Three Centuries of Curve Minimization: from the Brachistochrone to Modern Optimal Control Theory

Héctor J. Sussmann

Department of Mathematics - Rutgers University
Hill Center – Busch Campus 110 Frelinghuysen Road
Piscataway, NJ 08854-8019 – U.S.A.

sussmann@hilbert.rutgers.edu

Jan C. Willems

Department of Mathematics – University of Groningen
P. O. Box 800 – 9700 AV Groningen
The Netherlands

J.C.Willems@math.rug.nl

THIS IS A DRAFT.
DO NOT QUOTE WITHOUT PERMISSION.
Date of this draft: December 7, 2003.

AMS Subject classification: Primary 54C40, 14E20; Secondary 46E25, 20C20.

ACKNOWLEDGEMENTS

H. Sussmann's research was supported in part by NSF Grant DMS-9803411 and AFOSR Grant 0923.

A large part of his work was done in the Netherlands, while he was "Johann Bernoulli visiting professor" at the University of Groningen in May, June and July 1999. He is immensely grateful to the University for its generous hospitality and exciting intellectual atmosphere.

We are grateful to

- ***Andrei Agrachev*** for having pointed out to us in 1993 that the anniversary of Johann Bernoulli's challenge was be celebrated in three years, and that this was an excellent opportunity for reclaiming this event in writing as a significant step in the history of optimal control,

- ***Tom Archibald***, for his comments on the spurious solutions of the differential equation derived by Johann Bernoulli in for the brachistocrone,

- ***Franco Cardin***, for acquainting us witht he work of W. M. Tulczyjew,

- ***Francis H. Clarke***, for pointing out to us in September 1995 that the existence problem for the brachistochrone—which, as will be explained in §XIV.8, is fairly easy to handle in the optimal control framework—is a delicate question in the classcial calculus of variations, due to the noncoercivity and the singular denominator of the Lagrangian.

- ***Jean-Paul Gauthier*** and ***Roe Goodman***, for helping us understand the role of the diamond Lie algebra,

  ¡

- ***Hans W. Knobloch***, for bringing to our attention the passage in C. Carathéodory's book that we discuss in §V.9,

- ***Geert Jan Olsder***, for providing helpful information on Dutch discoveries in the South Pacific and for many other insightful comments,

- ***Annye K. Sussmann***, for her careful reading of the manuscript and her invaluable suggestions,

- ***M. Vidyasagar***, for bringing to our attention an inaccurate statement about Dr. Watson and Sherlock Holmes that appeared in an earlier draft.

Abstract

On the occasion of the tricentennial of the 1697 issue of *Acta Eruditorum*, containing work on the brachistochrone problem by Johann Bernoulli, his brother Jakob, Newton, Leibniz, l'Hôpital, and Tschirnhaus, we outline how the process that started 300 years ago has led to modern optimal control theory. We describe the circuitous path going through the Euler-Lagrange equations, the Legendre-second order condition, Hamilton's reformulation of the necessary conditions, and Weierstrass' side condition, to the modern maximum principle of optimal control. We emphasize the occasions when opportunities were missed, and the protagonists of our story—especially Hamilton and Weierstrass—came close to finding the modern general version of the necessary conditions, but failed to do so. We trace this failure to the fact that the Hamilton equations of motion were written in terms of a "Hamiltonian" which is a function of the positions and momenta only, a form which obscures the basic facts that are immediately revealed by using a different "Hamiltonian," which is a function of the velocities as well as the momenta and the positions. We then outline the recent evolution of the necessary conditions beyond the classical version of the maximum principle, and describe the connections with nonsmooth analysis and differential geometry. In the last part of the paper, we look at the brachistochrone problem and show how the various steps in the evolution of the necessary conditions for optimality provide new insights into the original problem. In particular, we show how the research program—sketched, for example, by J. L. Synge in 1944—to describe all properties of interest of a curve optimization problem in terms of "Hamilton's principal function," when applied to the brachistochrone problem, naturally leads to a "differential-geometric analysis," in which the problem is formulated in 4-space, and an interesting Lie-algebraic structure arises—involving a 4-dimensional Lie algebra (the "diamond Lie algebra") and a 3-dimensional distribution—as well as nontrivial analytic questions, due to the presence of unbounded controls, and resulting in solutions that contain impulse controls We conclude that the brachistochrone is a true optimal control problem, in the sense that only the optimal control formulation fully brings out the rich structure associated to what Leibniz called a "splendid problem."

COMMENT NO. 1. This abstract is a relic from the old days when we thought of this as a paper rather than a book. I suppose now that it's a book the abstract is no longer needed.

Contents

COMMENT NO. 2. We have to fix the subject classification footnote, by putting in the correct classification numbers. (The numbers that are there were put in for formatting purposes only. They come from a model style file we got, and have nothing to do with the topic of our book.)

Chapter I

Introduction

In the June 1696 issue of the journal *Acta Eruditorum*, Johann Bernoulli challenged mathematicians to solve the "problema novum" of finding the curve that—following Leibniz' suggestion—he named the "brachistochrone"[1]. The journal's May 1697 issue contains articles on this problem by six of the most renowned mathematicians of the time[2]. Leibniz, who was one of the six authors, 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[3] 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.

Today, it is clear that the events of 1696-97 represented a critical turning point in the history of mathematics, and it is widely[4] agreed that with them an important new field, later to be called the "calculus of variations," was born[5]. 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

---

[1]Cf. Chapter XIV below for a precise statement of Bernoulli's problem.

[2]Johann Bernoulli, Johann's elder brother Jakob, Leibniz, Newton, the marquis de l'Hôpital, and Tschirnhaus; cf. §XIV.1 in Page 296 below.

[3]We will speculate on these reasons in §XIV.2 (in Page 298) below.

[4]Though not unanimously, cf. *infra*, Chapter II, esp. Remark II.4.

[5]The words "calculus of variations" were first used by Euler in a 1760 paper.

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 (cf. page 14). And it is probably the only mathematical curve that has been deemed worthy of a monument in its honor (cf. Scholium I below, in Page 10).

With this book we wish to honor, in our own way, the brachistochrone and the mathematicians who studied it, on the occasion of the the 300th anniversary of the publication of the two issues of *Acta Eruditorum*.

Our original objective, when we set out to do the thinking and carry out the research that would eventually lead to this work, was to make a modest but very personal contribution to the celebration of the 300th anniversary of the events of 1696-97, by paying tribute to the mathematicians who participated in them and showing how, in the three centuries that followed, their ideas have been fruitful to an extent that would have made them proud. We wanted to refrain from retelling once more, from a slightly different angle, a story that has already been told many times in many nearly identical versions. Rather, we were determined to take full advantage of our specific difference within the community of mathematicians, and use our own specialized mathematical knowledge to look at the story from a new perspective that, we hoped, might yield truly novel insights on an old subject. Since for both of us *control theory* is the primary area of scientific interest, we found it natural to choose this field to guide us in our inquiry and provide us with a theme that would lend distinctness and freshness to our account. This led us naturally to explore the various ways in which the 300-year process of discovery that began in 1696-97 is related to ***optimal control theory***, the branch of control theory that deals with curve minimization problems and is most closely related to the calculus of variations.

When we started our investigation we felt that our decision to focus on optimal control was justified on the basis of

- our own experience with optimal control, which had taught us that this theory was richer and much more powerful than its immediate predecessor the calculus of variations, and had made us confident that this superior power would manifest itself if one applied both theories to an old problem and compared the results;

- the striking coincidence that the word "brachistochrone" actually means "shortest time" (cf. box below), so that 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;

- the fact that there was at least one example—the question of the existence of an optimum for the brachistochrone (cf. §XIV.8 in page 307 below)—where it was evident that the control approach yielded better results than calculus of variations methods[6].

Although we realized that these were not sufficient grounds for making a sweeping claim about the significance of optimal control—as opposed to the calculus of variations—for the brachistochrone problem, they seemed good enough to us as a point of departure for an inquiry that, we hoped, would yield additional evidence strengthening our case. We were persuaded that we would find good arguments to advocate the view that *the brachistochrone problem is closely related to optimal control theory for much more substantive reasons than the strikingly prescient choice of a word made by Leibniz and Johann Bernoulli, and the usefulness of optimal control as a technique for proving an existence result.*

[6]It was F. H. Clarke who brought our attention the fact that the existence problem for the brachistochrone was very delicate in the calculus of variations setting.

THE WORD "BRACHISTOCHRONE"

"Brachistochrone" means "shortest time." It is derived from the Greek words "brachistos" ($\beta\rho\acute{\alpha}\chi\iota\sigma\tau o\varsigma$), which means "shortest," and "chronos" ($\chi\rho\acute{o}\nu o\varsigma$), which means "time."
"Brachistos" is the superlative of "brachys" ($\beta\rho\alpha\chi\acute{\upsilon}\varsigma$), which means "short." Words such as brachylogy, brachycephalic, brachydactylic, brachyuran, brachypterous, are derived from "brachys."
Some authors[a] use the spelling "brachystochrone," but we choose the much more common spelling "brachistochrone" to reflect the fact that the "y" (Gr. "υ") is already changed into an "i" (Gr. "ι") in Greek, when forming the superlative.
It was Leibniz who suggested to Johann Bernoulli that he use this name for the "shortest time" curve that solves his "splendid" problem.

[a] Including ourselves in our earlier paper [138]!

Subsequently, our perspective evolved, as we repeatedly came across new and even more convincing reasons for thinking that our original decision to look at the 300-year evolution from an optimal control perspective had been fruitful beyond our best expectations. We realized that not only the brachistochrone problem, but also many other classical questions of the calculus of variations, were related to optimal control in a much more complex and interesting way than we had anticipated at first.

Moreover, the control point of view threw new light on a number of episodes in the 300-year history of the calculus of variations. As we retraced on our own the zigzagging path that had led from the brachistochrone to modern optimal control via the calculus of variations, we recognized that those who had followed it first had been part of an intellectually exciting adventure, full of surprising twists and interesting cases of ***missed opportunities***. Significant discoveries had been within the reach of giants such as Johann Bernoulli, Karl Weierstrass, and Constantin Carathéodory, but had failed to be made for reasons that to us, with our turn of the millenium hindsight and our optimal control perspective, look minor and perhaps even trivial. It is a story rich in connections with several other areas of mathematics, and full of lessons about the decisive importance of foundational issues—such as the choice of a correct formalism, or the precise specification of the set of variables on which a function depends—in the process of discovery.

We concluded that the march along the path of discovery, as it actually took place, is related to present-day optimal control control theory in a great variety of ways:

- First, there is a simple *kinship* relation, in the form of a direct line of descent going from the brachistochrone ancestor to the twentieth-century optimal control offspring, following the calculus of variations path.

- Second, *the history throws light on the present*, in that the meaning of the powerful results of today is much better understood by regarding them as the last steps in a long but simple sequence of conceptually natural improvements, each one of which corresponds to a point in the 300-year path.

- Third, *the present throws light on the history.* Several false steps and missed opportunities along the 300-year trek are best understood by looking at them from our privileged vantage point, which enables us to see what was there all along but was not seen by the original explorers. This is especially true of important "non-events," such as the explorers' failure to make certain discoveries that were clearly within their reach. There is a perfect analogy with the bizarre adventure of the 17th century explorer A. J. Tasman[7], whose achievements—such as the discovery of Tasmania, New Zealand, Tonga and the Fiji islands in 1642—pale in comparison with the much more amazing fact that he circumnavigated Australia but failed to discover it (cf. Scholium V in Page 15). The story of Tasman's trip takes on an extra dramatic dimension when one follows it while keeping in mind that Australia was there all along. In exactly the same way, the work by Johann Bernoulli on the brachistochrone-related optimization problem that gives rise to what we would now call the "Poincaré half-plane," the study by Weierstrass of the "excess function," and the reformulation by Carathéodory of Weierstrass's result, are seen in a new light when we realize how close these mathematicians had been to two major discoveries—hyperbolic geometry, the maximum principle of optimal control—and try to understand why they failed to notice the presence of these enormous "Australias" that were plainly there to be found.

- Fourth, *the results of the present enhance the value of the old problems.* If one "looks back at the past from a twentieth-century perspective" and views

[7]who, as befits our story, was Dutch, and in fact a native of the very same province of Groningen where Johann Bernoulli did his work on the brachistochrone and related problems, and came within a whisker of discovering hyperbolic geometry and the independence of Euclid's parallel postulate

the old problems in the light of the most recent results and techniques, they often yield new and interesting insights.

- Fifth, in particular, *the brachistochrone problem itself acquires new dimensions* when approached from the optimal control perspective. This perspective sheds new light on the old problem, and leads to new results that we believe would have pleased Johann Bernoulli and the other five contributors to the May 1697 *Acta Eruditorum*. We therefore feel justified in claiming, as we will later on, in Chapter XIV, that ***it is only within the optimal control framework that one can do full justice to Johann Bernoulli's "splendid problem."***

It became clear to us that there was a lot to the story that was not widely known and needed to be told, so we decided to tell it, entertainingly if possible, but supplying enough details to go beyond the purely anecdotal and clarify the conceptual issues, while at the same trying to reach a wide audience of mathematicians, pysicists and engineers.

We now understand that our original decision to tell the story from an optimal control perspective had been correct for many more reasons than those of which we had been aware at the beginning of our journey. The relationship between the 300-year path and modern optimal control has turned out to be multifaceted, and optimal control has shown itself to be much more than a twentieth-century derivative by-product of the calculus of variations. Among other things, modern optimal control theory becomes a powerful explanatory tool, whose mere existence—in the Platonic universe of ideas[8], even before it was discovered by humans—helps us makes sense of some past events, and renders them more dramatic, by revealing to us that these events happened *because the maximum principle of optimal control—like Australia—was really there all the time, waiting to be discovered.*

The links between the brachistochrone and optimal control have proved so strong, and optimal control has shown such a remarkable ability of to render meaningful and coherent many facts about the calculus of variations and its history, that we reached the somewhat radical conclusion that the best way to think of the events of 1696-97 was as marking *the birth of optimal control*. In other words:

- The formulation and solution of the brachistochrone problem was one of the first steps in a long series of results of increasing generality and mathematical elegance about the curve minimization problem.

[8]Or in a Popperian "third world'."

- With the maximum principle of optimal control theory, this evolution attains a high point, from which one can survey the surroundings and see how the winding calculus-of-variations path leading to the top had been all along a part of a meaningful and coherent landscape.

- It is therefore reasonable to view the 300-year history of the calculus of variations as a process of ascent towards the more general theory of optimal control, so that the history of optimal control can truly be said to start 300 years ago.

COMMENT NO. 3. Footnote 9 must be rewritten.

Actually, we eventually reached the conclusion that the brachistochrone was not the only old curve minimization problem that had a more natural home in the optimal control theory setting than in that of the classical calculus of variations. Other old questions, such as the ones of refraction and reflection of light, the famous—and ancient—isoperimetric problem[9] and even the most ancient of all such problems—that of finding the shortest path joining two points—turn out to lend themselves to a much smoother treatment by means of an optimal control approach than by calculus of variations methods[10]

COMMENT NO. 4. We must make sure that the issues discussed in Footnote 10 are indeed covered in a later section, and we must put the number of this section in Footnote 10. Also; we must add to footnote the remark that all these problems are covered by HJS's more recent nonsmooth results.

[9] The case of the isoperimetric problem is somewhat different from the others. The problem does not belong to the category of "standard calculus of variations problems," but it does belong to a larger class of calculus of variations problems with constraints, aptly named "isoperimetric problems." So one might surmise that at least in this case the calculus of variations framework is sufficient to do full justice to the problem. We will, however, argue that this is not quite so, but the reasons will be more subtle. Cf. §XV.1 for our discussion of this problem.

[10] Although, of course, they can be solved directly by *ad hoc* methods, as is well known. Our point here is not that optimal control theory is *needed* to solve these problems, but that until the advent of optimal control theory these problems required special methods and the main results of the theory were not were not directly applicable to them. For example, it is an elementary freshman calculus exercise to derive Snell's law of refraction from Fermat's minimum time principle, but this problem is *not* covered by the Euler-Lagrange equation or by the classical "smooth" maximum principle of optimal control. We feel it is inelegant for a problem that played such an important role in the birth of the calculus of variations and optimal control to be left outside the range of applicability of the main results of these theories, and relegated to the much less appealing class of "special problems that can be solved directly by special *ad hoc* methods." For another example, the classical Euler-Lagrange equation is not directly applicable to the problem of the shortest path joining two points because for that problem the Lagrangian is not differentiable; one therefore needs an *ad hoc* trick, such as giving a direct geometric argument, or proving that length minimization is equivalent to energy minimization and then solving the energy minimization problem instead of the original one, or even making mathematical mistakes in the analysis, as in some texbooks. These issues will be explored at length in §GNU below.

This had a radicalizing effect on our views, as we concluded that a good case can be made that optimal control is at least three thousand years old, and as old as geometry.

Our collaboration began in September 1995, when we exchanged ideas on the subject during a conference in Groningen, in the Netherlands. At that time, we intended that the final product of our effort—with our relatively modest initial claims—should appear in print on the occasion of the 300th anniversary of Johann Bernoulli's 1696 challenge or, in case further delays were unavoidable, on the 300th anniversary of the publication of the 1697 articles. And the pursuit of this limited objective met with success, for our paper ([138]) appeared on time, in June 1997. This paper, however, only gave an abrigded version of some parts of the story—explaining, for example, why we felt comfortable with the claim that the events of 1696-97 mark the birth of optimal control theory, but not presenting the complete argument. Moreover, by the time our manuscript was completed—and certainly much more so by the time the paper appeared—our findings had taken us a lot farther, and we were encouraged to go far beyond the humble goals of the original project and give a more detailed account of our story that would spell out our more ambitious claims.

Thus the writing continued, and the final result is this book, a belated contribution to what we now feel should really be the celebration of the 300th anniversary of the events of 1696-97. It is clear to us that these events constituted an important turning point in the history of the old subject of curve minimization, and can be regarded as either

- marking the birth of the calculus of variations *and optimal control*,

or, if one adopts the alternative view that optimal control is in fact much older,

- signaling the end of the first period in the history of optimal control, which had been characterized by the study of special problems solved by ad hoc methods, and ushering in a new era in which general techniques were developed to tackle broad classes of problems.

As we have already suggested, our retrospective journey along the path from the brachistochrone to the maximum principle will be unabashedly "control-centric"[11]

[11] Readers who hold views currently in vogue about the social construction of reality and the impossibility of an unbiased account of anything whatsoever might suspect that our strong claims about the importance of optimal control and the brachistochrone are merely reflections of the professional and nationalistic biases of the authors. We gladly plead guilty to most of this charge, except only for the word "merely," and state for the record that *we are both control theorists, and one of us teaches at Groningen, where Johann Bernoulli was a professor from 1695 to 1705*. These facts almost certainly explain why we became interested in our topic to begin with, and why we approcahed it from our particular point of view, but are of course irrelevant to the merit and validity of our claims.

and guided by our own turn of the millenium hindsights[12].

In our discussion, the three topics of (1) the brachistochrone problem, (2) optimal control, and (3) the 300-year path from the former to the latter, will be presented in a partly historical, partly logical order, selected so as to emphasize the main points outlined above. In particular, we will wait until Chapter XIV to embark upon our detailed analysis of the brachistochrone problem. By then, the reader will have become acquainted with the main steps in the 300-year evolution that led from the brachistochrone problem to the classical calculus of variations, from there to modern optimal control and, more recently, to differential-geometric control and the use of nonsmooth analysis techniques. We will then be in a position to show how these techniques yield new insights on the original 300-year old problem of Johann Bernoulli.

COMMENT NO. 5. The previous paragraph is incomplete. We need to put in more stuff to have a smooth ending for the chapter!

SCHOLIUM I. The brachistochrone monument

In 1994, on the occasion of the 375th anniversary of its foundation, the University of Groningen, in the Netherlands, 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.

COMMENT NO. 6. EXACT DATE? MONTH, AT LEAST.

In September 1995 a conference was held at the university to commemorate the 300th anniversary of the appointment of Johann Bernoulli. The unveiling of the monument in 1996 coincided with the 300th anniversary of the publication of Johann's challenge in *Acta Eruditorum.*

SCHOLIUM II. Johann Bernoulli, his family and his time

Since Johann Bernoulli is going to be a constant presence throughout our narrative, he deserves an early introduction, so we will now briefly discuss the person, his family, and the historical context that surrounded his life and work.

[12]In postmodernist parlance, it will be "deeply anchored in our own historicity."

The Bernoullis were a Protestant family originally from Antwerp in Flanders. They fled Antwerp in 1583 to escape religious oppression by the Spanish rulers and, after spending some time in Frankfurt, finally settled in Basel, Switzerland, early in the seventeenth century. In the Bernoulli family there were eight mathematicians in three consecutive generations, at least three of whom were outstanding and became deservedly famous. Most of them ended up as professors in Basel, but many spent extensive periods in other universities in Europe. The most prominent of the Bernoullis were the brothers Jakob (1654-1705) and Johann (1667-1748), and Johann's son Daniel (1700-1782), born in Groningen while his father was a professor there. Jakob Bernoulli made important contributions, in particular to probability theory (so that, for example, the "Bernoulli distributions" are named after him). Daniel is the discoverer of Bernoulli's law in hydrodynamics, one of the truly great laws in physics. "A majority of the mathematical members of this family in the second and third generations ... did not deliberately choose mathematics as a profession but drifted into it in spite of themselves as a dipsomaniac returns to alcohol." (Bell, [8], p. 132). For example, Johann started as a doctor of medicine, and Daniel, "starting out in law he became professor of eloquence in Basel before succeding his father in the chair of mathematics." ([8], p. 136.)

FIGURE 2. Bernoulli family logo

At the time that the young Johann Bernoulli came of age, mathematics was undergoing a revolutionary change. In 1684 Leibniz published his first article about differential calculus in the *Acta Eruditorum*. He illustrated the power of the *Nova Methodus*, as he called the new techniques, through a number of examples for determining maxima and minima in a much more effective way than had been possible before. Johann and Jakob Bernoulli were among the first to master Leibniz'

Nova Methodus, and in 1691 Johann achieved his first success by determining the catenary, the shape of a hanging chain, by means of the differential calculus (cf. §VI.3, in page 162). Still in his mid twenties, Johann was hired by the Marquis de l'Hôpital, one of the leading mathematicians of his time, to teach him differential calculus. Johann always claimed that he was the real discoverer of "L'Hôpital's rule," which appears in L'Hôpital's book. In 1992, documents were found which brought positive evidence for Johann's claim (cf. Scholium IV below).

Johann Bernoulli was not an easy person. He often quarreled with his colleagues, and had grievances about his salary, his health, and his work. In 1695, shortly after taking up the chair in Groningen, he complained in a letter to Leibniz, who had encouraged him to take the chair: *I have not met any of the practitioners of Algebra, which you consider present in Holland. To the contrary, I have not had the honor to meet one single person who would even deserve to be called a "mediocre mathematician."* In the same letter he complained that his teaching took too much of his time, and that *the more progress the students make, the less progress I make.* Such politically incorrect views were not only expressed in private letters to his friends. Johann Bernoulli did not hesitate to make his views public. While in Groningen he got into serious difficulties with the local Protestant theologians and clergy, who disapproved of the way new discoveries in the physical sciences cast doubt on the validity of revealed truth.

His disputes with his colleagues were often very open and unrelenting. He was perhaps the most abrasive contender in the bitter dispute between the English, Newtonian, and the continental, Leibnizian, schools, regarding the originality and rigor of the differential calculus. He "was a man of violent likes and dislikes: Leibniz and Euler were his gods; Newton he positively hated and greatly underestimated." ([8], p. 135.) His rivalry with his brother Jakob became an embarrassment to the scientific community and when in 1699 they were both elected to the Paris Academy, it was on the explicit condition that they promise to cease arguing, a promise that of course was not kept.

COMMENT NO. 7. Check date

Even more peculiar was Johann's rivalry with his own son Daniel, whom he criticized for being a Newtonian, and of whose success he was very jealous. 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. [8], p. 134. Probably with the help of his loyal disciple student 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. [67]). Daniel, however, remained dutifully respectful towards his father, although he frequently expressed his misgivings to his friend Euler

Below is a photograph of a glass-stained window of the Academy building (the main venue of the university) in Groningen. It shows Daniel Bernoulli sweetly clutching his father's robe, while Johann is holding a drawing of his brachistochrone.

COMMENT NO. 8. PUT PHOTOGRAPH OF STAINED GLASS HERE.

SCHOLIUM III. Tschirnhaus

Ehrenfried Tschirnhaus (1651-1708) studied mathematics, philosophy and medicine at the University of Leiden, visited various European countries, met Wallis, Leibniz and Huygens, and did scientific work, among other things, on new methods for making porcelain. A factory at Meissen started production of Tschirnhaus's porcelain in 1710, and the first sales of any consequence of porcelain took place at the Leipzig Fair in 1713. European governments competed to obtain his porcelain technique, but Tschirnhaus refused to let them have it, and ended up incurring huge debts. As a mathematician, he is most famous for the discovery of the "Tschirnhaus transformation," that can be used to eliminate the term of degree $n-1$ from a polynomial equation $x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$ (by setting $x = y - \frac{a_1}{n}$), and for the detailed study of the plane curve with equation $3ay^2 = x(x-a)^2$, known as "Tschirnhaus's cubic" since being so named by R. A. Archibald in 1900.

SCHOLIUM IV. The Marquis de L'Hôpital

Guillaume Francois Antoine Marquis de L'Hôpital (1661-1704) became a mathematician after having to resign as a cavalry office because of nearsightedness. In 1691 he hired Johann Bernoulli to teach him the new methods of calculus, which Johann did with lessons taught in Paris and at de l'Hôpital's country house at Oucques, and then later by correspondence, when Johann was living in Basel. In 1696, de l'Hôpital published his calculus textbook *Analyse des infiniment petits pour l'intelligence des lignes courbes*, in whose preface he stated that he was "obliged to the gentlemen Bernoulli for their many bright ideas, particularly to the younger Mr. Bernoulli who is now a professor in Groningen." The book contains the famous "l'Hôpital rule" for computing the limit of a quotient when the numerator and the denominator go to zero. After L'Hôpital's death in 1704, Johann claimed the he was the real author of the book, and in particular that L'Hôpital's rule about limits was due to him. He said that he had not disclosed this fact before because he was bound by his contract with L'Hôpital to let the marquis take credit for the discoveries which Johann may make during this teaching. His contemporaries tended to ignore this claim, since Johann was not known to be particularly generous to others or objective about his own achievements. However, in 1922 a copy of the notes made by his nephew Nicolaus Bernoulli of Johann' course was found in Basel, and this brought positive evidence for Johann's claim, since the course turned out to contain practically all the material found in the book.

SCHOLIUM V. A. J. Tasman's exploration of the South Pacific

Abel Janszoon Tasman (1603-1659), "the greatest of Dutch explorers, was the discoverer of Tasmania, New Zealand, Tonga and the Fiji Islands" ([56], p. 1070).

Tasman was born at Lutjegast, in the Dutch province of Groningen, in whose capital—the city of Groningen—Johann Bernoulli would live later from 1695 to 1705, do his work on the brachistochrone, and follow in Tasman's footsteps by coming close to making, but failing to make, an earth-shaking discovery for which he had all the tools, namely, that of a model of non-Euclidean geometry, which would have enabled him to settle the long-standing question of the independence of Euclid's Fifth Postulate (cf. §XI.6 in Page 271).

In August 1642 Tasman sailed from Batavia (now Jakarta, in the island of Java, in Indonesia) to Mauritius, and then southward and eastward, discovering Tasmania[13], exploring New Zealand, Tonga and the Fiji Islands, and then turning northwest, returning to Batavia in June 1643 after sailing along the coast of New Guinea.

Tasman's life, and in particular his spectacular ten-month voyage, are fascinating material for a biographer, historian or novelist. But a narrative of his life

[13]The discovery took place in late November 1642. The first European to see the island wasate Tasman, who sighted it from the *Heemskerck*, one of the two ships in his expedition, at 4pm on November 24. The first European to land in it was Pieter Jacobsz, a master carpenter who swam ashore from the Heemskerck and placed the Dutch flag. The island was named "Van Diemen's Land," after Anthoony Van Diemen, the Governor-General of the Dutch East India Company, who had commissioned Tasman's expedition. It was visited by French and British explorers, among them the famous James Cook in 1777, and the infamous William Bligh in 1788 and 1792. (Bligh's 1788 visit, on the HMS Bounty, consisted of a stop at Adventure Bay on August 2, 1788, to replenish water and firewood supplies before sailing on September 4 for Tahiti. The mutiny occured on April 28, 1789, while the Bounty was headed to the Caribbean and then back to England, after a long stay in Tahiti.) Van Diemen's Land was formally claimed for Britain in 1788. It became a penal colony for British convicts, dreaded and cursed by prisoners for its horrendously repressive system (cf. Hughes [75]), at a time when Britain had lost its American colonies and needed a new outlet for its large number of convicts, among them many political prisoners, in particular from Ireland. The island was renamed Tasmania in 1856, as part of an effort to give it a more appealing image. The convict system was completely abolished in 1886.

and travels cannot possibly omit mentioning the amazing fact that *he circumnavigated Australia but failed to discover it*. Australia was not really discovered by European explorers until about 100 years after Tasman's death. Before that date, and even during Tasman's life, there had been some sightings by Dutch explorers of the Northwest coast of Australia. Tasman himself explored in 1644 part of the northeast coast of Australia, from the Gulf of Carpentaria to 22^o S, but it was not until the late 18th century that the true dimensions of the island-continent were understood and its systematic exploration undertaken. The first recorded sighting of a kangaroo by a European was in 1770.

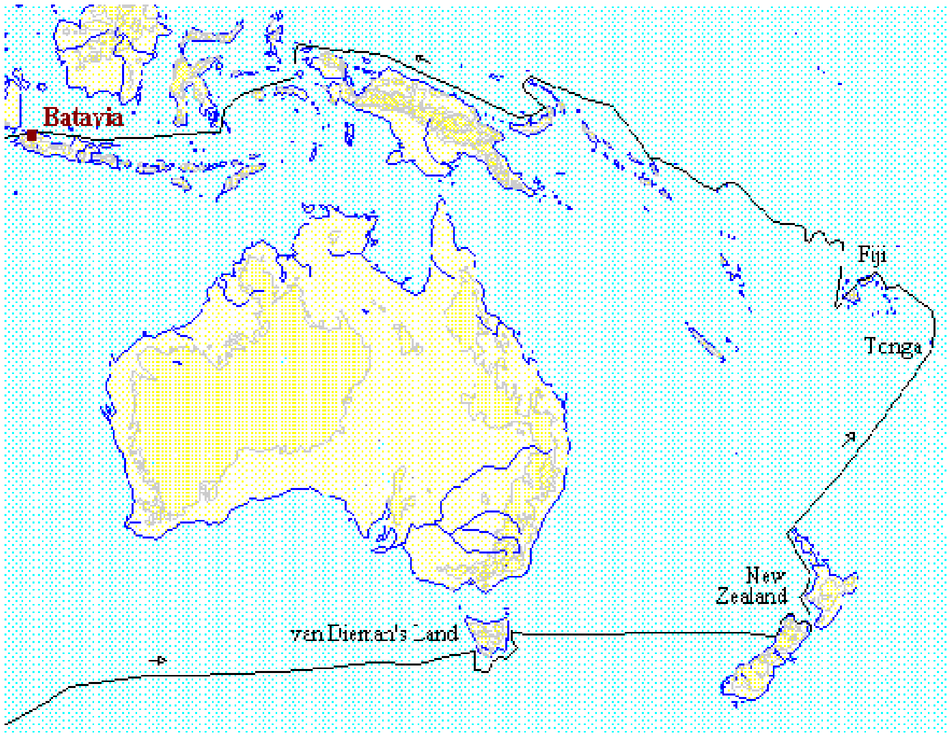


FIGURE 3. Tasman's 1642 trip

The Tasman example is quite relevant to the issue of the appropriatenes of talking about missed opportunities. It is often argued that it is pointless to pass judgement on the past and discuss events that did not happen but might, or perhaps should, have happened. We agree that this is so when the event that did not happen was so remote from the range of realistic possibilities that it is unimaginable that it might have happened. We would not take seriously a biographer who told us that Tasman failed to discover the planet Pluto, or Maxwell's equations, or the structure of DNA. But the fact that he missed Australia is clearly a central event in our perception of the story, to which it adds drama and emotion. It is hard to reflect on Tasman's journey without feeling that he could and should have found Australia, and without being tempted to speculate on what would have happened if he had.

Similarly, in the path that led from the brachistochrone to the maximum principle there were at least two Australias (the already mentioned maximum principle and also, as will be explained in Chapter XI, the independence of Euclid's parallel postulate) and at least three "Tasmans"—Johann Bernoulli, Hamilton and Weierstrass. No account of their adventures can be truly complete unless it points out their failure to make momentous discoveries to which they were very close[14]. □

I.1 A preview of our main themes

We begin with a rough outline of our main themes, which will be developed in more detail in subsequent chapters.

I.1.1 The setting: calculus of variations vs. optimal control, and optimal control vs. control

As a preliminary to our 300-year journey, and to set the stage for the discussion and comparisons that will follow, we will describe, in Chapter II, the subject matters of the calculus of variations and optimal control, and explain how the latter is much

[14]Another good example of the usefulness of hindsight to throw light on past scientific research is the history of blood transfusions. Experiments on transfusions were conducted as far back as the 1660s, by physicians such as Richard Lower (1631-1691). It was soon noticed that in some cases the patient would show a spectacular improvement, while in others there would be a severe reaction that could even result in death. Why this was so remained a mystery until blood types were discovered in 1900. The mystery is now explained, and with our twentieth-century hindsight we can make sense of the facts observed earlier. In his 1997 novel *An Instance of the Fingerpost* ([106]), Iain Pears tells the story of a physician who tried to help a patient with transfusions in 1663. In the first transfusion, the patient was given blood from her own daughter, and the patient's condition clearly improved. For the second transfusion the daughter was not available, so the physician used his own blood, and the patient got worse and eventually died. Blood types are not explicitly mentioned, but the dramatic effect of the scene when the physician decides to use his own blood is certainly enhanced by the fact that *the reader is aware of a serious danger that the physician knows nothing about*.

more general and naturally leads to the study of richer mathematical structures than the former. We then digress to point out that *control*, as opposed to *optimal control*, is an even larger and richer area. The presentation in Chapter II is informal, but we provide at the end of the book an appendix with some truly formal definitions, for the benefit of those readers who wish to see completely rigorous formulations.

I.1.2 Optimal control was born in the years 1696-97, or perhaps even earlier

We have already stated that "brachistochrone" means "shortest time" and that, as the name suggests, the brachistochrone is indeed the solution of a minimum time problem, i.e., of a genuine optimal control problem. We will argue that this is not a coincidence, and that there are profound reasons why the brachistochrone problem truly belongs to optimal control theory. These reasons are

a. that the problem involves an interplay between a *nontrivial dynamics* and the minimization of a Lagrangian cost functional, exactly as in modern optimal control theory (as explained in Chapter II), whereas in the classical calculus of variations one minimizes functionals over the space of all curves, so no interesting dynamical constraint is involved other than the one arising from the minimization,

b. that the optimal control approach, applied to the brachistochrone, is superior to other methods such as that based on the classical calculus of variations, and thus comes closest to doing full justice—as will be shown in Chapter XIV—to what Leibniz called a "splendid" problem.

I.1.3 The path from the brachistochrone problem to the maximum principle

Johann Bernoulli's challenge was the initial step along a continuous path that led, by successive incremental improvements, to modern optimal control.

In Chapter V we will retrace some of the main steps along this path, and show how the work on the brachistochrone was extended by Johann Bernoulli's successors Euler, Lagrange, Legendre, Hamilton, Weierstrass, Bolza and Carathéodory, who discovered necessary conditions for a minimum that were of increasing strength and applicable to successively larger classes of curve minimization problems. This process led from the study of the brachistochrone to the Euler-Lagrange equations, and then to Legendre's second order necessary condition, Hamilton's equations, Weierstrass' "excess function" side conditions, Bolza's analysis of abnormal extremals, and the effort by Carathéodory to reformulate Weierstrass' result in Hamiltonian form. Eventually, this resulted—in the 1950s—in the statement of the "maximum principle" of optimal control theory and the proof of its first version, stated here in Chapter VI.

I.1.4 The clarification of the role of the momentum

As we follow the evolution of the necessary conditions for an optimum, we will pay particular attention to the role of the "momentum." The classical Euler-Lagrange equations contain the vector

$$\pi(t) = \pi_{\xi}(t) \stackrel{\mathrm{def}}{=} \frac{\partial L}{\partial \dot{x}}(\xi(t), \dot{\xi}(t), t)\,, \tag{I.1}$$

known as the *momentum* along the trajectory $\xi(\cdot)$. (In modern terminology, π is a *field of covectors* along $\xi(\cdot)$. In control theory, π is often referred to as the "adjoint vector.") The momentum also occurs in Hamilton's equations and in Weierstrass' condition. As we shall show, the crucial discovery that Weierstrass and Carathéodory came close to making but missed, was that, once we have the Weierstrass condition, *the momentum no longer needs to be defined by (I.1)*. One can regard the momentum as an "independent" object, and interpret the necessary conditions discovered by Euler, Lagrange and Weierstrass as asserting the existence of a field of covectors π having certain properties. Under the specific technical assumptions made by Weierstrass throughout his work on the calculus of variations, these properties imply in particular that (I.1) holds. But these technical assumptions can be considerably weakened, and when this is done the necessary conditions remain valid, as assertions about the existence of a momentum field π with certain properties, but it no longer follows that π is given by the "classical momentum condition" (I.1).

I.1.5 The significance of the maximum principle

The maximum principle was a major achievement that, as we shall explain in Chapters V and VI, unified, simplified and clarified a very large part of the understanding of path optimization problems acquired in the three preceding centuries, by turning a collection of highly technical results with lengthy and complicated statements into a much simpler and more accessible formulation. Moreover, the maximum principle had the additional advantage of applying to a much broader class of problems, and pointing the way towards even newer and more general discoveries. This makes it a significant turning-point in the march that began in 1696-97, marking the culmination of 300 years of research as well as a new beginning. (Some of the work that has arisen as an outgrowth of the maximum principle, much of which is currently in progress, will be described in Chapters IX, IX.8, XI, and XII.)

COMMENT NO. 9. JAN: In the "missed opportunities" section we have to say something about Dyson's famous article [54]. I am planning to read it when I get back from Europe, but if you want you can go ahead and read it and write something about it.

