2}\Big(-\frac{d^2}{dx^2}+x^2\Big)\,, \end{aligned}$$

so that H is the Hermite operator.

6.6. Energy conservation. We now compute the *accessibility distribution* Λ generated by L. Formally, $q \mapsto \Lambda(q)$ is the map that assigns to each state q the linear space

$$\Lambda(q) = \{X(q) : X \in L\}\,.$$

To find Λ we first observe that, even though the vector fields F, G, $[F,G]$ and $[F,[F,G]]$ are linearly independent over $\mathbb{R}$, a simple computation shows that

$$G = \zeta_1 F + \zeta_2 [F,G] + \zeta_3 [F,[F,G]]\,, \tag{6.7}$$

where the ζ_i are smooth functions of q, given by

$$\zeta_1 = -\frac{v}{g}\,,\quad \zeta_2 = -\frac{w}{g}\,,\quad \zeta_3 = -\frac{v^2+w^2}{2g^2}\,. \tag{6.8}$$

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 Λ *is 3-dimensional at each point.* 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 (6.4). This integral of motion can be easily computed and turns out to be the energy E, given by

$$E(x,y,v,w) = \frac{v^2+w^2}{2} + gy\,. \tag{6.9}$$

Notice that E is a smooth function on $\mathbb{R}^4$, with nowhere vanishing gradient. Therefore the level hypersurfaces of E—i.e., the sets

$$S_{\bar{E}} \stackrel{\text{def}}{=} \{q : E(q) = \bar{E}\}\,,$$

for $\bar{E} \in \mathbb{R}$—are smooth 3-dimensional manifolds. (The set $S_{\bar{E}}$ is given by the equation

$$y = \frac{2\bar{E} - v^2 - w^2}{2g}\,,$$

which shows that each $S_{\bar{E}}$ is the graph of a smooth function $(x,v,w) \mapsto y$.)

6.7. 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.)

REMARK 6.7.1. The reason for **(#)** is as follows. The necessary condition given by the Maximum Principle for a trajectory-control pair (ξ, η) to be optimal is that (ξ, η) be either a "normal extremal" or an "abnormal extremal." Moreover, being an abnormal extremal is a necessary condition for the pair (ξ, η) to be such that the terminal point of ξ does not belong to the interior of the reachable set from the initial point. If a nontrivial constraint such as $E =$ constant is satisfied, then the reachable set from any initial point q is contained in a submanifold of the state space of positive codimension. Hence the reachable set has empty interior. Therefore all trajectory-control pairs are abnormal extremals. ♢

6.8. Controllability. We now investigate the controllability properties of our system. First of all, it is clear that two points q_i, q_t cannot be joined by a trajectory of (6.3) unless they belong to the same leaf S_E. We now show that the converse is true, that is, *any two points q_i, q_t belonging to the same leaf of (6.3) can be joined by a trajectory.* In other words, *the restriction of (6.3) to each leaf is completely controllable.*

To prove our statement, we use the technique of "Lie saturates" introduced by Jurdjevic and Kupka. Fix E and let $(6.3.E)$ denote the restriction of (6.3) to S_E. By construction, the system $(6.3.E)$ has the accessibility property. So to prove complete controllability it suffices to establish complete controllability of the Lie saturate family of vector fields $\mathcal{L}$. It is clear that F, G and $-G$ belong to $\mathcal{L}$. So the desired conclusion will follow if we show that $-F \in \mathcal{L}$. But this follows easily, because the identity $[G, [G, F]] = -F$ implies

$$e^{\pi G} F e^{-\pi G} = -F\,.$$

6.9. The existence problem. A precise definition of Hamilton's principal function V was given in Remark 6.3.1. It is clear that, if q_i and q_t belong to $\mathbb{R}^4$, then $V(q_i, q_t) < +\infty$ if and only if q_i and q_t belong to the same leaf S_E.

It is then reasonable to ask whether an optimal control steering q_i to q_t exists whenever $V(q_i, q_t) < +\infty$. Equivalently, we want to know if the infimum that occurs in the definition of V is in fact a minimum whenever it is finite.

The usual way to prove existence of an optimum is to pick a sequence $\{\eta_j\}_{j\in\mathbb{N}}$ of controls steering q_i to q_t with costs c_j that converge to the infimum, look at the corresponding trajectories ξ_j, and then try to produce a subsequence $\{\eta_{j(\ell)}\}_{\ell\in\mathbb{N}}$ of $\{\eta_j\}_{j\in\mathbb{N}}$ that converges in some sense to a limiting control η_∞ and is such that the $\xi_{j(\ell)}$ converge to a trajectory ξ_∞ of η_∞ and the costs $c_{j(\ell)}$ converge to the cost c_∞ of $(\xi_\infty, \eta_\infty)$.

This method may fail for various reasons. The simplest one is that the ξ_j may "go off to infinity." That is, there may not exist a subsequence $\{\xi_{j(\ell)}\}_{\ell\in\mathbb{N}}$ that stays in a fixed compact set. In the following definition, we give a name to the property that this particular obstruction to the existence of optimal controls does not occur.