I.1.6 Missed opportunities

Our twentieth-century control-centric perspective will enable us to tell some interesting stories of "missed opportunities," by identifying points along this path where the actors had all the knowledge that was needed to make a dramatic discovery and get close to the maximum principle, but failed to make it for trivial reasons, such as their having chosen one particular formalism rather than another that on the surface looked nearly identical, but would have had profoundly different implications. Specifically, we will argue that

> *The evolution of the calculus of variations reached a critical fork in the road when Hamilton's equations were written*,

and that

> *the way the equations were actually written was only one of two possible formulations, and may not have been the best one*.

Precisely, we will show that

> *If Hamilton's equations had been written in a slightly different way—using a Hamiltonian that is a function of all three sets of variables (positions, velocities and momenta) rather than a function of the positions and momenta only—then some important discoveries that were not made until the 1950s would almost certainly have been made much sooner*.

To make this point, we will explain how a number of facts from the classical calculus of variations, such as the Euler-Lagrange equations, Legendre's second order condition, and the Weierstrass side conditions, can be seen right away to fit together into a formulation which is shorter, simpler and more elegant than the conditions stated separately, provided only that the "correct" Hamiltonian formalism is used. We will then show that Weierstrass had essentially found this formulation, but failed to appreciate what it meant because he was not using the right Hamiltonian formalism. The damage done by the use of an inappropriate Hamiltonian formalism will become even more evident when we point out how Carathéodory came even closer, by attempting to write Weierstrass' condition in Hamiltonian form, but ended up getting a complicated expression that failed to reveal the simple underlying phenomenon of Hamiltonian maximization.

Furthermore, we will argue that the "correct" (i.e., control-theoretic) formulation leads directly to a natural conjecture, in the form of another statement which is even simpler, and in addition happens to be much more powerful and general. We will show that this conjecture actually turns out to be essentially true, and all that is needed for the statement to become completely true is one small extra step, namely, the addition of the "abnormal" multiplier, whose need had already been understood by Bolza as far back as 1913. One then gets the full force of the maximum principle of optimal control theory, which was only discovered in the 1950s and, as we indicated earlier, is not only ***stronger*** but also ***simpler*** than the combination of its predecessors.

I.1.7 The superior power of the control-theoretic formulation of Hamilton's equations

Those readers who are skeptical about this kind of historical speculation may choose to regard our "missed opportunities" stories as merely a pedagogical device, whose purpose is to show why the "correct" (control) formulation of Hamilton's equations is much more powerful than the classical one, even though both formulations look very similar and are about equally easy to understand.

This will be demonstrated in Chapters V, VI and VII in three ways:

1. In Chapter V we will give several very simple examples of minimization problems that at first glance do not appear to lend themselves to a calculus of variations approach but are easily solved when Hamilton's equations are reformulated in "control" form.

2. Also in Chapter V, we will show that the transition from Hamilton's equations to the maximum principle is a very small and completely natural step, *provided that the Hamilton equations are written in their control version*. We will speculate on the reason for Weierstrass' Tasman-like failure to discover

the simple Hamiltonian interpretation of his excess function result and to find the far-reaching generalization that this interpretation would have suggested.

Our proposed explanation is as follows:

a. Weierstrass was working with Lagrangians that are *positively homogeneous of degree one with respect to the velocities*.
b. Restricting oneself to such Lagrangians is something that can be done without loss of generality, because a problem with a more general Lagrangian is always equivalent—by a change of time parameter, as will be explained in Chapter V—to one of Weierstrass' special kind.
c. On the other hand, "without loss of generality" is not the same as "without loss of useful insights." Actually, it turns out that *for the positively homogeneous Lagrangians used by Weierstrass the reformulation of the Euler-Lagrange necessary conditions in classical Hamiltonian form is **never** possible*, because the classical Hamiltonian $(q,p,t) \mapsto H(q,p,t)$ can only be defined when, for each (q,t), the map

$$\dot{q} \mapsto p = \frac{\partial L}{\partial \dot{q}}(q,\dot{q},t)$$

   defined by the momentum condition can be inverted, so that $\dot{q}$ can be eliminated by expressing it as a function of q, p, and t.
d. Furthermore, the classical Hamiltonian formulation of the Euler-Lagrange conditions was the only one known to him.
e. It follows that, by choosing to work with his special class of Lagrangians, and given the limitations of the classical Hamiltonian formulation, Weierstrass unwittingly placed himself precisely in the case where there was no chance that he would see how his result could be formulated in Hamiltonian form.

We will show how, if Weierstrass had used the "control" Hamiltonian formulation instead of the classical one, his side condition would have been seen right away to be no more than the maximization of the Hamiltonian, and he would never have needed to introduce a new entity called the "excess function."

Additional supporting evidence for our argument about Weierstrass' missed opportunity will be presented in §V.9, where we will show how Carathéodory, in his work on the Calculus of Variations, came within striking distance to discovering the Hamiltonian maximization condition, but failed to find it, because of his insistence on using the classical Hamiltonian.

3. In Chapter VII, we will look at the proper way to derive a Hamiltonian formulation of relativistic mechanics, a topic that mechanics books usually avoid "like the plague" (in the words of H. Goldstein, the author of one of the leading textbooks on the subject; cf. his statement quoted in Chapter VII), or handle superficially and unconvincingly, and often in a noncovariant way. We will show that in this case all the difficulties disappear if one uses the control Hamiltonian. This is not surprising, because the difficulties arise precisely from the very same fact that had stood in Weierstrass' way before, namely, that the relativistic Lagrangians belong precisely to the class that Weierstrass had used.

I.1.8 Hamilton's value function: the eclipse and re-emergence of a great idea

Having mentioned Hamilton, we will briefly digress from our account of the evolution from the Euler-Lagrange equations to the maximum principle, and discuss an isolated episode whose connection to the rest of our story will only become apparent later, in Chapter VIII. Quoting from a 1944 account by J. L. Synge, we will describe Hamilton's idea of the "characteristic function" (or "principal function"). According to Synge, the principal function was "Hamilton's great central idea," but "Hamilton's plan to develop all the properties of the extremals from the characteristic function" is "hard to follow," so hard in fact that "few mathematicians in the ensuing century have given serious consideration" to it. As will be pointed out in Chapter VIII, Synge wrote this just a few years before his view of the importance of Hamilton's "great central idea" was completely vindicated. Soon after, the new theory of ***modern optimal control*** would be born, and in this theory Hamilton's function—renamed the "value funcion"—would play precisely the central role for which Synge thought it was destined.

This will provide us with a new argument in favor of our control-centered perspective, because it will show that the control approach reveals the true meaning of the momentum and makes it possible to carry out Hamilton's program as Synge had suggested.

COMMENT NO. 10. On mathematical rigor: we have to check carefully that we really do all the things promised here in §I.1.9!

I.1.9 The issue of mathematical rigor

The progression described in Chapter V will involve successively more general *formulations*, which in all cases will be imprecise and nonrigorous according to the standards of twentieth-century mathematics. In Chapters IX and IX.8 we will explain how these vague statements can be and have been translated into rigorous theorems. Our reason for waiting so long to address this question is that we feel that it is desirable to move as fast as possible towards the best possible formulation, and then—and *only then*—take a careful look at it and seek to state it and prove it rigorously. Naturally, one could do this at each intermediate step, and, for example, discuss in detail the rigorous proofs of the Euler-Lagrange equation or the Weierstrass' side condition. We will not do so, because all the information that could be derived from doing it will in any case be contained in the rigorous results arising from the most general version.

COMMENT NO. 11. Actually, we really should and probably will talk about proofs of intermediate results.

In some cases, however, we will discuss certain issues having to do with the rigorous justification of intermediate versions. This will be done to show how the later developments can also throw light on the true meaning of earlier results, e.g. by showing that these results can be proved under weaker technical conditions than had originally been thought necessary. (For example, we will briefly review the well-known textbook proof of the Euler-Lagrange equation, in order to show that, as usually presented, it requires much more differentiability of the Lagrangian than is really needed. With the maximum principle, or even with Hamilton's equations in control form, one can see rather easily that the Lagrangian need not be differentiable with respect to the velocity.)

By contrast with our formulation of the maximum principle in Chapter VI, where it will be stated as a *principle* (i.e., not yet a complete theorem, with precise hypotheses and conclusions), Chapters IX and IX.8 will focus on various completely rigorous *versions* of the principle. That is, we will state a number of rigorous mathematical theorems obtained by choosing precise technical hypotheses, beginning with the first version of the 1962 book [112] by Pontryagin, Boltyanskii, Gamkrelidze and Mischenko—for systems of class C^1—and proceeding from there to the nonsmooth versions—for locally Lipschitz systems—and to more recent versions for systems with a nonlipschitz and even a discontinuous right-hand side.

COMMENT NO. 12. Need a box on "Principles and Versions," either here or in the rigorous chapters.

I.1.10 The geometric connection, and how Johann Bernoulli missed an opportunity to discover non-Euclidean geometry

In Chapter XI we follow the evolution of a relationship of central importance in mathematics, namely, the multiple connections between the theory of curve minimization problems and geometry. Some of these go back to many centuries before 1696, as shown by the fact that in ancient civilizations it was understood that the straight-line segment is the shortest path between two points. Furthermore, their importance and practical usefulness was noticed, for example, by those who took advantage of this characterization of segments by using “rope-stretching” in order to build temples and altars with perfectly straight lines (cf. §XI.2). It is manifest in the study of the isoperimetric problem in classical Greece, and in Johann Bernoulli’s work on the brachistochrone and related problems, where a non-event of truly Tasmanian dimensions took place, when Bernoulli studied the light rays in a plane medium with speed of light proportional to $|y|$, and found what we would now call the “Poincaré half-plane,” but failed to notice that this provides a very simple model for non-Euclidean hyperbolic geometry. The linkage shows up again in Lagrange’s proof of the invariance of the Euler-Lagrange equations under arbitrary nonlinear coordinate changes, which was the first important step in the road to differential geometry. It then becomes a crucial aspect of the work of Bernhard Riemann, who created differential geometry by by inverting the traditional relationship between geometry and curve minimization, and regarding curve minimization not only as a collection of interesting problems to be *studied* using geometry, but as a source that could be used to *create new geometries*. As a preparation for §XII, which discusses the application of differential-geometric ideas to control, we include in Chapter XI a general discussion of the concepts of covariance and invariance.

I.1.11 Differential geometry and control

In Chapter XII we move beyond the maximum principle, by showing how the incorporation of differential-geometric ideas makes it possible to give coordinate-free descriptions of important properties of control systems, in a spirit very similar to that of Lagrange’s ideas about coordinate invariance.

COMMENT NO. 13. JOB F HJS: Expand and complete this section

I.1.12 Some open problems

In Chapter XIII we bring the reader to the current frontier of research, by stating three challenging open problems whose solution, in our view, would result in significant progress. Unlike Johann Bernoulli's 1696 challenge, ours consists of problems that we do not know how to solve and, to our knowledge, are open.

I.1.13 The brachistochrone

In Chapter XIV we return to our starting point, and look at Johann Bernoulli's problem from various perspectives. We review in succession (a) Bernoulli's own rather ingenious solution, based on a geometrical optics analogy, (b) the calculus of variations approach, based on the Euler-Langrange equation, (c) the optimal control method, and (d) the analysis of the problem based on ideas from differential-geometric control. In each case, we show how the new technique adds to the understanding of the problem.

In particular, we will argue in §XIV.12 that the brachistochrone problem, as commonly stated, exhibits some rather peculiar features, and a proper formulation leads to a problem which is quite singular from the point of view of general optimal control theory, because it has a nontrivial constant of motion (the energy) and unbounded controls. It will then turn out that the use of differential-geometric analysis is not only convenient but in fact necessary for analyzing the problem. We will carry out this analysis and show how it provides a considerably sharper understanding of the true underlying geometric structure. It is interesting to note that a very natural way to understand these peculiarities of the brachistochrone problem is to ask Hamilton's "central" question, i.e., to try to find the value function for the problem. This leads immediately to a four-dimensional state space, a three-dimensional distribution, the "diamond Lie algebra," and impulse controls,

We also include three more examples, involving the application of optimal control to three problems closely related to the brachistochrone. One of them is due to M. Hestenes, and can be solved using the classical maximum principle. The other two require that we use a recent "nonsmooth" version of the maximum principle (Sussmann [130, 131, 135, 137, 140, 141, 142]). Hestenes' problem is "an interesting variation of the brachistochrone" dealing with "an idealized rocket ship moving in a vertical plane (Hestenes [73], p. 64).

Hestenes proposed his problem in a effort "to speculate on what type of brachistochrone problem the Bernoulli brothers might have formulated had they lived in modern times." In a similar vein, we propose a second mathematically natural extension of the brachistochrone problem, namely, the problem of the ***reflected***

brachistochrone, in which one looks for the light rays in the whole (x,y) plane (rather than in a half-plane) with a speed of light proportional to $\sqrt{|y|}$.

With the third example we travel back in time even farther, and look at the derivation of ***Snell's law of refraction*** from Fermat's minimum time principle. Snell's law was a crucial tool in Johann Bernoulli's solution of the brachistochrone problem. The other two problems are very natural extensions of the original brachistochrone question, and we have no doubt that Johann Bernoulli, who was always eagerly looking for new problems to test his methods, would have found them interesting.

It turns out that all three problems fall outside the scope of standard necessary conditions of the calculus of variations. The Hestenes problem naturally belongs to "classical" optimal control, but the other two do not, because their right-hand sides are not Lipschitz continuous, and in the case of Snell' law they even fail to be continuous.[15] But all three problems are covered by the more recent nonsmooth result.

I.1.14 A speculative conclusion

Based on these examples, we argue that it is very likely that Johann Bernoulli himself would have appreciated the superiority of optimal control methods and would have accepted that *optimal control* is the general technique that he had been after for solving curve minimization problems, while the Euler-Lagrange equations and other methods of the calculus of variations prior to the 1950s were only incomplete and preliminary steps in the process of discovery of this powerful approach.

[15]Naturally, Snell' law of refraction can be derived from Fermat's principle by elementary calculus, but this is not our point, as we explained in Foonote 10.

Chapter II

Calculus of variations, optimal control, and control

Our first task will be to present our reasons for questioning the conventional wisdom that regards optimal control as a twentieth century newcomer to mathematics and a late offspring of the much older calculus of variations.

This will require that we explain in detail how optimal control differs from the calculus of variations. In particular, we will have to give precise definitions of the notions of a "standard calculus of variations problem" and an "optimal control problem." For this purpose, it will be necessary to agree on some notational conventions, and to be clear about the choices to be made of the "configuration spaces" for our problems.

II.1 Some notational conventions

In our discussion, it will be necessary to refer to the work of several authors from the seventeenth to the nineteenth century. Every time we do so, we will be confronted with the question whether it is better to cite these authors using their own mathematical notations, or to stick to a uniform set of notational conventions of our own.

In the interest of clarity and consistency, we will choose the latter alternative. ***We will always use our own mathematical notations and terminology rather than those of the various authors under consideration***. To do otherwise would make our presentation much- harder to follow, and would render difficult the comparison of the contributions of different authors, since their notations vary wildly, and often conflict with twentieth-century standards of mathematical rigor (e.g. by failing to distinguish between definite and indefinite integrals). So, for example,

- the letter L will always stand for the "Lagrangian,"
- the state variables will usually—but not always[1]–be called q,
- the independent variable—often called x or y in early papers on the subject—will usually be called t, and should be thought of as time,
- we will use dots—and on a few occasions primes, and also $\frac{d}{dt}$, when we want to differentiate a long expression—to denote differentiation with respect to time (cf. Equation (V.5) below for an example of the use of these notations). Thus, the derivative of a scalar- or vector-valued function $\mathbb{R} \ni t \mapsto \xi(t) \in \mathbb{R}^n$ will usually be called $\dot{\xi}$, although sometimes we may use ξ' or $\frac{d\xi}{dt}$ instead of $\dot{\xi}$.

In addition, we shall adopt the following convention:

- ***Roman letters*** *will always refer to the various "variables"—such as positions, velocities, momenta, controls, and time—related to a calculus of variations or optimal control problem.* ***Greek letters*** *will be used to denote* ***functions of time*** *with values in the spaces of the corresponding variables.*

Thus, for example, a curve in q-space will often be called ξ, and $\dot{\xi}$ will be its time-derivative. The Lagrangian will be a function $(q,u,t) \mapsto L(q,u,t)$ of a "position variable," a "velocity variable" u, and the "time variable" t. If $t \mapsto \xi(t)$ is a curve in q-space, then we can plug in $\xi(t)$ in the q-slot and $\dot{\xi}(t)$ in the u-slot, thereby obtaining a function $t \mapsto L(\xi(t),\dot{\xi}(t),t)$ of time. For this reason, calculus of variations books sometimes use the notation $\dot{q}$ for the velocity variable, and write the Lagrangian as a function

$$(q,\dot{q},t) \mapsto L(q,\dot{q},t)\,.$$

We will occasionally do so as well, but usually we will prefer using the less confusing name u for the velocity, which makes it clear that u is an "independent variable," a fact that may be less evident to some readers if this variable were to be called $\dot{q}$.

[1]Sometimes we will use x instead of q, or write $q = (x,y)$ or $q = (x,y,z)$ when q takes values in $\mathbb{R}^2$ or $\mathbb{R}^3$.

II.2 Calculus of variations and optimal control problems

The calculus of variations deals mainly with optimization problems in which one wants to minimize a functional

$$J(\xi) = \int_a^b L(\xi(t), \dot{\xi}(t), t)\, dt \tag{II.1}$$

of a curve ξ. The integrand of the above expression involves a real-valued function

$$(q, \dot{q}, t) \mapsto L(q, \dot{q}, t)$$

of "position, velocity, and time." Nowadays, this function is usually called the ***Lagrangian.*** Optimal control theory is more general, since it considers situations where the velocity $\dot{q}$ is allowed to depend on position and time, as well as on an extra "control" parameter[2] u. So in optimal control theory the dynamical law governing the system's behavior is of the form

$$\dot{q} = f(q, u, t)\,.$$

It is then natural to allow the Lagrangian L to depend on u as well. Moreover, since $\dot{q}$ is determined by q, u and t, there is no need to think of L as a function $(q, \dot{q}, u, t) \mapsto L(q, \dot{q}, u, t)$, since one could always rewrite such a function as $L(q, f(q, u, t), u, t)$, i. e., a function of q, u, t only. So the Lagrangian in optimal control theory will be a function

$$(q, u, t) \mapsto L(q, u, t)\,.$$

To make all this more precise, we start by discussing the nature of the "configuration spaces" Q where the variable q takes values.

II.3 Configuration spaces

In contemporary mathematical writing, it is often customary to represent the "position"—or "configuration," or "state"—of a system, at least locally, by a vector[3]

$$q = (q^1, q^2, \ldots, q^n) \tag{II.2}$$

in n-dimensional space $\mathbb{R}^n$. The velocity is then also represented by a vector

$$\dot{q} = (\dot{q}^1, \dot{q}^2, \ldots, \dot{q}^n) \in \mathbb{R}^n\,. \tag{II.3}$$

[2] The letter u is traditionally used to denote the control because the Russian word for control is "upravleniya."

[3] We follow the modern convention of using superscripts for "contravariant indices," i.e., components of vectors.

We should think of the members of $\mathbb{R}^n$ as *column vectors*, so that in fact we really should have written

$$q = \begin{bmatrix} q^1 \\ q^2 \\ \vdots \\ q^n \end{bmatrix}, \qquad \dot{q} = \begin{bmatrix} \dot{q}^1 \\ \dot{q}^2 \\ \vdots \\ \dot{q}^n \end{bmatrix},$$

or $q = (q^1, q^2, \ldots, q^n)^\dagger$, $\dot{q} = (\dot{q}^1, \dot{q}^2, \ldots, \dot{q}^n)^\dagger$ (where the symbol "†" stands for "transpose") rather than (II.2) and (II.3).

The *momentum* is also a vector, usually represented as p, which should be thought of as a member of the *dual space* of $\mathbb{R}^n$, i.e., as a *covector*, or *linear functional* on $\mathbb{R}^n$. We represent these linear functionals as *row vectors*, and the space of all n-dimensional real row vectors as $\mathbb{R}_n$, so the momentum is[4] $p = (p_1, \ldots, p_n) \in \mathbb{R}_n$, and the result of applying such a p as a linear functional on a vector (II.2) is the ordinary matrix product[5]

$$p \cdot q = p_1 q^1 + p_2 q^2 + \ldots + p_n q^n = \sum_{j=1}^{n} p_j q^j\,. \tag{II.4}$$

We will also write this product as $p(q)$ (to emphasize the fact that p is alinear functional), or as an inner product $\langle p, q \rangle$.

In our discussion, we will use Q to denote the "configuration space," that is, the set of all points q. Often, we will simply think of the points $q \in Q$ as n-tuples $q = (q^1, q^2, \ldots, q^n)$, i. e., as members of $\mathbb{R}^n$. But in all interesting physical or engineering applications, it would not be accurate to say that a point q of configuration space really ***is*** an n-tuple of numbers. The n-tuples $(q^1, q^2, \ldots, q^n)$ are only *coordinate representations* that we use to talk about the points, but are not the points themselves.

For example, suppose Q is "physical space." Then a point of Q is not *really* a triple of numbers, although it can be *represented* as such a triple by choosing a *coordinate system*, i. e., a bijective map $\kappa : Q \mapsto \mathbb{R}^3$. The choice of a coordinate system is not unique, and in general there is no natural way to single out a canonical κ. For example, in classical mechanics[6] one can single out ***linear*** coordinate

[4]Here we are following standard conventions once again, and using subscripts for "covariant indices," i.e., components of covectors. "Covectors" are dual objects to "vectors."

[5]The "Einstein covention" prescribes that one should sum over pairs of repeated indices, one of which is contravariant—i.e., a superscript— while the other one is covariant—i.e., a subscript. It also customary to omit the summation symbol from such sums, and write, for example, $p_j q^j$ for the summation of (II.4), but we will not go that far.

[6]We are talking here, naturally, of "classical mechanics with absolute space," in which there is "absolute rest," so physical space-time S is the product $S = Q \times T$ of three-

systems (specified by choosing an origin O of coordinates and three noncoplanar straight lines passing through O, known as the "coordinate axes"; such systems are distinguished, for example, by the property that, with respect to them, the motion of a particle on which no external force is acting takes place with constant velocity), and ***Cartesian*** coordinate systems (that is, linear coordinate systems for which the three coordinate axes are orthogonal). But it is also possible to use other, non-linear, coordinate systems, such as, for example, the well known spherical and cylindrical coordinates. Such systems often are not "valid for the whole space." (For example, if Q is a two-dimensional space, and we represent Q as $\mathbb{R}^2$ by choosing Cartesian coordinates, then we can also use *polar coordinates*, but the corresponding map $Q \ni (x,y) \to \kappa(x,y) = (r,\theta)$ will be defined on Q minus a "cut-off" C—often taken to be the half-line $C = \{(x,y) : y = 0, x \leq 0\}$—and will map the remaining region $Q\backslash C$ to the product $]0,+\infty[\times I$, where I is an open interval of length 2π).

The simplest and most common examples of configuration spaces are the spaces $\mathbb{R}^n$ themselves, as well as the line $\mathbb{E}^1$, the plane $\mathbb{E}^2$, and the three-dimensional space $\mathbb{E}^3$ of Euclidean Geometry. We emphasize that the spaces $\mathbb{E}^k$ of Euclidean geometry are ***not*** the same objects as their close relatives $\mathbb{R}^k$. For example, $\mathbb{R}^3$ has a special point called "the origin," whereas $\mathbb{E}^3$ does not. The true mathematical structure of $\mathbb{E}^3$ is that of an ***abstract three-dimensional Euclidean space***. That is, $\mathbb{E}^3$ ***is a three-dimensional real affine space***[7] ***endowed with a Euclidean metric***[8].

Physical space P serves as the configuration space for a point particle. For a

---

dimensional "physical space" and the one-dimensional time axis T. Classical mechanics really is more complicated, because there is no absolute rest, so physical space-time S is four-dimensional and is not endowed with a canonical product decomposition $S = Q \times T$. There is a natural class of coordinate systems $\kappa : S \mapsto \mathbb{R}^3 \times \mathbb{R}$, called ***inertial systems***. They are related to one another by the so-called ***Galilean transformations***, but these transformations do not preserve the decomposition of S a a product of "space" and "time."

[7]A ***real affine space*** is like a real linear space but without a privileged origin. More precisely, a real affine space is a set A endowed, for each choice of an "origin" a of A, with a structure L_a of a linear space having a as its origin. The structures L_a have to satisfy some obvious compatibility conditions. For example, one must require that the midpoint relative to the linear structure L_a of the segment joining two points P_1 and P_2 be independent of a, that is, that $\frac{1}{2}(P_1 +_a P_2) = \frac{1}{2}(P_1 +_b P_2)$ for all $a,b \in A$. More generally, the sums $\alpha_1 P_1 +_a \ldots +_a \alpha_m P_m$ such that $\sum_{i=1}^m \alpha_k = 1$ are called ***affine combinations***, and one requires that all such sums be independent of a.

[8] A ***Euclidean metric*** on an affine space A is a way of measuring distances between points of A (that is, a "metric" on A, i.e., a map $A \times A \ni (a,b) \mapsto d(a,b) \in [0,+\infty[$ that satisfies $d(a,b) = 0 \Leftrightarrow a = b$, $d(a,b) = d(b,a)$, and $d(a,c) \leq d(a,b) + d(b,c)$), that in addition possesses three properties that relate it to the affine structure of A, namely, (1) the identity $d(b, r.a + (1-r).b) = rd(b,a)$, (2) the "parallellogram law" $4d(a, \frac{1}{2}b + \frac{1}{2}c)^2 + d(b,c)^2 = 2(d(a,b)^2 + d(a,c)^2)$, and, (3) the translation invariance property $d(b+a-a', c+a-a')) = d(b,c)$, which have to be valid whenever $a,a',b,c \in A$, $r \in \mathbb{R}$ and $r \geq 0$.

long time[9] it was believed that P could be accurately modelled as a three-dimensional Euclidean space E, which could then be identified (in a non-canonical way) with $\mathbb{R}^3$ by choosing Cartesian coordinates. This view has changed with the advent of Einstein's theory of relativity, although mathematical pioneers such as Gauss had already envisioned the possibility that the geometry of physical space might be non-Euclidean. Here, we will stick to the traditional point of view of classical mechanics, in which P is regarded as a three-dimensional Euclidean space. Similarly, to study a system of N point particles the natural choice of configuration space is the product $Q = P^N$ of N copies of P, which can be identified with $\mathbb{R}^{3N}$ by choosing Cartesian coordinates on P.

There are, in addition, many important configuration spaces that cannot be naturally identified with a space $\mathbb{R}^n$ or even with with an open subset of $\mathbb{R}^n$. For example, there are many systems (such as wheels or compass needles) whose natural configuration space consists of *directions* rather than numbers or points of some Euclidean space. For such systems, the natural choice of a configuration space Q is the unit circle in a two-dimensional Euclidean space E. By choosing Cartesian coordinates on E, one can identify this configuration space with $\mathbb{S}^1$, the unit circle in $\mathbb{R}^2$, given by

$$\mathbb{S}^1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}.$$

(This identification of Q with $\mathbb{S}^1$ depends on choosing coordinates. In general Q does not have a natural, or distinguished point, whereas $\mathbb{S}^1$ does. For example, $\mathbb{S}^1$ can be regarded as the set of complex numbers z such that $|z| = 1$, and complex multiplication turns $\mathbb{S}^1$ into a group, having the complex number $z = 1$ as identity element. A better way to view the configuration space of a wheel is as a *principal space* over $\mathbb{S}^1$, cf. Footnote 12.)

A configuration space such as $\mathbb{S}^1$ is an example of a *differentiable manifold* that cannot be represented[10] as an open subset of $\mathbb{R}^n$ for any n.

For a second, more sophisticated example of a configuration space which is

[9]With modern hindsight, one could perhaps attempt to interpret the Earth-centered physical space of pre-Copernican science as a ***linear*** rather than an affine space, because it had a natural center. The Copernican revolution did not substantially alter this geometric structure, although it put the Sun rather the Earth as the center. The truly modern view that the Universe has no privileged center was expressed, for example, by Giordano Bruno, who speculated that the universe is infinite, that the stars we see in the night sky are suns like ours, and that they probably have planets inhabited by intelligent civilizations. For espousing these heretical ideas, Bruno was sentenced to die at the stake by representatives of the best minds of the allegedly intelligent civilization that inhabits the planet Earth, and was killed on February 17, 1600. Due to the seriousness of Bruno's offense, his judges denied him the milder form of execution often used for other crimes, in which the victim was strangled first so that he would be spared the extreme suffering of being burnt alive.

[10]Precisely: there does not exist a triple (n, Ω, κ) such that $n \in \mathbb{N}$, Ω is an open subset of $\mathbb{R}^n$, and $\kappa : \mathbb{S}^1 \mapsto \Omega$ is a homeomorphism. The reason is, simply, that $\mathbb{S}^1$ is compact whereas a nonempty open subset of $\mathbb{R}^n$ cannot be compact unless $n = 0$, in which case Ω would consist of a single point.

naturally a differentiable manifold, let us suppose that we want to study the motion of a rigid body B in three-dimensional space. Then we need three coordinates (x,y,z) for the position of the center of mass c of B, and in addition we have to specify the ***orientation*** of B, by indicating how B has been rotated with respect to some fixed system F of Cartesian coordinates in space. This can be done by choosing a Cartesian coordinate system F' whose origin is c, that moves rigidly with B (so that, for instance, the x axis of F' goes from c to some fixed point P_x on the boundary of B). At each time t, the F' system will be related to the F system by a rotation matrix[11] $R(t)$. Therefore the contribution of the orientation to the configuration variable q of B is a rotation matrix, i. e., a member R of the three-dimensional orthogonal group $O(3)$. Hence a natural choice[12] of a configuration space for rigid body motion is

$$Q = \mathbb{R}^3 \times O(3)\,.$$

In our subsequent discussion, we will always deal with configuration spaces that are either (a) open subsets of $\mathbb{R}^n$ for some n, or (b) differentiable manifolds. The readers who are not familiar with differentiable manifolds can limit themselves to the case of open subsets of $\mathbb{R}^n$, but those who are used to working with differentiable manifolds are advised to pay attention to this more general setting, because the manifold formulation is mathematically more natural and elegant.

Remark II.1. Considering configuration spaces of dimensions higher than 3 is natural in today's mathematics, but the idea that one might as well write $(q^1,q^2,\ldots,q^n)$ rather than x, or (x,y), or (x,y,z), and "do everything in n-dimensions, with n

[11]Recall that an n-dimensional ***rotation matrix***, or ***orthogonal matrix***, is a real $n\times n$ matrix $R=(r_{ij})_{1\le i,j\le n}$ such that $\sum_{j=1}^n r_{ij}r_{kj}=\delta_{ik}$ for all $i,k\in\{1,\ldots,n\}$, where $\delta_{ik}=1$ if $k=i$, and $\delta_{ik}=0$ if $k\neq i$.

[12]To be precise, this already involves a previous choice of coordinates for physical space, as well as the choice of the frame F. To make a truly "intrinsic" choice of Q one needs some more mathematical machinery. Physical space is not really $\mathbb{R}^3$ but an ***abstract three-dimensional Euclidean space*** E (cf. Footnote 8). And the set $\mathscr{F}$ of "frames" should not be $O(3)$ itself but a ***principal space over*** $O(3)$, that is, a space S that can be "canonically identified" with $O(3)$ after a particular member F of S has been selected, but which—unlike $O(3)$ itself—does not have a privileged "origin," or "identity element." In other words, S is an object such that, given any two members F_1, F_2 of S, there is one and only one rotation matrix $R\in O(3)$ such that $F_1\cdot R=F_2$. The mathematically precise notion is that S is a manifold endowed with a an action $S\times O(3)\ni(F,R)\mapsto F\cdot R\in S$ of $O(3)$ which is transitive (that is, given any two members F_1, F_2 of S, there is an $R\in O(3)$ such that $F_1\cdot R=F_2$) and faithful (that is, R is unique). So the correct choice of Q is $E\times S$.

completely arbitrary," was by no means natural to the founders of the subject and took a long time to evolve. Lagrange, for example, devotes a lot of space in his *Mécanique Analytique* to persuading the reader that a system of N point particles in 3-dimensional space can be thought of as a single point evolving in a space of $3N$ coordinates. □

Remark II.2. One can also consider problems in which q is a vector in an infinite-dimensional space, and a lot of the current activity deals with such questions. However, *in this book we will only discuss finite-dimensional problems.* □

II.4 Spaces of curves

Once we have chosen our configuration space Q, the problem of minimizing the functional $\xi \mapsto J(\xi)$ only makes precise sense once we have specified the set Ξ of curves ξ in Q within which the minimization is to be carried out.

For most interesting problems, the minimization in the space of all curves, with no restrictions whatsoever, would be devoid of interest. For example, the shortest paths in the plane are the single points, i. e., the paths of length zero. To get a truly interesting "shortest path" problem one needs to impose a *boundary condition*. The simplest such condition consists of *fixing the endpoints* $\bar{q}$, $\hat{q}$ of the path, and looking for the shortest path in Q going from $\bar{q}$ to $\hat{q}$.

Moreover, even after we have fixed the endpoints $\bar{q}$, $\hat{q}$, it still does not make sense to work in the space of "all paths in Q going from $\bar{q}$ to $\hat{q}$." Indeed, if we take the word "path" to mean "continuous curve," as is often done, then the functional J given by (II.1) will not have a well defined value for most paths ξ, because in general a continuous path ξ will not have a derivative $\dot{\xi}(t)$ defined at all values of t. So one has to work with a more restricted class of paths. It then turns out that some of the most obvious choices are not the right ones. For example, one may think of taking Ξ to be the space of all paths ξ in Q going from $\bar{q}$ to $\hat{q}$ and such that $\dot{\xi}$ exists everywhere, or one might limit the paths ξ even further by requiring in addition that $\dot{\xi}$ be bounded, or continuous. These spaces of paths turn out to be too restrictive, since it is very easy to give examples of problems for which "the solution" has points where it is not differentiable. (See, for instance, Example V.7 below, where the solution has "corners.")

In twentieth century mathematics, it has become clear that one cannot talk about "all curves," and that one has to work curves in some suitably chosen function space, selected with great care, such as that of all absolutely continuous curves, or that of all Lipschitz curves. Furthermore, it is now understood that the choice of function space can sometimes be crucial, as can be seen from our discussion of the Lavrentiev phenomenon in §IX.4 below.

The founders of the calculus of variations did not explicitly address this question, and usually talked about a minimization in the class of "all curves." This is understandable, because the modern mathematical concept of a function—which need not be continuous or even measurable, and *a fortiori* need not be everywhere differentiable, and in fact might even be *nowhere differentiable*[13]—was only developed in the late nineteenth century.

Here, we shall adopt the following compromise solution: as long as we are discussing the formal aspect of the necessary conditions for an optimum, we will work with an unspecified space Ξ of curves, in the classical calculus of variations, or with an unspecified set $\mathscr{C}$ of trajectory-control pairs, in optimal control. Later, when we get to the rigorous statement of the results and their proofs, we will be careful about the precise choice of Ξ and $\mathscr{C}$.

II.5 The conventional wisdom

The calculus of variations is supposed to be approximately three hundred years old, whereas optimal control theory is said to be a much younger subject, born about four decades ago in the now extinct Soviet Union, with the work on the "Pontryagin maximum principle" by L. S. Pontryagin and his group (cf. [112]), or perhaps a few years earlier, with the work of McShane and Hestenes, cf. [101, 72]. Moreover, there is a widespread belief among mathematicians that optimal control was no more than a minor addition to the classical calculus of variations, essentially involving the incorporation of inequality constraints.

Remark II.3. The article by L. Markus in [55] gives a nice account of the unenthusiastic reaction at the 1958 ICM (International Congress of Mathematicians) to the announcement of the maximum principle by the Soviet group. Several nonmathematical factors may have contributed to the negative reaction. Among these, two clearly stand out: first of all, Pontryagin's personality and, in particular, his notorious antisemitism, and second, the impression—undoubtedly enhanced by the obvious overtones of the word "control"—that the result was primarily intended for military applications. □

Within the almost unanimous consensus that the calculus of variations was born approximately three hundred years ago, it is a very broadly held view that the birth

[13] Weierstrass showed that there are continuous functions $[0,1] \ni t \mapsto \mathbb{R}$ that are not differentiable at any value of t.

took place in 1696-97. For example, G. A. Bliss wrote in 1925 (cf. [17]):

> It is fair to say that *the theory of the calculus of variations had its beginnings in the interesting brachistochrone problem of John Bernoulli.* One should not infer from this that no problems of the calculus of variations were known earlier, for ... Newton had proposed such a problem and described a characteristic property of its solution. Furthermore, the brachistochrone problem itself was definitely in the mind of Galileo in 1630 and 1638 ... A still older problem of the calculus of variations is the isoperimetric problem of the ancient Greeks. *None of these, however could rightly be regarded as the starting point of the theory of the calculus of variations*, for in the early references to them there were no indications of other problems of similar type, or of methods of solution possessing generality of application. [Emphasis ours.]

A similar assessment is made by J. Stillwell in his recent textbook (cf. [122], p. 174), where it is asserted of "James[14] Bernoulli's solution" of the brachistochrone problem—rather than Johann's—that "it is now considered to be the first major step in the development of the calculus of variations." D. J. Struik, in [124], p. 392, says of the articles published in the May 1697 issue of *Acta Eruditorum* that "these papers opened the history of a new field, the calculus of variations."

Remark II.4. Not all historians of mathematics agree with this view. For some, it is Newton's earlier (1685) investigation of the shape of a body with minimal drag that marks the birth the calculus of variations. For others (for example, M. Kline, cf. [81]), this momentous event did not take place until much later, when Euler, in 1760, officially baptized the newborn by using for the first time the words "calculus of variations." □

II.6 An unconventional alternative

We will claim that

> *The brachistochrone problem truly belongs to optimal control theory, and therefore optimal control is at least three hundred years old, and possibly much older.*

[14] that is, Jakob

This is, of course, a much more controversial thesis than merely claiming that 1696-97 marks the birth of the calculus of variations.[15] Since extraordinary claims must be supported by strong evidence, we set out to present our reasons. But first we must give a precise characterization of the subject matters of the calculus of variations and optimal control.

II.7 Standard calculus of variations problems

The calculus of variations deals mainly with optimization problems involving a function $(q,\dot{q},t) \mapsto L(q,\dot{q},t)$ of "position, velocity, and time. Nowadays, it is customary to refer to this function as the ***Lagrangian.***

The objective of a typical calculus of variations problem is to find a curve $t \mapsto \xi(t)$ in q-space that minimizes the integral

$$I = \int_a^b L(\xi(t),\dot{\xi}(t),t)dt$$

of the Lagrangian, among all curves that satisfy some extra conditions. As explained in §II.4, "all curves" really means "all curves in some given space of curves Ξ".

For example, one could specify a time interval $[a,b]$, an initial point $\bar{q}$, and a terminal point $\hat{q}$, and seek to minimize I in the class of all curves $\xi : [a,b] \to \mathbb{R}^n$ sucht that $\xi(a) = \bar{q}$ and $\xi(b) = \hat{q}$. This is a "standard calculus of variations problem":

$$\boxed{\begin{array}{rl} \text{Given} & Q, L, a, b, \bar{q}, \hat{q}, \Xi\,, \\[2ex] \text{minimize} & I = \int_a^b L(\xi(t),\dot{\xi}(t),t)dt\,, \\[2ex] \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\[2ex] \text{and} & \xi \in \Xi\,. \end{array}} \qquad \text{(II.5)}$$

Alternatively, one may prefer to avoid using the symbol $\dot{q}$ as variable, and seek to make it transparently clear that L is "a function of three independent variables"— namely, the configuration point q, the velocity vector u, and time t— by using a

[15] A view closely related to ours is expressed by J. L. Casti, an author who has written extensively on systems, models, paradigms, reality, nature, and man. In his remarkable book [37] on alternate realities, Casti indeed observes that "classical optimal control theory dates back at least to the well-known 'brachistochrone' problem."

letter such as u rather than $\dot{q}$ as the name for the velocity variable. This leads to regarding the Lagrangian as a function

$$(q,u,t) \mapsto L(q,u,t)$$

and rewriting (II.5) in the equivalent form

$$\boxed{\begin{array}{rl} \text{Given} & Q,\, L,\, a,\, b,\, \bar{q},\, \hat{q},\, \Xi\,, \\ \text{minimize} & I = \int_a^b L(\xi(t),\eta(t),t)dt\,, \\ \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\ & \dot{\xi}(t) = \eta(t) \text{ for } a \le t \le b\,, \\ \text{and} & \xi \in \Xi\,. \end{array}} \tag{II.6}$$

The vector variable q is the *state*, or *configuration*, of the problem (II.5) or (II.6). The set Q of all possible values of q is the *configuration space*, or *state space*, of the problem. One can allow Q to be a subset of a Euclidean space $\mathbb{R}^n$, or a more general differentiable manifold, such as a spherical surface. So we will stipulate that

- the configuration space Q is either $\mathbb{R}^n$, or an open subset of $\mathbb{R}^n$, or, more generally, a smooth manifold;
- if Q is open in $\mathbb{R}^n$, then the variable u—the "velocity"—takes values in $\mathbb{R}^n$, and L is a real-valued function on $Q \times \mathbb{R}^n \times [a,b]$;
- if Q is a manifold then the values of u are, for each $q \in Q$, the members of the tangent space T_qQ of Q at q—so that in fact the pair (q,u) should be thought of as the name of a variable, say V, with values in TQ, the tangent bundle of Q—and L is a real-valued function on $TQ \times [a,b]$;
- Ξ is the space of all curves $[a,b] \ni t \mapsto \xi(t) \in Q$ that have some extra regularity property, such as continuous differentiablity, piecewise continuous differentiablity, absolute continuity, or Lipschitz continuity.

Remark II.5. A minimization problem of the form (II.5)—or (II.6)—is a *point-to-point* calculus of variations problem, because the endpoint constraint on the trajectory $\xi(\cdot)$ specifies that $\xi(\cdot)$ has to start at a given point $\bar{q}$ and end at another given point $\hat{q}$. One can of course consider, both for the calculus of variations and for optimal control, more general *set-to-set* problems, in which the endpoint constraints are

$$\xi(a) \in S_1 \quad \text{and} \qquad \xi(b) \in S_2\,, \tag{II.7}$$

where S_1, S_2 are two given subsets of Q. Even more generally, one can consider "mixed" endpoint constraints of the form

$$(\xi(a),\xi(b)) \in S\,, \tag{II.8}$$

where S is a given subset of $Q \times Q$. Such more general problems have been studied since the very beginning of the calculus of variations, and give rise to interesting questions about "transversality conditions." In this book, however, *we will only deal with calculus of variations (and optimal control) point-to-point problems*, since all the points we wish to make can already be made in this special situation. This means that we are leaving out the whole area of transversality conditions. $\square$

II.8 Isoperimetric problems

In addition to the standard problems (II.5) or (II.6), the more general class of "isoperimetric problems" was considered from the early days of the calculus of variations. These are problems of the form

$$\boxed{\begin{array}{rl}
\text{Given} & Q, L^0, L^1, \ldots, L^m, \ell^1, \ldots, \ell^m, a, b, \bar{q}, \hat{q}, \Xi\,, \\[2ex]
\text{minimize} & \int_a^b L^0(\xi(t),\dot{\xi}(t),t)\,dt\,, \\[2ex]
\text{subject to} & \int_a^b L^1(\xi(t),\dot{\xi}(t),t)\,dt = \ell^1\,, \\
& \qquad\vdots \\
& \int_a^b L^m(\xi(t),\dot{\xi}(t),t)\,dt = \ell^m\,, \\[2ex]
& \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\[2ex]
\text{and} & \xi \in \Xi\,.
\end{array}} \tag{II.9}$$

We will show later—in §II.10— that *once one makes the transition from the classical calculus of variations to optimal control, the distinction between "standard" and "isoperimetric" problems disappears*, because the addition of extra integral constraints to a "standard optimal control problem" P just leads to another "standard optimal control problem" P'.[16]

II.9 Optimal control problems

The distinctive feature of classical calculus of variations problems is that the minimization of (II.5) or (II.6) takes place in the space of "all curves" (but cf. §II.4), so nothing interesting happens on the level of the set of curves under consideration, and all the nontrivial features of the problem arise because of the Lagrangian L. In other words, for calculus of variations problems the ***kinematics*** (that is, the specification of which motions in q-space are permitted, or "feasible") is trivial and uninteresting, amounting to no more than the statement that one can move in all possible directions at all possible speeds, subject only to some regularity conditions (such as differentiablity) of the resulting path.

Optimal control problems, by contrast, involve a minimization over a set $\mathscr{C}$ of curves which is itself determined by ***dynamical constraints*** of a particular kind. Precisely, we assume that an "agent," or "controller," can choose freely the value of a parameter u belonging to a "control set" U. This choice can actually be made separately at each time, so the controller is in fact selecting a *control function* $I \ni t \mapsto \eta(t) \in U$ on some time interval I. We also assume that the control affects the evolution of q in such a way that q obeys a differential equation

$$\boxed{\dot{q} = f(q,u,t)} \tag{II.10}$$

Then $\mathscr{C}$ would consist of all curves $t \mapsto \xi(t)$ that satisfy the differential equation

$$\dot{\xi}(t) = f(\xi(t), \eta(t), t) \tag{II.11}$$

for some choice of the control function $t \mapsto \eta(t) \in U$.

Even more precisely: it may happen that a member ξ of $\mathscr{C}$ does not uniquely determine the control η for which (XVI.1) holds (cf. Example II.10 below). It is therefore better to take $\mathscr{C}$ as a set of *trajectory-control pairs* $(\xi(\cdot), \eta(\cdot))$.

An ordinary differential equation such as (XVI.1) defines a *control system*, or *controlled dynamics*, or *controlled dynamical system*[17].

As will be explained below, controlled dynamical systems of the form (XVI.1) are interesting objects to study even in the absence of a minimization problem.

[16]Although P' will in general not be a "standard" calculus of variations problem, even if P is.

[17] Precise definitions of the basic control concepts, and a discussion of the issue of the adequate choice of the set $\mathscr{C}$ of "admissible trajectory-control pairs," will be given in the appendix.

Thus ***control theory is a much richer subject than optimal control***, which deals with only one of many kinds of important control problems.

A standard "point-to-point" optimal control problem has the form

$$\boxed{\begin{array}{ll} \text{Given} & Q, U, f, L, a, b, \bar{q}, \hat{q}, \mathscr{C}\,, \\[2ex] \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)\,dt\,, \\[2ex] \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\[2ex] & (\xi, \eta) \in \mathscr{C}\,, \\[2ex] \text{and} & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \le t \le b\,. \end{array}} \tag{II.12}$$

One can also consider "variable time" problems, in which a and b are not given, and finding them is part of the problem. As for the classical calculus of variations, the endpoint conditions can also be more general. All these possibilities are subsumed in the following general form of an optimal control problem:

$$\boxed{\begin{array}{ll} \text{Given} & Q, U, f, L, S, \mathscr{C}\,, \\[2ex] \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)\,dt\,, \\[2ex] \text{subject to} & (\xi(a), a, \xi(b), b) \in S\,, \\[2ex] & (\xi, \eta) \in \mathscr{C}\,, \\[2ex] \text{and} & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \le t \le b\,. \end{array}} \tag{II.13}$$

Here S is a subset of $Q \times \mathbb{R} \times Q \times \mathbb{R}$.

Example II.6. Consider the following controlled dynamics in $\mathbb{R}^3$:

$$\begin{array}{rcll} \dot{x}_1 & = & u_1 \cdot \cos\theta\,, & \\ \dot{x}_2 & = & u_1 \cdot \sin\theta\,, & \qquad |u_1| \le 1\,,\ |u_2| \le 1 \\ \dot{\theta} & = & u_2\,. & \end{array} \tag{II.14}$$

This represents a "vehicle" such as a ***unicycle***. The vehicle has a *position* (x_1,x_2) in the plane $\mathbb{R}^2$ (for example, the position of its center of mass) as well as an *orientation* θ, which is an angle[18], thought of as the direction towards which the vehicle is pointing. The first two equations of the system (II.14) express the fact that the vehicle moves in the direction of the vector $\mathbf{v} = (\cos\theta, \sin\theta)$, with a speed u_1 that can be selected by the driver. The third equation captures the fact that the driver can also steer the vehicle by changing its direction. The variables u_1 and u_2 represent the two controls available to the driver, namely, "speed" and "steering."

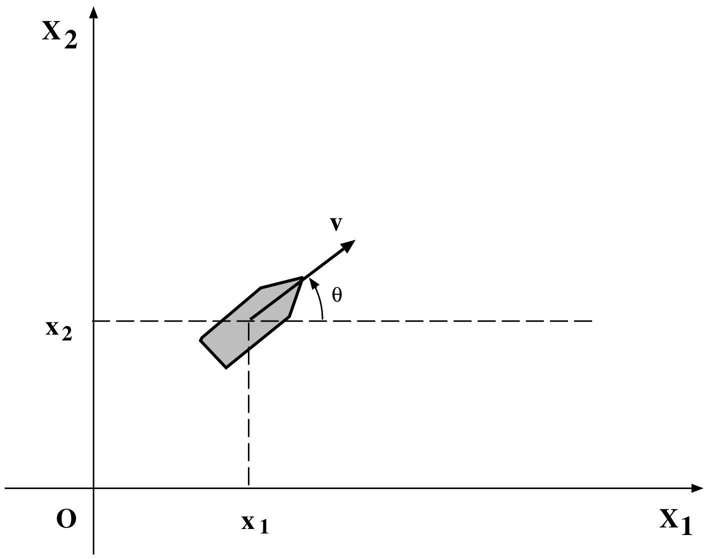


FIGURE 4. The unicycle

The dynamical behavior described by (II.14) is already an object worth studying even before we associate to it an optimal control problem by asking to optimize something. For example, we could ask the following interesting questions.

The complete controllability question: Given any two vehicle configurations $\bar{q} = (\bar{x}_1,\bar{x}_2,\bar{\theta})$, $\hat{q} = (\hat{x}_1,\hat{x}_2,\hat{\theta})$, can the driver always steer the vehicle from $\bar{q}$ to $\hat{q}$?

The parking question: Can the driver make the vehicle move sideways, that is,

[18]Which means that, if we want to be completely accurate, we should really take the configuration space Q to be $\mathbb{R}^2 \times \mathbb{S}^1$ rather that $\mathbb{R}^3$.

go from an initial configuration $\bar{q} = (\bar{x}_1, \bar{x}_2, \bar{\theta})$ to a terminal configuration $\hat{q} = (\hat{x}_2, \hat{x}_2, \bar{\theta})$ such that $(\hat{x}_1, \hat{x}_2) = (\bar{x}_1, \bar{x}_2) + R(-\sin\bar{\theta}, \cos\bar{\theta})$ for some nonzero number R?

The effective complete controllability question: Assuming that the answer to the complete controllability question is "yes," can we produce an effective algorithm enabling the driver to compute, for any given $\bar{q}$, $\hat{q}$, a control strategy $t \mapsto (u(t), v(t))$ that will drive the vehicle from $\bar{q}$ to $\hat{q}$?

The effective parking question: How can the driver effectively compute a strategy for making the vehicle move sideways?

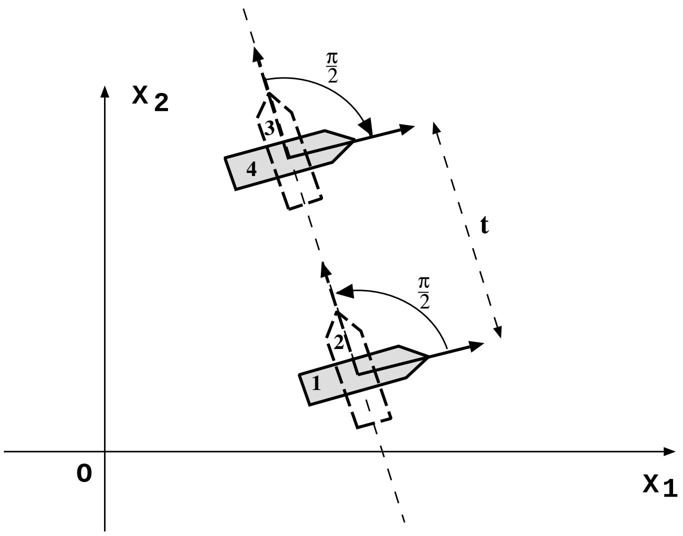


FIGURE 5. Moving sideways

In this case, it turns out that the answers to the above questions are all very simple. For example, a solution to the effective parking question is as follows: to move sideways to the left a distance **t**, from a given initial configuration 1, move to a configuration 2 by making a 90^o turn to the left (that is, using $u = 0$, $v = 1$ during $\frac{\pi}{2}$ units of time), then move forward **t** units of length at full speed to configuration 3 (that is, use $u = 1$, $v = 0$) during **t** units of time), and then reorient the vehicle into the desired configuration 4 by making a 90^o turn to the right. Once the effective parking question is solved, the effective complete controllability question becomes trivial, because the dynamical equations (II.14) already tell us how to move in two of the three linearly independent directions that exist at each point of $\mathbb{R}^3$ (namely, forward and rotating), and motion in the third "sideways" direction is achieved by our solution of the parking problem.

Example II.7. Consider the following controlled dynamics in $\mathbb{R}^3$:

$$\begin{aligned} \dot{x}_1 &= \cos\theta\,, \\ \dot{x}_2 &= \sin\theta\,, \qquad |u| \leq 1\,, \\ \dot{\theta} &= u\,. \end{aligned} \tag{II.15}$$

This represents a vehicle similar to that of the preceding example, except that now the driver is only allowed to steer by changing the orientation, the position (x_1, x_2) only varies by forward motion at a fixed speed.

The four questions of Example II.6 can still be asked, and the answers are still positive, even though now it is much harder to control the vehicle, because backwards motion is not available, and the vehicle is not even allowed to "stay in place while turning," as in the first and third steps of Figure 5.

Example II.8. For the control systems described in Examples II.6 and II.7, we could consider the ***minimum time problem***: given two vehicle configurations $\bar{q} = (\bar{x}_1, \bar{x}_2, \bar{\theta})$, $\hat{q} = (\hat{x}_1, \hat{x}_2, \hat{\theta})$, steer the vehicle from $\bar{q}$ to $\hat{q}$ in minimum time.

These are optimal control problems of the form (II.13). The dynamical law f is given by the right-hand sides of the equations of (II.14) and (II.15). The Lagrangian L is just identically equal to 1. The set S is equal, for given $\bar{q}$ and $\hat{q}$, to $\{\bar{q}\} \times \mathbb{R} \times \{\hat{q}\} \times \mathbb{R}$, expressing the fact that the initial time a and the terminal time b are not prescribed. (One could also fix the initial time to be 0, and take S to be $\{\bar{q}\} \times \{0\} \times \{\hat{q}\} \times \mathbb{R}$.)

Remark II.9. The minimum time problem for the system (II.15), in which the vehicle can only move forwards, is known as ***Dubins' problem***. It was solved by L. Dubins in his 1957 paper [52]. The problem for the system (II.14), in which the vehicle can move forwards and backwards, is known as the ***Reeds-Shepp problem***. It was solved by J.A. Reeds and L.A. Shepp in their 1990 paper [115].

The purpose of the following two utterly trivial examples is to illustrate the fact that giving a curve $t \mapsto \xi(t)$ in q-space that arises from a control does not suffice to determine the control. It is for this reason that we have to talk about "trajectory-control pairs" rather than "trajectories."

Example II.10. Consider a control system Σ defined by the dynamical equation

$$\dot{q} = uq\,, \tag{II.16}$$

with $q \in \mathbb{R}$, $u \in \mathbb{R}$. Then the trajectory $t \mapsto \xi_*(t)$ given by

$$\xi_*(t) \equiv 0$$

does not arise from a unique control function. Indeed, $[a,b] \ni t \to \eta(t) \in \mathbb{R}$ is any function, then ξ_* satisfies $\dot{\xi}_*(t) = \eta(t)\xi_*(t)$ for every t, so ξ_* is a trajectory of Σ for the control η. □

Example II.11. Consider a one-dimensional optimal control problem whose dynamical behavior is defined by the same equation (II.16) as in Example II.10. Suppose we wish to minimize the integral cost $\int_0^1 \eta(t)^2\,dt$, subject to the end-point constaints $\xi(0) = \xi(1) = 0$ and the dynamical constaint $\dot{\xi}(t) = \eta(t)$ for all $t \in [0,1]$. Clearly, the minimum value of the integral is 0, which is achieved by taking $\eta(t) \equiv 0$. So the solution of our problem is the trajectory-control pair $(0,0)$ (that is, $\xi(t) \equiv 0$ and $\eta(t) \equiv 0$). It would not be correct, however, to say that "the solution of our problem is the trajectory ξ given by $\xi(t) \equiv 0$," because this curve is also a trajectory for any other control function η, and only the choice of $\eta(t) \equiv 0$ minimizes the cost. □

II.10 Isoperimetric calculus of variations problems are standard optimal control problems

Exactly as we did in §II.8 for classical calculus of variations problems, we could consider enlarging the class of optimal control problems by adding to it the "isoperimetric optimal control problems." That is, given a standard optimal control problem (II.13) and a collection of m additional integral constraints $I^1 = 0, \dots, I^m = 0$, where

$$I^j = \int_a^b L^j(\xi(t), \eta(t), t)\,dt\,, \tag{II.17}$$

we could consider the corresponding "isoperimetric problem":

$$\boxed{\begin{array}{ll} \text{Given} & Q, U, f, L^0, L^1, \ldots, L^m, \ell^1, \ldots, \ell^m, S, \mathscr{C}, \\ \text{minimize} & I = \int_a^b L^0(\xi(t), \eta(t), t)\, dt, \\ \text{subject to} & (\xi(a), a, \xi(b), b) \in S, \\ & (\xi, \eta) \in \mathscr{C}, \\ & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \le t \le b, \\ \text{and} & \int_a^b L^j(\xi(t), \eta(t), t)\, dt = \ell^j, \text{ for } j = 1, \ldots, m. \end{array}} \tag{II.18}$$

It turns out, however, that there is no need to introduce "isoperimetric optimal control problems" as a class separate from that of "standard optimal control problems." In fact, *"isoperimetric optimal control problems" are again standard optimal control problems.*

To see this, suppose we are given a problem (II.18). We can then augment the state space Q of (II.12) by adding a new m-vector $r = (r^1, \ldots, r^m)$ of state variables, subject to the dynamical laws

$$\dot{r}^j = L^j(q, u, t), \quad j = 1, \ldots, m. \tag{II.19}$$

If we let

$$Q^* = Q \times \mathbb{R}^m, \qquad L^*(q, r, u, t) = L^0(q, u, t), \tag{II.20}$$

$$q^* = \begin{bmatrix} q \\ r^1 \\ \vdots \\ r^m \end{bmatrix}, \; f^*(q^*, u, t) = \begin{bmatrix} f(q, u, t) \\ L^1(q, u, t) \\ \vdots \\ L^m(q, u, t) \end{bmatrix}, \tag{II.21}$$

and

$$S^* = \left\{ \left(\begin{bmatrix} \bar{q} \\ \bar{r} \end{bmatrix}, a, \begin{bmatrix} \hat{q} \\ \hat{r} \end{bmatrix}, b \right) : (\bar{q}, a, \hat{q}, b) \in S, \bar{r} = 0, \hat{r} = \begin{bmatrix} \ell^1 \\ \vdots \\ \ell^m \end{bmatrix} \right\}, \tag{II.22}$$

then we see that (II.18) is equivalent to a standard optimal control problem (II.13), with Q, q, f, L, S, replaced by their starred versions.

In particular,

Isoperimetric problems arising from standard classical calculus of variations or from standard optimal control problems are standard optimal control problems. On the other hand, isoperimetric problems arising from standard classical calculus of variations problems need not be standard classical calculus of variations problems.

II.11 Optimal control is much more general and richer than the calculus of variations

An optimal control problem has at least ***two*** ingredients that give the situation interesting structure, namely,

- the ***controlled dynamics*** f,

- the ***Lagrangian*** L, which gives rise to the ***cost functional*** I to be minimized.

This makes it richer than the calculus of variations, where the dynamics f is trivial—since $f(q,u,t) = u$—and the only source of interesting structure is the Lagrangian L.

In particular, optimal control theory contains, at the opposite extreme from the classical calculus of variations, problems where the "Lagrangian" L is identically equal to 1, i.e., completely trivial, and all the interesting action occurs because of the dynamics f. Such problems, in which it is desired to minimize time—i.e., the integral I of (II.13) with $L \equiv 1$—are called *minimum-time problems*. It is in these problems that the difference between optimal control and the calculus of variations is most clearly seen, and it is no accident that these were the problems that propelled the development of optimal control in the early 1960s, and that time-optimal control is prominently represented in today's research and in modern optimal control textbooks.

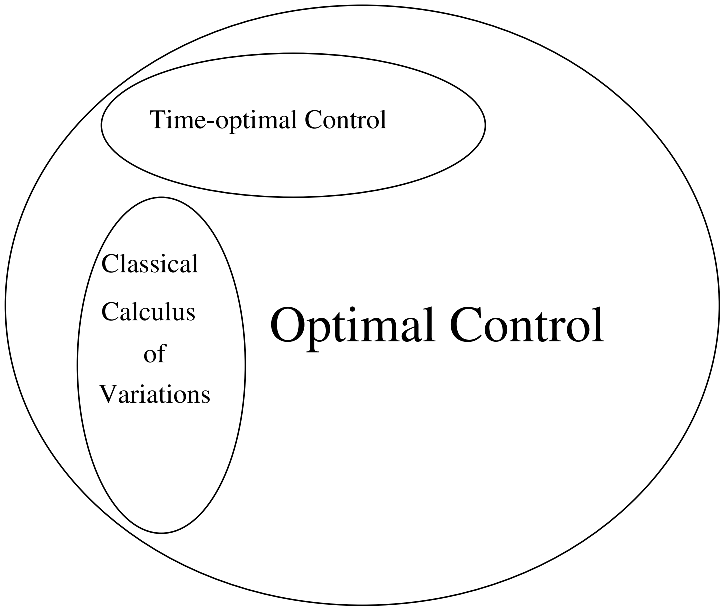


FIGURE 6.

II.12 To a controlled dynamics there is associated an interesting Lie-algebra structure

More recently, it has become apparent that, as R. Hermann (cf. [70]) and others (cf. Brockett [31], Hermes [71], Krener [83], Lobry [93], Jurdjevic-Sussmann [127, 77]) had argued since the 1960s and early 70s, a controlled dynamics such as (XVI.1) is, really, a *family of vector fields*, and such an object has to be studied using the machinery of differential geometry, which allows us to associate an *algebraic structure* to a family of vector fields, by looking at their Lie brackets. *This rich and interesting algebraic structure, arising from a controlled dynamics even in the absence of an optimization problem, is completely absent from the classical calculus of variations.*

II.13 Control is much more general than optimal control

In modern terms, one can think of a controlled dynamical system (XVI.1) as defining a generalization of the concept of a "continuous-time dynamical system," in which the future evolution of the system state q depends not only on the current state but also on the action of a "controlling agent," or "controller," who chooses the control function $t \mapsto \eta(t)$. One might think that a system defined by (XVI.1) is no more than a "dynamical system with parameters," i.e., an object not significantly different from an ordinary dynamical system, in the sense that, once the function $t \mapsto \eta(t)$ is chosen (XVI.1) reduces to an ordinary dynamical system. This, however, would miss the profound conceptual difference that exists between the study of dynamical systems and the kind of problems that interest control theorists.

For example, control theorists want to understand the natural questions of ***reachability*** and ***controllability***: for what pairs q_1, q_2 is it possible to choose $t \mapsto u(t)$ so as to transfer the state q of (XVI.1) from q_1 to q_2? For what systems f can this be done for all pairs q_1, q_2? Given a starting point $\hat{q}$, what does the set of all q such that $\hat{q}$ can be transferred to q look like?

And, most importantly, they are primarily interested in ***design questions***.

II.14 Design questions

Design problems are problems in which the objective is to "choose the control law u" so as to satisfy some requirements, or "specifications."

One possible design question is that of choosing the function $t \mapsto \eta(t)$ so as to minimize an integral I as before, and the study of this particular question is what ***optimal control theory*** is about. But this is just one example of a large number of possible design questions that are important for the engineer. Usually, more information is available at each time t than just the value of t, and one wants to *feed* all this information *back* into the control, so that the control law will in fact take the form $u = u(t, \iota)$, where ι represents the other information available at time t.

II.15 Feedback

Feedback control laws are laws in which the control is chosen as a "function" of the state of the system, or of that part of the state that is known through measurements.

The simplest form of feedback is "static state feedback," in which u is chosen as a function of the state q. (When you drive your car, you are applying a control u, consisting of the pressure exerted by your foot on the brake, the pressure on the gas pedal, the motion of the steering wheel, and other parameters. After you have driven from a point A to another point B, you can look back at your trajectory, and observe that the control u that you have applied was a certain function of time. This, however, does not mean that the control function $t \mapsto \eta(t)$ was specified by you *a priori*, and the motion of the car was then determined by solving (XVI.1). Your driving strategy involved choosing u as function of the **state** of the car (for example, "turn left when you get the corner of Broadway and 57th," or "slow down when your distance to B is smaller than a certain threshold and your speed is larger than a certain value"). This is **feedback**. A state feedback law, when applied with a given initial condition, will give rise *a posteriori* to an open-loop control. Conceptually, hwever, a state feedback is a very different kind of object.

We will explain below that the general concept of a feedback control is much broader than this, but before we do that we present a trivial example of state feedback.

Example II.12. Consider the dynamical behavior described by the equation

$$\dot{q} = u\,, \qquad q \in \mathbb{R}^n\,. \tag{II.23}$$

Here the vector u is arbitrary but subject to a bound $||u|| \le c$, where $c > 0$ is a constant. Suppose you want to reach the origin in minimum time. The obvious solution, starting at a point $\bar{q}$, is to take

$$\eta(t) = -c\frac{\bar{q}}{||\bar{q}||}\,. \tag{II.24}$$

Now think of a boat in a still lake, which can move in any direction with speed not greater than c, and suppose that your objective is to get to some point—say a small island—no matter where you may have started. An obvious physical solution would be to tie a rope to a pole located at the target point, and pull. Mathematically, this corresponds to choosing

$$u = \omega(q) = -c\frac{q}{||q||}\,. \tag{II.25}$$

There is an enormous difference between the solutions given by (II.24) and (II.25). In the former case, you compute the function $t \mapsto \eta(t)$ once and for all, and then follow it. If you made a small error in your initial estimation of $\bar{q}$, or if there is a small error in the dynamics—e.g., if the motion is really governed by

$$\dot{q} = u + v(q)\,,$$

where v is a small perturbation term, corresponding, for example, to the fact that the lake is not quite still—then you will miss the target. With the feedback solution (II.25) you will not. Whether or not the feedback solution "makes sense"—i.e., can be implemented—for a particular problem, is an engineering question, whose answer depends on such issues as whether there really is a pole in the target island, and enough rope is available. □

As we can see from this simple example, a control design problem may have different kinds of solutions, whose relevance depends on specific engineering aspects of the situation being modeled.

Example II.13. For a mathematically richer example, consider the equation

$$\ddot{x} + ax = u\,, \tag{II.26}$$

which governs the motion of a frictionless pendulum under the action of an external force u. (The coefficient a is a parameter, and we assume that $a > 0$.) Suppose our objective is "to drive x and $\dot{x}$ to 0." We can rewrite this as a 2-dimensional system

$$\dot{q} = f(q,u)\,,$$

by letting

$$q = \begin{bmatrix} q^1 \\ q^2 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix},$$

and

$$f(q,u) = \begin{bmatrix} q^2 \\ u - aq^1 \end{bmatrix},$$

so the equations of motion read:

$$\begin{cases} \dot{q}^1 & = & q^2\,, \\ \dot{q}^2 & = & u - aq^1\,. \end{cases}$$

It is easy—and completely uninteresting and useless—to see that

> *given any initial condition $(\bar{q}^1, \bar{q}^2)$, there exists a control function $t \mapsto \eta(t)$ that gives rise to a solution $t \mapsto \xi(t)$ such that $\xi(t) \to 0$ as $t \to +\infty$.*

(PROOF: Let $[0,\infty[\ni t \mapsto x(t) \in \mathbb{R}$ be any function of class C^2 that goes to 0 together with its first derivative, and satisfies $x(0) = \bar{q}^1$, $\dot{x}(0) = \bar{q}^2$. Then define $\eta(t) = \ddot{x}(t) + ax(t)$.)

Much more interesting is the question of choosing u as a *feedback control law*, that is, a "function" of the state variables q^1, q^2 (i.e., x and $\dot{x}$) so as to *stabilize* the system, i.e., to guarantee that all the solutions go to 0. One obvious way to do that is to *create a friction force*, for example by taking

$$u = -c\dot{x}, \quad \text{with} \quad c > 0.$$

This solution is realistic, as long as one can "have access" to the velocity $\dot{x}$. (For example, a possible physical implementation of the control law would be to add friction by just dipping the pendulum in a container full of water.) □

Example II.14. Let us consider the system of the previous example, but supposing now that only the position x is available to us, i.e., we want u to "depend on x only." How can we stabilize the pendulum? We recommend that the reader try answering this question as an exercise. The answer will be given below in Scholium VI, but we provide here a long hint.

First, a **WARNING**. Almost certainly, your first guess will be to look for a function

$$u = u(x)$$

of the position x such that all the solutions of the resulting equation

$$\ddot{x} + ax = u(x)$$

go to zero. But if you try this then you will not get very far, because *such a function cannot exist.* The reason is quite simple: define the *energy* $E(x,\dot{x})$ by

$$E(x,\dot{x}) = \dot{x}^2 + ax^2 - 2U(x),$$

where

$$U(x) = \int_0^x u(y)\,dy.$$

Then $E(x,\dot{x})$ *is constant along all trajectories*, because its derivative along a trajectory $t \mapsto x(t)$ is given by

$$\begin{aligned} \frac{d}{dt}E(x(t),\dot{x}(t)) &= 2\dot{x}(t)\ddot{x}(t) + 2ax(t)\dot{x}(t) - 2u(x(t))\dot{x}(t) \\ &= 2\dot{x}(t)\Big(\ddot{x}(t) + ax(t) - u(x(t))\Big) \\ &= 0. \end{aligned}$$

It follows that, if you start with an initial condition $q_{in} = (x(0), \dot{x}(0))$ such that $E(q_{in}) \neq 0$, then $x(t)$ and $\dot{x}(t)$ cannot go to zero.

This would seem to suggest that the answer to our question is that what we were asked to do cannot be done, i.e., that "the frictionless pendulum described by equation (II.26) cannot be stabilized by means of a feedback control law that depends on the position x alone." This is quite true *if by*

(I) a feedback control law that depends on the position x alone

we mean

(II) a law of the form $u = u(x)$.

But, if you read carefully the question as it was actually stated, you will notice that we never said that (I) is supposed to mean (II). When we talked about u "depending on x only," we did *not* say that this meant "choosing $u = u(x)$."[19] There are many other enginnering ways to implement a relation of "dependence of u on x." For example, in a dynamic setting u might be a *functional*, depending on past values of x as well as present ones. That is, at any time t the quantity u might depend on the whole collection $\{x(s) : s \leq t\}$ of values of x at all times prior to t.

This observation is the key to solving our the problem. We have to allow u at time t to be a function not just of x at time t but also of *past values* of x. A nice way to do that is to have a functional relation

$$u = h(x,y)\,,$$

where y is a new variable that obeys some dynamical equation

$$\dot{y} = g(x,y)\,.$$

One then looks for functions $g(x,y)$, $h(x,y)$, such that, for the full system

$$\begin{aligned} \dot{x} &= w \\ \dot{w} &= u - ax\,, \\ \dot{y} &= g(x,y)\,, \\ u &= h(x,y)\,, \end{aligned}$$

[19] Actually, we did not say what (I) meant at all. This is in fact a typical example of the problems that arise every day when "applied mathematicans" and users of mathematics—in this case engineers—try to communicate. Mathematicians will often hurry and attach to a not very precise question what they think is "its obvious mathematical meaning," without realizing that the only reason why this meaning seems obvious to them is that it was the first interpretation they could think of. They will then go back to the users with an impossibility theorem, showing that what the users wanted cannot be done. The users will then think it over again, and point out that what they wanted should not have been interpreted that way anyhow. At this point, the mathematicians may go back to the drawing board and conclude that, with a reasonable reinterpretation of the question's meaning, there is an answer after all. Or perhaps the users will give up all hope of getting help from mathematicians, and solve the problem by themselves. And, in either case, each side will blame the other for the misunderstanding.

that is,

$$\begin{cases} \dot{x} &= w \\ \dot{w} &= h(x,y) - ax\,, \\ \dot{y} &= g(x,y)\,, \end{cases} \tag{II.27}$$

all the trajectories converge to 0 as $t \to +\infty$. (Notice that we have introduced a new variable w to stand for $\dot{x}$, so that (II.27) becomes a system of first-order differential equations. The precise technical requirement imposed on the system (II.27) is that 0 be a globally asymptotically stable equilibrium.)

One should think of y as the state variable of a "controller" or "regulator" system that we, the controllers, are running, for example by solving the equation $\dot{y}(t) = g(x(t), y(t))$ on line with a digital computer, or by feeding x as a physical input —e.g., a voltage—into some physical device such as an electric circuit which evolves according to $\dot{y}(t) = g(x(t), y(t))$. The value of $x(t)$ is supposed to be available to us at each time t, and $y(t)$ is something we can compute or measure. The equation $u = h(x,y)$ produces our control u as a function of x and the new variable y. Naturally, we do not allow u to depend on w, because w—i.e., $\dot{x}$—is not available.

In engineering terms, if we think of the dynamical law given by (II.26) as defining an "input-to-output" relation Σ, with "input" u, "output" x, and internal—or "state"—variables x and $\dot{x}$, and use the notation

$$\xrightarrow{\quad u \quad} \boxed{\Sigma} \xrightarrow{\quad x \quad} \tag{II.28}$$

to represent such a correspondence, then we are asking to connect this to another regulator system

$$\xrightarrow{\quad x \quad} \boxed{\mathscr{R}} \xrightarrow{\quad u \quad}\,, \tag{II.29}$$

where we will feed x as input, and obtain the desired control u as output. The new system $\mathscr{R}$ should have an internal, or "state" variable y (which can be a scalar or a vector), a dynamical law $\dot{y} = g(x,y)$, and output given by $u = h(x,y)$. The desired objective is that the full "closed-loop system"

$$\begin{tikzcd} \arrow[r, "u"] & \boxed{\Sigma} \arrow[r, dash, "x"] & {} \arrow[d, dash] \\ {} \arrow[u, dash] & \boxed{\mathscr{R}} \arrow[l, dash, "u"'] & {} \arrow[l, "x"'] \end{tikzcd} \tag{II.30}$$

be such that all the trajectories go to zero.

One possible choice of g and h is given by

$$\begin{aligned} h(x,y) &= y\,, \\ g(x,y) &= ax - y\,. \end{aligned}$$

This will give

$$\dot{u} = \dot{y} = ax - y = -\dot{w} = -\ddot{x}\,,$$

i.e.,

$$u = -\dot{x} + \text{constant}\,,$$

which is close to what we need, since we know that if we had $u = -\dot{x}$ then the equation $\ddot{x} + ax = u$ reduces to $\ddot{x} + \dot{x} + ax = 0$, all whose solutions $t \mapsto x(t)$ decay exponentially as $t \to +\infty$.

But we have not yet truly solved the problem! To do so, we would have to get rid of the constant! The readers should try to do this, and if they conclude that the question is hard, then they may turn to Scholium VI.

For those readers who can answer this question easily but are not control theorists, we propose a few harder questions:

1. Suppose you know that $a > 0$ but other than that you do not know the value of a. How would you then design a feedback control to stabilize the system for all $a > 0$?

More generally,

2. How would you design a stabilizing feedback if you don't know that $a > 0$?

Even worse,

3. How would you design a stabilizing feedback if all you know is that x obeys an equation
$$\ddot{x} + b\dot{x} + ax = u\,,$$
but you don't know anything about the values of a and b, not even their signs?

These questions are examples of ***adaptive control*** problems. We refer the reader to textbooks such as [35, 87, 118] for details. □

COMMENT NO. 14. JAN: At this point, it would be great if we could include a description of the behavioral approach, and how it differs from and improves upon the "inputs-and-outputs" approach. Naturally, there is no obvious upper bound for how long this part could be. (Anywhere from 1 to 100 pages will do.) It could take the form of an entirely new chapter, if you want. The only constraint is that whatever you write should blend well with the stuff we already have.