DEFINITION 6.9.1. We say that an optimal control problem with dynamics (6.2) and Lagrangian cost functional with Lagrangian L satisfies the *properness condition* if for every pair q_i, q_t of states such that $V(q_i, q_t) < +\infty$ there exist a number $c \in \mathbb{R}$ and a compact subset K of Q such that

(1) $V(q_i, q_t) < c$,
(2) whenever η is a control steering q_i to q_t with cost $< c$, and ξ is the corresponding trajectory, it follows that ξ is entirely contained on K. $\diamondsuit$

When a problem satisfies the properness condition, it still does not follow that optimal controls exist, because the minimizing sequences $\{\eta_j\}_{j\in\mathbb{N}}$ need not have subsequences that converge in any reasonable sense. On the other hand, it is commonly agreed that in this case optimal controls really exist if one enlarges the original class $\mathcal{U}$ of admissible controls by adding to it some "generalized controls." We will show that this is indeed what happens for our brachistochrone problem, and that in this case the generalized controls can be easily described in concrete terms, and turn out to be none other than "impulse controls."

To begin with, we prove the following.

PROPOSITION 6.9.2. *The minimum time problem for the brachistochrone system (6.3) satisfies the properness condition.*

PROOF. Let q_i, q_t be two points of $\mathbb{R}^4$ having the same energy E. Fix an arbitrary $c \in \mathbb{R}$ such that $c > V(q_i, q_t)$. We have to find a constant A such that $\|\xi(t)\| \le A$ for all $t \in [0, \tau]$ whenever $\tau < c$ and $\xi : [0, \tau] \mapsto \mathbb{R}^4$ is a trajectory of (6.3) for some control $t \mapsto \eta(t)$ such that $\xi(0) = q_i$ and $\xi(\tau) = q_t$.

Write $\xi(t) = [x(t), y(t), v(t), w(t)]^\dagger$. Let

$$\theta(t) = [v(t), w(t)]^\dagger\,, \quad \zeta(t) = [v(t), w(t)]^\dagger\,.$$

Then (6.3) implies that

$$\frac{d}{dt}\,\|\zeta(t)\|^2 = -2g\cdot w(t)\,.$$

Therefore

$$\left|\frac{d}{dt}\,\|\zeta(t)\|^2\right| = 2g|w(t)| \le g^2 + w(t)^2 \le g^2 + \|\zeta(t)\|^2\,,$$

and then, if we write

$$\psi(t) = \Big|\,\|\zeta(t)\|^2 - \|\zeta(0)\|^2\Big|\,,$$

we find that

$$\psi(t) \le \int_0^t (g^2 + \|\zeta(s)\|^2)\,ds \le c(g^2 + \|\zeta(0)\|^2) + \int_0^t \psi(s)\,ds\,.$$

Then Gronwall's inequality implies that

$$\psi(t) \le c(g^2 + \|\zeta(0)\|^2)e^t \le c(g^2 + \|\zeta(0)\|^2)e^c\,.$$

It then follows that

$$\|\zeta(t)\|^2 \le \|\zeta(0)\|^2 + c(g^2 + \|\zeta(0)\|^2)e^c\,,$$

which gives an *a priori* pointwise bound for $\zeta(t)$ in terms of q_i and c.

To conclude the proof, we have to find an *a priori* pointwise bound for $\theta(t)$ in terms of q_i and c. But this is trivial, since $\theta(t) = \theta(0) + \int_0^t \zeta(s)\,ds$. $\diamondsuit$ $\square$

6.10. 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. [**37**]) can be applied to the problem restricted to $S_{\bar{E}}$. The conclusion turns out to be exactly the same as that for the unrestricted problem, except that the nontriviality condition for the momentum is slightly stronger. This extra strength is actually fundamental for, as explained in §6.7, without it the conclusion would be completely uninformative.

If $[a, b] \ni t \mapsto \xi_*(t) \in \mathbb{R}^4$ is an optimal trajectory for our problem, corresponding to a control $[a, b] \ni t \mapsto \eta_*(t) \in \mathbb{R}$, then the usual version of the Maximum Principle yields the existence of a nontrivial momentum vector field

$$[a, b] \ni t \mapsto \pi(t) \in \mathbb{R}^4$$

(often called an "adjoint vector") that satisfies the adjoint equation, the Hamiltonian maximization condition, and the additional condition that the maximized value of the Hamiltonian is a nonnegative constant[15]. The principle applied to the restriction to the leaves yields the stronger conclusion that $\pi(t)$ cannot be orthogonal to the leaf $S_{\bar{E}}$ that contains ξ_*. (In differential-geometric terms, $\pi(t)$ is really a *covector* on $S_{\bar{E}}$ at $\xi_*(t)$, and has to be nontrivial as such. Equivalently, if one insists on regarding $\pi(t)$ as a vector in $\mathbb{R}^4$, then it should not be orthogonal to the tangent space to $S_{\bar{E}}$ at $\xi_*(t)$.) (Naturally, one could also avoid using the Maximum Principle on manifolds by invoking instead the fact that $S_{\bar{E}}$ is identified with $\mathbb{R}^3$ in an obvious way, since it is the graph of a smooth function $(x, v, w) \mapsto y(x, v, w)$. 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 the function

$$\mathbb{R}^4 \times \mathbb{R} \times \mathbb{R}^4 \ni (q, u, p) \mapsto \mathbb{R}$$

defined by

$$H(q, u, p) = \varphi(p, q) + u\,\psi(q, p)\,, \tag{6.10}$$

where the functions φ, ψ are defined by

$$\begin{aligned} \varphi(p, q) &= \langle p, F(q)\rangle\,, \\ \psi(q, p) &= \langle p, G(q)\rangle\,. \end{aligned}$$