II.16 The brachistochrone problem truly belongs to optimal control

We have already pointed out that "brachistochrone" means "shortest time" (cf. Page 4), and sortest time problems naturally belong to optimal control. This is an interesting, though somewhat fortuitous, coincidence. However, our claim the brachistochrone problem truly belongs to control is based on much more serious considerations. We will show in Chapter XIV that

1. the best solutions of Bernoulli's problem are those based on optimal control,
2. in its most natural formulation, the problem is defined by an interesting dynamical law $\dot{x} = f(x,u)$ in four dimensions, which gives rise to a highly non-trivial Lie-algebraic structure (involving the "diamond Lie algebra"), while its Lagrangian L is $\equiv 1$, i.e., completely trivial and uninteresting.

SCHOLIUM VI. Solution of the stabilization problem of §II.15

The regulator system $\mathscr{R}$ is taken to be of the form

$$\left\{\begin{array}{lcl} \dot{y}_1 & = & y_2 + \alpha_1(x - y_1)\,, \\ \dot{y}_2 & = & u - ay_1 + \alpha_2(x - y_1)\,, \\ u & = & ay_1 - y_1 - y_2\,. \end{array}\right. \tag{II.31}$$

The interpretation of this is that y_1 and y_2 are estimated values of x and $\dot{x}$ [20]. These estimated valued are subject to the same evolution equations as x and $\dot{x}$, that is, $\dot{y}_1 = y_2$ and $\dot{y}_2 = u - ay_1$—except that at each time we compare our observed value of x with that of y_1, and make "error corrections" by adding to dy_i and extra term $\alpha_i(x - y_1)dt$. A good stabilizing control law would be

$$u = ax - x - \dot{x}\,, \tag{II.32}$$

[20] Never mind that x is directly observed, so there really is no need to estimate it.

which would transform our equation (II.26) into the obviously stable dynamics given by

$$\ddot{x} + \dot{x} + x = 0\,. \tag{II.33}$$

The law (II.32) cannot, however, be implemented, because $\dot{x}$ is not directly available, so we use instead the law

$$u = ay_1 - y_1 - y_2\,, \tag{II.34}$$

in which x and $\dot{x}$ are replaced by our estimates for their values.

The intuition is then that, if our error-correcting mechanism works, then the estimation errors e_1, e_2 given by

$$e_1 = y_1 - x\,, \quad e_2 = y_2 - \dot{x}\,,$$

will go to zero as $t \to +\infty$. Then u will be close to the value given by (II.32), so x will essentially evolve according to (II.33).

To be able to turn this into a rigorous argument, we have to make an adequate choice of the coefficients α_1, α_2. Clearly, the equation $\dot{y}_1 = y_2 + \alpha_1(x - y_1)$ implies

$$\dot{e}_1 = \dot{y}_1 - \dot{x} = y_2 - \dot{x} + \alpha_1 e_1\,,$$

so

$$\dot{e}_1 = e_2 + \alpha_1 e_1\,. \tag{II.35}$$

Also,

$$\begin{aligned} \dot{e}_2 &= \dot{y}_2 - \ddot{x} \\ &= u - ay_1 + \alpha_2(x - y_1) - (u - ax) \\ &= a(x - y_1) + \alpha_2(x - y_1) \\ &= (a + \alpha_2)e_1\,. \end{aligned}$$

Therefore

$$\ddot{e}_1 = \dot{e}_2 + \alpha_1\dot{e}_1 = (a + \alpha_2)e_1 + \alpha_1\dot{e}_1\,,$$

so

$$\ddot{e}_1 - \alpha_1\dot{e}_1 - (a + \alpha_2)e_1 = 0\,.$$

Therefore we can choose

$$\alpha_1 = -1\,, \qquad \alpha_2 = -1 - a\,,$$

and guarantee that e_1 satisfies

$$\ddot{e}_1 + \dot{e}_1 + e_1 = 0\,, \tag{II.36}$$

so e_1 and $\dot{e}_1$ decay exponentially. With this choice of the α_i it also follows—for example, because of (II.35) and (II.36)—that e_2 and $\dot{e}_2$ decay exponentially. If we now rewrite (II.31) as

$$\begin{cases} \dot{y}_1 = y_2 + \alpha_1 e_1 \,, \\ \dot{y}_2 = -y_1 - y_2 + \alpha_2 e_1 \,, \end{cases} \tag{II.37}$$

we see that

$$\ddot{y}_1 = \dot{y}_2 + \alpha_1 \dot{e}_1 = -y_1 - y_2 + \alpha_1 \dot{e}_1 + \alpha_2 e_1 \,,$$

that is,

$$\ddot{y}_1 + y_1 + y_2 + \alpha_1 \dot{e}_1 + \alpha_2 e_1 = 0 \,.$$

But $\dot{y}_1 = y_2 + \alpha_1 e_1$, so

$$\ddot{y}_1 + \dot{y}_1 + y_1 + y_2 + \alpha_1 (\dot{e}_1 - e_1) + \alpha_2 e_1 = 0 \,.$$

Since e_1 and $\dot{e}_1$ decay exponentially, and the equation $\ddot{y} + \dot{y} + y = 0$ is exponentially stable, we see that y_1 and $\dot{y}_1$ decay exponentially. Since $x = y_1 + e_1$ and $\dot{x} = \dot{y}_1 + \dot{e}_1$. we see that x and $\dot{x}$ decay exponentially as well. □

COMMENT NO. 15. JAN: I have written down in full detail the solution of the adaptive control problem stated as No. 3 in page 57. Do you think we should put it in? It's fairly long, and I see reasons for having it (it's far from obvious for mathematicians who don't know about control) and reasons against (it would be straying too far from our main subject). Wha do you think?

Chapter III

Adaptive stabilization

The plant we want to stabilize is given by

$$\ddot{x} + c\dot{x} + kx = u\,.$$

Here the parameters c and k are unknown, are we are only allowed to use position measurements.

The adaptive stabilizer is going to be a 6-dimensional system with state vector

$$Z = (X, U, \xi, \eta, \gamma, \kappa)\,.$$

This system will obey a dynamical law $\dot{Z} = F(Z,x)$, driven by x as an input, and will produce as an output a function $u = \omega(Z,x)$, which will then be fed back to the plant, yielding an 8-dimensional Plant+Regulator closed-loop system

$$\begin{aligned}
\dot{x} &= y\,, \\
\dot{y} &= -kx - cy + \omega(Z,x)\,, \\
\dot{Z} &= F(Z,x)\,.
\end{aligned}$$

We want to choose F and ω so that for every initial condition the trajectories of the closed-loop system stay bounded for all positive times, and the x and y components go to zero. We will actually get all the cooordinates of the vector $(x, y, X, U, \xi, \eta, \gamma - c, \kappa - k)$ to go to zero, except for the last two, which will just stay bounded but might not go to zero.

You should think of

X as a filtered version of x,

U as a filtered version of u,

ξ as an estimate for X,
η as an estimate for $\dot{X}$,
γ as an estimate for c,
κ as an estimate for k.

The dynamical equations for the adaptive stabilizer are:

$$\begin{array}{rcl}
\dot{X} & = & x - X\,, \\
\dot{U} & = & \omega(Z,x) - U\,, \\
\dot{\xi} & = & \eta + \lambda(Z,x)(X - \xi)\,, \\
\dot{\eta} & = & -\kappa\xi - \gamma\eta + \sigma(Z,x)(X - \xi) + U\,, \\
\dot{\gamma} & = & \varphi(Z,x)\,, \\
\dot{\kappa} & = & \psi(Z,x)\,.
\end{array}$$

The five functions ω, λ, σ, φ, ψ will be chosen later.

We now look at how the "estimation error" $E = X - \xi$ evolves along a trajectory $t \to \Xi(t)$ of the full closed-loop system. We have

$$\dot{\xi} = \eta + \lambda E$$

and

$$\dot{\eta} + \kappa\xi + \gamma\eta = \sigma E + U\,.$$

So

$$\dot{\eta} + \kappa\xi + \gamma(\dot{\xi} - \lambda E) = \sigma E + U\,,$$

and then

$$\dot{\eta} + \gamma\dot{\xi} + \kappa\xi = (\gamma\lambda + \sigma)E + U\,.$$

Also

$$\ddot{\xi} = \dot{\eta} + \dot{\lambda}E + \lambda\dot{E}\,,$$

so

$$\dot{\eta} = \ddot{\xi} - (\dot{\lambda}E + \lambda\dot{E})\,,$$

and then

$$\ddot{\xi} - (\dot{\lambda}E + \lambda\dot{E}) + \gamma\dot{\xi} + \kappa\xi = (\gamma\lambda + \sigma)E + U\,,$$

so that

$$\ddot{\xi} + \gamma\dot{\xi} + \kappa\xi = (\gamma\lambda + \sigma + \dot{\lambda})E + \lambda\dot{E} + U\,.$$

On the other hand, x satisfies

$$\ddot{x}+c\dot{x}+kx=\omega\,,$$

i.e. $Ax=\omega$, where A is the differential operator D^2+cD+k. So, if we write $x=\dot{X}+X$, i.e. $x=BX$—where B is the differential operator $D+1$—we find

$$BAX=ABX=Ax=\omega\,.$$

So, if we let $W=AX$, we have $BW=\omega$, that is

$$\dot{W}=\omega-W\,.$$

If we compare this with $\dot{U}=\omega-U$, we see that the difference $H=W-U$ satisfies $\dot{H}=-H$. So H will decay exponentially as $t\uparrow+\infty$. We now have $AX=W=U+H$, i.e.

$$\ddot{X}+c\dot{X}+kX=U+H\,.$$

Then

$$(\ddot{X}+c\dot{X}+kX)-(\ddot{\xi}+\gamma\dot{\xi}+\kappa\xi)=U+H-(\gamma\lambda+\sigma+\dot{\lambda})E-\lambda\dot{E}-U\,,$$

so that

$$\ddot{E}+\gamma\dot{E}+\kappa E+(c-\gamma)\dot{X}+(k-\kappa)X=U+H-(\gamma\lambda+\sigma+\dot{\lambda})E-\lambda\dot{E}-U\,.$$

Therefore

$$\ddot{E}+\gamma\dot{E}+\kappa E=H-(\gamma\lambda+\sigma+\dot{\lambda})E-\lambda\dot{E}+(\gamma-c)\dot{X}+(\kappa-k)X\,.$$

We now choose

$$\lambda=1-\gamma,$$

and

$$\sigma=1-\kappa-(\gamma\lambda+\dot{\lambda})\,.$$

(Notice that $\dot{\lambda}=-\dot{\gamma}$, which is a perfectly well defined function of Z and x thanks to the dynamical equations.)

With these choices, the equation for E becomes

$$\ddot{E}+\dot{E}+E=H+(\gamma-c)\dot{X}+(\kappa-k)X\,.$$

We now define

$$V=E^2+E\dot{E}+\dot{E}^2+\frac{1}{2}\Big((\gamma-c)^2+(\kappa-k)^2\Big)\,,$$

and observe that V is a function of x and Z (that is, a function of the full state (x,y,Z) of the closed-loop system which does NOT depend on y), because $E = X-\xi$, $\dot{X} = x - X$, and

$$\dot{E} = \dot{X} - \dot{\xi} = x - X - \eta - \lambda(Z,x).(X-\xi) = x - X - \eta - (1-\gamma).(X-\xi)\,.$$

Moreover,

$$E^2 + E\dot{E} + \dot{E}^2 = \Big(E + \frac{\dot{E}}{2}\Big)^2 + \frac{3}{4}\dot{E}^2\,,$$

so V is ≥ 0 everywhere, and $V = 0$ iff $E = \dot{E} = \gamma - c = \kappa - k = 0$.

NOTE: The most obvious "Liapunov function" for the E, $\dot{E}$ part would be $E^2 + \dot{E}^2$. However, the derivative of this expression along trajectories is $2\dot{E}(E+\ddot{E})$, i.e. $-2\dot{E}^2$+extra stuff. This is not good enough because, even after we take care of the extra stuff, we will only be controlling the behavior at ∞ of $\dot{E}$, and we will not be able to deal with E itself. That's why I am using $E^2 + E\dot{E} + \dot{E}^2$ instead. The derivative now is $-(E^2 + E\dot{E} + \dot{E}^2)$+extra stuff, and $E^2 + E\dot{E} + \dot{E}^2$ is strictly positive definite in E and $\dot{E}$.

Differentiation of V along trajectories yields

$$\dot{V} = 2E\dot{E} + \dot{E}^2 + \tilde{E}\ddot{E} + \dot{\gamma}(\gamma - c) + \dot{\kappa}(\kappa - k)\,,$$

where

$$\tilde{E} = E + 2\dot{E}\,.$$

But

$$\ddot{E} = -E - \dot{E} + H + (\gamma - c)\dot{X} + (\kappa - k)X\,.$$

So

$$\dot{V} = -E^2 - E\dot{E} - \dot{E}^2 + \tilde{E}\Big(H + (\gamma - c)\dot{X} + (\kappa - k)X\Big) + \dot{\gamma}(\gamma - c) + \dot{\kappa}(\kappa - k)\,.$$

Therefore

$$\dot{V} = -E^2 - E\dot{E} - \dot{E}^2 + \tilde{E}H + (\gamma - c)(\tilde{E}\dot{X} + \dot{\gamma}) + (\kappa - k)(\tilde{E}X + \dot{\kappa})\,.$$

We now choose

$$\varphi = -\tilde{E}\dot{X}\,, \qquad \psi = -\tilde{E}X\,.$$

Notice that φ and ψ are indeed functions of Z and x (and in particular do NOT involve the derivative $\dot{x}$!).

So now we have

$$\dot{V} = -E^2 - E\dot{E} - \dot{E}^2 + \tilde{E}H = -\frac{\tilde{E}^2}{4} - \frac{3E^2}{4} + \tilde{E}H\,.$$

This almost says that V is decreasing along trajectories. To make this rigorous, we use

$$\tilde{E}H \leq \frac{\tilde{E}^2}{8} + 2H^2\,,$$

and conclude that

$$\dot{V} \leq -\frac{\tilde{E}^2}{8} - \frac{3\dot{E}^2}{4} + 2H^2\,.$$

Since H decays exponentially along each trajectory, V has a finite limit as $t \uparrow \infty$, and then $E, \dot{E}$, γ, and κ stay bounded as $t \uparrow \infty$. Moreover, if $t \to \Xi(t)$ is a trajectory of the full closed-loop system, then

$$\int_0^T \mathscr{E}(\Xi(t))dt \leq V(\Xi(0)) - V(\Xi(T)) + 2\int_0^T H(t)^2 dt\,,$$

where

$$\mathscr{E} = \frac{\tilde{E}^2}{8} + \frac{3\dot{E}^2}{4}\,.$$

So the function $t \to \mathscr{E}(\Xi(t))$ is integrable on $[0,\infty[$, and this implies that $t \to E(\Xi(t))$ and $t \to \dot{E}(\Xi(t))$ are square-integrable on $[0,\infty[$. It then follows that $E(\Xi(t))$ goes to 0 as $t \to +\infty$.

So far, the boundedness of V only implies the boundedness of four of the eight variables of the closed-loop system. That is fine, because we have not yet said anything about how to choose our feedback ω so as to stabilize the system.

The idea now would be to use the "certainty equivalence" stabilizer

$$U = (\gamma - 1)\eta + (\kappa - 1)\xi\,,$$

which has the property that, if γ and κ actually did have the "right values" c, k, and ξ, η were indeed X and $\dot{X}$, then you would get the equation

$$\ddot{X} + \dot{X} + X = H\,,$$

which is clearly stable. We cannot, however, choose U this way, because U is a state variable. But we can choose u, because the function ω is still at our disposal.

To get $U = (\gamma - 1)\eta + (\kappa - 1)\xi$ we should have

$$\dot{U} = \dot{\gamma}\eta + (\gamma - 1)\dot{\eta} + \dot{\kappa}\xi + (\kappa - 1)\dot{\xi}\,.$$

This would yield

$$\dot{U} + U = \dot{\gamma}\eta + (\gamma - 1)(\eta + \dot{\eta}) + \dot{\kappa}\xi + (\kappa - 1)(\xi + \dot{\xi})\,.$$

So we define

$$\omega(Z,x) = \dot{\gamma}\eta + (\gamma - 1)(\eta + \dot{\eta}) + \dot{\kappa}\xi + (\kappa - 1)(\xi + \dot{\xi}),$$

and observe that this is acceptable because ξ, η, γ, κ and their derivatives are just functions of Z and x.

Now, if we define

$$W = (\gamma - 1)\eta + (\kappa - 1)\xi,$$

we see that $\dot{W} + W = \omega(Z,x)$. Since $\dot{U} + U = \omega(Z,x)$ by our dynamical laws, the difference $G = U - W$ satisfies $\dot{G} = -G$, so G decays exponentially along trajectories.

So we now have

$$U = (\gamma - 1)\eta + (\kappa - 1)\xi + G,$$

which is essentially the feedback we wanted, up to an exponentially decaying term.

The two equations for ξ and η now become

$$\begin{aligned} \dot{\xi} &= eta + \lambda(Z,x)E, \\ \dot{\eta} &= \xi - \eta + \sigma(Z,x).E + G. \end{aligned}$$

Therefore

$$\begin{aligned} \ddot{\xi} &= \dot{\eta} + \dot{\lambda}E + \lambda\dot{E}, \\ &= -\xi - \eta + \sigma E + G + \dot{\lambda}E + \lambda\dot{E}, \\ &= \xi - \dot{\xi} + \lambda E + \sigma E + G + \dot{\lambda}E + \lambda\dot{E}, \end{aligned}$$

so that

$$\ddot{\xi} + \dot{\xi} + \xi = (\lambda + \sigma + \dot{\lambda})E + G + \lambda\dot{E}.$$

Our choice of λ and σ is such that

$$\sigma + \dot{\lambda} = 1 - \kappa - \gamma\lambda.$$

So

$$\lambda + \sigma + \dot{\lambda} = 1 - \kappa - (\gamma - 1)\lambda,$$

and the equation for ξ becomes

$$\ddot{\xi} + \dot{\xi} + \xi = (1 - \kappa - (\gamma - 1)\lambda)E + G + \lambda\dot{E}.$$

Since $\lambda = 1 - \gamma$, this says that

$$\ddot{\xi} + \dot{\xi} + \xi = (1 - \kappa + (\gamma - 1)^2)E + G + (1 - \gamma)\dot{E}.$$

Now we are ready to conclude the stabilization proof. Let us work along a fixed trajectory. We know that E and $\dot{E}$ are square-integrable, and G decays exponentially. So ξ and $\dot{\xi}$ are bounded (because s^2+s+1 is stable). We also know that E and $\dot{E}$ are bounded. So X and $\dot{X}$ are bounded, because $X = \xi + E$. If we now go back to the equation

$$\ddot{E}+\dot{E}+E = H+(\gamma-c)\dot{X}+(\kappa-k)X\,,$$

and recall that H decays exponentially and γ and κ are bounded, we can conclude that $\ddot{E}$ is bounded. Since we know that $\dot{E}$ is square-integrable, we can now conclude that $\dot{E}$ goes to 0.

Now use again the equation $\ddot{\xi}+\dot{\xi}+\xi = (\kappa-1+(\gamma-1)^2)E+G+(1-\gamma)\dot{E}$. Since G decays exponentially, γ and κ are bounded, and E and $\dot{E}$ go to 0, we conclude that ξ and $\dot{\xi}$ go to zero, and then the equation implies that $\ddot{\xi}$ goes to zero. Since $X = \xi + E$, and ξ, $\dot{\xi}$, E and $\dot{E}$ go to 0, we can conclude that X and $\dot{X}$ go to zero. Going back to $\ddot{E}+\dot{E}+E = H+(\gamma-c)\dot{X}+(\kappa-k)X$, we can now infer that $\ddot{E}\to 0$. Since $\ddot{\xi}\to 0$, we get $\ddot{X}\to 0$. Now $x = \dot{X}+X$, and $y = \dot{x} = \ddot{X}+\dot{X}$. So x and y go to zero. Finally, since $\ddot{X}+c\dot{X}+kX = U+H$, and we already know that X, $\dot{X}$ and $\ddot{X}$ go to 0, we see that $U\to 0$ as well.

For completeness, we now write the complete system of equations for the adaptive stabilizer and the closed-loop system.

FINAL FORM OF THE CLOSED-LOOP SYSTEM

$$
\begin{aligned}
\dot{x} &= y,\\
\dot{y} &= -kx - cy - 2X - U + \xi - 2\eta + 3X\gamma + 2X\kappa + U\gamma - 3\xi\gamma\\
&\quad +2\eta\gamma + \eta\kappa\\
&\quad +4xX\xi + 3xX\eta + x\xi\eta - 5xX^2 - x\xi^2 + 2x\eta^2 - 2x^2\eta + 2x^2X - 2x^2\xi\\
&\quad -2X\xi^2 + X\xi\eta - 2X\eta^2 - 2X\gamma\kappa - 2X\gamma^2 - X^2\xi - X^2\eta\\
&\quad +3X^3 + 2\xi\gamma^2 + \xi\gamma\kappa - \eta\gamma^2\\
&\quad -10xX\xi\gamma + 7xX^2\gamma + 3x\xi^2\gamma\\
&\quad -2x^2X\gamma + 2x^2\xi\gamma + 3X\xi^2\gamma + X\gamma^3\\
&\quad +4X^2\xi\gamma - 2X^2\xi\gamma + 4X^2\xi\gamma\\
&\quad +2X^2\eta\gamma - 2X^2\eta\gamma - 3X^3\gamma - 2X^3\gamma - \xi\gamma^3\\
&\quad +4xX\xi\gamma^2 - 2xX^2\gamma^2 - 2x\xi^2\gamma^2 - 4X^2\xi\gamma^2 + 2X^3\gamma^2\\
\dot{X} &= x - X,\\
\dot{U} &= -2X - 2U + \xi - 2\eta + 3X\gamma + 2X\kappa + U\gamma - 3\xi\gamma + 2\eta\gamma + \eta\kappa\\
&\quad +4xX\xi + 3xX\eta + x\xi\eta - 5xX^2 - x\xi^2 + 2x\eta^2 - 2x^2\eta + 2x^2X - 2x^2\xi\\
&\quad -2X\xi^2 + X\xi\eta - 2X\eta^2 - 2X\gamma\kappa - 2X\gamma^2 - X^2\xi - X^2\eta\\
&\quad +3X^3 + 2\xi\gamma^2 + \xi\gamma\kappa - \eta\gamma^2\\
&\quad -10xX\xi\gamma + 7xX^2\gamma + 3x\xi^2\gamma\\
&\quad -2x^2X\gamma + 2x^2\xi\gamma + 3X\xi^2\gamma + X\gamma^3\\
&\quad 4X^2\xi\gamma - 2X^2\xi\gamma + 4X^2\xi\gamma\\
&\quad +2X^2\eta\gamma - 2X^2\eta\gamma - 3X^3\gamma - 2X^3\gamma - \xi\gamma^3\\
&\quad +4xX\xi\gamma^2 - 2xX^2\gamma^2 - 2x\xi^2\gamma^2 - 4X^2\xi\gamma^2 + 2X^3\gamma^2\\
\dot{\xi} &= X - \xi + \eta - X\gamma + \xi\gamma,\\
\dot{\eta} &= +X - \xi - X\kappa - X\gamma + \xi\gamma - \eta\gamma\\
&\quad -6xX\xi + 2xX\eta + 5xX^2 - 2x\xi\eta + x\xi^2 - 2x^2X + 2x^2\xi\\
&\quad +2X\xi\eta + X\xi^2 + X\gamma^2 + 4X^2\xi - 2X^2\eta - 3X^3 - \xi\gamma^2\\
&\quad +4xX\xi\gamma - 2xX^2\gamma - 2x\xi^2\gamma - 4X^2\xi\gamma + 2X^3\gamma,\\
\dot{\gamma} &= 5xX - x\xi + 2x\eta - 2x^2 + X\xi - 2X\eta - 3X^2 - 2xX\gamma + 2x\xi\gamma\\
&\quad -2X\xi\gamma + 2X^2\gamma,\\
\dot{\kappa} &= -2xX - X\xi + 2X\eta + 3X^2 + 2X\xi\gamma - 2X^2\gamma.
\end{aligned}
$$

The adaptive stabilizer is the 6-dimensional system described by the last six equations, with state variable $Z = (X, U, \xi, \eta, \gamma, \kappa)$.

The feedback control law is $u = \omega(x, Z)$, where $\omega(Z, x)$ is the expression that, according to the right-hand side of the second equation, is equal to $\dot{y} + cy + kx$.

Chapter IV

The behavioral approach

Control is a multifaceted subject. The aim of this chapter is to put the problems discussed in this book in a broad control and systems theory perspective. We will discuss a number of main themes, in particular, open loop and feedback control, and the various mathematical views of the notion of a dynamical system that have emerged from control and from the modelling of open systems.

The main subject of this book is *trajectory optimization.* It is about how to transfer the state of a dynamical system from one point to another. As we have explained, this is a subject that has roots going all the way back to antiquity.

A second aspect of control is as a process of decision making. Assume that continuous observations are made on a dynamical systems, and on the basis of these observations, actions are taken to guide the system, for example to steer the state of a system along a particular path. This idea of choosing control variables on the basis of measurements, is referred to as *feedback control*. This way of feedback decision making is, of course, very much part of daily life. It guides the actions of individuals, of societies, of economic agents, etc. It is also very much part of technology. Feedback control devices have also been constructed since antiquity, in water wheels, to control the speed of windmills, temperature controlling devices, etc. Feedback control mechanisms became widespread during the age of industrialization in the seventeenth century, starting with various controllers, as the Watt regulator, that were used to control the speed and pressure in steam engines. However, with the advent of electronics, starting with the feedback amplifier, and certainly with the present day computers, feedback control is an increasingly important part of industrial processes. A good source for further reading about the history of feedback control is [102, 12].

A third way of viewing control is as design of subsystems aimed at achieving good performance of the overall system. These subsystems may or may not act as feedback devices. Examples of such passive controllers that do not act as feedback processors are mechanical dampers, pressure valves, heat fins, etc. Such devices are, of course, part of our daily environment. We will refer to this more general way of viewing control as *behavioral control,* or *control by interconnection*.

In this chapter we will briefly explain these various aspects of control. However, these developments have found a fall-out in various mathematical concepts, which form the core of an area of applied mathematics called *Systems Theory*. We will often find it necessary and advantageous to put the developments in this broader perspective.

IV.1 Open loop versus feedback control

We start by sketching the important distinction between what is called *open loop control* and what is called *feedback control*. We will do this first in the context of state feedback. Later on more general situations will be discussed. Consider a system described by the first order differential equation

$$\dot{x} = F(x,u)\,. \tag{IV.1}$$

In (IV.1), $x \in \mathbb{R}^n, q \in \mathbb{R}^m$, and $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$.

The variable x is called the *state* of the system (in a mechanical system, for example, the state consists of the vector of the positions and velocities in a multi-body system), u is an external input variable that is imposed by the environment (for example, the input could be external forces that act on the multi-body system), and F is the *vector field;* $F(x,u)$ tells how the state x of the system will involve when the system is in state x and input u is applied to it. Note that it may be more accurate to call F a *family* of vector fields, since for each u, $F(\cdot,u)$ defines a vector field in the usual sense of the term. Throughout this chapter we will assume that we are dealing with systems in which the time-signals take their values in finite dimensional real vector spaces. More general situations in which the variables, for example the state x in (IV.1), belong to a manifold, or, for example the input u in (IV.1), satisfy inequality constraints, are very meaningful in applications, but will not be considered in this chapter, since they would unnecessarily encumber the exposition.

For the purposes of this and the following sections, u is a variable that can be chosen in order to control the system, and will therefore often be called the *control (input).* It is in the way in which u is generated that the rather subtle distinction between open loop and feedback control emerges.

The simplest way of choosing the input is through a function $\omega : \mathbb{R} \to \mathbb{R}^m$, resulting in the input signal $u = \omega(t)$ at time t. Thus, if ω is the control input that is chosen, the state of the system will evolve according to

$$\frac{d}{dt}x(t) = F(x(t), \omega(t)), \tag{IV.2}$$

with a right hand side which, due to the presence of $\omega(t)$, is time-varying. Under reasonable conditions of F and ω, there will be a unique solution $x : \mathbb{R} \to \mathbb{R}^n$ to (IV.2) emerging form any initial state $x(t_0) = x_0$. This leads to *open loop control*. Different choices of $\omega(\cdot)$ lead to different paths $x(\cdot)$. Some resulting trajectories $x(\cdot)$ may be bad, some may be good, some may be optimal (in some well-defined sense). Many problems in trajectory optimization come down to choosing the optimal $\omega(\cdot)$. The discussion of mathematical techniques for finding these optimal $\omega(\cdot)$'s form the core of this book.

However, in many applications the control input is not chosen as a function of time directly, but as a function of the state. That is, through a function $\alpha : \mathbb{R}^n \to \mathbb{R}^m$. This function α tells that if the state variable equals x, then the control input will be $u = \alpha(x)$. The system state is then be governed by the differential equation

$$\frac{d}{dt}x(t) = F(x(t), \alpha(t)). \tag{IV.3}$$

Under reasonable conditions on F and α, there will again be a unique solution $x : \mathbb{R} \to \mathbb{R}^n$ to (IV.3) emerging form any initial state $x(t_0) = x_0$. This leads to *feedback control*. Different choices of α lead to different paths $x(\cdot)$. Some resulting trajectories $x(\cdot)$ may be bad, some may be good, some may be optimal (in a well-defined sense). Many control problems in feedback control come down to choosing the optimal $\alpha(\cdot)$.

Note, very importantly, that for a given feedback law $\alpha : \mathbb{R}^n \to \mathbb{R}^m$, the trajectory $x(\cdot)$ that emerges from $x(t_0) = x_0$, can also be generated by applying the open loop control law $\omega(t) = \alpha(x(t))$. Thus the distinction between open loop and feedback control does not lie in the fact that different trajectories are actually generated, but by the way in which these trajectories are generated. In open loop control, the controller is *clock driven:* the clock time $t \in \mathbb{R}$ determines that the control input $u = \omega(t) \in \mathbb{R}^m$ will be used at time t. In state feedback control, the controller is *state driven:* the state variables $x \in \mathbb{R}^n$ determine that the control input $u = \alpha(x)$ will be used when the system is in state x.

We now explain one of the main features for using feedback control: robustness. Assume that system (IV.1) is in state $x(t_0) = x_0$ at time t_0. Now consider for example the problem of transferring the state x_0 at time t_0 to a desired state x_{final}

in minimum time over all open loop controls $\omega : \mathbb{R} \to \mathbb{U}$, and over all feedback controls $\alpha : \mathbb{R}^n \to \mathbb{U}$, with $\mathbb{U} \subset \mathbb{R}^m$ some give constraint set (e.g., $\mathbb{U} = [-1, +1]^{\mathtt{m}}$). Of course, in order for such optimal controls to exist, $\mathbb{U}$ will typically have to be a compact subset of $\mathbb{R}^m$, but the precise nature of $\mathbb{U}$ is not important for the present exposition. Assume that we have computed both the 'optimal' open loop $\omega^* : \mathbb{R} \to \mathbb{U}$, and the 'optimal' state feedback control law $\alpha^* : \mathbb{R}^n \to \mathbb{U}$. The resulting trajectories, i.e., the solutions of

$$\frac{d}{dt}x'(t) = F(x'(t), \omega^*(t))), \quad x'(t_0) = x_0,$$

and

$$\frac{d}{dt}x''(t) = F(x''(t), \alpha^*(x''(t)), \quad x''(t_0) = x_0,$$

are thus both "optimal." It is easy to see that this must lead to $x' = x''$, which we will denote by x^*. Also, $\omega^*(t) = \alpha^*(x^*(t))$.

So far, it looks like that there is no difference between using an open loop or a feedback law. The difference emerges if we consider disturbances, which leads to the main practical advantage of feedback versus open loop control. Assume that for some reason, the trajectory x^* becomes disturbed during the execution of the optimal path. Say that a disturbance causes the system state at some time $t_1 \in (t_0, t_{\text{final}})$ to jump from the expected state $x^*(t_1)$ to the unexpected $x(t_1) = x_1$. Then the feedback control will execute the trajectory

$$\frac{d}{dt}x(t) = F(x(t), \alpha^*(x(t)), \quad x(t_1) = x_1,$$

which, if α^* has been computed well, may now execute the optimal state transfer from x_1 to x_{final}. However, the open loop trajectory, generated by

$$\frac{d}{dt}x(t) = F(x(t), \omega^*(t))), \quad x(t_1) = x_1,$$

unaware that a disturbance has happened, will in general not even execute the desired transfer. This shows an important form of *robustness* of feedback control, robustness against disturbances. Feedback also gives robustness against variations in the system parameters. Two main examples of such controllers are feedback amplifiers and PID-controllers. See [111] for a discussion of these controllers.

IV.2 Output feedback

The use of state feedback instead of open loop control comes at a price: indeed, it requires that all the state variables are continuously measured, whereas open loop control only requires knowledge of clock time.

In a typical application, the state dimension `n` is large. For a multi-body system, the state dimension equals twice the number of degrees of freedom, and measuring the state would require measuring the position and velocity of each of the components. For a flexible system, for example, the state space even becomes in principle an infinite dimensional space. So, it is very desirable to consider situations in which only part of the state, or, more generally, a function of the state is measured. This leads to the following generalization of (IV.1):

$$\dot{x} = F(x,u), \quad y = H(x,u)\,. \tag{IV.4}$$

Here $x \in \mathbb{R}^n, u \in \mathbb{R}^m, y \in \mathbb{R}^p$, and $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n, H : \mathbb{R}^n, \mathbb{R}^m \to \mathbb{R}^p$. The 'physical' meaning of y is that it is the output of the sensors, and H expresses the law by which the state and output generates the sensor outputs. Usually `m,p` $<<$ `n`. The system (IV.4) is driven by the input u, which causes by the vector field $F(x,u)$ the evolution of the state x. The *read-out map* $H(x,u)$ then gives the output. The system (IV.4) is called an *input/state/output system.*

An *output feedback control law* is now also an input/state/output system in its own right, but one that acts in the opposite 'direction' as (IV.4), i.e.

$$\dot{z} = K(z,y); \quad u = L(z,y)\,: \tag{IV.5}$$

it takes the outputs of (IV.4) as its inputs, and yields the inputs to (IV.4) as its outputs. In applications, the controller is often implemented as a microprocessor based computing device that accepts the sensor outputs and returns the actuator inputs. An *actuator* is typically a *transducer* that translates electronic signals into forces, currents, or similar physical variables that are used to control a system.

Present-day control theory is completely centered around the signal flow graph shown in figure IV.1. The system to be controlled (called the *plant* in the engineering literature) has four terminals (supporting variables which are typically vector-valued). There are two input terminals, one for the control, one for the disturbances (or other exogenous variables, as set-points, reference signals, tracking inputs, etc.) and there are two output terminals, one for the measurements, and one for the to-be-controlled variables. By using feed-through terms in the plant equations this configuration accommodates, by incorporating these variables in the measurements, for the possible availability to the controller of set-point settings, reference signals, or disturbance measurements for feed-forward control, and, by incorporating the control input in the to-be-controlled outputs, for penalizing excessive control action. The control inputs are generated by means of actuators and the measurements are made available through sensors.

Usually, the dynamics of the actuators and of the sensors are considered to be part of the plant. On the other hand, for the *feedback controller*, an input/output

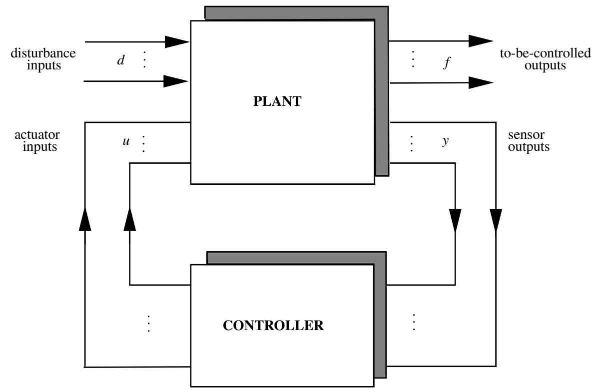


Figure IV.1: Intelligent control

system in its own right, the inputs are the measured variables, and the outputs the control variables. We call this structure *intelligent control*, since it consists of a controller that uses observations to make decisions, which we view as one of the essential features of thinking.In intelligent control, the controller is thought of as a micro-processor type device which is driven by the sensor outputs and which produces the actuator inputs through an algorithm involving a combination of feedback, identification, and adaptation. Also, often loops expressing model uncertainty are incorporated in the above as well. Of course, many variations, refinements, and special cases of this structure are of interest, but the basic idea is that of supervisor reacting in an intelligent way to observed events and measured signals. The aim of feedback control is to design, for a given plant, the feedback controller such that the closed loop system meets certain specifications. For example, stability, disturbance attenuation, optimality, robustness (for example, preservation of stability robustness against parameter variations), etc.

In equations, this leads to

$$\dot{x} = F(x,u,d), \quad y = H(x,u,d), \quad f = G(x,u,d)\,. \tag{IV.6}$$

Here $x \in \mathbb{R}^n, u \in \mathbb{R}^m, u \in \mathbb{R}^{\mathtt{d}}, y \in \mathbb{R}^p, f \in \mathbb{R}^{\mathtt{f}}$, and $F : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{\mathtt{d}} \to \mathbb{R}^n, H : \mathbb{R}^n \times \mathbb{R}^{\mathtt{d}} \mathbb{R}^m \to \mathbb{R}^p, G : \mathbb{R}^n \times \mathbb{R}^{\mathtt{d}} \mathbb{R}^m \to \mathbb{R}^{\mathtt{f}}$. Together with a controller

$$\dot{z} = K(z,y), \quad u = L(z,y)\,, \tag{IV.7}$$

this leads to the *closed loop system* obtained by combining (IV.6) with (IV.7). In the closed loop system, d should be considered the input, and f the output. Note that u and y become internal variables in the closed loop system.

IV.3 Input/output systems

We have seen in the previous section how feedback control leads to input/output structures. The input/output framework can be seen as a *cause/effect* structure, with the inputs as the variables which act as causes for the evolution of the system variables, and the outputs the variables that are the effects. Note that 'feedback' results a blurring of cause and effect, since what is the cause in the plant, is the effect in the controller, and vice-versa.

In the present section, we discuss briefly input/output systems. Since, we do no need to distinguish between different types of inputs and outputs, we discuss

$$\dot{x} = F(x,u), \quad y = H(x,u)\,. \tag{IV.8}$$

The study of such systems is a very well-developed area of research whose development strongly emerged since the 1960's, in the context of Control Theory, and which resulted in an independent field called *Mathematical Systems Theory.*

The main aim of (IV.8) is to describe the relation between input and output trajectories. We use the term relation, since usually (IV.8) does not determine a mapping from the input trajectory to the output trajectory. Indeed, the output trajectory depends not only on the input trajectory, but also on the initial state. The state usually plays an ancillary role: it merely serves to specify the input/ouput relation. Questions that spring up are: *When do two systems as (IV.8) define the same input/output relation? When and how can a given input/output relation be expressed in the form (IV.8)? When do we want to call (IV.8) a stable system?* Etc., etc.

The theory is particularly rich, which explicit, effectively computable results, when the maps F and H in (IV.8) are linear, leading to the system

$$\dot{x} = Ax + Bu, \quad y = Cx + Du\,, \tag{IV.9}$$

with $A \in \mathbb{R}^{\mathtt{n}\times\mathtt{n}}, B \in \mathbb{R}^{\mathtt{n}\times\mathtt{m}}, C \in \mathbb{R}^{\mathtt{p}\times\mathtt{n}}$, and $D \in \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$. This system is commonly denoted as $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$. In this case, and assuming that the initial state is zero at $t = 0$, i.e., $x(0) = 0$, it is possible to obtain the input/output relation as the convolution integral

$$y(t) = Du(t) + \int_0^t H(t-t')u(t')\,dt,$$

with $H : \mathbb{R}_+ \to \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$ given by

$$H(t) = Ce^{At}B.$$

$D\delta + H$ is called the *impulse response matrix*. Equivalently, assuming again $x(0) = 0$, the input/output relation can also be expressed in terms of the Laplace transforms $\hat{u}$ of the input trajectory and $\hat{y}$ of the output trajectory as

$$\hat{u} = G\hat{y},$$

with

$$G(s) = D + C(Is - A)^{-1}B.$$

The matrix of rational functions $G \in \mathbb{R}(s)^{\mathtt{p}\times\mathtt{m}}$ is called the *transfer function* of $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$.

These input/output representations lead to some interesting inverse problems that go under the name *realization theory: Given an impulse response $H : \mathbb{R}_+ \to \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$, or a transfer function $G \in \mathbb{R}(s)^{\mathtt{p}\times\mathtt{m}}$, does there exists a 'realization' $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$ with the given impulse response or transfer function? To what extent is it unique? What is the relation between two realizations? Etc., etc.* The study of the systems $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$ is very deep and varied. This is in sharp contrast with the rather limited range of issues involved in the study of the analogous flows

$$\dot{x} = Ax.$$

The system (IV.9) is said to be *input/output stable* if $H \in \mathfrak{L}_1(\mathbb{R}_+, \mathbb{R}^{\mathtt{p}\times\mathtt{m}})$, equivalently if G has all its poles in the open left half of the complex plane. This is the case if A has all its eigenvalues in the open left half of the complex case. When $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$ is state controllable (see section IV.9) and state observable (see section IV.10) this 'if' is in fact 'if and only if'. Two important norms for systems $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$ are the $\mathscr{H}_2$-norm and the $\mathscr{H}_\infty$-norm (the $\mathscr{H}$ refers to Hardy spaces). They are most easily defined in terms of the impulse response or the transfer function. The $\mathscr{H}_2$-norm is finite if and only if $D = 0$ and the system is stable, and equals

$$\begin{aligned} \left|\left| \left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right] \right|\right|_{\mathscr{H}_2} &= \sqrt{\int_0^\infty \operatorname{trace}(H^\top(t)H(t))\,dt} \\ &= \sqrt{\frac{1}{2\pi}\int_0^\infty \operatorname{trace}(G^\top(-i\omega)G(i\omega))\,d\omega}. \end{aligned}$$

The $\mathscr{H}_\infty$-norm is finite if and only if the system is stable and equals

$$\left|\left| \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \right|\right|_{\mathscr{H}_\infty} = \sup_{\omega\in\mathbb{R}} ||G(i\omega)||,$$

with $||G(i\omega)||$ the Euclidean induced matrix norm, i.e. square root of the maximal eigenvalue of $G^\top(-i\omega)G(i\omega)$. These system norms measure the gain of the system, the degree to which the input is amplified by the system. Especially the $\mathscr{H}_\infty$-norm is quite natural, since it is the induced norm of (IV.3) viewed as a mapping from $\mathfrak{L}_2(\mathbb{R}_+(\mathbb{R},\mathbb{R}^m)$ to $\mathfrak{L}_2(\mathbb{R}_+(\mathbb{R},\mathbb{R}^p)$. In the next section, we will see how these norms enter in the design of output feedback controllers.

IV.4 Optimal output feedback control

Consider again the plant

$$\dot{x} = F(x,u,d), \quad y = H(x,d), \quad f = G(x,u,d)\,. \tag{IV.10}$$

Note that we have assumed that H does not involve u directly. This will allow us to write explicit equations for the closed loop system. The controller

$$\dot{z} = K(z,y), \quad u = L(z,y)\,, \tag{IV.11}$$

leads to the closed loop system

$$\dot{\begin{bmatrix} x \\ z \end{bmatrix}} = \begin{bmatrix} F(x,L(z,H(x,d))) \\ K(z,H(x,d)) \end{bmatrix}, \quad f = G(x,d,L(z,H(x,d)))\,. \tag{IV.12}$$

This is a system with input d, output f and state (x,z). The problem of optimal feedback control is to find, for a given functions F,H,G, defining the plant (IV.10), functions K,L, defining the controller (IV.11), such that it is optimal in some well-defined sense. The performance criterion, of course, usually refers to the closed loop system (IV.12). Note that finding the optimal $\dim(z)$ is part of the problem.

The most important and explicit results in the area of optimal feedback control have been obtained for linear time-invariant systems

$$\dot{x} = Ax + Bu + Gd, \quad y = Cx + Dd, \quad f = Hx + Ju + Qd,$$

with A,B,G,C,D,H,J,Q real matrices of suitable dimension. The two main problems that have been solved are the minimization of the $\mathscr{H}_2$- and the $\mathscr{H}_\infty$-norm of the closed loop system viewed as a system with the disturbances d as input, and the to-be-controlled output f as output. Actually, in both cases, it is possible to prove that the optimal controller is also a finite dimensional linear time-invariant system. Note however that linearity was actually built in in our definition of the $\mathscr{H}_2$ and the $\mathscr{H}_\infty$-norm, but these definitions are readily generalized to more general systems.

The optimal feedback controller in the $\mathscr{H}_2$ case has the same order as the plant $\dim(z) = \dim(x)$. The computation of the optimal gains (K, L, M, N) of the optimal feedback controller

$$\dot{z} = Kz + Ly, \quad u = Mz + Ny$$

involves the solution of two uncoupled Riccati equations [149]. The optimal $\mathscr{H}_2$ controller can also be given a stochastic interpretation in the case that the disturbance input d is modelled as a vector of stationary gaussian stochastic processes, and the expected value of the norm squared of the to-be-controlled output is to be minimized. It was, in fact, under this form that this problem was first appeared and was solved in the 1960's, under the name of the LQG (linear-quadratic-gaussian) problem. This problem has in fact the well-known Kalman filter as a special case.

The optimal feedback controller in the $\mathscr{H}_\infty$ case is not obtained directly, but first conditions are derived for the existence of a feedback controller such that the $\mathscr{H}_\infty$-norm of the closed loop system is less that some number γ^2. Then by iterating on γ, the optimal controller is obtained. Checking the existence of a feedback controller such that the $\mathscr{H}_\infty$-norm of the closed loop system is less than γ^2 involves solving two Riccati equations, and checking a coupling condition. If the required controller exists, then its state dimension is again the same as that of the plant, $\dim(z) = \dim(x)$, and the gains (K, L, M, N) of the optimal feedback controller

$$\dot{z} = Kz + Ly, \quad u = Mz + Ny$$

follow from the solution of two Riccati equations. This result has something spectacular about it, and was first obtained in [51]. It is also covered in [149]. The ramifications of the $\mathscr{H}_\infty$ problem dominated control theory in the 1990's.

IV.5 Open versus closed systems

Control theory, together with signal processing, are typical areas of applied mathematics that study *'open'* systems, i.e. dynamical systems whose evolution does not only depend on the initial conditions and the system laws, but also on the environment. In the input/output setting, this influence of the environment is captured by the inputs. Dynamical systems whose evolution depend on the environment do not only occur in control and signal processing. Indeed, they are the rule, not the exception. Many of the basic laws encountered in physics and elsewhere are basically open systems. Newton's second law gives the relation between position and force, but it does not explain the force. Modelling the force is not part of Newton's second law. It should be viewed as imposed by the environment. In Maxwell's equations, relating the electric field, magnetic field, current density, and charge density, there are also free variables. In economic models, it may be natural to view things as consumer demand, or exchange rates as exogenous inputs. The study of open systems is indeed a much more cogent framework in which to cast the study of dynamical systems than the rather restricted framework of flows (on manifolds),

$$\dot{x} = F(x),$$

which has carried the day in mathematics as a formalization of dynamics.

The question remains what a suitable mathematical framework is for the study of open systems. We have argued that flows are not, but are input/output systems the panacea? In signal processing, and in control using actuators and sensors, the input/output framework is very natural indeed. However, most physical and economic systems do not behave as signal processors, and this has led to the need for a broader framework, behavioral systems.

IV.6 Behavioral systems

In order to capture clearly the notion of a behavioral system, we start this section with some introductory remarks about the nature of a mathematical models. We view a model as follows. Assume that we have a phenomenon (that is, a set of outcomes) that we try to model. Nature (that is, the reality that governs this phenomenon) can produce certain outcomes. The totality of these possible outcomes (*before* we have modelled the phenomenon) forms a set $\mathbb{U}$, called the *universum*. A *mathematical model* restricts the outcomes that a model declares possible to a subset $\mathfrak{B}$ of $\mathbb{U}$: according to the model, outcomes in $\mathfrak{B}$ are possible, those not in $\mathfrak{B}$ are not possible. This subset $\mathfrak{B}$ is called the *behavior* of the model.

As an example, assume that we wish to model the relation between the temperature T, the pressure P, and the volume V of a gas in a container. Before we have modelled this phenomenon, every $(T,P,V) \in \mathbb{R}^3_+ = \mathbb{U}$ is possible. After we have modelled (for example for an ideal gas), we obtain the behavior $\mathfrak{B} = \{(T,P,V) \in \mathbb{R}^3_+ \mid \frac{PV}{T} = \text{ a constant}\}$.

In the study of dynamical systems we are, more specifically, interested in situations where the outcomes of the phenomenon are signals, i.e., maps with independent variables (time for the case at hand) and dependent variables (the space where the time-signals that are being modelled take on their values). In this case the universum is therefore the collection of all maps from the set of independent variables to the set of dependent variables. It is hence convenient to distinguish these sets explicitly in the notation: $\mathbb{T}$ (suggesting 'time') for the set of independent, and $\mathbb{W}$

for the set of dependent variables. Whence we define a *dynamical system* as a triple

$$\Sigma = (\mathbb{W}, \mathbb{T}, \mathfrak{B})$$

with $\mathbb{T} \subseteq \mathbb{R}$, and $\mathfrak{B}$, the behavior, a subset of $\mathbb{W}^{\mathbb{T}}$, ($\mathbb{W}^{\mathbb{T}}$ is the standard mathematical notation for the set of all maps from $\mathbb{T}$ to $\mathbb{W}$). In the behavioral framework, therefore, a dynamical system is simply identified with a family of trajectories.

For example, *Newton's second law* imposes a restriction that relates the position $\vec{q}$ of a point mass and the force $\vec{F}$ acting on it. This relation is $\vec{F} = m\frac{d^2}{dt^2}\vec{q}$, with m the mass. This is a dynamical system with $\mathbb{T} = \mathbb{R}, \mathbb{W} = \mathbb{R}^3 \times \mathbb{R}^3$ (typical elements of $\mathfrak{B}$ are maps $w = (\vec{q}, \vec{F}) : \mathbb{R} \to \mathbb{R}^3 \times \mathbb{R}^3$) and behavior $\mathfrak{B}$ consisting of all maps $t \in \mathbb{R} \mapsto (\vec{q}, \vec{F})(t) \in \mathbb{R}^3 \times \mathbb{R}^3$ that satisfy the differential equation $\vec{F} = m\frac{d^2}{dt^2}\vec{q}$.

First principles laws in physics always state that some outcomes can happen (those satisfying the model equations) while others cannot happen (those violating the model equations). This is a far distance from specifying a system as being driven by free inputs which together with an initial state specifies the other variables, the outputs. The behavioral framework treats a model for what it is: an exclusion law.

IV.7 Latent variables

In the basic equations describing mathematical models, very often other variables appear in addition to those whose behavior the model aims at describing. The origin of these auxiliary variables varies from case to case. They may be state variables (as in automata and input/state/output systems); they may be potentials (as in the well-known expressions for the solutions of Maxwell's equations); they may be interconnection variables that express how the subsystems of an interconnected system interact. It is important to incorporate these auxiliary variables in our basic modeling language *ab initio*, and to distinguish clearly between the variables whose behavior the model aims at, and the auxiliary variables introduced in the modeling process. We call the former *manifest* variables, and the latter *latent* variables.

A *mathematical model with latent variables* is defined as a triple $(\mathbb{W}, \mathbb{L}, \mathfrak{B}_{\text{full}})$ with $\mathbb{W}$ the universum of manifest variables, $\mathbb{L}$ the universum of latent variables, and $\mathfrak{B}_{\text{full}} \subseteq \mathbb{W} \times \mathbb{L}$ the *full behavior*. It induces the *manifest model* $(\mathbb{W}, \mathfrak{B})$, with

$$\mathfrak{B} = \{w \in \mathbb{W} \mid \text{there exists } \ell \in \mathbb{L} \text{ such that} (w, \ell) \in \mathfrak{B}_{\text{full}}\}.$$

A *dynamical system with latent variables* is defined completely analogously as $\Sigma_{\text{latent}} = (\mathbb{T}, \mathbb{W}, \mathbb{L}, \mathfrak{B}_{\text{full}})$ with $\mathfrak{B}_{\text{full}} \subseteq (\mathbb{W} \times \mathbb{L})^{\mathbb{T}}$. It induces the manifest dynamical system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$, with

$$\mathfrak{B} = \{w : \mathbb{T} \to \mathbb{W} \mid \text{ there exists } \ell : \mathbb{T} \to \mathbb{L} \text{ such that } (w, \ell) \in \mathfrak{B}_{\text{full}}\}.$$

The notion of a system with latent variables is usually the natural end-point of a modeling process, for example by tearing and zooming, and hence a very natural starting point for the analysis and synthesis of systems. We shall see that latent variables enter also very forcefully in representation questions.

Examples:

1. In modeling the port behavior of an electrical circuit, the manifest variables are the voltage, or vector of voltages, V across the port(s) and the current, or vector of currents, I into the circuit through the port(s). However, it is usually not possible to come up directly with a model (say in the form of a differential equation) that involves only the port variables (V, I). In order to model the port behavior, we usually need to look at the internal structure of the circuit, and introduce the currents through, and the voltages across, the internal branches as latent variables. Using Kirchhoff's laws and the constitutive equations of the elements in the branches, this readily leads to a latent variable model for the port variables. Similar situations occur in other systems, for example in mechanical systems with subcomponents, and, more generally, in any type of interconnected system.

2. Assume that we want to model the relation between the temperatures and the heat flows radiated at the ends of a heat conducting bar of length 1. The bar is assumed to be isolated, except at the ends. Denote the manifest variables by T_0, T_1, q_0, q_1. These are the temperatures and heat flows at both ends. In order to obtain a model, it is convenient to introduce the temperature distribution $T(x,t), 0 \leq x \leq 1$, in the bar as latent variables. The full behavior is then described by the partial differential equation

$$\frac{\partial}{\partial t} T = \frac{\partial^2}{\partial x^2} T$$

combined with the boundary conditions

$$T(0,t) = T_0(t), T(1,t) = T_1(t), \frac{\partial}{\partial x} T(0,t) = -q_0(t), \frac{\partial}{\partial x} T(1,t) = q_1(t).$$

This defines a latent variable system with $\mathbb{T} = \mathbb{R}, \mathbb{W} = \mathbb{R}^4, \mathbb{L} = \mathbb{R}^{[0,1]}$, and $\mathfrak{B}_{\text{full}}$ consisting of all maps $((T_0, T_1, q_0, q_1), T) : \mathbb{R} \to \mathbb{R}^4 \times \mathbb{R}^{[0,1]})$ such that the above PDE and boundary conditions are satisfied.

3. In a state model for a dynamical system, the input/output behavior is specified through a system of differential equations as

$$\dot{x} = F(x, u), \quad y = H(x, u).$$

This defines a latent variable system $(\mathbb{R}, \mathbb{U} \times \mathbb{Y}, \mathbb{X}, \mathfrak{B}_{\text{full}})$ with $\mathfrak{B}_{\text{full}}$ all trajectories $((\omega, \eta), \xi) : \mathbb{R} \to (\mathbb{U} \times \mathbb{Y}) \times \mathbb{X}$ that satisfy

$$\frac{d}{dt}\xi(t) = F(\xi(t), \omega(t)), \quad \eta(t) = H(\xi(t), \omega(t),$$

for all $t \in \mathbb{R}$. The manifest behavior is the input/output behavior, that is all trajectories $(\omega, \eta) : \mathbb{R} \to (\mathbb{U} \times \mathbb{Y})$ that are 'supported' (in the sense made apparent by the full behavior) by a state trajectory $\xi : \mathbb{R} \to \mathbb{X}$.

Situations in which models use latent variables either for mathematical reasons or in order to express the behavioral constraints abound: internal voltages and currents in electrical circuits, momenta in mechanics, chemical potentials, entropy and internal energy in thermodynamics, prices in economics in order to explain production, state variables, the wave function in quantum mechanics in order to explain observables, the basic probability space Ω in probability, etc.

In conclusion, the behavior, combined with manifest and latent variables forms a very attractive framework for discussing mathematical models in general, and dynamical systems in particular. They have the standard structures studied traditionally (flows, input/output systems, input/state/output systems, etc.) as special cases. But, more importantly, they formalize what is usually the end result of a modelling procedure, and that can definitely not be said from any of these special cases.

IV.8 Differential systems

The 'ideology' that underlies the behavioral approach is the belief that in a model of a dynamical (physical) phenomenon, it is the behavior $\mathfrak{B}$, i.e., a set of trajectories $w : \mathbb{T} \to \mathbb{W}$, that is the central object of study. However, as we have seen, in first principles modeling, also latent variables enter *ab initio*. But, the set $\mathfrak{B}$ or $\mathfrak{B}_{\text{full}}$ of trajectories must be specified somehow, and it is here that differential (and difference) equations enter the scene. Of course, there are important examples where the behavior is specified in other ways (for example, in Kepler's laws for planetary motion), but differential equations are certainly the most prevalent specification of the behavior of a dynamical system encountered in applications. For $\mathbb{T} = \mathbb{R}$, and in the case of systems without latent variables, $\mathfrak{B}$ then consists of the solutions of a system of differential equations as

$$F_1(w(t), \frac{d}{dt}w(t), \ldots, \frac{d^N}{dt^N}w(t)) = F_2(w(t), \frac{d}{dt}w(t), \ldots, \frac{d^N}{dt^N}w(t)).$$

We call these *differential* systems.

Of particular interest (at least in control, signal processing, circuit theory, etc.) are systems with a signal space that is a finite-dimensional vector space and behavior described by linear constant-coefficient differential equations. A *linear time-invariant differential system* is a dynamical system $\Sigma = (\mathbb{R}, \mathbb{W}, \mathfrak{B})$, with $\mathbb{W}$ a finite-dimensional (real) vector space, say $\mathbb{R}^{\mathtt{w}}$, whose behavior consists of the solutions of

$$R(\frac{d}{dt})w = 0,$$

with R a real polynomial matrix of suitable dimensions. Of course, the number of columns of R equals the dimension of $\mathbb{W}$. The number of rows of R, which represents the number of equations, is arbitrary. In fact, when the row dimension of R is less than its column dimension, as is usually the case, $R(\frac{d}{dt})w = 0$ is what is sometimes called an *under-determined* system of differential equations which is typical for models in which the influence of the environment is taken into account. Indeed, in general, some of the components of w are left free by the differential equation $R(\frac{d}{dt})w = 0$. They are specified by the environment. The initial conditions are in general insufficient to determine the solution $w : \mathbb{R} \to \mathbb{R}^{\mathtt{w}}$.

The precise definition of what we consider a solution of $R(\frac{d}{dt})w = 0$ is an issue that we will slide over. But, to be precise, we take it to be the set of $\mathfrak{C}^\infty$ solutions. Then the behavior defined by $R(\frac{d}{dt})w = 0$ equals

$$\mathfrak{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \mid R(\frac{d}{dt})w = 0\}.$$

Hence $\mathfrak{B} = \ker(R(\frac{d}{dt}))$, with $\ker(R(\frac{d}{dt}))$ viewed as a map from $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$ to $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathrm{rowdim}(R)})$. We hence call $R(\frac{d}{dt})w = 0$ a *kernel representation* of the associated dynamical system. There are many other useful system representations, and we shall meet some later on in this chapter.

The dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \ker(R(\frac{d}{dt}))$ is *linear* (the behavior is a linear space), *time-invariant* (the behavior is invariant under time-shifts) and *differential* (the behavior is the solution set of a set of differential equations). We denote this set of dynamical systems by $\mathfrak{L}^{\mathtt{w}}$, when the signal space is $\mathbb{R}^{\mathtt{w}}$, i.e., the number of manifest variables is $\mathtt{w}$. Denote further $\mathfrak{L}^\bullet = \cup_{\mathtt{w} \in \mathbb{Z}_+} \mathfrak{L}^{\mathtt{w}}$.

A differential system with latent variables, is described by

$$\begin{aligned} F_1(w(t), \frac{d}{dt}w(t), \ldots, \frac{d^N}{dt^N}w(t), \ell(t), \frac{d}{dt}\ell(t), \ldots, \frac{d^N}{dt^N}\ell(t)) \\ = F_2(w(t), \frac{d}{dt}w(t), \ldots, \frac{d^N}{dt^N}w(t), \ell(t), \frac{d}{dt}\ell(t), \ldots, \frac{d^N}{dt^N}\ell(t)), \end{aligned}$$

relating the (vector of) manifest variables w to the (vector of) latent variables ℓ. In the linear time-invariant case this becomes

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell,$$

with R and M real polynomial matrices of suitable dimension. Define the *manifest* behavior of this system as

$$\{w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \mid \text{ there exists } \ell \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\ell}) \text{ such that } R(\frac{d}{dt})w = M(\frac{d}{dt})\ell\}.$$

The question occurs whether $\mathfrak{B}$ itself can be described by a linear constant coefficient differential equation. In other words: *Is the projection of the set of solutions of a system of linear time-invariant differential equations itself the set of solutions of a system of linear constant coefficient equations?* This is the case indeed. In fact: For any real polynomial matrices (R,M) with $\text{rowdim}(R) = \text{rowdim}(M)$, there exists a real polynomial matrix R' such that the manifest behavior of $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$ has the kernel representation $R'(\frac{d}{dt})w = 0$.

The above result is called the *elimination theorem* for linear differential systems. Its relevance in object-oriented modeling is as follows. As can already be seen from simple electrical circuits, a model obtained in the usual way involves very many variables and equations, among them many algebraic ones. The elimination theorem tells us that the latent variables may be eliminated and (in the case of linear time-invariant differential systems) that the number of equations can be reduced to no more than the number of manifest variables. Of course, the order of the differential equation goes up in the elimination process.

The elimination theorem also holds for linear constant coefficient partial differential equations [109], but, we emphasize, not for nonlinear differential systems. In other words, the fact that the components of an interconnected system are all described by smooth nonlinear differential equations is no guarantee for the manifest behavior to be described by a differential equation. In fact, only in exceptional cases will the manifest behavior of an interconnected system be described by a system of differential equations involving manifest variables only.

In the next three sections, we will show how some classical notions from mathematical systems theory become generalized to behavioral systems.

IV.9 Controllability

An important property in the analysis and synthesis of dynamical systems is controllability. Controllability refers to be ability of transferring a system from one mode of operation to another. By viewing the first mode of operation as undesired and the second one as desirable, the relevance to control and other areas of applications becomes clear. However, the notion of controllability has found many other applications, for example in stochastics, in the theory of state representations, in model reduction, etc.

The concept of controllability has originally been introduced in the context of state space systems. The classical definition runs as follows. The system described by the controlled vector-field

$$\dot{x} = F(x,u),$$

with $u \in \mathbb{R}^{\mathtt{m}}$, $x \in \mathbb{R}^{\mathtt{n}}$, and $F : \mathbb{R}^{\mathtt{n}} \times \mathbb{R}^{\mathtt{m}} \to \mathbb{R}^{\mathtt{n}}$ is said to be *controllable* if for all $a, b \in \mathbb{R}^{\mathtt{n}}$ there exists $T \geq 0$ and $\omega : [0,T] \to \mathbb{R}^{\mathtt{m}}$, such that the solution to

$$\frac{d}{dt}x(t) = F(x(t), \omega(t))$$

with initial $x(0) = a$ yields $x(T) = b$. One of the elementary results of system theory states that the finite-dimensional linear system

$$\dot{x} = Ax + Bu$$

is controllable if and only if the composite matrix

$$[B \quad AB \quad A^2B \quad \cdots \quad A^{\dim(x)-1}B]$$

has linearly independent rows. Various generalizations of this result to time-varying, to nonlinear (with controllability conditions involving Lie brackets), and to infinite-dimensional systems exist.

A disadvantage of this classical notion of controllability is that it refers to a particular representation of a system, notably to a state space representation. Thus a system may be uncontrollable either for the intrinsic reason that the control has insufficient influence on the system variables, or because the state has been chosen in an inefficient way. It is clearly not good to confuse these reasons. It is obviously desirable to have a notion of controllability that can handle other system representations than state representations. In the context of behavioral systems, a definition of controllability has been put forward that involves the system variables directly, and that can be applied to any system representation.

Let $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ be a dynamical system with $\mathbb{T} = \mathbb{R}$ or $\mathbb{Z}$, and assume that is time-invariant, that is $\mathfrak{B}$ is invariant under time shift, is said to be *controllable* if for all $w_1, w_2 \in \mathfrak{B}$ there exists $T \in \mathbb{T}$, $T \geq 0$ and $w \in \mathfrak{B}$ such that $w(t) = w_1(t)$ for $t < 0$ and $w(t) = w_2(t-T)$ for $t \geq T$. Thus controllability refers to the ability to steer the system from any one trajectory in the behavior to any other one, allowing some time-delay.

Two questions that occur are the following: *What conditions on the parameters of a system representation imply controllability? Do controllable systems admit a particular representation in which controllability becomes apparent?* For linear time-invariant differential systems, these questions can be answered very explicitly, as follows. Let $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \mathfrak{B})$ be a linear time-invariant differential system. The following are equivalent:

1. $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \mathfrak{B}) \in \mathfrak{L}^{\mathtt{w}}$ is controllable;
2. The polynomial matrix R in a kernel representation $R(\frac{d}{dt})w = 0$ of $\mathfrak{B}$ satisfies $\mathrm{rank}(R(\lambda)) = \mathrm{rank}(R)$ for all $\lambda \in \mathbb{C}$;
3. The behavior $\mathfrak{B}$ is the image of a linear constant-coefficient differential operator, that is, there exists a real polynomial matrix M of suitable dimension such that $\mathfrak{B} = \{w \in \mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \mid w = M(\frac{d}{dt})\ell \text{ for some } \ell \in \mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathrm{coldim}(\ell)})\}$.

There exist effective algorithms for verifying controllability of a system starting from the coefficients of the polynomial matrix R in a kernel (or a latent variable) representation of Σ, but we will not enter into these algorithmic aspects.

A point of the above result that is worth emphasizing is that controllable systems admit a representation as the manifest behavior of the latent variable system of the special form

$$w = M(\frac{d}{dt})\ell.$$

We call this an *image* representation of the w-behavior. It follows from the elimination theorem that every system in image representation can be brought in kernel representation. But not every system in kernel representation can be brought in image representation: it is precisely the controllable ones for which this is possible.

The controllability issue has been pursued for many other classes of systems. In particular (more difficult to prove) generalizations have been derived for differential-delay, for linear *n-D* systems, for linear constant coefficient *PDE*'s, and for limited classes of nonlinear differential systems. Systems in an image representation, and close relatives of it, have received much attention recently for nonlinear differential-algebraic systems, where they are referred to as *flat* systems.

IV.10 Observability

A second notion that has become central in systems theory is that of observability. The notion of observability is always introduced hand in hand with controllability.

In the context of the input/state/output system

$$\dot{x} = F(x,u), \quad y = H(x,u)$$

it refers to the possibility of deducing, using the laws of the system, the state from observation of the input and the output trajectories. For the linear system

$$\dot{x} = Ax + Bu, \quad y = Cx + Du,$$

state observability holds if and only if the composite matrix

$$[C^\top \quad A^\top C^\top \quad (A^\top)^2 C^\top \quad \cdots \quad (A^\top)^{\dim(x)-1} C^\top]$$

has linearly independent rows.

The definition that is used in the behavioral context is more general in that the variables that are observed and the variables that need to be deduced from the observed variables are kept general.

Let $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ be a dynamical system, and assume that $\mathbb{W}$ is a produt space, $\mathbb{W} = \mathbb{W}_1 \times \mathbb{W}_2$. Hence elements of $\mathfrak{B}$ are pairs $(w_1, w_2) : \mathbb{T} \to \mathbb{W}_1 \times \mathbb{W}_2$. Then w_1 is said to be *observable* from w_2 in Σ if $(w_1, w_2') \in \mathfrak{B}$ and $(w_1, w_2'') \in \mathfrak{B}$ imply $w_2' = w_2''$. Observability thus refers to the possibility of deducing the trajectory w_1 from observation of w_2 and from the laws of the system ($\mathfrak{B}$ is assumed to be known).

The theory of observability runs parallel to that of controllability. Let $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}_1} \times \mathbb{R}^{\mathtt{w}_2}, \mathfrak{B})$ be a linear time-invariant differential system. The following are equivalent:

1. In $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}_1} \times \mathbb{R}^{\mathtt{w}_2}, \mathfrak{B})$, w_2 is observable from w_1;
2. The polynomial matrix R_2 in a kernel representation $R_1(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$ of $\mathfrak{B}$ satisfies $\operatorname{rank}(R_2(\lambda)) = \dim(w_2)$ for all $\lambda \in \mathbb{C}$;
3. there exists a set of differential equations satisfied by ('consequences of') the behavior, of the form $w_1 = R_2'(\frac{d}{dt})w_2$, which puts observability into evidence.

IV.11 State representations

As shown in the previous sections, there are many ways to represent a system, as kernels, as images, with latent variables, etc. The fact that systems are often modelled with latent variables and this obviously induces a multitude of descriptions of a given manifest behavior. One very important system representation is the state space representation, as (IV.4). It is the introduction of this class of system representations, combined with optimal control, that propelled and greatly enhanced the mathematical and applications scope of systems and control theory in the early 1960's. The state property is a consequence of the fact that this is a first order differential equation. In this section, we will formalize the state concept. We will now explain how the state can be viewed in the context of latent variable systems in which the latent variable has the additional property that it separates the past from the future.

Let $\Sigma = (\mathbb{T}, \mathbb{W}, \mathbb{L}, \mathfrak{B}_{\text{full}})$ be a latent variable dynamical system. Then this system is said to be a *state system* if $\mathfrak{B}_{\text{full}}$ has the following property: if $(w_1, \ell_1) \in \mathfrak{B}_{\text{full}}$, $(w_2, \ell_2) \in \mathfrak{B}_{\text{full}}$ and $\ell_1(t_0) = \ell_2(t_0)$, then the concatenation of (w_1, ℓ_1) and (w_2, ℓ_2) at t_0 must also belong to $\mathfrak{B}_{\text{full}}$. Concatenation is defined as

$$(w_1, \ell_1) \wedge_{t_0} (w_2, \ell_2)(t) = \begin{cases} (w_1, \ell_1)(t) & for \quad t < t_0 \\ (w_2, \ell_2)(t) & for \quad t \geq t_0 \end{cases}$$

This property must hold for all such (w_1, ℓ_1), $(w_2, \ell_2) \in \mathfrak{B}$, and all $t_0 \in \mathbb{T}$. Henceforth, we will use the notation x instead of ℓ when a latent variable representation has, or is expected to have, the state property.

It is possible to prove, under some mild regularity conditions, that a differential system is a state space system if and only if $\mathfrak{B}_{\text{full}}$ can be described by a differential equation that is *first order* in x and *zero-th order* in w, i.e., a differential equation of the form

$$F_1(w, x, \dot{x}) = F_2(w, x, \dot{x}) .$$

In particular, for linear time-invariant differential systems, this yields the representation

$$Kw + Lx + M\frac{d}{dt}x = 0 ,$$

with K, L, M matrices of suitable dimensions.

It can be shown that for every linear time-invariant differential dynamical system (and hence, by the elimination theorem, for every system described by linear constant-coefficient differential equations with latent variables) it is possible to define latent variables $x = X(\frac{d}{dt})w$ through the polynomial matrix X ($X(\frac{d}{dt})$ is then called a *state map*) such that the full behavior defined by

$$\{(w, x) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}+\dim(x)}) \mid R(\frac{d}{dt})w = 0, x = X(\frac{d}{dt})w\}$$

satisfies the state property. In other words, such that the variables w, x are related by a linear constant coefficient differential equation that is *first order* in x and *zero-th* *order* in w. Moreover, if the dimension of x (equivalently, the number of rows of X) is as small as possible (under the constraint that the relation between x and w is a state system), then this first order state equation can actually be written as

$$\frac{d}{dt}x = Ax + Bw_1, \quad w_2 = Cx + Dw_1,$$

with (w_1, w_2) a component-wise partition of w, i.e., $(w_1, w_2) = w$ up to a permutation of the components.

This shows how the popular input/state/output representation emerges as a special system representation that involves (i) a suitable, component-wise, partition of the manifest variables, and (ii) a suitable choice of the latent variables to represent the behavior in state form. Hence such an input/state/output representation displays both the partition of its manifest variables in its free (input) components and its bound (output) components, and also its memory (as parameterized by the state variables). The usefulness of such representations for the analysis and synthesis is uncontested. However, as a *starting point* for practical and theoretical work, it is not suitable. It requires a great deal of manipulations and computation to bring a first principles model into this form. For example, first principles models invariably contain algebraic equations, and more latent variables that only the state variables.

IV.12 Control as interconnection

The behavioral point of view is quite compelling as an approach for modelling (open) dynamical systems. It is reasonable that when modelling a physical dynamical system it makes no sense to prejudice oneself (as one would be forced to do in an input/output setting) as to which variables should be viewed as inputs and which variables should be viewed as outputs. We have argued this point extensively throughout the previous sections. This is not to say, by any means, that there are no situations where the input/output structure is natural. Quite to the contrary. In fact, whenever logic devices are involved, in information processing, the input/output structure is often a must.

On first sight, the behavioral approach seems less apt in the context of control. Firstly, there is something very natural in viewing control variables as inputs and measured variables as outputs. Control then becomes decision making on the basis of observations. When subsequently a controller is regarded as a feedback processor, one ends up with the feeling that the input/output structure is in fact an essential feature of control. Secondly, since, as we have seen, it is possible to prove that every linear time-invariant differential system admits a component-wise input/output representation, one gets the impression that the input/output framework can be adopted without second thoughts, that nothing is lost by taking it as a universal starting point. However, this partition in inputs and outputs may not be possible for smooth nonlinear differential systems.

The intelligent control paradigm embodied in figure IV.1 has been universally prevalent ever since the beginning of the subject, from our interpretation of the Watt regulator, Black's feedback amplifier, and Wiener's cybernetics, to the ideas underlying modern adaptive and robust control. It is indeed a deep and very appealing paradigm, which will undoubtedly gain in relevance and impact as logic devices become ever more prevalent, reliable, and inexpensive. This paradigm has a number of features which are important for considerations which will follow. Some of these are:

1. There is an asymmetry between the plant and the controller: it remains apparent what part of the system is the plant and what part is the controller. This asymmetry disappears to some extent in the closed loop.

2. The intelligent control paradigm tells us to be wary of errors in the measurements. Thus it is considered as being ill-advised to differentiate measurements, presumably, because this will lead to noise amplification.

3. The plant and the controller are dynamical systems which can be interconnected at any moment in time. If for one reason or another the feedback controller temporarily fails to receive an input signal, then the control input can be set to a default value, and later on the controller can resume its action.

However, in many practical control problems, for example, the damper of an automobile, pressure control valves, heat radiation fins, and similar passive controllers, simply do not function as signal processors. The signal-flow-graph interpretation of figure IV.1 is untenable. In these examples, there is no asymmetry between plant and controller, there are no sensors, no actuators, no measurements, and the state of the plant needs to be prepared before the controller can be put in place.

In this section, we will sketch a broader framework for control. A framework that incorporates the above examples, and in which the feedback control appears as a special case. The solution which we propose is the following. We will keep the distinction between the plant and the controller. However, we will abandon the intelligent control signal flow graph as a paradigm for control. Instead, we will view *interconnection of a controller to a plant* as the central paradigm for control theory. However, we by no means claim that the intelligent control paradigm is without merits. To the contrary, it is extremely useful and important. Claiming that the input/output framework is *not always* the suitable framework to approach a problem does not mean that one claims that it is *never* the suitable framework. However, a good universal framework for control should be able to deal both with interconnection, with designing subsystems, and with intelligent control. *The behavioral framework does, the intelligent control framework does not.*

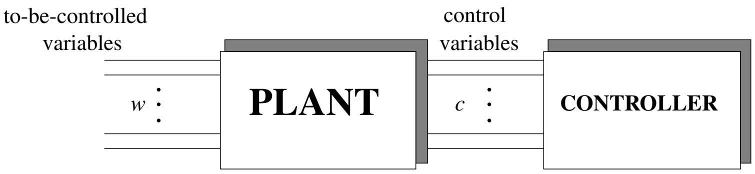


Figure IV.2: Control as interconnection

The nature of control that we would like to transmit is illustrated in the system configuration depicted in figure IV.2. In the top part of the figure, there are two systems, shown as black-boxes with terminals. It is through their terminals that systems interact with their environment. The black-box imposes relations on the variables that 'live' on its terminals. These relations are formalized by the behavior of the system in the black-box. The system to the left in figure IV.2 is called the *plant*, the one to the right the *controller*. The terminals of the plant consist of *to-be-controlled variables* w, and *control variables* c. The controller has only terminals with the control variables c. Before interconnection, the variables w and c of the plant have to satisfy the laws imposed by the plant behavior. After interconnection, the variables c also have to satisfy the laws imposed by the controller. Thus, after interconnection, the restrictions imposed on the variables c by the controller will be transmitted to the variables w. Choosing the black-box to the right so that the variables w have a desirable behavior in the interconnected black-box is, in our view, the basic problem of control.