Since u is completely unrestricted, H cannot have a maximum as a function of u unless $\psi(q, p) = 0$. So

$$\psi(\xi_*(t), \pi(t)) = 0 \quad \text{for all} \quad t\,. \tag{6.11}$$

Differentiation of (6.12) yields

$$\rho(\xi_*(t), \pi(t)) = 0 \quad \text{for all} \quad t\,, \tag{6.12}$$

where

$$\rho(q, p) \stackrel{\text{def}}{=} \langle p, [F, G](q)\rangle\,.$$

---

[15]The Maximum Principle is usually stated as yielding a pair (π, π_0) consisting of a momentum field π and a nonnegative "abnormal multiplier" π_0. The value of the maximized Hamiltonian is supposed to vanish, and the nontriviality condition says that $(\pi(t), \pi_0) \neq (0, 0)$ for all t. These conditions are in terms of the Hamiltonian $\mathcal{H}$ given by $\mathcal{H} = H - \pi_0$, where H is the Hamiltonian defined by (6.10). It then follows that $\pi(t) \neq 0$— since $\pi(t) = 0$ would imply $\pi_0 = 0$, contradicting the nontriviality condition— and one recovers the form of the necessary conditions stated in the text.

A further differentiation—together with the fact that $[G,[F,G]] = F$—yields

$$\theta(\xi_*(t),\pi(t)) + u(t)\varphi(\xi_*(t),\pi(t)) = 0\,, \tag{6.13}$$

where

$$\theta(q,p) \stackrel{\text{def}}{=} \langle p, [F,[F,G]](q)\rangle\,.$$

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 state space feedback law $u = u(q)$, as we now show.

First of all, (6.7) implies that

$$\psi = \zeta_1\varphi + \zeta_2\rho + \zeta_3\theta \tag{6.14}$$

along our trajectory. Next, we use the crucial fact that

$$\varphi(\pi(t),\xi_*(t)) \neq 0\,, \tag{6.15}$$

which follows from the stronger version of the Maximum Principle. (The precise argument is as follows. The covector $\pi(t)$ must be nontrivial as a linear functional on the tangent space at $\xi_*(t)$ of the leaf containing $\xi_*(t)$. Since this space is spanned by the three vectors $F(\xi_*(t))$, $[F,G](\xi_*(t))$ and $[F,[F,G]](\xi_*(t))$, one of the numbers $\varphi(\pi(t),\xi_*(t))$, $\rho(\pi(t),\xi_*(t))$, $\theta(\pi(t),\xi_*(t))$ must be nonzero. We know that $\rho(\pi(t),\xi_*(t)) = 0$. If $\varphi(\pi(t),\xi_*(t)) = 0$, then (6.13) would imply that $\theta(\pi(t),\xi_*(t)) = 0$, and we would get a contradiction. Hence (6.15) holds.)

Therefore (6.13) implies

$$u(t) = -\frac{\theta(\xi_*(t),\pi(t))}{\varphi(\xi_*(t),\pi(t))}\,, \tag{6.16}$$

that is

$$u(t) = -\frac{\langle\pi(t),[F,[F,G]](\xi_*(t))\rangle}{\langle\pi(t),F(\xi_*(t))\rangle}\,. \tag{6.17}$$

Moreover, it follows from (6.14), together with the fact that $\psi = \rho = 0$ along our trajectory, that

$$\zeta_1(\xi_*(t))\varphi(\pi(t),\xi_*(t)) + \zeta_3(\xi_*(t))\theta(\pi(t),\xi_*(t)) = 0\,. \tag{6.18}$$

Then (6.16) implies

$$u(t) = \frac{\zeta_1(\xi_*(t))}{\zeta_3(\xi_*(t))}\,, \tag{6.19}$$

as long as $\zeta_3(\xi_*(t)) \neq 0$. Therefore

$$u(t) = \omega(\xi_*(t))\,, \tag{6.20}$$

where

$$\omega(q) = \omega(x,y,v,w) \stackrel{\text{def}}{=} \frac{2vg}{v^2+w^2}\,. \tag{6.21}$$

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

$$\Sigma = \{(x,y,v,w) : v^2+w^2 = 0\}\,. \tag{6.22}$$

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

$$Z(q) = F(q) + \omega(q)G(q)\,, \tag{6.23}$$

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 resulting trajectories are easily computed. First, observe that, if we use the feedback $u = \omega(q)$, the dynamic equations become

$$\left\{\begin{array}{rcl} \dot{x} & = & v\,, \\ \dot{y} & = & w\,, \\ \dot{v} & = & -2g(v^2+w^2)^{-1}vw\,, \\ \dot{w} & = & 2g(v^2+w^2)^{-1}v^2 - g \\ & = & (v^2+w^2)^{-1}(2gv^2 - g(v^2+w^2)) \\ & = & g(v^2+w^2)^{-1}(v^2-w^2)\,. \end{array}\right. \tag{6.24}$$

If we introduce a complex variable z defined by $z = v + iw$, then the last two equations of (6.24) say that

$$\dot{z} = \frac{giz^2}{|z|^2}\,, \quad \text{that is,} \quad \dot{z} = \frac{giz}{\bar{z}}\,. \tag{6.25}$$

The formulas of (6.25) define a well known dynamical system of the complex plane $\mathbb{C}$. The right-hand side of the equations is a smooth vector field on $\mathbb{C}\backslash\{0\}$. The integral curves are circles passing through the origin and having center on the real axis (plus the imaginary axis, parametrized by $\mathbb{R} \ni t \to (0, -gt)$, which is also a solution since along it (6.25) just says that $\dot{z} = -gi$). The parametric equation of the circles is

$$z(t) = r(1 + \sigma\, e^{\frac{igt}{r}})\,, \tag{6.26}$$

where $\sigma \in \{+1, -1\}$ and r is a positive real number.

Once it has been shown that the velocity vector (v, w) evolves according to (6.26), a trivial integration shows that the trajectories described by the position vector (x, y) are cycloids.

6.11. 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.

Let q_i, q_t be two points of $\mathbb{R}^4$. It is clear that $V(q_i, q_t) = +\infty$ unless both q_i and q_t belong to 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 the points q_i and q_t belong to the same level hypersurface $S_{\bar{E}}$, then we know that they can be joined by a trajectory, so $V(q_i, q_t)$ will be finite. Moreover, it is reasonable to expect that such a curve will exist, and if it does not then we ought to be able to explain why.

Write $q_i = (A_i, a_i)$, $q_t = (A_t, a_t)$, where A_i, a_i, A_t, a_t, belong to $\mathbb{R}^2$. The classical analysis carried out earlier seems to suggest that, once A_i, A_t and $\bar{E}$ are specified, there will be only one optimal arc joining them. So, if we specify A_i and A_t 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 accommodate all possible pairs (a_i, a_t) (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_i, θ_t, of the velocity vectors a_i and a_t in a completely arbitrary way.* (Their magnitudes are determined by the fact that the energy at q_i and q_t must equal $\bar{E}$.) But *given A_i, A_t and $\bar{E}$ there is a unique cycloid from A_i to A_t corresponding to the energy $\bar{E}$*, and in particular the directions $\bar{\theta}_i$, $\bar{\theta}_t$ at A_i, A_t of this cycloid are completely determined by A_i, A_t and $\bar{E}$.

In other words, *we ought to be able to choose freely the two angles θ_i, θ_t, 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_i, the set $C(A_i,\bar{E})$ of those a's such that $(A_i,a)\in S_{\bar{E}}$ is either empty or a circle, except for the "borderline" case when $(A_i,0)\in S_{\bar{E}}$ and $C(A_i,\bar{E})$ consists of a single point. If $C(A_i,\bar{E})$ is a circle, then through each point of $\{A_i\}\times C(A_i,\bar{E})$ 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 $\{A_t\}\times C(A_t,\bar{E})$ at just one point. This means that there is a unique pair $(\bar{a}_i,\bar{a}_t)$ such that $(A_i,\bar{a}_i)$ and $(A_t,\bar{a}_t)$ lie on the same integral curve of Z. The portion of that curve joining $(A_i,\bar{a}_i)$ to $(A_t,\bar{a}_t)$ is, of course, our famous cycloid.

What is then the optimal trajectory joining (A_i,a_i) to (A_t,a_t) for general a_i, a_t? What is the corresponding value of $V(A_i,a_i,A_t,a_t)$? The answer is quite simple: the "optimal trajectory" is obtained by

1. applying an *impulse control* to rotate a_i to $\bar{a}_i$ instantaneously,
2. then following the cycloid from $(A_i,\bar{a}_i)$ to $(A_t,\bar{a}_t)$,
3. finally applying another impulse control to rotate $\bar{a}_t$ to a_t.

(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 $V(A_i,a_i,A_t,a_t)$ is therefore equal to $V(A_i,\bar{a}_i,A_t,\bar{a}_t)$.

7. Five modern variations on the theme of the brachistochrone

7.1. Hestenes' "variation of the brachistochrone problem". In 1956, M. Hestenes wrote (in [**18**], 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*
>
> $$x=x(t)\qquad y=y(t)\qquad \varphi=\varphi(t)\qquad 0\le t\le T$$
>
> *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

$$\mathbf{u}(t) = \frac{\omega(t)}{\|\omega(t)\|}\,. \tag{7.1}$$

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 (7.1) 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.

7.2. The reflected brachistochrone problem. In §3, 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 (5.1)). For a recent study of the reflected brachistochrone problem from the dynamic programming perspective, see Malisoff [**28, 29**].

Solving this more general problem amounts to finding the light rays when the medium is *the whole plane*, and the speed of light is $\sqrt{|y|}$. In other words, *we want to find the minimum time paths for the control system (5.1) with control set U given by (5.2) and state space equal to the whole plane.*

Notice that this problem is "completely controllable," in the sense that

(C) *any two points A, B of $\mathbb{R}^2$ can be joined by a feasible path, even if they lie on opposite sides of the x axis.*

To prove (C), observe first that the desired conclusion is trivial if A and B both lie in the open upper half plane or in the open lower half plane. So it clearly suffices to prove the conclusion if A lies in the open upper half plane and B belongs to the x axis. Moreover, we can assume that A lies on the same vertical line as B, for otherwise we can pick A' on the same vertical as B but strictly above B, and join A to A' first in finite time. So let us suppose that $A = (a, \alpha)$ and $B(a, 0)$. Then the curve $[0, \alpha] \ni s \mapsto (a, \alpha - s)$ certainly goes from A to B. Along this curve, $dy = -\sqrt{y}dt$, so

$$dt = -\frac{dy}{\sqrt{y}} = \frac{ds}{\sqrt{\alpha - s}}\,.$$

Therefore the total time T required to move from A to B along this curve is given by

$$T = \int_0^{\alpha} \frac{ds}{\sqrt{\alpha - s}} = 2\sqrt{\alpha}\,,$$

so $T < +\infty$, and (C) is proved.

The crucial point here is that *the right-hand side of (5.1) vanishes along the x axis, but this does not prevent the existence of feasible paths crossing the x-axis in finite time, because the "speed of light function"*

$$(x, y) \mapsto c(x, y) = \sqrt{|y|}$$

is not Lipschitz near the x axis.