This leads to the following mathematical formulation. The *plant* and the *controller* are both dynamical systems $\Sigma_{\text{plant}} = (\mathbb{T}, \mathbb{W} \times \mathbb{C}, \mathfrak{B}_{\text{plant}})$ and $\Sigma_{\text{controller}} = (\mathbb{T}, \mathbb{C}, \mathfrak{B}_{\text{controller}})$ where $\mathbb{W}$ denotes the signal space of the to-be-controlled variables, $\mathbb{C}$ denotes the signal space of control variables, and both systems are assumed to have the same time axis $\mathbb{T}$. The interconnection of Σ_{plant} and $\Sigma_{\text{controller}}$ leads to the system $\Sigma_{\text{latent}} = (\mathbb{T}, \mathbb{W}, \mathbb{C}, \mathfrak{B}_{\text{full}})$ which is a system with latent variables ($\mathbb{C}$) and full behavior defined by

$$\mathfrak{B}_{\text{full}} = \{(w,c) : \mathbb{T} \to \mathbb{W} \times \mathbb{C} \mid (w,c) \in \mathfrak{B}_{\text{plant}} \text{ and } c \in \mathfrak{B}_{\text{controller}}\}\mathbb{N}$$

The manifest system defined by Σ_{latent} is the *controlled system* and is hence defined as $\Sigma_{\text{controlled}} = (\mathbb{T}, \mathbb{W}, \mathfrak{B}_{\text{controlled}})$ with

$$\mathfrak{B}_{\text{controlled}} = \{w : \mathbb{T} \to \mathbb{W} \mid \text{ there exists } c : \mathbb{T} \to \mathbb{C} \text{ such that}\mathbb{N} \\ (w, c) \in \mathfrak{B}_{\text{plant}} \text{ and } c \in \mathfrak{B}_{\text{controller}}\}\mathbb{N}$$

The notion of a controller put forward by the above view considers interconnection as the basic idea of control. The special controllers that consist of sensor-outputs to actuator-inputs signal processors emerge as (very important) special cases. However, our view of control as the design of suitable subsystems greatly enhances the applicability of control. It views control as integrated subsystem design.

A question that arises in this context is the following. Assume that Σ_{plant} is given. What systems $\Sigma_{\text{controlled}}$ can be obtained by suitably choosing $\Sigma_{\text{controller}}$? This question can be answered very explicitly, at least for linear time-invariant differential systems.

Consider the plant given by $\Sigma_{\text{plant}} = (\mathbb{R}, \mathbb{R}^{\mathtt{w}} \times \mathbb{R}^{\mathtt{c}}, \mathfrak{B}_{\text{plant}}) \in \mathbb{L}^{\mathtt{w}+\mathtt{c}}$ with $\mathfrak{B}_{\text{plant}}$ described by a system of linear constant differential equations $R(\frac{d}{dt})w = R(\frac{d}{dt})c$. Let $\Sigma_{\text{controller}} = (\mathbb{R}, \mathbb{R}^{\mathtt{c}}, \mathfrak{B}_{\text{controller}}) \in \mathfrak{L}^{\mathtt{c}}$, and assume that $\mathfrak{B}_{\text{controller}}$ is similarly described by a system of linear constant coefficient differential equations $C(\frac{d}{dt})c = 0$. Then, by elimination theorem $\Sigma_{\text{controlled}} = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \mathfrak{B}_{\text{controlled}})$ has also a behavior that is described by a system of linear constant coefficient differential equations. It turns out that the behaviors $\mathfrak{B}_{\text{controlled}}$ that can be obtained this way can be characterized very nicely.

Define therefore two behaviors, $\mathfrak{P}$ and $\mathfrak{N}$; $\mathfrak{P}$ is called the *realizable (plant) behavior* and $\mathfrak{N}$ the *hidden behavior*. They are defined as follows: $\mathfrak{P}$ is the manifest behavior of the plant, with c viewed as a latent variable, i.e.,

$$\mathfrak{P} = \{w : \mathbb{R} \to \mathbb{R}^{\mathtt{w}} \mid \text{ there exists } c : \mathbb{R} \to \mathbb{R}^{\mathtt{c}} \text{ such that } (w, c) \in \mathfrak{B}_{\text{full}}\},$$

and $\mathfrak{N}$ is defined as

$$\mathfrak{N} = \{w : \mathbb{R} \to \mathbb{R}^{\mathtt{w}} \mid (w, 0) \in \mathfrak{B}_{\text{plant}}\}.$$

Hence $\mathfrak{N}$ is the behavior of the plant variables that are compatible with the control variables equal to zero. Thus, whatever be the controller, only signal in $\mathfrak{P}$ will be possible, and the signals in $\mathfrak{N}$ will remain possible.

We say that $\mathfrak{B}_{\text{controller}}$ *implements* $\mathfrak{B}_{\text{controlled}}$ if there exists a controller such that the controlled behavior after interconnecting the controller with behavior $\mathfrak{B}_{\text{controller}}$ to the plant, yields $\mathfrak{B}_{\text{controlled}}$ as the controlled behavior. The controller implementability problem asks what behaviors $\mathfrak{B}_{\text{controlled}}$ can be obtained this way. For linear time-invariant systems it is possible to prove that $\mathfrak{B}_{\text{controlled}}$ is implementable if and only if

$$\mathfrak{N} \subseteq \mathfrak{B}_{\text{controlled}} \subseteq \mathfrak{P}.$$

Hence the behaviors that are implementable by means of a controller are precisely those that are wedged in between the hidden behavior $\mathfrak{N}$ and the realizable plant $\mathfrak{P}$. The problem of control can therefore be reduced to finding such a behavior. Of course, the issue of how to construct $\Sigma_{\text{controller}}$ (for example, as a signal processor from the sensor outputs to the actuator inputs) must be addressed as well, but this can be done.

We believe that the point of view of control that emerges from this, as designing a subsystem (with feedback control as a special case) greatly enhances the scope and applicability of control as a discipline. In this setting, control comes down to sub-system design. More details on the behavioral approach to dynamical systems and control can be found in [157, 111, 158].

IV.13 Conclusions

In this chapter, we have attempted to give a broad overview of control and systems theory.

We started by contrasting open loop control with feedback control. Whereas there are many important and interesting problems of open loop control in the form of trajectory optimization, feedback control is in a sense a more central principle of control engineering. Feedback control leads straight to the study of input/output systems.

We subsequently asked whether the input/output framework is the most logical setting for the study of 'open' dynamical systems. We came to the conclusion that it is not. Most physical systems simply do not function as signal processors. We then introduced a more cogent framework, the behavioral approach. We also discussed that this leads to a view of control as interconnecting subsystems.

Chapter V

From ancient times to the maximum principle

In this chapter we outline the main steps in the evolution that led to the discovery of general necessary conditions for an optimum, focusing on the search for the correct ***form*** of the conditions. We will not worry at this stage about the questions of rigorous mathematical interpretation and proof techniques. These are, of course, important issues, and we shall return to them later.

The evolution can roughly be divided into three periods:

1. ***The era of special problems solved by ad hoc methods,*** extending from ancient times until the birth of the calculus of variations in the 17th century;
2. ***The era of the calculus of variations,*** lasting from the beginning of the 17th century until the 1950s;
3. ***The era of optimal control,*** which began in the 1950s and continues to the present time.

Naturally, the division is only approximate since, for example, some calculus of variations problems truly belong to optimal control, as we shall try to argue later for the case of the brachistochrone.

V.1 The era of special problems solved by ad hoc methods

Like all mathematical theories, the study of general curve minimization problems was preceded by a long period where particular problems were tackled by special

ad hoc methods. In some cases, the effort to solve these problems led to the invention of new methods of greater generality. Eventually, after many special problems had been solved, and the new methods developed to solve them had in turn been applied to other problems, it became clear that there had to exists a systematic theory based on general principles.

However, in their eager search for new tools that would apply to ever larger classes of problems, the creators of the calculus of variations were guilty of some sloppiness, in that they sometimes forgot to acknolewdge their indebtedness to many of the original special problems without which the field might never have been born, by incorporating them into their new theories.

This is easily seen by looking at a representative list of these special problems and asking how well they fit within the framework of the classical calculus of variations. Here we will look at six such problems.

V.1.1 Finding the shortest path between two points.

This is of course the oldest of all curve minimization problems, and the fact that the answer is a straight-line segment has been known for at least 3,000 years. Yet, a careful analysis—to be carried out in §VI.1—will show that this problem does not quite fit the framework of the calculus of variations in a direct and natural way. One can, of course,

a. prove directly that the solution is a segment, without using the calculus of variations,

or

b. use some clever transformation to reduce the problem to one where the Euler-Lagrange equation is applicable,

or

c. give a logically questionable argument[1].

But, as we shall show in §VI.1, it is not possible to apply the necessary conditions of Euler-Lagrange directly to this problem without following one of the three strategies listed above.

[1] that is, cheat

V.1.2 The classical isoperimetric problem.

The oldest curve optimization problem after the shortest-path problem is the *isoperimetric problem* (henceforth abbreviated as "IP"), also known as *Queen Dido's problem*, inspired by the mythical story[2] told by Virgil (70-19 BCE) in the *Aeneid* about the foundation of Carthage (*ca.* 850 BCE).

The question is

(I) *To find the closed plane curve of a given length that encloses the largest possible area.*

It is quite easy to see that this is equivalent[3] to the problem

(II) *To determine the shortest of all closed plane curves that enclose a given area.*

It is a fairly obvious guess that the solution must be a circle. (In modern terms, it is almost evident that the solution—modulo translations—has to be rotationally symmetric, since the question, as stated, treats all directions in the plane equally. This, however, is not a rigorous argument, since it might conceivably happen that there is "symmetry breaking," so that the solution modulo translations is not unique, and each solution individually is not rotationally symmetric, even though the *set* of all solutions is[4]. To transform the rotational symmetry argument into a fully rigorous proof that the solution of the IP is a circle, one would have to establish, for example, uniqueness of the solution modulo translations.)

---

[2]According to Virgil, "when Dido landed in North Africa, she persuaded the local chief to sell her as much land as an oxhide could contain. She then cut a hide into very narrow strips, and joined them to make a continuous thread more than two-and-a-half miles long. She was then able to enclose between the thread and the sea the land on which Carthage was built" (Pars [105]).

[3]*Proof of the equivalence*: Suppose a closed curve γ solves Problem I. This means that γ has length L, encloses an area A, and every closed curve ξ of length L encloses an area $\mathscr{A}(\xi)$ such that $\mathscr{A}(\xi) \le A$. We claim that γ solves Problem II, that is, if ξ is any closed curve that encloses a region with area A, then the length $\mathscr{L}(\xi)$ of ξ satisfies $\mathscr{L}(\xi) \ge L$. Suppose this conclusion was not true. Let ξ be a closed curve that encloses a region with area A and is such that $\mathscr{L}(\xi) < L$. Let $\kappa = \frac{L}{\mathscr{L}(\xi)}$. Let τ be the transformation $\mathbb{R}^2 \ni (x,y) \mapsto (\kappa x, \kappa y) \in \mathbb{R}^2$. Then $\tau(\xi)$ has length L, and encloses an area $\kappa^2\mathscr{A}(\xi) = \kappa^2 A > A$. This contradicts the fact that γ was a solution to Problem I. So γ solves Problem II. The converse implication is easily proved by a similar argument.

[4]This is exactly what happens, for example, for the problem of *minimizing* rather than maximizing an enclosed area. In this case, the set S of solutions is very large, and consists of *all* loops of length P that enclose a zero area. (For example, S contains all "degenerate rectangles" with two opposite sides of length $P/2$ and the other two of length zero.) None of these solutions is rotationally symmetric, but of course the family of all the solutions is symmetric.

The Greeks knew that the solution was be the circle, and realized that this needed proof. ***Zenodorus***[5] (*ca.* 200-140 BCE) showed that

(Z1) among all polygons with a given perimeter P and given number N of sides, the regular polgyon has the largest area,

and

(Z2) the area enclosed by a circle is larger than that enclosed any regular polygon of the same perimeter.

These observations essentially[6] solve the isoperimetric problem and establish that the solution is the circle, although it took until the 19th century for this to be proved in a way that meets our contemporary standards of rigor.

The isoperimetric problem will be discussed in §XV.1. Here, we will just observe that this is one of the problems for which the classical calculus of variations—as presented, for example, in Euler's 1744 book—works relatively well, provided, of course, that one broadens the setting to include problems with integral constraints. (According to Struik [124], p. 399, it is with Euler's 1744 *Methodus inveniendi* ... that "the isoperimetric problem enter[ed] into the calculus.")

We remark, however, that even Euler, in the title of his 1744 book (cf. §V.2 in page 103 below), felt that he had to distinguish between ***two separate cases***, namely, the problems of finding curves "with a maximum or minimum property" and the "isoperimetric problems." This should be constrasted with the fact that in optimal control the distinction between "standard problems" and "problems with integral constraints" disappears, as was shown in §II.10.

V.1.3 Light reflection.

Hero (or Heron) of Alexandria[7] showed in his *Catoptrics* that when a light ray emitted by an object is reflected by a mirror it follows a path from the object to the eye which is the shortest of all possible such paths. Once again, this fact can easily

[5]For more information on Zenodorus's work, cf. Heath [68], pp. 382-383.

[6]A 20th century mathematician might want to add that if ξ is an arbitrary simple rectifiable closed curve, then ξ can be approximated by polygons. Supplemented by this observation, the two results of Zenodorus add up to a complete solution. Naturally, the concept of an "arbitrary rectifiable curve" only makes sense within the context of modern analysis.

[7]Exact dates unknown. For a long time, historians were divided, some of them believing that he had lived around 150 BCE while others thought he had lived on 250 CE. O. Neugebauer in 1938 identified a description by Hero of an eclipse as corresponding to an eclipse that took place in 62 CE. Based on this, it is now believed that Hero was probably born in 10 CE and died in 75 CE. Some historians attribute his optics work to a "Hero the Younger," who may have lived in the 7th or 8th century CE.

be established by a direct geometric argument: suppose a path π goes from point C on one side of the mirror MM to a point D_* on the mirror and then to a point E on the same side as C. Suppose, in addition, that π is the shortest of all such paths going from C to E. Then the parts π_-, π_+ of π that go from C to D_* and from D_* to E, respectively, must obviously be straight-line segments. Let $\tilde{E}$ be the point on the other side of MM such that the segment σ from E to $\tilde{E}$ is perpendicular to MM and intersects MM at the midpoint of σ. (That is, $\tilde{E}$ is the "mirror image" of E, or the result of "reflecting" E with respect to MM.) For any choice of the point D on MM, let π_D be the path consisting of the segment from C to D followed by the segment from D to E. Then we can associate to π_D a path $\tilde{\pi}_D$ going from C to $\tilde{E}$, obtained from π_D by substituting the segment $D\tilde{E}$ for the segment DE. The new path has exactly the same length as π_D. Since the shortest path joing two points is the segment, we see that the length of π_D is minimized when D is chosen such that $\tilde{\pi}_D$ is a segment. This will happen if and only if

$$\theta_i = \theta_r\,,$$

where θ_i is the *angle of incidence* (that is, the angle between the segment CD_* and the line L through D_* perpendicular to MM), and θ_r is the *angle of reflection*, (that is, the angle between CD_* and L). This is the well know law of reflection.

V.1.4 Light refraction.

Pierre de Fermat (1601-1665) formulated the general principle that *light rays follow the fastest—i. e., minimum-time—paths*, and showed how this explained not only Hero's observation about reflection but also Snell' law of refraction. Snell's law is about the light rays in plane medium where the speed of light has a constant value c_+ on one side S_+ of a line L and a different constant value c_- on the other side S_-. To derive the law one must find the path joining two points C_+, C_+ lying on S_+ and S_-, respectively, that minimizes travel time. This can easily be done using calculus, and was done, for example, by Leibniz in his 1684 ***Nova methodus*** paper (cf. Scholium VII, in page 135).

In modern terms, if we choose Cartesian coordinates x, y such that L is the x axis and C_+ lies on the y axis and above L, we can let $(0, y_+)$ be the coordinates of C_+ and (x_-, y_-) be the coordinates of C_-. It is obvious that the minimizing path will consist of a segment from C_+ to a point $\bar{C}$ on L, followed by the segment from $\bar{C}$ to C_-. Let $(\bar{x}, 0)$ be the coordinates of C_-. Then the travel time is

$$T = \frac{1}{c_+}\sqrt{\bar{x}^2 + y_+^2} + \frac{1}{c_-}\sqrt{(x_- - \bar{x})^2 + y_-^2}\,.$$

To minimize T as a function of $\bar{x}$, we must have

$$\frac{\partial T}{\partial \bar{x}} = 0,$$

which yields

$$\frac{\bar{x}}{c_+\sqrt{\bar{x}^2 + y_+^2}} + \frac{\bar{x} - x_-}{c_-\sqrt{(x_- - \bar{x})^2 + y_-^2}} = 0.$$

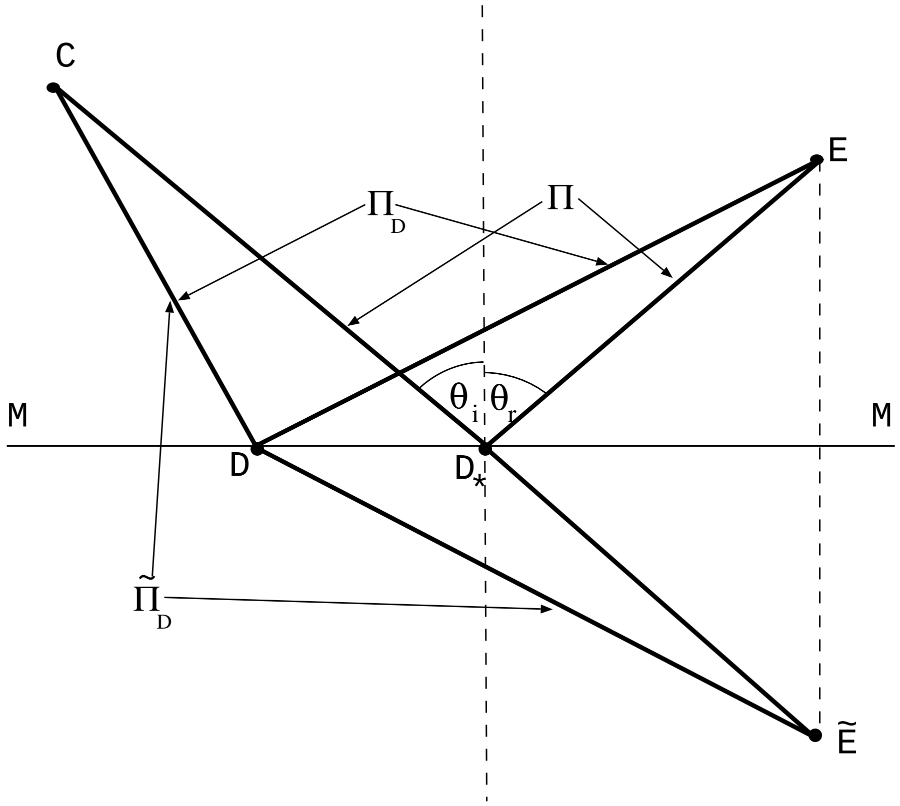


FIGURE 7. Reflection of light

Therefore

$$\frac{\bar{x}}{c_+\sqrt{\bar{x}^2 + y_+^2}} = \frac{x_- - \bar{x}}{c_-\sqrt{(x_- - \bar{x})^2 + y_-^2}} = 0. \tag{V.1}$$

Now, the quotient $\frac{\bar{x}}{\sqrt{\bar{x}^2+y_+^2}}$ is the sine of the "angle of incidence," that is, the angle θ_i between the line $C_+\bar{C}$ and a line perpendicular to L. Similarly, $\frac{\bar{x}}{\sqrt{\bar{x}^2+y_+^2}} = \sin\theta_r$, where θ_r is the "angle of refraction," that is, the angle between the line $\bar{C}C_-$ and a line perpendicular to L. Hence (V.1) says that

$$\frac{\sin\theta_i}{c_+} = \frac{\sin\theta_i}{c_-}. \tag{V.2}$$

V.1.5 The catenary and the brachistochrone.

These problems were solved by Johann Bernoulli in 1691 and 1696, respectively, and will be discussed in Example III of §VI.3 and in Chapter XIV.

V.2 Euler

Leonhard Euler (1707-1793) entered the University of Basel in 1720 at the age of 13[8]. There he studied the methods developed by the Bernoullis for the study of isoperimetric problems, but soon became convinced that there was a need for a general theory, that would go beyond the solutions of specific problems. He began to develop his approach in 1731, and by 1740 he had achieved his goal and come up with a general theory. He published his results in 1744, in a book (Euler [57]) entitled

> *Methodus inveniendi maximi minime propriate gaudentes sive solutio problematis isoperimetrici latissimo sensu accepti* (A method for discovering curved lines having a maximum or minimum property or the solution of the isoperimetric problem taken in its widest sense).

[8]Paul Euler, Leonhard's father, had studied theology at the University of Basel and also attended Jakob Bernoulli's lectures there. Paul and Johann Bernoulli lived in Jakob's house while they were students in Basel. Paul then became a Protestant minister, but he was also interested in mathematics and taught some to his son. Paul wanted Leonhard to become a minister, and sent him to the University of Basel for that reason. At the University, Leonhard was introduced to Johan Bernoulli, and asked him to give him private lessons. Johann refused, but recommended some readings, and allowed Leonhard to visit him on Sundays to ask questions. Thus Johann soon discovered that this young student was a mathematical prodigy. Leonhard completed his Master's degree in philosophy in 1723, and began the study of theology, but he really wanted to pursue mathematics, and Johann was able to persuade his friend Paul to give his consent for this change. Euler completed his studies in Basel in 1726, and in November of that year he accepted a position in St. Petersburg that had become vacant in July, when Nicolaus(II) Bernoulli died. Euler arrived in St. Petersburg in May 1727.

There he gave a general procedure for writing down what became known as *Euler's equation*.

COMMENT NO. 16. Must discuss precise relation between this and Euler-Lagrange

The words "calculus of variations" do not yet appear in Euler's *Methodus inveniendi*. They were coined by him later, in 1760, as a name for Lagrange's method that Euler had decided to adopt in lieu of his own discretization approach, and appeared in print in 1766.

V.3 Lagrange

In H. Goldstine's words ([66], p. 110.):

> On 12 August 1755 one Ludovico de la Grange Tournier, a 19-year-old from Turin, wrote Euler a brief letter to which was attached an appendix containing mathematical details of a very beautiful and revolutionary idea. He saw how to eliminate from Euler's methods the tedium and need for geometrical insight and to reduce the entire process to a quite analytic machine or apparatus, which could turn out the necessary condition of Euler and more, almost automatically. This basic idea of Lagrange ushered in a new epoch in the calculus of variations. Indeed, after seeing Lagrange's work, Euler dropped his own method, espoused that of Lagrange, 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."

Joseph-Louis Lagrange (1736-1813) derived the necessary condition for a curve $t \to \xi_*(t)$ to be a solution of the problem (II.5), known today as the "Euler-Lagrange equation." This condition says[9] that the identity

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial L}{\partial q} \tag{V.3}$$

[9] We use the notations that are common today, not Lagrange's. The symbol ∂ for partial derivative was first used by Legendre in 1786.

must be satisfied along the curve ξ_*.

Equation (V.3) makes perfect sense and is a necessary condition for optimality for a vector-valued variable q as well as for a scalar one. It can be written as a system:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) = \frac{\partial L}{\partial q^i}\,, \quad i = 1,\ldots,n\,. \tag{V.4}$$

Alternatively, we can regard Equation (V.3) as a vector identity, in which

$$q = (q^1,\ldots,q^n)$$

denotes an n-dimensional vector, and the expressions

$$\frac{\partial L}{\partial q}\,, \qquad \frac{\partial L}{\partial \dot{q}}$$

stand for the n-tuples

$$\left(\frac{\partial L}{\partial q^1},\ldots,\frac{\partial L}{\partial q^n}\right), \qquad \left(\frac{\partial L}{\partial \dot{q}^1},\ldots,\frac{\partial L}{\partial \dot{q}^n}\right).$$

A modern mathematician might be troubled by the use of $\dot{q}$ both as an "independent variable" and as a function of time evaluated along a trajectory, and might prefer to write (V.3) as

$$\boxed{\frac{d}{dt}\left[\frac{\partial L}{\partial u}\Big(\xi(t),\dot{\xi}(t),t\Big)\right] = \frac{\partial L}{\partial q}\Big(\xi(t),\dot{\xi}(t),t\Big)\,,\ i=1,\ldots,n,} \tag{V.5}$$

where the Lagrangian $L(q,u,t)$ is a function on $\mathbb{R}^{2n+1}$, i.e., a function of $q\in\mathbb{R}^n$, $u\in\mathbb{R}^n$, $t\in\mathbb{R}$. This makes it clear that to compute the left-hand side of (V.3) one first evaluates $\frac{\partial L}{\partial \dot{q}}$ "treating $\dot{q}$ as an independent variable," then plugs in $\xi(t)$ and $\dot{\xi}(t)$ for q, $\dot{q}$, and finally differentiates with respect to t.

V.4 Legendre

The Euler-Lagrange system (V.3)—or (V.5)—was derived by looking at the "first variation" δI of the cost functional I, and requiring that $\delta I = 0$, as we shall see in Chapter IX. So in fact (V.3) is a necessary condition for the *stationarity* of I.

The next natural step was to look at the second variation of I, and this was done in 1786 by ***Adrien-Marie Legendre*** (1752-1833), who found an additional necessary condition for a minimum. His condition, derived by him for the scalar case, is

$$\boxed{\frac{\partial^2 L}{\partial \dot{q}^2}\Big(\xi_*(t),\dot{\xi}_*(t),t\Big) \geq 0 \ \ (\text{i.e., } \frac{\partial^2 L}{\partial u^2}\Big(\xi(t),\dot{\xi}(t),t\Big) \geq 0)\,.} \tag{V.6}$$

With an appropriate reinterpretation, Legendre's condition (V.6) is also necessary in the vector case: all we have to do is read (V.6) as asserting that the Hessian matrix $\left\{\frac{\partial^2 L}{\partial u^i \partial u^j}\Big(\xi_*(t),\dot{\xi}_*(t),t\Big)\right\}_{1\leq i,j\leq n}$ has to be nonnegative definite.

V.5 Hamilton misses an opportunity

At this point, we have reached the first and most critical fork in the road, involving the work of ***William Rowan Hamilton*** (1805-1865), published in 1834 and 1835. In a sense, the issue at stake will seem rather trivial, just a matter of rewriting the Euler-Lagrange system in a different formalism. However, sometimes formalisms can make a tremendous difference. To understand what happened and what could have happened but did not, let us see if we can make sense of the two necessary conditions for a minimum that have been presented so far, namely, the Euler-Lagrange equation (V.3) and the Legendre condition (V.6), and look for a single simple property that would imply both conditions.

The Legendre condition is clearly the second-order necessary condition for a minimum of a function, namely, $L(\xi_*(t),u,t)$ regarded as a function of u. But (V.3) does not look at all like the first-order condition for a minimum of that same function. It is natural to ask whether there might be a way to relate the two conditions. Is it possible that both could be expressed as necessary conditions for a minimum of one and the same function? The answer, as we now show, is "yes," and understanding how this function can be defined leads straight to optimal control theory, the maximum principle, and far-reaching generalizations of the classical theory.

First, let us look at an alternative way of writing (V.3). Define the "momentum" $\pi_{\xi,L}$ corresponding to L along a curve ξ by letting[10]

$$\pi_{\xi,L}(t) \stackrel{\text{def}}{=} \frac{\partial L}{\partial u}\Big(\xi(t),\dot{\xi}(t),t\Big)\,. \tag{V.7}$$

Define a function $H(q,u,p,t)$ of the three sets of variables q,u,p, and of $t\in\mathbb{R}$, by letting[11]

[10] In modern terminology, $\pi_{\xi,L}$ is a *field of covectors* along ξ.

[11] Geometrically, H is defined on the fiber product of the tangent and cotangent bundles of the configuration space Q.

$$\boxed{H(q,u,p,t) = \langle p,u\rangle - L(q,u,t)\,.} \tag{V.8}$$

Now suppose that a curve $t \mapsto \xi_*(t)$ is a solution of (V.3).

It is then clear that

$$\frac{\partial H}{\partial p} = u\,.$$

Therefore, along our curve ξ_*, the equality

$$\frac{d\xi_*}{dt}(t) = \frac{\partial H}{\partial p}\Big(\xi_*(t), \dot{\xi}_*(t), \pi(t), t\Big) \tag{V.9}$$

holds, if we let $\pi(t) = \pi_{\xi_*,L}(t)$. Also,

$$\frac{\partial H}{\partial q} = -\frac{\partial L}{\partial q}\,.$$

It follows that (V.5) says that

$$\frac{dp}{dt}(t) = -\frac{\partial H}{\partial q}\Big(\xi_*(t), \dot{\xi}_*(t), \pi(t), t\Big)\,. \tag{V.10}$$

Finally,

$$\frac{\partial H}{\partial u} = p - \frac{\partial L}{\partial u}\,,$$

so (V.7) says:

$$\frac{\partial H}{\partial u}\Big(\xi_*(t), \dot{\xi}_*(t), \pi(t), t\Big) = 0\,. \tag{V.11}$$

The system of equations (V.9), (V.10), (V.11), usually written more concisely as

$$\boxed{\frac{dq}{dt} = \frac{\partial H}{\partial p}\,, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q}\,, \quad \frac{\partial H}{\partial u} = 0\,,} \tag{V.12}$$

is ***exactly equivalent*** to (V.3), provided that H is defined as in (V.8).

We will call the function H the "control Hamiltonian," and refer to (V.12) as the ***control Hamiltonian form*** of the Euler-Lagrange equations. In our view, Formula (V.8) is the definition that Hamilton should have given for the Hamiltonian, and Equations (V.12) are *"Hamilton's equations as he should have written them."*

What Hamilton actually wrote was (in our notation, not his)

$$\frac{dq}{dt} = \frac{\partial \mathscr{H}}{\partial p}\,, \;\; \frac{dp}{dt} = -\frac{\partial \mathscr{H}}{\partial q}\,, \tag{V.13}$$

where $\mathscr{H}(q,p,t)$ is a function of p,q and t alone, defined by the formula

$$\mathscr{H}(q,p,t) = \langle p,\dot{q}\rangle - L(q,\dot{q},t)\,, \tag{V.14}$$

which resembles (V.8) but is not at all the same. The difference is that in Hamilton's definition $\dot{q}$ is supposed to be treated *not* as an independent variable, but as a function of q,p,t, defined implicitly by the equation

$$p = \frac{\partial L}{\partial \dot{q}}(q,\dot{q},t)\,. \tag{V.15}$$

It is easy to see that

> **If the map $(q,\dot{q},t) \mapsto (q,p,t)$ defined by (V.15) can be inverted,** *i.e., if we can "solve (V.15) for $\dot{q}$ as a function of q,p,t," then (V.13) is equivalent to (V.12).*

Indeed, it is clear that

$$\mathscr{H}(q,p,t) = H(q,u(q,p,t),t)\,,$$

where $u = u(q,p,t)$ satisfies $p = \frac{\partial L}{\partial u}(q,u,t)$. So

$$\frac{\partial \mathscr{H}}{\partial q} = \frac{\partial H}{\partial q} + \frac{\partial H}{\partial u}\cdot\frac{\partial u}{\partial q}\,. \tag{V.16}$$

Since

$$\frac{\partial H}{\partial u}(q,u,t) = 0 \quad \text{if} \quad u = u(q,p,t)\,,$$

we see that

$$\frac{\partial \mathscr{H}}{\partial q} = \frac{\partial H}{\partial q}$$

along solutions of (V.12), and then the first equation of (V.13) holds as well. Similarly, the second equation of (V.13) also holds. The converse is also easily proved.

V.6 The "control Hamiltonian" versus the classical Hamiltonian

It should be clear from the above discussion that the Hamiltonian reformulation of the Euler-Lagrange equations in terms of the "control Hamiltonian" is at least as natural as the classical version (V.13)-(V.14)-(V.15), and perhaps even simpler. Moreover, the control formulation has at least one obvious advantage, namely,

(A1) *the control version of the Hamilton equations is* ***exactly equivalent*** *to the Euler-Lagrange system under completely general conditions, whereas the classical version (V.13)-(V.14)-(V.15) only makes sense when the transformation (V.15) can be inverted, at least locally, to solve for $\dot{q}$ as a function of q, p, t.*

We now show that (A1) is not the only advantage of the control view over the classical one. To see this, we must take another look at Legendre's condition (V.6). Since $H(q,u,p,t)$ is equal to $-L(q,u,t)$ plus a linear function of u, (V.6) is completely equivalent to

$$\frac{\partial^2 H}{\partial u^2}\Big(\xi_*(t),\dot{\xi}_*(t),\pi(t),t\Big) \le 0\,, \quad \text{i.e.,} \quad \frac{\partial^2 H}{\partial \dot{q}^2}\Big(\xi_*(t),\dot{\xi}_*(t),\pi(t),t\Big) \le 0\,. \tag{V.17}$$

Now let us write (V.17) side by side with the third equation of (V.12):

$$\boxed{\frac{\partial H}{\partial u} = 0 \quad \text{and} \quad \frac{\partial^2 H}{\partial u^2} \le 0}\,, \tag{V.18}$$

and examine the result for a few seconds.

These equations unmistakably suggest something! They show that

the Euler-Lagrange and Legendre conditions have to do with critical points—with respect to the velocity variable u—of a function H. Precisely, Euler-Lagrange says that the gradient $\nabla_u H$ of this function has to vanish, and then Legendre adds to this the extra requirement that the second u-derivative of H—i.e., the Hessian matrix $\{\frac{\partial^2 H}{\partial u_i \partial u_j}\}_{1\le i,j\le n}$—must be nonpositive definite.

The message should be clear: ***what must be happening here is that H attains a maximum as a function of u.***

We state this formally as a conjecture.

CONJECTURE M1: *Besides (V.12) (or the equivalent form (V.3)), an additional necessary condition for optimality of an arc ξ_* is that $H(\xi_*(t),u,\pi(t),t)$, as a function of u, have a maximum at $u = \dot{\xi}_*(t)$ for each t.*

We contend that

1. Conjecture M1 is an extremely natural consequence of rewriting Hamilton's equations "as they should originally have been written," in terms of the "control Hamiltonian."

2. It is reasonable to guess that, if "Hamilton's equations" had been written from the veru beginning in this alternative way, then Hamilton himself, or some other nineteenth century mathematician, would have written (V.18) and been led by it to Conjecture M1.

3. On the other hand, it is only by using the Hamiltonian of (V.8), as opposed to Hamilton's own form of the Hamiltonian, that one can see that the Euler-Lagrange and Legendre condition really are first- and second-order conditions about critical points **of the same function.** This function cannot be L itself, because the first order conditions do *not* say that $\frac{\partial L}{\partial u} = 0$. Nor can it be the classical Hamiltonian $\mathscr{H}$, which isn't even a function of u. *Only the use of the "control" Hamiltonian leads naturally to Conjecture M1.*

Remark V.1. Assuming the validity of the invertibility condition needed to define the velocity as a function of position, momentum and time, it is possible to rewrite the Legendre condition in terms of the classical Hamiltonian $\mathscr{H}$. Indeed, the Legendre condition is equivalent, in to his case, to the statement that

$$\text{the matrix } \left\{\frac{\partial^2 \mathscr{H}}{\partial p_i \partial p_j}(\xi_*(t), \pi_{\xi_*,L}(t), t)\right\}_{1\leq i,j\leq n} \text{ is positive definite.} \tag{V.19}$$

(The proof of (V.19) is essentially contained in our discussion of Carathéodory's Hamiltonian formulation of Weierstrass' condition, cf. §V.9. Indeed, we show there that

$$\frac{\partial \mathscr{H}}{\partial p_i}(q,p,t) = \varphi_i(q,p,t). \tag{V.20}$$

Moreover,

$$\sum_{\ell=1}^{n} \frac{\partial \chi_j}{\partial \dot{q}_\ell}(q, \varphi(q,p,t), t)\frac{\partial \varphi_\ell}{\partial p_i}(q,p,t) = \delta_{j,i}, \tag{V.21}$$

since $(q,\dot{q},t) \mapsto (q,\chi(q,\dot{q},t),t)$ and $(q,p,t) \mapsto (q,\varphi(q,p,t),t)$ are inverse maps. But

$$\chi_j(q,\dot{q},t) = \frac{\partial L}{\partial \dot{q}_j}(q,\dot{q},t). \tag{V.22}$$

It follows from (V.20), (V.21) and (V.22) that

$$\sum_{\ell=1}^{n} \frac{\partial^2 L}{\partial \dot{q}_\ell \partial \dot{q}_j}(q,\varphi(q,p,t),t)\frac{\partial^2 \mathscr{H}}{\partial p_i \partial p_\ell}(q,p,t) = \delta_{j,i}\,. \tag{V.23}$$

So the matrix

$$M = \left\{ \frac{\partial^2 \mathscr{H}}{\partial p_i \partial p_\ell}(\xi_*(t),\pi_{\xi_*,L}(t),t) \right\}_{1\le i,j\le n}$$

is the inverse of the Hessian matrix N of L with respect to the velocities. The Legendre condition says that N is nonnegative definite, and the invertibility hypothesis implies that N is nonsingular. Hence N is strictly positive definite, and M is positive definite as well.)