Remark 7.2.1. If the function c was Lipschitz-continuous, then the well known uniqueness theorem for ordinary differential equations would imply that every solution passing through a point on the x axis is a constant curve. This would obviously render it impossible for feasible curves starting at a point not on the x axis to reach the x axis in finite time. An important example of this situation is the case when the speed o light is $|y|$ rather than $\sqrt{|y|}$. That case was also studied by Johann Bernoulli, as is clearly seen in the text we quoted at the end of §2.2, where he says that *if, for example, the velocities were as the altitudes, then both curves* [HJS & JCW: that is, the brachistochrone and the tautochrone] *would be algebraic, the one a circle, the other one a straight line.* It turns out that the "brachistochrones" for the case when the speed of light is $|y|$ are *half circles whose center lies on the x axis.* Moreover, *these circles take infinite time to reach the x axis*, precisely because the function $y \mapsto |y|$ is Lipschitz. $\diamondsuit$

We have seen that the complete controllability of (5.2) depends essentially on the non-Lipschitzian character of its speed of light function. However, *this same non-Lipschitzian character also renders the Maximum Principle inapplicable, in its classical and nonsmooth versions (including the so-called "Lojasiewicz version," cf. Sussmann* [**35**]*), since all these results require a Lipschitz reference vector field.*

Can we still use necessary conditions to find the solution, perhaps by invoking a more general version of the Maximum Principle that would be applicable to this nonsmooth, non-Lipschitzian case?

Suppose, for example, that we want to find an optimal trajectory ξ going from A to B, where A lies in the upper half-plane and B is in the lower half-plane. Then one can easily show, first of all, that an optimal trajectory ξ exists, using the Ascoli-Arzelà theorem as before (cf. §5.2). 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 (3.5), 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 is a very straightforward consequence of Lemma 5.2.1.)

Therefore ξ must consist 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.

Naturally, X is the only point on ξ in a neighborhood of which the usual smooth and nonsmooth versions of the Maximum Principle are inapplicable. On the other hand, it is easy to produce a nonrigorous formal argument yielding the missing condition, as we now show.

Let us apply the Maximum Principle formally, even though the technical conditions required for the validity of the usual versions are not verified. The maximized value of the Hamiltonian is clearly equal to $\|\vec{p}\|\sqrt{|y|} - p_0$, and we know that this number is in fact equal to zero. So

$$\|\vec{p}(t)\|\sqrt{|y|} = p_0 \quad \text{for all} \quad t\,. \tag{7.2}$$

If we multiply (5.6) by $|y(t)|$, and use (7.2), we find

$$|y(t)|\Big(1 + y'(x)^2\Big) = \frac{|y(t)| \cdot \|\vec{p}\|^2}{p^2} = \frac{p_0^2}{p^2}\,. \tag{7.3}$$

The adjoint system (5.4) tells us that p is constant on each time interval that does not contain the point τ such that $\xi(\tau) = X$. If we make the *Ansatz* that p is actually constant on the whole time interval, i.e., that the constant values of p on both sides of τ are equal, we find that *the cycloids Γ_1, Γ_2 must be such that the* *quantity*

$$C = |y(x)|(1 + y'(x)^2)\,. \tag{7.4}$$

which is constant along each Γ_i, *actually has the same value for both curves.* This extra condition determines X.

The above argument is, of course, not rigorous. It turns out, however, that *the recent "very nonsmooth" versions of the Maximum Principle developed by one of us—cf. Sussmann* [**39, 40, 41, 42, 43, 44**]*—apply to this problem and provide a rigorous justification for our formal argument.* The reason for this is that the new results do not require Lipschitz continuity—or even continuity—of the reference vector field. All that is needed is that the *flow maps* generated by the reference flow be differentiable in an appropriate sense[16].

Remarkably, our "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

(C) *The rolling circles that generate the upper and lower cycloids must have equall radii.*

7.3. 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

1. *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

2. *they would have found elegant the geometric condition of our solution of the reflected brachistochrone problem.*

This seems clear to us but, since *de gustibus non est disputandum*, we shall not pursue this particular argument further.

7.4. The brachistochrone with friction. In the problem studied in §6, we assumed that there was no friction. This implied that energy was conserved and that the four-dimensional state space was foliated by three-dimensional orbits of the system's accessibility distribution.

It is then natural to consider the more general case of the *brachistochrone with friction*, and to expect that for this problem there will be no nontrivial conserved functions of the state, so the accessibility distribution will be four-dimensional, and the system will have the accessibility property.

We now outline some of the results, without carrying out a complete analysis.

[16]For the specific case of the reflected brachistochrone, classical differentiability suffices. For more general problems, one needs other theories of "generalized differentiation," as explained in [**39, 40, 41, 42, 43, 44**]

The dynamical behavior of the system is given by the equations

$$\left\{\begin{array}{rcl} \dot{x} & = & v\,, \\ \dot{y} & = & w\,, \\ \dot{v} & = & -uw - \rho v\,, \\ \dot{w} & = & uv - \rho w - g\,, \end{array}\right. \tag{7.5}$$

where, as before, u is an unbounded scalar control. The parameter ρ is a real number, the friction coefficient. The problem with friction corresponds to $\rho > 0$. When $\rho = 0$ we get the frictionless case considered earlier.

If we write

$$q = \begin{bmatrix} x \\ y \\ v \\ w \end{bmatrix}, \qquad F(q) = \begin{bmatrix} v \\ w \\ -\rho v \\ -\rho w - g \end{bmatrix}, \qquad G(q) = \begin{bmatrix} 0 \\ 0 \\ -w \\ v \end{bmatrix}.$$

then (7.5) takes the form

$$\dot{q} = F(q) + uG(q)\,. \tag{7.6}$$

From now on, we assume that $\rho \neq 0$, since the case when $\rho = 0$ has already been discussed. Let L be the Lie algebra of vector fields generated by F and G. In order to analyze the structure of L, it will be convenient to introduce five new vector fields Z, V, W, X, and Y. The following eight formulas express F, G, and $[G, F]$ in terms of the usual differential operator notation for vector fields, and define Z, V, W, X, and Y:

$$\begin{array}{rcrcrcrcr} F & = & v\partial_x & + & w\partial_y & - & \rho v\partial_v & - & (\rho w + g)\partial_w\,, \\ G & = & & & & - & w\partial_v & + & v\partial_w\,, \\ [G, F] & = & -w\partial_x & + & v\partial_y & - & g\partial_v & & , \\ Z & = & & & & & v\partial_v & + & w\partial_w\,, \\ V & = & & & & & \partial_v & & , \\ W & = & & & & & & & \partial_w\,, \\ X & = & \partial_x & & & & & & , \\ Y & = & & & \partial_y & & & & . \end{array}$$

Let S be the linear span of the eight vector fields F, G, $[G, F]$, Z, V, W, X, and Y. It then turns out that $S \subseteq L$, because

$$\begin{array}{rcl} Z & = & \sigma(F + [G, [G, F]])\,, \\ W & = & \gamma([F, Z] + F + \rho Z)\,, \\ V & = & [G, W]\,, \\ X & = & \rho V - [F, V]\,, \\ Y & = & \rho W - [F, W]\,, \end{array}$$

where

$$\sigma = \frac{1}{\rho} \text{ and } \gamma = \frac{1}{2g}\,.$$

On the other hand, the formulas

$$\begin{aligned}
[F,[G,F]] &= -\rho[G,F] + 2gX - 2\rho gV\,,\\
[G,[G,F]] &= -F - \rho Z\,,\\
[F,Z] &= -F - \rho Z - 2gW\,,\\
[F,V] &= -X + \rho V\,,\\
[F,W] &= -Y + \rho W\,,\\
[F,X] &= [F,Y] = 0\,,\\
[G,Z] &= [G,X] = [G,Y] = 0\,,\\
[G,V] &= -W\,,\\
[G,W] &= V\,,
\end{aligned}$$

show that $[F,S] \subseteq S$ and $[G,S] \subseteq S$. So $L = S$.

A straightforward calculation shows that *F, G, $[G,F]$, Z, V, W, X, and Y are linearly independent over $\mathbb{R}$.* So *L is an 8-dimensional Lie algebra.*

Let Λ be the accessibility distribution of our system, so $\Lambda(q) \stackrel{\text{def}}{=} \{h(q) : h \in L\}$ if $q \in \mathbb{R}^4$. Since the vector fields X, Y, V and W belong to L, it is clear that $\Lambda(q) = \mathbb{R}^4$ for every q. So *our system has the accessibility property.*

An analysis similar to the one carried out in the previous section for the frictionless problem yields an optimal "feedback control"

$$u = \frac{2gv}{v^2+w^2} + 2\rho\frac{\zeta v^2 - vw}{v^2+w^2}\,, \tag{7.7}$$

involving an additional parameter ζ, given by

$$\zeta = \frac{\pi_Y}{\pi_X}\,.$$

(Here π_X, π_Y are the X and Y components of the momentum, which are easily seen to be constant along extremals.)

In other words, *through every point $q \in \mathbb{R}^4$ there passes a one-parameter family of optimal trajectories, depending on the parameter ζ.*

7.5. The classical brachistochrone as a degenerate problem. The family of brachistochrone problems with friction coefficient ρ constitutes an "unfolding" of the classical frictionless problem. For $\rho \neq 0$, the situation is generic, in the sense that the accessibility distribution is four-dimensional.

The generic behavior of the time-optimal trajectories for such problems is precisely the one described here: the control is given in "feedback" form, depending on the state and an extra parameter. Formula (7.7) shows that the extra parameter only occurs as part of a term that is multiplied by ρ. It follows that when $\rho = 0$ the extra parameter disappears, and we get a state space feedback as determined earlier.

7.6. Deriving Snell's law from the Maximum Principle. Our third "modern variation" will involve going back to ancient times, long before 1696, and considering Fermat's minimum time principle and Snell's law of refraction.

We have already explained how Snell's law is intimately related to the brachistochrone problem, since it played a crucial role in Johann Bernoulli's solution. So, if we pursue Hestenes' line of inquiry further, and speculate on other questions that the Bernoulli brothers might have asked "had they lived in modern times," it is easy to imagine them asking the following.

(Q) *Given that in these modern times powerful general theories have been developed, going far beyond the particular examples that were dealt with by ad hoc methods up to the 17th and early 18th centuries, do these general theories actually cover* **all** *these special examples whose study preceded and begat them?*

As an important special case of this question, the brothers would certainly want to know whether the necessary conditions of the calculus of variations and optimal control theory apply, in particular, to the derivation of Snell's law.

So let us consider a medium such as the one we described when discussing Snell's 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's law be derived from the Maximum Principle.