Notice, however, that this does not solve the problem of expressing both the first-order condition and the Legendre condition as conditions on a critical point of the same function. Indeed, (V.20) implies that

$$\nabla_p \mathscr{H}(\xi_*(t),\pi_{\xi_*,L}(t),t) = \dot{\xi}_*(t)\,,$$

so $\pi_{\xi_*,L}(t)$ is *not* in general a critical point of the function $p \mapsto \mathscr{H}(\xi_*(t),p,t)$. $\square$

It will turn out that

> **Conjecture M1 is true. Moreover, once its truth is known then vast generalizations are possible.**

But before we get there, we must move to the next three chapters in our tale, and discuss

(i) Hamilton's idea of the "principal function,"

(ii) the work of Weierstrass, who essentially discovered and proved Conjecture M1, but did it using a language that obscured the simplicity of his own result,and for that reason missed some of its most profound implications,

and

(iii) Carathéodory's attempt to make sense of the Weiesrtrass discovery in Hamiltonian terms, and how this effort led nowhere because Carathéodory used the wrong Hamiltonian.

V.7 J. L. Synge on Hamilton's principal function and the 100-year eclipse of a great idea

At this point, we digress to discuss another of Hamilton's ideas, 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 [147] by J. L. Synge. For a discussion of what this idea becomes in the optimal control setting, we will have to wait until Chapter VIII. For the specific case of the brachistochrone, we will show in XIV.12 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
>
> $$V = \int v\Big(x,y,z,\frac{dx}{du},\frac{dy}{du},\frac{dz}{du}\Big)du\,. \tag{V.24}$$
>
> 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, is is so hard that few mathematicians in the ensuing century have given serious consideration to this plan.

Synge goes on to say that

> Hamilton belonged to the heroic age of mathematics,

by which he means, among other things, that while

> the modern mathematician weaves an intricate pattern of microscopic precision

and

> to him, a false statement ... is an unforgivable sin,

the attitude of "heroic mathematicians" was quite different:

> The heroic mathematician painted with broad splashes of color, with a grand contempt for singular cases until they could no longer be avoided.

Synge denies that "Hamilton lived in the clouds and spurned details," but goes on to add that

> the details with which he busied himself were not the details that a modern mathematician considers important. Few modern mathematicians like to work with a function without a careful statement regarding its general properties (continuity and existence of derivatives), or without assurance of existence under stated conditions. Hamilton, in the spirit of the heroic age, goes straight ahead. That is the feature which doubtless repels modern mathematicians from a more careful study of his work. It explains in large measure the eclipse of Hamilton.

As will be shown in Chapter VIII below, Synge's reference to Hamilton's "eclipse" during "the ensuing century" was made at a time when a new theory was about to be born, in which Hamilton's profound concept of the principal function—now renamed the "value function," or "Bellman function"—occupies precisely the central role that Hamilton and Synge thought it deserved. This theory is none other than modern optimal control, and its rise has given a new impetus to the study of Hamilton's principal function, confirming Synge's assessment of the significance of Hamilton's insight. Under the name of "dynamic programming," the method of following "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.

At the same time, the rise of optimal control has shown that Synge's complaint against modern mathematicians for their unwillingness to ignore certain types of "details" was not entirely justified. Indeed, as will be argued in Chapter VIII, modern optimal control theory shows that a rich and exciting theory of the value function exists, in which the question of the appropriate regularity properties of the value function plays a crucial role. In many cases, this theory is the best if not the only tool for the analysis of an optimal control problem. Moreover, the new theory is completely rigorous, and the effort to make it rigorous has turned out to be worthwhile, in the sense that the existence and continuity issues mentioned by Synge are quite interesting in their own right, as we will show for the case of the brachistochrone problem.

So the issues that Hamilton failed to "busy himself" with, "in the spirit of the heroic age," have in fact turned out to be of the utmost importance. It is now clear that, in a sense, the instincts of those mathematicians who were "repelled" by Hamilton's inattention to "details" such as "a careful statement regarding [the] general properties [of the value function] (continuity and existence of derivatives)" were not completely wrong, since the value function often lacks such properties, and this fact is significant and cannot be ignored. Therefore both Synge and the people whom he criticized have been fully vindicated: Synge, because his assessment of the importance of Hamilton's "central idea" was right on target, and the mathematicians who asked for more rigor and precision, because this rigor and precision have turned out to be necessary. With hindsight, it can now be seen that the "eclipse," however regrettable it may have appeared to Synge, was perhaps unavoidable, because only with the tools of late-20th-century mathematics has it become possible to do full justice to Hamilton's idea.

V.8 Weierstrass comes close, but fails to discover Hamiltonian maximization

The contributions of Karl Weierstrass (1815-1897) to the calculus of variations are contained in the notes of his three series of lectures on the topic, given in 1875, 1879 and 1882, and not formally published as books or journal articles.

Weierstrass considered the problem of minimizing an integral I of the form $I = \int_a^b L(q(s),\dot{q}(s))\,ds$ for Lagrangians L such that

> ***(W)*** *$L(q,\dot{q})$ is positively homogeneous with respect to the velocity $\dot{q}$ (that is, $L(q,\alpha\dot{q}) = \alpha L(q,\dot{q})$ for all $q,\dot{q}$ and all $\alpha \geq 0$) and does not depend on time.* □

As will become clear from Remark V.2, we have a good reason for using s rather than t as the "time" variable in the expression for I.

Remark V.2. In a sense, one can always assume (W) "without loss of generality." Indeed, suppose we are interested in a completely arbitrary problem (II.5). We can always define a new function $\Lambda(q,t,u,\tau) = \tau L(q,u/\tau,t)$, and think of t as a new q variable, say q^0, and of τ as $\frac{dq^0}{ds}$, were s is a new time variable, or "pseudotime," not to be confused with the true time variable t. Then, using q_{new} for (q,q^0) (so that $q_{new} = (q,t)$), we have

$$\int L(q,\dot{q},t)dt = \int L\left(q,\left(\frac{dt}{ds}\right)^{-1}\frac{dq}{ds},t\right)\frac{dt}{ds}ds = \int \Lambda\left(q_{new},\frac{dq_{new}}{ds}\right)ds,$$

so the minimization problem for Λ is equivalent to (II.5). Moreover, if $\alpha>0$, then $\Lambda(q,t,\alpha u,\alpha\tau) = \alpha\Lambda(q,t,u,\tau)$, i.e., $\Lambda(q_{new},\alpha\dot{q}_{new}) = \alpha\Lambda(q_{new},\dot{q}_{new})$, where we are now using the dot to denote differentiation with respect to s rather than t. So the new Lagrangian Λ is positively homogeneous with respect to the new velocity variable $\dot{q}_{new}$ and does not depend on the new time s. $\square$

So when Weierstrass imposed Condition (W) on his Lagrangians, he was doing so "without loss of generality." However, "without loss of generality" is a dangerous phrase, and does not at all entail "without loss of insight." We shall argue below that this restriction, in conjunction with the dominant view that Hamilton's equations had to be written in the form (V.13), may have served to conceal from Weierstrass the true meaning and the far-reaching implications of the new condition he discovered.

Weierstrass introduced the "excess function"

$$\mathscr{E}(q,\bar{u},u) = L(q,u) - \frac{\partial L}{\partial u}(q,\bar{u})\cdot u\,, \tag{V.25}$$

depending on three sets of independent variables q, $\bar{u}$ and u. He then proved his *side condition*:

> ***(SC)*** *For a curve $s \mapsto q_*(s)$ to be a solution of the minimization problem, the function $\mathscr{E}$ has to be ≥ 0 when evaluated for $q = q_*(s)$, $\bar{u} = \dot{q}_*(s)$, and a completely arbitrary u.*

Notice that Lagrangians with Property (W) satisfy the identity

$$L(q,u) = \frac{\partial L}{\partial u}(q,u)\cdot u\,. \tag{V.26}$$

Therefore Weierstrass could equally well have written his excess function as

$$\mathscr{E}(q,\bar{u},u) = L(q,u) - \frac{\partial L}{\partial u}(q,\bar{u})\cdot u - \left(L(q,\bar{u}) - \frac{\partial L}{\partial u}(q,\bar{u})\cdot\bar{u}\right), \tag{V.27}$$

or as

$$\mathscr{E}(q,\bar{u},u,t) = L(q,u,t) - \frac{\partial L}{\partial u}(q,\bar{u},t)\cdot u - \left(L(q,\bar{u},t) - \frac{\partial L}{\partial u}(q,\bar{u},t)\cdot \bar{u}\right), \quad \text{(V.28)}$$

if he had allowed his Lagrangians to depend on time.

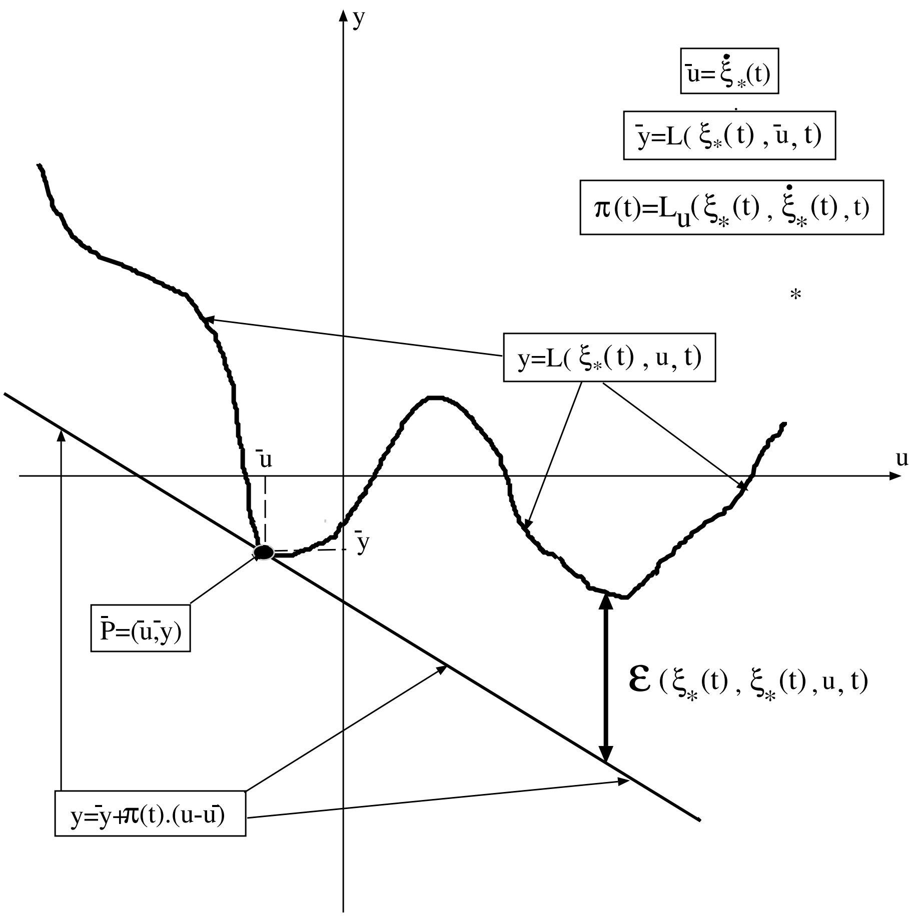


FIGURE 8. The Weierstrass side condition

Remark V.3. Formula (V.27) has a simple geometric interpetation. For a given point q, let G_q be the graph of

$$y = L(q,u) \stackrel{\text{def}}{=} L_q(u)$$

as a function of u. Fix a value $\bar{u}$ of u, and let

$$\bar{y} = L(q,\bar{u})\,,$$

so the point

$$\bar{P} = (\bar{u},\bar{y})$$

lies on G_q. Let $\Gamma_{q,\bar{u}}$ be the tangent hyperplane to G_q at $\bar{P}$. Then the equation of $\Gamma_{q,\bar{u}}$ is

$$y = \bar{y} + \frac{\partial L}{\partial u}(q,\bar{u}) \cdot (u - \bar{u})\,,$$

that is

$$y = L(q,\bar{u}) + \frac{\partial L}{\partial u}(q,\bar{u}) \cdot (u - \bar{u})\,.$$

Given any u, the ordinates $y_{graph}(u)$, $y_{plane}(u)$ of the points of G_q, $\Gamma_{q,\bar{u}}$, that lie on the line $\{u\} \times \mathbb{R}$ are given by

$$y_{graph}(u) = L(q,u)\,, \quad y_{plane}(u) = L(q,\bar{u}) + \frac{\partial L}{\partial u}(q,\bar{u}) \cdot (u - \bar{u})\,.$$

Therefore

$$\mathscr{E}(q,\bar{u},u) = y_{graph}(u) - y_{plane}(u)\,.$$

In other words (cf. Figure 8):

> ***the number $\mathscr{E}(q,\bar{u},u)$ tells us by how much the actual value of the function $y = L_q(u)$ at a particular u is larger than the value at u of the linearization (i.e., the tangent approximation) of this function at the point $(\bar{u}, L(q,\bar{u}))$.***

Therefore the Weierstrass side condition says (reverting now to the more general time-varying fomulation) that

> ***(SC)*** *If a curve $s \mapsto \xi_*(s)$ is a solution of the minimization problem, then for every t, if we draw the tangent hyperplane to the graph of the function $u \mapsto L(\xi_*(t),u,t)$ at the point of this graph corresponding to $u = \dot{\xi}_*(t)$, then the graph must lie entirely above the hyperplane.* $\square$

Using

$$\pi(t) = \frac{\partial L}{\partial u}(\xi_*(t), \dot{\xi}_*(t), t)\,,$$

as in (V.7), we see that

$$\mathscr{E}(\xi_*(t), \dot{\xi}_*(t), u, t) = \Big(L(\xi_*(t), u, t) - \langle \pi(t), u\rangle\Big) - \Big(L(\xi_*(t), \dot{\xi}_*(t)) - \langle \pi(t), \dot{\xi}_*(t)\rangle\Big)\,, \tag{V.29}$$

which the reader will immediately recognize as saying that

$$\mathscr{E}(\xi_*(t), \dot{\xi}_*(t), u, t) = H(\xi_*(t), \dot{\xi}_*(t), \pi(t), t) - H(\xi_*(t), u, \pi(t), t)\,, \tag{V.30}$$

where H is our "control Hamiltonian." So Weierstrass' condition, expressed in terms of the control Hamiltonian, simply says that

(MAX) *Along an optimal curve $t \mapsto \xi_*(t)$, if we let $\pi(t) = \pi_{\xi_*,L}(t)$, then for every t the value $u = \dot{\xi}_*(t)$ must maximize the (control) Hamiltonian $H(\xi_*(t), u, \pi(t), t)$ as a function of u.*

COMMENT NO. 17. JAN: In the earlier versions I talked about configurations (or states) q and curves $t \mapsto q(t)$ in q-space. Then the momentum was $p(t)$, and so on. I then decided that it would be better to be more systematic and use Latin letters for points and Greek lettters for functions of time, so the curves are $\xi(t)$, the momentum is $\pi(t)$, the open-loop controls are $\eta(t)$, etc. I have done the conversion UP TO THIS POINT, but from here on it's not done. I don'y have time to do it now. I'll do it when I get back from Europe in August.

In Weierstrass' formulation, the condition was stated in terms of the excess function, for the special Lagrangians that satisfy (W). (In that case the resulting H is independent of time, as in our equation (V.30)). But, if one rewrites Weierstrass's condition as we have done, in terms of H, then one can take a general Lagrangian, not satisfying (W), transform the minimization problem into one in Weierstrass's form, as explained in Remark V.2, write the Weierstrass condition in the form (MAX) (so in particular H is independent of time) and then undo the transformation and go back to the original problem. The result is (MAX), as written, with the control Hamiltonian of the original problem. So the Weierstrass condition, if reformulated as in (MAX), is valid for all problems, with exactly the same statement. For comparison, notice that if one applies the same procedure to the formulation in terms of $\mathscr{E}$, then one also ends up with a statement valid for all problems, but the expression for $\mathscr{E}$ is now much more complicated, since one has to use (V.27) and cannot substitute for it the simpler expression (V.25).

Moreover, (MAX) can be considerably simplified. Indeed,

the requirement that $p(t)$ ***be defined via (V.7) is now redundant***

because, if $H(q_*(t),u,p(t),t)$, regarded as a function of u, has a maximum at $u = \dot{q}_*(t)$, then $\frac{\partial H}{\partial u}(q_*(t),\dot{q}_*(t),p(t),t)$ has to vanish, so $p(t)$ has to be given by (V.7). Moreover, the vanishing of $\frac{\partial H}{\partial u}(q_*(t),\dot{q}_*(t),p(t),t)$ is also one of the conditions of (V.12). So we can just drop (V.7) altogether, and state (V.12) and (MAX) together:

(NCO) *If a curve $t \mapsto q_*(t)$ is a solution of the minimization problem (II.5), then there has to exist a function $t \mapsto p(t)$ such that (p is absolutely continuous and) the following three conditions hold for all t:*

$$\dot{q}_*(t) = \frac{\partial H}{\partial p}(q_*(t),\dot{q}_*(t),p(t),t), \tag{V.31}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial q}(q_*(t),\dot{q}_*(t),p(t),t), \tag{V.32}$$

$$H(q_*(t),\dot{q}_*(t),p(t),t) = \max_u H(q_*(t),u,p(t),t). \tag{V.33}$$

Remark V.4. The parentheses for the absolute continuity condition refer to the fact that

(i) this is a natural technical condition to impose on p from the 20th century point of view, since (V.32) is a differential equation for $p(\cdot)$,

while, on the other hand,

(ii) the concept of an absolutely continuous function obviously belongs to a circle of ideas about functions of a real variable that were not yet available to Weierstrass. □

COMMENT NO. 18. JAN: The “Weierstrass comes close” section was written before I looked at Carathéodory. When I read Carathéodory I was fascinated by the extra term in Carathéodory’s formula for the excess function, because this term shows that Carathéodory really didn’t get it. So I put in some stuff about Carathédory. But now I should look at the “Weierstrass comes close” section again, because some things that are there should be moved and put AFTER the Carathéodory section.

V.9 Carathéodory gets even closer, but misses too

C. Carathéodory (1873-1950) published his work on the calculus of variations in a 1935 book, cf. [36]. His discussion of the excess function, and of how Weierstrass' condition can be expressed in Hamiltonian form, is worth quoting in detail[12], because it shows that the author understood the importance of expressing the Weierstrass condition in Hamiltonian form, but was unable to find its simple Hamiltonian meaning because he was using an inappropriate Hamiltonian formalism.

Carathéodory uses x_i for the position variables (that is, our q^i) and y_i for the momenta (that is, our p_i). He works in the neighborhood of a "regular line element" $e = (t, x_i, \dot{x}_i)$, where the "regularity" assumption is the requirement that the Hessian matrix $\{L_{\dot{x}_i,\dot{x}_j}\}_{1\leq i,j\leq n}$ be positive definite at e. This condition guarantees that, near e, one can express everything in "canonical coordinates" (t, x_i, y_i), i.e., that the map $(t,x,\dot{x}) \to (t,x,y)$ is invertible, if we let $y_i = \chi_i(t,x,\dot{x})$, where $\chi_i(t,x,\dot{x}) = \frac{\partial L}{\partial \dot{x}_i}(t,x,\dot{x})$.

He then expresses the Weierstrass condition, which involves the "line element $(t,x_i,\dot{x}_i)$" together with a nearby "line element" (t,x_i,x'_i), in terms of the corresponding canonical coordinates (t,x_i,y_i) and (t,x_i,y'_i). He uses H to represent what he calls "the Hamiltonian" (that is, our classical Hamiltonian), given by the formula $H = \sum_{i=1}^n y_i\dot{x}_i - L(t,x,\dot{x})$, where $\dot{x}_i = \varphi_i(t,x,y)$, and φ is the inverse function of χ (that is, $\varphi(t,x,\chi(t,x,u)) = u$ and $\chi(t,x,\varphi(t,x,y)) = y$). In addition, he uses H' to represent the Hamiltonian evaluated at the other point (t,x,y') in phase space, that is, $H' = \sum_{i=1}^n y'_i x'_i - L(t,x,x')$, where $x'_i = \varphi_i(t,x,y')$, $y'_i = \chi_i(t,x,x')$.

With these notations, the Weierstrass excess function becomes

$$
\begin{aligned}
\mathscr{E}(t,x,\dot{x},x') &= L(t,x,x') - L(t,x,\dot{x}) - \frac{\partial L}{\partial \dot{x}}(t,x,\dot{x}) \cdot (x' - \dot{x}) \\
&= L(t,x,x') - L(t,x,\dot{x}) - y \cdot (x' - \dot{x}) \\
&= y \cdot \dot{x} - L(t,x,\dot{x}) - \Big(y \cdot x' - L(t,x,x')\Big) \\
&= y \cdot \dot{x} - L(t,x,\dot{x}) - \Big(y' \cdot x' - L(t,x,x')\Big) - (y - y') \cdot x' \\
&= H - H' - (y - y') \cdot x' \,.
\end{aligned}
$$

[12] We follow [36], pages 210-212.

Since

$$\begin{aligned}
\frac{\partial H'}{\partial y'_j} &= \frac{\partial}{\partial y'_j}\Big(\sum_{i=1}^n y'_i x'_i - L(t,x,x')\Big) \\
&= x'_j + \sum_{i=1}^n y'_i \frac{\partial \varphi_i}{\partial y'_j}(t,x,y') - \frac{\partial}{\partial y'_j}\Big(L(t,x,\varphi(t,x,y'))\Big) \\
&= x'_j + \sum_{i=1}^n y'_i \frac{\partial \varphi_i}{\partial y'_j}(t,x,y') - \sum_{i=1}^n \frac{\partial L}{\partial x'_i}(t,x,x')\frac{\partial \varphi_i}{\partial y'_j}(t,x,x') \\
&= x'_j + \sum_{i=1}^n y'_i \frac{\partial \varphi_i}{\partial y'_j}(t,x,y') - \sum_{i=1}^n y'_i \frac{\varphi_i}{\partial y'_j}(t,x,x') \\
&= x'_j ,
\end{aligned}$$

Carathéodory concluded that "the equation

$$\begin{aligned}
\mathscr{E}(t,x,\dot{x},x') &= L' - L - L_{\dot{x}_j}(x'_j - \dot{x}_j) \\
&= H - H' - H'_{y'_j}(y_j - y'_j)
\end{aligned}$$

therefore holds, by which the $\mathscr{E}$ function is represented in the desired form."

V.10 A PUZZLE: What has disappeared?

(NCO) display a ***remarkable new fact*** *that makes it fundamentally different from all our previous necessary conditions. We urge the reader to stare at (NCO) and try to figure out what this fact may be before reading on.*[13] ***The answer will be given below.*** □

[13] *A cryptic hint for playful quiz-loving folks.* In many situations, you can gain a lot by focusing on omissions, voids, and vanishing acts, by thinking about what is missing and ignoring what is visibly in front of you. This obvious point might possibly call to mind a story, probably familiar to you, about a dog that did not bark (cf. [48], pp. 347 and 349). Such a curious fact had no significant implications for Dr. Watson's simplistic mind, but his famous companion soon saw how crucial it was. Similarly, it is as plain as daylight that in your analysis of Condition (NCO) **you should look for what you cannot find in its formulas** and should stop worrying about things within your sight. If you think that all this sounds too cryptic and hard to grasp, you may go back—without loss—to our discussion of how an amazingly circuitous path, in many ways astonishingly similar to that of A. J. Tasman's circumnavigation of Australia, took us from our old calculus of variations conditions of a distant past to today's work on optimal control and dynamic programming. Or, if this hint is too baffling for you, thanks to your inability to find what is obviously missing in it, you may look at Scholium VIII for a good indication of what this is all about and why our words may sound so odd to you. You will find in that scholium two quotations from a translation, by G. Adair, of a truly unusual and amazing work of fiction by a famous Gallic author ([108]), and as soon as you grasp what that book is about all your doubts will finally go away. But you must insist anyhow on doing your own thinking, and in no way allow that odd book to put wrong thoughts into your mind.

As a version of the necessary conditions for optimality, (NCO) encapsulates in one single statement the combined power of the Euler-Lagrange necessary conditions and the Weierstrass side condition as well, of course, as the Legendre condition, which obviously follows from (MAX). Notice the elegance and economy of language achieved by this unified statement: there is no need to bring in an extra entity called the "excess function." Nor does one need to include a formula specifying how $p(t)$ is defined, since (V.33) does this automatically. So the addition of the new Weierstrass condition to the three equations of (V.12) results in a new set of three, rather than four, conditions, a set much "simpler than the sum of its parts." Notice moreover that (MAX)—or, more precisely, the Weierstrass side condition part of (MAX)—is exactly Conjecture M. So we can surmise at this point that (MAX), as stated, could probably have been discovered soon after the work of Hamilton, since it is strongly suggested by (V.18), and almost certainly by Weierstrass, if only Hamilton's equations had been written in the form (V.8), (V.12).

So we can now add two new items to our list of advantages of the "control formulation" of Hamilton's equations over the classical one:

(A2) *Using the control Hamiltonian, it would have been an obvious next step to write Legendre's condition in "Hamiltonian form," as in (V.18), and this would have led immediately to the formulation of Conjecture M, a proof of which would then have been found soon after.*

(A3) *In terms of the control Hamiltonian, the Weierstrass side condition has a much simpler statement, not requiring the introduction of an "excess function," and can be combined with the Hamilton equations into an elegant unified formulation (NCO) of the necessary conditions for optimality.*

But this is by no means the end of our story. There is much more to the new formulation (NCO) than just elegance and simplicity. If you compare (NCO) with all the other necessary conditions that we had written earlier, a remarkable new fact becomes apparent. Quite amazingly, ***the derivatives with respect to the*** u ***variable are completely gone.*** (*This is the answer to our puzzle.*) All the earlier equations involved u-derivatives of L or of H, and even if we use the classical version (V.13) of Hamilton's equations—which involves no functions of u and therefore no u-derivatives—the fact remains that in order to get to (V.13) we first have to solve

In particular, you should avoid following that book's unjustifiably critical opinion about Pontryagin. (In p. 47 of Adair's translation, [108] wrongly says that "Though it took him from 1935 to 1955, working at it virtually nonstop, Pontryagin finally had to admit to accomplishing nothing." What a ridiculous thing to say! It's totally absurd! Naturally, you could think, if you wish, that Pontryagin was an awful man, but it is obvious that his work was important and rich in major innovations.)

(V.15), which does involve a u-derivative.

Now, if our necessary conditions for optimality can be stated without any references to u-derivatives, it is reasonable to guess that the u-derivatives of L are not needed. Also, the minimization that occurs in (V.33) makes sense over any subset U of $\mathbb{R}^n$, so there is no longer any reason to insist that the range of values of u be the whole space. This leads us to

> ***CONJECTURE M2:*** *(NCO) should still be a necessary condition for optimality even for problems where the derivative $\dot{q}$ is restricted to belong to some subset U of $\mathbb{R}^n$, and $L(q,u,t)$ is not required to be differentiable or even continuous with respect to u.* □

Naturally, once u is restricted to a subset U of $\mathbb{R}^n$, and $L(q,u,t)$ is not required to be differentiable with respect to u, the maximization condition (V.33) no longer implies (V.15). So we see that ***(V.33) is more fundamental than (V.15)***.

Notice also, that there has been a subtle change in the formulation of the necessary conditions for optimality. In the classical conditions of the calculus of variations (including Weierstrass' own version of the side condition), the vector p—the "momentum"—is defined by (V.15), and the necessary conditions are equations happening along the trajectory q_*. These equations can be formulated somewhat more simply using p, but p can also be completely eliminated, since one can always substitute for it its value given by (V.15). The new version of Conjecture M2 is of a different form: rather than stating that certain equations hold along q_*, the conditions are now an *existential statement*, asserting the existence of a vector-valued function $t \mapsto p(t)$ with certain properties. The elimination of p, by expressing $p(t)$ in terms of $q_*(t)$, $\dot{q}_*(t)$ and t, is only possible when $U = \mathbb{R}^n$ and L is differentiable with respect to u, in which case (V.15) follows from (V.33). Otherwise, p is a "new," "independent" object.

Since Weierstrass was working under assumptions that imply (V.15), he missed the chance to see that his condition was really about p, not about $\frac{\partial L}{\partial u}$ evaluated along q_*, and that, stated this way, his condition was valid for much more general problems, for which $p(t)$ no longer equals $\frac{\partial L}{\partial u}(q_*(t),\dot{q}_*(t),t)$.

V.11 Three toy problems

To test Conjecture M2, let us look at three toy problems where the answer is reasonably easy to guess directly, and see whether using (NCO) gives the same result.

Example V.5. To test the differentiability part of Conjecture M2, let us suppose that we are given $L > 0$, and we want to find a real-valued (Lipschitz continuous) function $t \mapsto x(t)$ on the interval $[0,L]$ that satisfies

$$x(0) = 1\,, \;\text{ and }\quad x(L) = 1\,, \tag{V.34}$$

and minimizes the integral $\int_0^L |\dot{x}(t)|\,dt$ among all the functions on $[0,L]$ satisfying the constraints (V.34). This looks exactly like a calculus of variations problem of the classical sort, with a Lagrangian $L(x,\dot{x},t) = |\dot{x}|$ (i.e., $L(x,u,t) = |u|$), except that L is no longer everywhere differentiable with respect to u. Moreover, it is quite obvious that the solution of this minimization problem is given by $u(t) \equiv 0$, i .e., by $x(t) \equiv 1$. Therefore the nondifferentiability issue truly matters for the applicability of the necessary conditions, since $\{(x,u,t) : u = 0\}$ is precisely where $\frac{\partial L}{\partial u}$ fails to exist. In particular, the classical Euler-Lagrange equation cannot even be written, let alone solved.

On the other hand, if we apply (NCO) as stated, there is no problem whatsoever. We have $H = p.u - |u|$ (so H in this case does not depend on x and t), and then (V.31) and (V.32) become $\dot{x} = u$ and $\dot{p} = 0$. Therefore the function $t \mapsto p(t)$ is a constant $\bar{p}$. Now, given any value of $\bar{p}$, one can easily check that the set of values of u where the function $u \mapsto \bar{p}u - |u|$ has a maximum is always a subset of $[0,\infty[$ or of $]-\infty,0]$. (If $|\bar{p}| < 1$, then the maximum is attained at $u = 0$ only. If $|\bar{p}| > 1$, then there is no maximum. If $\bar{p} = 1$ then the maximum is attained at all values of u such that $u \geq 0$. If $\bar{p} = -1$ then it is attained at all u's such that $u \leq 0$.) Since $u(t)$ has to be chosen, for each t, so as to maximize $\bar{p}u - |u|$, we see that either $u(t) \geq 0$ for all t, or $u(t) \leq 0$ for all t. In either case, the only way that $x(0)$ and $x(L)$ can be equal is if $u(t) \equiv 0$. (Of course, what we should really say is "$u(t) = 0$ a. e.".) So (NCO) works perfectly, and singles out the solution we had already found. □

Example V.6. We now look at a two-dimensional analogue of the previous example. We try to find a Lipschitz continuous curve $t \mapsto \xi(t) = (\xi_1(t),\xi_2(t)) \in \mathbb{R}^2$, defined on the interval $[0,1]$, that satisfies

$$\xi(0) = (0,0)\,, \;\text{ and }\quad \xi(1) = (1,2)\,, \tag{V.35}$$

and minimizes the integral

$$J(\xi) = \int_0^1 \Big(|\dot{\xi}_1(t)| + |\dot{\xi}_2(t)|\Big)\,dt \tag{V.36}$$

in the set Ξ of all the Lipschitz functions $[0,1] \ni t \mapsto \xi(t) \in \mathbb{R}^2$ satisfying the constraints (V.35). Once again, this looks like a calculus of variations problem of the classical sort, with a Lagrangian

$$\mathbb{R}^2 \times \mathbb{R}^2 \times [0,1] \ni (x_1,x_2,u_1,u_2,t) \mapsto L(x_1,x_2,u_1,u_2,t) \stackrel{\text{def}}{=} |u_1| + |u_2|\,.$$

which is not everywhere differentiable with respect to u. It is easy to see that

$$J(\xi) \geq 3 \quad \text{for every} \quad \xi \in \Xi\,.$$

Indeed, if $\xi \in \Xi$ then

$$1 = \xi_1(1) - \xi_1(0) = \int_0^1 \dot{\xi}_1(t)\,dt \leq \int_0^1 |\dot{\xi}_1(t)|\,dt\,, \tag{V.37}$$

and a similar argument shows that

$$2 \leq \int_0^1 |\dot{\xi}_2(t)|\,dt\,. \tag{V.38}$$

Therefore $J(\xi) = \int_0^1 \Big(|\dot{\xi}_1(t)| + |\dot{\xi}_2(t)|\Big)\,dt \geq 3$, as stated.

So any arc $\xi \in \Xi$ for which $J(\xi) = 3$ is a solution of our minimization problem. One example is the arc $\xi_* : [0,1] \mapsto \mathbb{R}^2$ given by

$$\xi_*(t) = (t, 2t) \quad \text{for} \quad 0 \leq t \leq 1\,.$$

Another example is the arc $\xi_\# : [0,1] \mapsto \mathbb{R}^2$ given by

$$\xi_\#(t) = \begin{cases} (2t, 0) & \text{if} \quad 0 \leq t \leq \frac{1}{2} \\ (1, 4t-2) & \text{if} \quad \frac{1}{2} \leq t \leq 1\,. \end{cases}$$

Notice that ***the arc $\xi_\#$ is such that for every t the vector $\dot{\xi}_\#(t)$ belongs to the set of (u_1,u_2) where L is not differentiable***. Therefore, exactly as in the previous example, the classical Euler-Lagrange equation cannot even be written, let alone solved.

Let us now apply (NCO) and see what we get. The Hamiltonian is now given by

$$H(x_1,x_2,u_1,u_2,p_1,p_2,t) = p_1u_1 + p_2u_2 - |u_1| - |u_2|\,,$$

(so in particular H does not depend on x and t). Then (V.31) and (V.32) tell us that

$$\dot{\xi} = u \quad \text{and} \quad \dot{\pi} = 0\,.$$

Therefore the function $t \mapsto \pi(t)$ is a constant $\bar{\pi} = (\bar{\pi}_1, \bar{\pi}_2) \in \mathbb{R}^2$.

As in the previous example, the set of values of u_1 where the function $u_1 \mapsto \bar{\pi}_1 u_1 - |u_1|$ has a maximum is always a subset of $[0,\infty[$ or of $]-\infty,0]$. Since the function $t \mapsto \dot{\xi}_1(t)$ has to be chosen so that for each t the value $u_1 = \dot{\xi}_1(t)$ maximizes $\bar{\pi}_1 u_1 - |u_1|$, we see that either $\dot{\xi}_1(t) \geq 0$ for all t, or $\xi_1(t) \leq 0$ for all t. Since $\int_0^1 \dot{\xi}_1(t)\,dt = 1$, the possibility that $\xi_1(t) \leq 0$ for all t is excluded. So ξ_1 is an arbitrary Lipschitz function such that

$$\xi_1(0) = 0\,, \quad \xi_1(1) = 1\,, \quad \text{and} \quad \xi_1 \text{ is monotonically nondecreasing}\,. \tag{V.39}$$

Similarly, ξ_2 is an arbitrary Lipschitz function such that

$$\xi_2(0) = 0\,, \quad \xi_2(1) = 2\,, \quad \text{and} \quad \xi_2 \text{ is monotonically nondecreasing}\,. \tag{V.40}$$

Conversely, *every Lipschitz curve* $[0,1] \ni t \mapsto \xi(t) = (\xi_1(t), \xi_2(t)) \in \mathbb{R}^2$ *such that (V.39) and (V.40) hold is a minimizer.* (This is easy to prove. It suffices to observe that under those conditions the inequalities occurring in (V.37) and (V.38) become equalities, so $J(\xi) = 3$.)

So we see once again that ***(NCO) works perfectly, and singles out exactly all the solutions of our minimization problem***. □

Example V.7. We now test the other part of Conjecture M2, by looking at a problem where the range of u is no longer the whole space.

Suppose that we want to find a real-valued (Lipschitz) function $t \mapsto x(t)$ on the interval $[0,L]$ that satisfies

$$x(0) = 1\,,\ \ x(L) = 1\,, \text{ and } |\dot{x}(t)| \leq 1 \text{ for all } t \in [0,L]\,, \tag{V.41}$$

and minimizes the integral $\int_0^L x(t)^2 dt$ among all the functions satisfying the constraints (V.41). Once again, this looks exactly like a calculus of variations problem of the classical type—with a Lagrangian $L(x,\dot{x},t) = x^2$—except that the derivative $\dot{x}$ is required to satisfy an "inequality constraint" $|\dot{x}| \leq 1$. If we apply the classical Euler-Lagrange equations formally, we get $\frac{\partial L}{\partial x} = 0$ (since L does not depend on $\dot{x}$, so $\frac{\partial L}{\partial \dot{x}} = 0$). So $x(t) \equiv 0$, contradicting the endpoint constraints. So there is no way at all to satisfy Euler-Lagrange together with the boundary conditions.

On the other hand, we can apply (NCO) formally, making the sensible guess that in this case the maximization with respect to u should be made over the set $U = [-1,1]$ of permissible values of u. Then (V.31) and (V.32) become the equations $\dot{x} = u$, $\dot{p} = 2x$. This suggests, among other things, that p as a function of t should be of class C^1, since $\dot{p}$ has to be continuous. In addition, (V.33) says that $u = 1$ when $p > 0$, and $u = -1$ when $p < 0$. (When $p = 0$ all values of $u \in U$ are solutions.)

So when $p > 0$ we have $\ddot{p} = 2$, and when $p < 0$ we have $\ddot{p} = -2$. This shows that $p(\cdot)$ is a convex function of t when $p(t) > 0$ and a concave function when $p(t) < 0$. Now, a continuous function on $[0,L]$ which is convex wherever it is > 0 and concave wherever it is < 0 can only be of one of the 13 types P, N, Z, PZ, PN, ZP, ZN, NP, NZ, PZP, PZN, NZP or NZN, where "Z" means "identically zero," "P means "> 0," "N" means "< 0," and, for example, a symbol such as PZN means "a positive piece followed by a zero piece and then by a negative piece." (This is easy to prove. For example, ZPZ is impossible because the P part would have to vanish at two points and be convex and > 0 in between, which clearly cannot happen. The reader should analyze all possibilities and verify that only our 13 cases can arise.) Clearly, P corresponds to $u = 1$, and N to $u = -1$. As for Z, it corresponds to $p(t) \equiv 0$, which yields $x(t) \equiv 0$ (since $\dot{p} = 2x$) and then $u(t) \equiv 0$ (since $\dot{x} = u$). So Z corresponds to $x = u = 0$. (Notice that (V.33) does not directly determine the value of u when $p = 0$, since in this case every value of u satisfies (V.33), but by suitably combining the three conditions of (NCO) we have still been able to find u for $p = 0$.) With this in mind, it is clear that the only way to satisfy the endpoint conditions $x(0) = 1$, $x(L) = 1$, is with a PN or an NP or an NZP curve. Of these three possibilities, PN is obviously not optimal, so we are left with NP and NZP, and then a direct analysis shows that the minimum is NP if $L \leq 2$ and NZP if $L > 2$. This result can of course be proved directly, and corresponds to the obvious intuition that to minimize $\int x^2$ one should try to move from the initial value $x = 1$ towards $x = 0$ as fast as possible, stay at $x = 0$ as long as as possible, and then move back towards $x = 1$, choosing the departure time from $x = 0$ so as to arrive to $x = 1$ exactly at time L. $\square$

Examples V.5 and V.7 show that (NCO), which is none other than the combination of the classical Euler-Lagrange condition and the Weierstrass side condition, is a surprisingly powerful tool, provided only that it is properly reinterpreted as we have done. The formal application of the new version of (NCO) has been shown to yield a complete solution for at least some problems that do not fit within the framework of the classical calculus of variations, either because of nondifferentiability of L with respect to u, or because $\dot{q}$ is subject to an inequality constraint. All that is really needed is to use the Hamiltonian formalism in its "control" version, and to apply the conditions with their full force, expressed in (NCO). In addition, it helps to be flexible and be willing to depart from the classical way of "solving problems analytically" since, as our two examples show, one has to be ready to combine calculus arguments of a classical sort with a little bit of "piecewise smooth analysis" and qualitative reasoning.