It is easy to see that there is no way to formulate this question 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 the usual versions of the Maximum Principle, it is not required that f be differentiable, or even continuous, with respect to the control u. But here the situation is different, since we need to be able to deal with an f which is *discontinuous with respect to the state variable* q.

Once again, it turns out that *the recent versions the Maximum Principle developed in Sussmann* [**39, 40, 41, 42, 43, 44**] *apply to this problem as well, and lead to Snell's law.* The reason is that in these version all that is needed is that the reference trajectory ξ_* arise from a flow $\{\Phi_{t',t}\}$ consisting of differentiable maps, and there is no need for this flow to be generated by a continuous vector field.

7.7. Was Johann Bernoulli close to discovering non-Euclidean geometry? Given a "speed of light function" $(x, y) \mapsto c(x, y)$ on some plane region, the time along a curve is given by

$$dt = \frac{\sqrt{dx^2 + dy^2}}{c(x, y)},$$

since $\sqrt{\frac{dx}{dt}^2 + \frac{dy}{dt}^2} = c(x, y)$. This means that, if we adopt a slightly different point of view, and think of time as "length," then the function c defines a *Riemannian metric* on the set $\Omega_c = \{(x, y) : c(x, y) > 0\}$, provided that c is sufficiently smooth. Let us assume that $c(x, y) > 0$ whenever $y > 0$, as was the case in our two examples. Then c gives rise to a Riemannian metric

$$ds^2 = \frac{dx^2 + dy^2}{c(x, y)^2}$$

on the upper half plane. The "brachistochrones" are then the *minimum length* curves, that is, the *geodesics*. In the case of the "true brachistochrone," when $c(x, y) = \sqrt{|y|}$, these geodesics are cycloids (plus vertical segments) but *maximal* *geodesics have finite length.* On the other hand, in the case when $c(x, y) = |y|$, so that the Riemannian metric on the upper half plane is given by

$$ds^2 = y^{-2}(dx^2 + dy^2)\,, \tag{7.8}$$

the geodesics are, as we have seen, *half circles of infinite length.*

So, using modern terminology, the circles that occur as solutions of this modified version of Johann Bernoulli's brachistochrone problem are the geodesics of a Riemannian metric on the upper half plane given by (7.8). We now know that this Riemannian metric defines the so called *Poincaré half plane*, which is the simplest model of non-Euclidean geometry.

This raises an obvious question: *how close did Johann Bernoulli get to solving the famous problem of the provability of Euclid's fifth postulate?* What renders the question particularly interesting is that

1. 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 [**32**], entitled *Euclides ab omni naevo vindicatus* (that is, "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."
2. The cycloids that occur in the original brachistochrone problem have many of the incidence properties of Euclidean lines, 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.
3. 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.
4. Johann Bernoulli did study these half circles, as we pointed out in Remark 7.2.1. So in a sense he did discover the Poincaré model.
5. Johann Bernoulli was definitely interested in the question, since he made significant contributions to the early stages of the development of the idea of geodesics on curved surfaces, by studying the curves of minimum length on a surface, which later[17] came to be known as "geodesics."
6. The solution of the problem of the independence of Euclid's postulate is extremely simple by today's standards, even though mathematicians took about 2200 years to find it. Actually, the solution is so simple that it can be explained to any undergraduate with a moderate understanding of plane analytic geometry. All that is required is to take any reasonable way to compute "time" along curves, rename "time" by calling it "length," and then call the length-minimizing curves (that is, the "brachistochrones") "lines." For example, the case when $c = |y|$ leads us to taking the half circles with center on the x axis (plus the vertical half lines) and calling them "lines." Once this trivial renaming has been carried out, it is easy to verify that all of Euclid's postulates other than the fifth hold, but the fifth postulate fails.

[17]The word "geodesic" was coined much later by Liouville (1809-1892).

So *Johann Bernoulli had essentially all the tools he needed to solve the problem and prove that Saccheri was wrong, by showing that Euclid's fifth postulate does not follows from the other ones. Why didn't he do it?*

One can surmise that he lacked the understanding of the *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.

In modern terms, this is a trivial remark, namely, that one can turn around the well known fact that the segment is the shortest path between two points, and make it the *definition of the notion of segment*, while at the same time using a more general prescription to measure "length." But the idea that minimization problems can be the source of new geometries evolved very slowly over two millennia, and took a long time to crystallize, reaching final form in the nineteenth century, in the era of Gauss and Riemann, when it revolutionized mathematics and then moved on, at the beginning of the twentieth century, to bring a a fundamental change in our understanding of the physical universe.

One can already find the idea in a rudimentary form in Fermat's least time principle[18], which can be interpreted as saying that light rays follow the shortest paths, with the proviso that "length" must now be defined in terms of a new and more refined geometry, in which the physical properties of space influence the geometry, and shortest paths can be curved or bent. This interpretation is suggested, for example, by G. Lochak, in [**27**], 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.

In fact, Leibniz definitely uses the word "shortest" in his analysis of the refraction problem in his 1684 *Nova methodus* paper, showing that he was at least entertaining the possibility that light rays in a medium with variable speed of light truly were "shortest paths" in some new geometry.

But it would appear that, in spite of these earlier glimpses of what would become one of the greatest scientific ideas of all times, the era of Johann Bernoulli was not yet ready to take the trivial step of turning around the lenght minimization property of segments.

Johann Bernoulli passionately wanted to succeed and be famous, and was obsessed by the desire to surpass his contemporaries, especially his brother Jacob and his son Daniel[19]. It is thus sadly ironic that he, of all people, got so close to solving an important 2000-year old problem but missed because he failed to take a small step, a step so simple that today it can easily be explained to beginners.

---

[18]which, as we have seen, was a direct ancestor of Johann Bernoulli's solution of the brachistochrone problem

[19]Johann was very jealous of Daniel's success. He once threw Daniel out of the house for having won a French Academy of Sciences prize for which Johannn had also been a candidate (cf. Bell [**1**], p. 134). Probably with the help of his loyal disciple Euler (a student of Johann in Basel and a colleague of Daniel in Saint Petersburg), Johann plagiarized Daniel's path-breaking work on hydrodynamics (cf. Guillen [**17**]).

Had Johann Bernoulli "lived in modern times," he probably would have been torn between the curiosity to find out about non-Euclidean geometry and the desire not to hear about the tragedy of his missed opportunity to reach eternal fame and glory.

References

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

[2] Beltrami, E., *Saggio di interpretazione della geometria non-euclidea.* Giornale di Matematiche **VI**, 1868, pp. 284-312 (in Italian). Eng. transl. in Stillwell, [**33**], pp. 7–34.

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

[4] Berna, P., N. Conze, M. Duflo, M. Lévy-Nahas, M. Rais, P. Renouard, M. Vergne, *Représentations des groupes résolubles.* Dunod, Paris, 1972.

[5] 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.)

[6] 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.)

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

[8] Bonnard, B., V. Jurdjevic, I. A. Kupka and G. Sallet, "Transitivity of families of invariant vector fields on the semidirect products of Lie groups." *Trans. Amer. Math. Soc.* **271**, no. 2, 1982, pp. 525–535.

[9] Bonnard, B., V. Jurdjevic, I. A. Kupka and G. Sallet, "Systèmes de champs de vecteurs transitifs sur les groupes de Lie semi-simples et leurs espaces homogènes" (French). In *Systems Analysis Conf. (Bordeaux, 1978)*, *Astérisque*, Soc. Math. France, Paris, 1980, pp. 75-76.

[10] 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.

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

[12] Clarke, F. H., "The maximum principle under minimal hypotheses." *SIAM J. Control Optim.* **14**, 1976, pp. 1078–1091.

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

[14] 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.

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

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

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

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

[19] Isidori, A., *Nonlinear Control Systems.* Springer-Verlag, Berlin, 1995.

[20] Jakubczyk, B., and W. Respondek eds., *Geometry of Feedback and Optimal Control.* Marcel Dekker, New York, 1997.

[21] Jurdjevic, V., and I. A. K. Kupka, "Control systems on semisimple Lie groups and their homogeneous spaces." *Ann. Inst. Fourier (Grenoble)* **31**, no. 4, vi, 1981, pp. 151–179.

[22] Jurdjevic, V., and I. A. K. Kupka, "Control systems subordinated to a group action: accessibility." *J. Differential Equations* **39**, no. 2, 1981, pp. 186–211.

[23] Jurdjevic, V., and I. A. K. Kupka, "Linear systems with degenerate quadratic costs." *Compt. Rend. Math. Rep. Acad. Sci. Canada* **10**, no. 2, 1988, pp. 101–106.

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

[25] Kupka, I. A. K., "Degenerate linear systems with quadratic cost under finiteness assumptions." In *New trends in nonlinear control theory (Nantes, 1988)*, Lecture Notes in Control and Inform. Sci. **122**, Springer, Berlin, 1989, pp. 231–239.

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

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

[28] Malisoff, M., "On the Bellman equation for optimal control problems with exit times and unbounded cost functionals." In *Proc. 38th IEEE Conference on Decision and Control, Phoenix, AZ, Dec. 1999*, IEEE Publications, New York, 1999, pp. 23-28.

[29] Malisoff, M., "A remark on the Bellman equation for optimal control problems with exit times and noncoercing dynamics." In *Proc. 38th IEEE Conference on Decision and Control, Phoenix, AZ, Dec. 1999*, IEEE Publications, New York, 1999, pp. 877-881.

[30] Nijmeijer, H., and A. van der Schaft, *Nonlinear Dynamical Control Systems.* Springer-Verlag, New York, 1990.

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

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

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

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

[35] 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.

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

[37] 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.

[38] Sussmann, H. J., and J. C. Willems, "300 years of optimal control: from the brachystochrone to the maximum principle." *IEEE Control Systems Magazine*, June 1997, pp. 32-44.

[39] 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.

[40] 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.

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

[42] Sussmann, H. J., "Résultats récents sur les courbes optimales." In *Quelques aspects de la théorie du contrôle, Journée annuelle de la Société Mathématique de France*, Paris, June 2000, pp. 1-50.

[43] Sussmann, H. J., "A nonsmooth hybrid maximum principle." In *Stability and Stabilization of Nonlinear Systems (Proceedings of the First European Community Nonlinear Control Network Workshop on "Stability and Stabilization" held in Ghent, Belgium, on March 15-16, 1999)*, D. Aeyels, F. Lamnabhi-Lagarrigue and A. J. van der Schaft Eds. Lect. Notes Control and Information Sciences **246**, Springer-Verlag, 1999, pp. 325–354.

[44] Sussmann, H. J., "New theories of set-valued differentials and new versions of the maximum principle of optimal control theory." To appear in *Nonlinear Control in the Year 2000 (Proceedings of the European Control Network Workshop on Nonlinear Control, Paris, France, June 5-7, 2000)*, A. Isidori, F. Lamnabhi-Lagarrigue and W. Respondek Eds., Springer-Verlag, 2000, pp. 487-526.

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