V.12 Two more toy problems, and the "abnormal" multiplier

So far we have shown that Conjecture M2 is almost forced on us if one looks at the classical condition from the right perspective and with the right formalism. But we have not yet said if it is actually true, nor have we given any indication as to how one might go about proving it.

So it may not be a bad idea to look at a few more toy problems. We will consider two of them.

Example V.8. Suppose $L > 0$ is given, and we want to *maximize* the integral

$$J = \int_0^L \dot{\xi}(t)^2 dt$$

among all Lipschitz functions $\xi(\cdot) : [0,L] \mapsto \mathbb{R}$ that satisfy the velocity constraint

$$\dot{\xi}(t) \geq 0 \qquad \text{for almost all} \quad t\,,$$

as well as the endpoint conditions

$$\xi(0) = \xi(L) = 1\,.$$

In this case the set U of possible velocities and the Lagrangian are given by

$$U = [0, +\infty]\,,\; L(q,u,t) = -u^2\,.$$

Therefore the Hamiltonian is

$$H(q,u,p,t) = pu + u^2\,.$$

We could go on and write (V.31) and (V.32) for our problem, but there is no need to do so. Just by looking at H, we can see that for every value of p the function $u \mapsto H(q,u,p,t)$ (which is in fact independent of q and t) does not have a maximum U. So no trajectories at all satisfy the conditions of (NCO). On the other hand, our optimization problem does have a solution, since $t \mapsto \xi(t) = 1$ is the only curve that satisfies our endpoint conditions as well as the constraint $\dot{\xi}(t) \geq 0$. So (NCO) is *not* satisfied by the solution. □

Example V.9. We want to minimize the integral

$$J = \int_0^1 \xi_1(t)\xi_2(t)\,dt$$

among all Lipschitz functions $\xi(\cdot):[0,1]\mapsto\mathbb{R}^2$ that satisfy the velocity constraint

$$\dot{\xi}_1(t)^2+\dot{\xi}_2(t)^2\leq 1 \qquad \text{for almost all} \quad t\,,$$

as well as the endpoint conditions

$$\xi(0)=(0,0)\,,\ \ \xi(1)=(1,0)\,.$$

In this case the set U of possible velocities and the Lagrangian are given by

$$U=\{(u_1,u_2)\in\mathbb{R}^2 : u_1^2+u_2^2\leq 1\}\,,\ L(q_1,q_1,u_1,u_2,t)=q_1q_2\,.$$

It is clear that there exists only one feasible path ξ, namely, the arc given by $\xi(t)=(t,0)\,,\quad \text{for} \quad 0\leq t\leq 1\,.$ Let us see whether this arc satisfies (NCO).

The Hamiltonian is

$$H(q_1,q_2,u_1,u_2,p_1,p_2,t)=p_1u_1+p_2u_2-q_1q_2\,.$$

The momentum $t\mapsto\pi(t)=(\pi_1(t),\pi_2(t))$, if it exists, would have to satisfy Hamilton's equations

$$\dot{\pi}_1(t)=\xi_2(t)\,,\ \ \dot{\pi}_2(t)=\xi_1(t)\,,$$

because

$$\frac{\partial H}{\partial q_1}=-q_2 \quad \text{and} \quad \frac{\partial H}{\partial q_2}=-q_1\,.$$

Since $\xi_1(t)=t$, we find

$$\dot{\pi}_2(t)=\xi_1(t)\,. \tag{V.42}$$

On the other hand, the Hamiltonian maximization condition says that for almost all t the function

$$U\ni(u_1,u_2)\mapsto\pi_1(t)u_1+\pi_2(t)u_2-\xi_1(t)\xi_2(t) \tag{V.43}$$

must be maximized by $u_1=\dot{\xi}_1(t)=1$, $u_2=\dot{\xi}_2(t)=0$. On the other hand, if $\pi(t)\neq 0$, then the unique value u^{max} of $u=(u_1,u_2)$ that maximizes (V.43) is

$$u^{max}=\frac{\pi(t)}{\|\pi(t)\|}\,.$$

It follows that $\pi(t)\neq 0\Rightarrow\pi_2(t)=0$. Since $\pi(t)=0\Rightarrow\pi_2(t)=0$ as well, we have shown that

$$\pi_2(t)\equiv 0\,. \tag{V.44}$$

Clearly, (V.44) contradicts (V.42). Therefore π does not exist, and (NCO) is *not* satisfied by the solution. □

Once again, as in Example V.10, we seem to be in trouble. But, as opposed to what happened in Example V.10, this time there is no mistake. The problem of Example V.8 is for a fixed time interval, and we have applied (NCO) correctly. So we have no choice, and have to acknowledge that the conjecture that seemed so natural is in fact false.

It turns out, however, that only a minor modification is needed to make it true. All we have to do is introduce a new p-variable p_0, called the "abnormal multiplier,"[14] and write the Hamiltonian as

$$H(q,u,p,p_0,t) = \langle p, f(q,u,t)\rangle - p_0 L(q,u,t)\,. \tag{V.45}$$

Everything we had done until now corresponded to taking $p_0 = 1$. We now impose, instead, the weaker requirement that $\dot{p}_0 = 0$ (i.e., p_0 is a constant) and $p_0 \geq 0$. We then observe that, if we use this new H rather than the old one, then the three conditions of (NCO) are always satisfied if we make the trivial choice $p(t) \equiv 0$, $p_0 = 0$. So our new conditions will give nothing interesting unless we impose a further *nontriviality condition*, stating that this possibility is excluded.

In our case, it turns out that, if we choose $p(t) \equiv -1$ and $p_0 = 0$, then the adjoint equation $\dot{p} = -\frac{\partial H}{\partial x}$ holds along our candidate trajectory, and the Hamiltonian $H = pu + p_0 u^2$ is maximized, on the set $U = [0, +\infty[$, by choosing $u(t) \equiv 0$. So our trajectory satisfies the modified version of the necessary conditions, with an "abnormal multiplier" equal to zero.

V.13 A stronger conjecture and a third toy problem

Conjecture M2 says that, now that we have liberated ourselves from the constraint that L be differentiable with respect to u, it ought to be possible for u—i.e., $\dot{q}$—to be ***anything***, and (NCO) will still work. Once this is understood, the next natural step is to allow $\dot{q}$ to be even "more arbitrary," for example a general function of some other variable u, and of q and t. So instead of letting $\dot{q}$ be u, we can write $\dot{q} = f(q,u,t)$ for a general function $f(q,u,t)$. The natural conjecture now is

> ***CONJECTURE M3:*** *(NCO) should still be a necessary condition for optimality even for problems where q is restricted to satisfy a differential equation $\dot{q} = f(q,u,t)$, with the "control function" $t \mapsto u(t)$ taking values in some set U and allowed to be a "completely arbitrary" U-valued function of t.* □

[14] The need for the abnormal multiplier had already been noticed by Bolza in 1913, cf. [18].

There is a problem, though. If u is going to take values in an arbitrary set U, whereas p obviously will still take values in $\mathbb{R}^n$, what should we make of the term $\langle p,u\rangle$ that occurs in (V.8)?

There is a fairly natural guess as to what should replace u in the expression $\langle p,u\rangle$. In the classical context, u was $\dot{q}$, so $\langle p,u\rangle$ was $\langle p,\dot{q}\rangle$ More precisely, $H(q,u,p,t)$ was the sum of $-L(q,u,t)$ plus p times "$\dot{q}$ as a function of u," whatever that means. In our new setting, "$\dot{q}$ as a function of u" is $f(q,u,t)$, so one can supplement Conjecture M3 with

SUBCONJECTURE M3.a: *The Hamiltonian H that must be used to apply (NCO) under the conditions of Conjecture M3 is given by*

$$H(q,u,p,t) = \langle p, f(q,u,t)\rangle - L(q,u,t)\,. \qquad \text{(V.46)}$$

As before, we can test this with a simple toy problem where the answer can easily be found directly.

Example V.10. Suppose we have a point in $\mathbb{R}^2$ with coordinates x,y, moving according to $\dot{x}=y$, $\dot{y}=u$, $|u|\leq 1$. (Think of x as the position of a point moving in the real line $\mathbb{R}$. Then y is the velocity, u is the acceleration, and we are constraining the motion by imposing a bound on the acceleration.) Starting at $x=4$, $y=1$, we want to get to $x=0$, $y=0$ in minimum time. (That is, we want to get to the target position $x=0$ and arrive with zero velocity, i.e., we have a "soft landing in minimum time" problem.) The reader should be able to solve this directly. (Clearly, since we are starting off with a velocity vector pointing away from the target, we should decelerate, i.e., use $u=-1$, at least initially. If we just use $u=-1$ all the time we will have $y=1-t$, $x=4+t-\frac{t^2}{2}$, so we get to $x=0$ at time $t=4$, but then $y=-3$, so we do not soft-land. So we should use $u=-1$ for a certain time T and then switch to $u=1$, i.e., slow down, so as to guarantee that $y=0$ when $x=0$. The time T can be computed exactly, and then one can give a direct proof that the resulting motion is indeed optimal.) If we use (NCO) with the Hamiltonian defined according to Subconjecture M3.a, we get (since the "Lagrangian" is now equal to 1) $H=\xi y+\eta u-1$, and then $\dot{x}=y$, $\dot{y}=u$, $\dot{\xi}=0$, $\dot{\eta}=-\xi$, and, finally, $u=1$ if $\eta>0$, $u=-1$ if $\eta<0$. (Here (ξ,η) are the components of the vector p. Recall that in this problem q and p are in $\mathbb{R}^2$.) So $\xi(t)$ is a constant, $\eta(t)$ is a linear function of t, and then our curve must be of one of the types P, N, NP or PN, ***unless*** $\eta(t)\equiv 0$. (Here P means $u\equiv 1$, N means $u\equiv -1$, and we are using symbols such as PN to denote concatenations, as we did in Example V.7.) A P curve can never go from $(4,1)$ to $(0,0)$, since along such a curve y will steadily increase. We have already computed the N curve and shown that it gets us to $x=0$ with $y=-3$, so this possibility is also excluded. A *PN* curve where u switches values at time t_0 will get us to $(4+t_0+h+t_0h+\frac{t_0^2-h^2}{2}, 1+t_0-h)$ in time t_0+h, and it is easy to see that this curve does not go through $(0,0)$. One can compute an *NP* curve going from $(4,0)$ explicitly, and this curve turns out to be unique.

Naturally, the curve we have just determined ought to be the solution of our minimization problem, and our argument seems to have led us to this solution by applying Conjecture M3, which has thus been verified.

All is not well, however, since ***we have not yet disposed of the possibility that*** $\eta \equiv 0$. And, unfortunately for us, this could indeed happen. In fact, if we take $\eta \equiv 0$ and $\xi \equiv 0$, and a completely arbitrary curve $t \mapsto (x(t), y(t))$ satisfying our constraints, we can easily see that *all the conditions of (NCO) hold.* So, after all our hard work, we end up making the disappointing discovery that ***every curve satisfies the necessary condition of (NCO), so (NCO) appears to be completely useless as a tool to single out the solution of our problem.***

At this point it may seem that we ought to abandon all hope and give up on Conjecture M3. It turns out, however, that Conjecture M3 is quite sound, if properly interpreted, and the apparent setback was caused by our mistaken understanding of its meaning. In fact, we contend that *Weierstrass would not have been deterred by this particular setback*, because Weierstrass knew a trick that takes care of our problem, a trick that we have already described before, in Remark V.2. This is how Weierstrass would have reasoned: "it was *our* mistake that we sought to apply (NCO) to a problem in which the time interval should be allowed to vary, even though all our necessary conditions for optimality, including (NCO), were for a *fixed* time interval. It is of the very essence of a minimum-time problem that the time interval not be fixed, for a minimum-time problem with a fixed time interval $[a,b]$ is obviously silly. More precisely, *all* arcs are solutions of such a problem, since they all take the same time, namely, $b-a$. So it was to be expected that applying (NCO) would have given us "all curves" as the answer, because that is indeed the right answer for a minimum time problem when the time interval is fixed. (NCO) is still fine, and Conjecture M3 has not yet been refuted. It was our fault that we asked the wrong question. To apply the fixed-time-interval condition (NC0) to a variable-interval problem such as ours, we first have to transform our problem to one with a fixed time interval! ... Actually, we know very well how to do that, because this is exactly what we were doing in Remark V.2 when we transformed a general problem to one with a Lagrangian satisfying (W). The thing to do is to introduce a 'pseudotime' parameter s, not to be thought of as true physical time, just making sure that $dt/ds > 0$. So if we write $dt/ds = v$, we have a new system of equations $dx/ds = vy$, $dy/ds = vu$, where $v > 0$ and $|u| \leq 1$. With such a reparametrization, we can always work on a fixed pseudotime interval, for example $[0,1]$. Now the functional to be minimized is the integral $\int_0^1 v(s)ds$, so our new Lagrangian is v."

Let us pursue this suggestion of our imaginary Weierstrass. Our equations are $\dot{x} = vy$, $\dot{y} = vu$, with two "controls" u, v, restricted to $|u| \leq 1$, $v > 0$. (The dot now stands for differentiation with respect to s.) Now the "control Hamiltonian" is $H = \xi vy + \eta vu - v$, and (NCO) gives $\dot{x} = vy$, $\dot{y} = vu$, $\dot{\xi} = 0$, $\dot{\eta} = -\xi v$. The tricky part now is to maximize H. Since $H = hv$, where $h = \xi y + \eta u - 1$, we see that h must vanish, for if $h \neq 0$ then H, as a function of v, would not have a maximum on the set $\{v : 0 < v < \infty\}$. So $h \equiv 0$. Now, given s, we still have to make sure that $u(s)$ maximizes H as a function of u, with $v = v(s)$. Since $v(s) > 0$, maximizing H amounts to maximizing h. Now h is exactly the Hamiltonian of our first, failed attempt to solve the problem. So as before we get that our curve must be P, N, NP or PN. The argument changes slightly: ξ is still a constant, but $\dot{\eta} = -\xi v$ no longer implies that η is linear. However, $\dot{\eta} = -\xi v$ does imply that η is strictly increasing—if $\xi < 0$—or strictly decreasing—if $\xi > 0$—since $v > 0$. Then our curve has to be P, N, NP or PN. In the degenerate case when $\xi = 0$ we get $\eta = c =$ constant, which leads to a P or N curve as before, unless $c = 0$. ***But now the possibility that*** $\xi = \eta = 0$ ***is automatically excluded***, because $\xi = \eta = 0$ would imply $h = -1$, which is out of the question since $h \equiv 0$. So we get exactly the same conclusion as in our failed attempt, except that now the "degenerate case" is excluded. So Conjecture M3 seems to work after all. □

Those readers who are familiar with optimal control theory will of course have recognized Conjecture M3 as being essentially[15] the same thing as the celebrated "Pontryagin maximum principle."

And we hope to have convinced *all* readers, even those who are not control theorists, that (NCO) is a very natural condition. It should be clear from our discussion that (NCO) could have been guessed almost immediately from "Hamilton's equations as Hamilton should have written them," together with the Legendre condition, and would have been an almost obvious conjecture to make once the Weierstrass side condition is known, *provided that one uses the "correct" Hamiltonian formalism, as in (V.8) and (V.12).*

Remark V.11. Notice that what we did in Example V.10—i.e., transforming a problem into one with a fixed time interval and a Lagrangian that satisfies Condition (W), and then using (NCO)—*could not possibly have been done with Hamilton's Hamiltonian.* Indeed, the condition needed for the classical Hamilton equations to be possible is the solvability of (V.15) for $\dot{q}$ as a function of q, p, and t. This

[15] We say "essentially" for a good reason. We will get to it soon.

condition *never* holds when (W) holds because, if L is positively homogeneous of degree 1 with respect to $\dot{q}$, then $\frac{\partial L}{\partial \dot{q}}$ is positively homogeneous of degree 0, i.e., constant along rays in $\dot{q}$ space. So the map $\dot{q} \mapsto \frac{\partial L}{\partial \dot{q}}(q, \dot{q}, t)$ is not one-to-one locally, in the neighborhood of any point, for any q, t. Therefore Weierstrass' use of Lagrangians with property (W), while involving "no loss of generality," had the effect of putting him in a situation where the use of a Hamiltonian formalism, as was known at the time, was *a priori* excluded. The "control" Hamiltonian formalism, on the other hand, would have worked perfectly. □

SCHOLIUM VII. Snell, Descartes, Fermat, Leibniz, Huygens on refraction

The famous "Snell law" of refraction was discovered by Snell[16] in 1621, but he did not publish it at the time. The law only become and only in 1703 did it become known when Huygens published Snell's result in Dioptrica. Snell also discovered the sine law.

Descartes published the law of refraction in 1637 in his ***Dioptrique***, together with his own explanation. Fermat was strongly opposed to this law at first, claiming that Descartes' own deduction was incorrect, and indeed Descartes' argument was wrong, because his theory required that light travel *faster* in a denser medium.

Subsequently, Fermat accepted the law, of which he gave a derivation based on the minimum-time principle: ***in a medium with variable speed of light, the light rays are paths that minimize travel time from the intial point to the terminal point***. Fermat's principle yielded the correct law of refraction only if light was assumed to travel more slowly in a denser medium. (Later, in 1690, C. Huygens (1629-1695) provided a justification of Fermat's principle by deriving it from wave optics.)

Descartes was very angry at Fermat's criticism, and strongly attacked Fermat's methods of maxima, minima and tangents. After Roverbal, Etienne Pascal[17] and Desargues intervened in the dispute, Descartes eventually admitted that Fermat was right, and wrote

---

[16]The Dutch mathematician ***Willebrord van Roijen Snell*** (1580-1626), studied law at the University of Leiden, the Netherlands, where his father, Rudolph Snell (1546-1613), was professor of mathematics. Willebrord's true interests were mathematics and astronomy, and he became a professor of mathematics at the University of Leiden in 1613, succeeding his father. He devised techniques for for measuring the Earth, and in particular the method of triangulation. He developed new approaches for calculating approximate values of π. Using his method with polygons of 96 sides he got a correct value of π to 7 decimal places. He studied navigation, and described the curve on the sphere —called the ***loxodrome***—that intersects the meridians at equal angles.

[17]The father of Blaise Pascal

> … ... seeing the last method that you use for finding tangents to curved lines, I can reply to it in no other way than to say that it is very good and that, if you had explained it in this manner at the outset, I would have not contradicted it at all.

Nevertheless, Descartes went on attacking Fermat and trying to damage his reputation, for example by writing a letter to Mersenne saying that Fermat's determination of the tangent to a cycloid was incorrect, and that Fermat was not a good mathematician, even though Fermat's work was correct and Descartes had said so in a letter to Fermat.

A readable summary of the early history of the refraction problem (not mentioning the contributions of Snell and Descartes) is given by none other than Johannm Bernoulli in his 1967 brachytochrone paper [15]. We quote from Struik [124], p. 393':

> Fermat, in a letter to De la Chambre, has shown that a light ray passing from a thin to a more dense medium is bent toward the perpendicular in such a way that, under the supposition that the ray moves continuously from the light to the illuminated point, it follows the path that requires the shortest time. With the aid of these principles he showed that the sine of the angle of incidence and the sine of the angle of refraction are in inverse proprtion to the densities of the media, hence directly as the velocities with which the light penetrates those media. Later Leibniz … and soon afterward the famous Huygens in his ***Treatise on light***, p 40, have demonstrated this more comprehensively and, by most valid arguments, have established the physical, or better the metaphysical, principle which Fermat seems to have abandoned at the insistence of Clerselier, remaining satisfied with his geometric proof and giving up his rights all too lightly.

The work of Leibniz alluded to by Johann is his his 1683 ***Acta Eruditorum*** paper [90], on a unified foundation for optics, catoptrics and dioptrics, where the great philosopher announced for the first time that he had a method for solving problems of maxima and minima, of such power that, according to Leibniz' own assessment in his later paper [91],

> Other very learned men have sought in many devious ways what someone versed in this calculus can accomplish in these lines as by magic. ... This is the beginning of much more sublime Geometry, pertaining to even the most difficult and most beautiful problems of applied mathematics, which without our differential calculus or something similar noone could attack with any such ease. [Quoted from Struik [124], p. 279.]

Subsequently, in 1684, Leibniz published a much longer paper [91], with the title

> A NEW METHOD FOR MAXIMA AND MINIMA AS WELL A TANGENTS, WHICH IS NEITHER IMPEDED BY FRACTIONAL NOR IRRATIONAL QUANTITIES, AND A REMARKABLE TYPE OF CALCULUS FOR THEM.

This paper gave a more detailed account of the method, as well as many examples. Here we shall focus on one example of particular interest to us, because it involves a ***curve minimization problem.***

Leibniz proposes to solve the following problem:

> Let two points C and E be given and a line SS in the same plane. It is required to find a point F on SS such that when C and E are connected the sum of the rectangle of CF and a given line h and the rectangle of FE and a given line r are as small as possible. In other words, if SS is a line separating two media, and h represents the density of the medium on the side of C (say water), r that of the medium on the side of E (say air), then we ask for the point F such that the path from C to E via F is the shortest possible. [Quoted from Struik [124], p. 278.]

By "the rectangle of CF and a given line h" Leibniz means "the area of a rectangle with base CF and height h," so the problem he is solving is that of making the sum

$$T = h|CF| + r|FE|$$

as small a possible, subject to the condition that F belong to the given line SS. Leibniz then applies his result to dioptrics and finds that

> the sines of the angles of incidence and of refraction … are in inverse ratio to r and h, the densities of the media in which the incidence and the refarction take place.

Since these densities represent "the resistance with which the light rays meet," they are inversely proportional to the speed of propagation of light in the two media, so Leibniz' solution is in fact the familiar law of Snell about refraction.

SCHOLIUM VIII. Two excerpts from G. Adair's English translation of Georges Perec's "La Disparition"

Georges Perec's remarkable 285-page novel "La Disparition" does not contain a single letter "e." The book has been translated into Engish by G. Adair, under the title"A Void." Here are two excerpts from the English translation, the second one of which contains the passage about Pontryagin cited in Footnote 13.

> Call him Ishmail, too. Having had to sustain many, almost inhuman hardships, this Ishmail drops anchor on an island that is, at first sight, totally without inhabitants. Initially, things go badly, with his almost dying of cold and starvation, burrowing into a sand pit, shaking, suffocating, coming down with malaria, curling up both day and night in a tight ball, simply waiting to pass away.
>
> But six days go past, and his unusually robust constitution allows him finally to sit up and look around him. Our protagonist is now disturbingly skinny but can still, if with difficulty, climb out of his pit, a pit that was almost his tomb. So Ishmail starts living again by first luxuriantly shaking his thirst; now swallows an acorn but instantly spits it out in disgust; and soon works out which fruits and mushrooms won't do him any harm. A particular fruit (akin to an apricot) brings him out in a rash of itchy, purplish spots, but bananas, avocados, nuts and kiwi fruit abound.
>
> At dusk, using a sharp rock, Ishmail cuts a notch on a stick (to tick off his first night as a castaway); and, with a total of 20 such cuts, constructs a hut, a sort of impromptu shack, with a door, four walls, and flooring and roofing built out of mud. With no matchsticks at hand, his food is invariably raw. Constant, too, is his panic that an animal might attack his camp; but, as it turns out (such, anyway, is his supposition), no lynx or puma, no jaguar, cougar or bison, stalks his island. At most, comfortably far off, a solitary orangutan prowling around but not caring (or daring?) to approach him.
>
> ([108], page 17)

Who first had this particular notion, who brought it to its maturity, who found a solution to it? Was is Gauss or Galois? Nobody could say. Nowadays, though, all of us know. But it's said that, just prior to dying, at night, at about 4 or 5 p.m., Galois put in writing on his jotting pad ... a long, continuous chain of factors in his own form of notation.. As part of his manuscript is missing, nobody knows to this day what conclusions Galois was hoping to draw from his calculations.

Cantor, Douady and Bourbaki thought, on many and various grounds (from algorithms to topoi, from Möbius strips to C-star, from Shih's K-functor to Thom's □s, and including all sorts of distributions, convolutions, Schwartz, Koszul, Cartan and Giorgiutti) of following a hunch that would surmount such an abrupt hiatus. It was, alas, all in vain.

Though it took him from 1935 to 1955, working at it virtually nonstop, Pontryagin finally had to admit to accomplishing nothing.

([108], pages 46-47)

Chapter VI

The maximum principle of optimal control theory

With Conjecture M3 adjusted[1] with the introduction of the new variable p_0 we have, finally, reached the justly celebrated ***Maximum Principle***:

(MP) *Consider a standard point-to-point optimal control problem*

$$\boxed{\begin{array}{ll} \text{Given} & Q, U, f, L, a, b, \bar{q}, \hat{q}, \mathscr{C}\,, \\ \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)dt\,, \\ \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\ & (\xi, \eta) \in \mathscr{C}\,, \\ \text{and} & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \leq t \leq b\,, \end{array}} \tag{VI.1}$$

where

1. *Q is an open subset of $\mathbb{R}^n$,*
2. *U is a set,*
3. *$a, b \in \mathbb{R}$ and $a \leq b$,*
4. *$\bar{q} \in Q$ and $\hat{q} \in Q$,*
5. *$Q \times U \times [a,b] \ni (q,u,t) \mapsto (f(q,u,t), L(q,u,t)) \in \mathbb{R}^n \times \mathbb{R}$ is a map,*

Then, under suitable technical conditions, if a pair $(\xi_*, \eta_*) \in \mathscr{C}$ is such that $[a,b] \ni t \mapsto \eta_*(t) \in U$ is a control and the curve $[a,b] \ni t \mapsto \xi_*(t) \in Q$ is a corresponding solution of (XVI.1), a necessary condition for (ξ_*, η_*) to solve the minimization

[1]It is because Conjecture M3 is false unless this adjustment is made that we used the word "essentially" in Footnote 15.

problem is that there exist a "momentum" function $[a,b] \ni t \mapsto \pi(t) \in \mathbb{R}_n$ and a constant $\pi_0 \geq 0$ such that

(NT) $(\pi(t), \pi_0) \neq (0,0)$ *for all* $t \in [a,b]$*;*

(HS) $\dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Xi_*(t))$ *and* $-\dot{\pi}(t) = \frac{\partial H}{\partial q}(\Xi_*(t))$ *for* $t \in [a,b]$*;*

(MC) $H(\Xi_*(t)) = \max_{u \in U} H(\xi_*(t), u, \pi(t), \pi_0, t)$ *for* $t \in [a,b]$*,*

where we write $\Xi_*(t) = (\xi_*(t), \eta_*(t), \pi(t), \pi_0, t)$*, and the Hamiltonian function* $Q \times U \times \mathbb{R}_n \times \mathbb{R} \times [a,b] \ni (q,u,p,p_0,t) \mapsto H(q,u,p,p_0,t) \in \mathbb{R}$ *is given by*

$$H(q,u,p,p_0,t) = \langle p, f(q,u,t) \rangle - p_0 L(q,u,t)\,. \tag{VI.2}$$

The three statements (NT), (HS) and (MC) are known, respectively, as the ***nontriviality condition***, the *Hamiltonian system*, and the ***Hamiltonian maximization condition***. Notice that (HS) is just a restatement of (V.31) and (V.32), with the new H, and (MC) is a restatement of (V.33). The second equation of (HS) is called the *adjoint equation*.

We will refer to pairs $(q(\cdot), u(\cdot))$—where $t \mapsto u(t)$ is a U-valued function on $[a,b]$ and $q(\cdot)$ is a solution of (XVI.1)—as *trajectory-control pairs*. (Many authors use the much less informative phrase "admissible pairs.") A trajectory-control pair that satisfies the condition of (MP)—i.e., is such that there exist p and $p_0 \geq 0$ for which (NT), (HS) and (MC) hold—will be called an *extremal*. So the maximum principle says that a necessary condition for a trajectory-control pair to be a solution of the minimization problem is that it be an extremal.

For a given trajectory-control pair $\gamma_* = (q_*, u_*)$, a vector-valued function $t \mapsto p_*(t)$ for which Conditions (NT), (HS) hold for some $p_0 \geq 0$ will be called an *adjoint vector*[2] for γ_*. If, in addition, p_0 can be chosen so that (MC) holds as well, then p_* will be called a *minimizing adjoint vector* for γ_*. Often, the whole pair (p_*, p_0) is also referred to as an adjoint vector, or a minimizing adjoint vector. We will refer to a triple (q_*, u_*, p_*) (or a quadruple (q_*, u_*, p_*, p_0)) such that (q_*, u_*) is a trajectory-control pair and p_* (or (p_*, p_0)) is a minimizing adjoint vector for (q_*, u_*) as an *extremal lift* of (q_*, u_*). A extremal lift (q_*, u_*, p_*, p_0) of a pair (q_*, u_*) is *normal* if $p_0 \neq 0$, and *abnormal* if $p_0 = 0$. An extremal is *normal* if it has a normal extremal lift, and *abnormal* if it has an abnormal extremal lift.

Some authors use "extremal" for what we are calling "extremal lift" here. Naturally, for those authors the categories of normal and abnormal extremal are mutually exclusive. This is not so with our terminology: it can happen that an extremal is both normal and abnormal, since the extremal lift need not be unique.

[2]For differential geometers: p is really a ***field of covectors*** along q_*.

When the abnormal multiplier p_0 is $\neq 0$, i.e., > 0, it can always be set equal to 1, since multiplication of both p_* and p_0 by a strictly positive constant leaves all the conditions of (MP) unchanged. As long as $p_0 = 1$, (MP) is exactly the same as our previous Conjecture M3. So the only possibility that has been added now is the case $p_0 = 0$. In other words, what the maximum principle asserts is that a solution of the minimization problem problem always satisfies either (i) the necessary condition given by Conjecture M3 as stated, or (ii) the necessary condition given by Conjecture M3 with $L \equiv 0$ and the extra requirement that p be $\neq 0$.

The reader will notice that in Examples V.5, V.7 and V.10 one can always either exclude the possibility that $p_0 = 0$ or show that it leads to nothing new. Indeed, in Examples V.5 and V.7 the Hamiltonian for $p_0 = 0$ is $H = p.u$, so the adjoint equation is $\dot{p} = 0$. Hence p is a constant, which cannot equal 0 by (NT). So $u(t) \equiv 1$ or $u(t) \equiv -1$, and in either case the endpoint conditions have to be violated. In Example V.10 (properly modified as suggested by our imaginary Weierstrass), the abnormal Hamiltonian is $H = \xi vy + \eta vu$, and $H = hv$, with $h = \xi y + \eta u$. Exactly as in our analysis of Example V.10, we see that h has to vanish, and this implies that our curve is of one of the forms we have already found, unless we are in the singular case $\xi \equiv \eta \equiv 0$. In the discussion of Example V.10, the singular case was excluded because it led to $h = -1$. Though this no longer works now, the singular case is still excluded anyhow, because of (NT).

As for Example V.8, it is easy to see that the trajectory x_* given by $x_*(t) = 1$—which is the only one that satisfies the boundary conditions and the constraint on $\dot{x}$—is an abnormal extremal. Indeed, the abnormal Hamiltonian is $H = pu$, and then the adjoint equation is $\dot{p} = 0$. We can then choose $p_*(t) \equiv -1$, and with this choice $u \mapsto H(x_*(t), u, p_*(t), 0, t)$ is maximized by $u = 0$.

Finally, we remark that for classical calculus of variations problems (MP) yields exactly the same conclusion as (NCO). Indeed, in this case we can exclude the possibility that $p_0 = 0$, because the abnormal Hamiltonian is $\langle p, u \rangle$, which is a linear function of u. Such a linear functional, if its domain is the whole space, can never have a maximum, unless $p = 0$. But $p = 0$, together with $p_0 = 0$, is excluded by (NT).

So (MP), as stated, is a true generalization of the necessary conditions (NCO), which covers many cases that cannot be handled by means of the classical calculus of variations, as has been seen in Examples V.5, V.7, V.10 and V.8.

We conclude by presenting the analogue of (MP) for problems with a variable time interval:

(MP') *For a minimization problem of the kind discussed in (MP), but with the time interval* $[a,b]$ *not fixed in advance, assuming that* f *and* L *do not depend on* t*, the necessary conditions are exactly the same as those of (MP), plus the extra requirement that* $H(q_*(t),u_*(t),p_*(t),p_0) \equiv 0$*.* □

This is easily derived from (MP), by means of the method that we used in our analysis on Example V.10. One has to transform the given problem to a fixed-time problem by means of a time reparametrization. What was done in Example V.10 works in general, so we will not repeat the argument here.

Statement (MP') applies in particular to *minimum-time problems*, i.e., problems where $L \equiv 1$.

VI.1 Example I: the straight-line segment

COMMENT NO. 19. JAN: This section still has to be written. If you want an idea of what I want to put in, look at the corresponding section in my French paper (item no. 58 in the "papers available on the Web" link of my Web page), where I dicuss the "rope-stretching problem" and related stuff.

VI.2 Example II: the isoperimetric problem

.

As explained in §V.1.2, the question is

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

or, equivalently,

(IP') Of all plane closed curves that enclose a given area A, find those of minimum length.

To formulate the problem mathematically, we have to make precise sense of the concepts of "closed curve," the "length" of a curve, and the area "enclosed" by a closed curve.

A *closed curve* in $\mathbb{R}^2$ is a continuous map $\xi : [a,b] \mapsto \mathbb{R}^2$, defined on some interval $[a,b]$, such that $\xi(a) = \xi(b)$.

Curves that have a well defined length are called "rectifiable." (This concept is defined precisely in §GNU below.) The "area enclosed by a closed curve ξ" is the area of the region enclosed by the curve. The concept of "area" of a "region" Ω is relatively unproblematic, at least if Ω is an open bounded set, since we can define the area of Ω to be its Lebesgue measure, which could also be characterized in a more elementary way as the supremum of all the sums $\sum_{j=1}^{N} \text{Area}(R_j)$, taken over all N and over all finite sequences $R_1, \ldots, R_N$ of pairwise disjoint rectangles contained in Ω.

On the other hand, the concept of "region enclosed by a closed curve" turns out to be suprisingly delicate. The curves we are interested in for the IP are *simple* closed curves, that is, closed curves ξ without double points. (This means that, if $\xi : [a,b] \mapsto \mathbb{R}^2$, then ξ is a one-to-one map on the interval $[a,b[$; in other words, if $t_1, t_2 \in [a,b]$, $t_1 \le t_2$ and $\xi(t_1) = \xi(t_2)$, then either (1) $t_1 = t_2$ or (2) $t_1 = a$ and $t_2 = b$.) It is intuitively clear that a simple closed curve ξ divides the plane into two regions, the "inside" $\text{Ins}(\xi)$ and the "outside" $\text{Out}(\xi)$. However, it is not completely trivial how to make this seemingly evident statement precise, because to do so one has to make sense first of the word "divides." But this is relatively easy: one defines a *connected component* of an open set Ω to be a subset C of Ω which is, for some $c \in \Omega$, the set of all points of Ω that can be joined to c by some continuous path in Ω; it is then easy to see that two connected components of Ω are either disjoint or equal. The statement about "dividing" the plane can then be reformulated by saying that

(JCT) *If $\xi : [a,b] \mapsto \mathbb{R}^2$ is a simple closed curve, and $|\xi|$ denotes the set $\{\xi(t) : a \le t \le b\}$, then $\mathbb{R}^2 \backslash |\xi|$ has exactly two connected components.*

Assertion (JCT) is the *Jordan curve theorem.* Although it is intuitively obvious, as we have explained, the proof is suprisingly difficult. (We present a complete self-contained proof in §GNU below.)

If we grant the truth of (JCT), then we can define $\text{Ins}(\xi)$ and $\text{Out}(\xi)$ precisely. Indeed, if r is large enough (such that $\|\xi(t)\| \le r$ for all t), then one of the two connected components of $\mathbb{R}^2 \backslash |\xi|$ must contain the set $S_r = \{(x,y) : x^2 + y^2 > r^2\}$, since $S_r \cap |\xi| = \emptyset$, and any two points of S_r can be joined by a continuous path in S_r; the component that contains S_r is obviously unbounded; the other component will then not intersect S_r, so it will be bounded; we define $\text{Ins}(\xi)$ and $\text{Out}(\xi)$ to be, respectively, the bounded and unbounded components $\mathbb{R}^2 \backslash |\xi|$.

Now that we know what "$\text{Ins}(\xi)$" means, we can use use the well-known "Green's formula"

$$\text{Area}\Big(\text{Ins}(\xi)\Big) = |A(\xi)|\,, \qquad \text{where} \qquad A(\xi) = \frac{1}{2}\int_{\xi} (x\,dy - y\,dx) \qquad \text{(VI.3)}$$

for the area of $\mathrm{Ins}(\xi)$. (Actually, the area of $\mathrm{Ins}(\xi)$ is $A(\xi)$ if the simple closed curve ξ is "oriented counterclockwise." If ξ is "oriented clockwise," then $A(\xi)$ is negative, and the area of $\mathrm{Ins}(\xi)$ is $-A(\xi)$, i.e., $|A(\xi)|$. The issue of orientation will be discussed in detail in §GNU.)

Naturally, (VI.3) requires proof, and the proof is not trivial[3]. We will give a rigorous proof in §GNU, but here we will just accept the validity of (VI.3).

The last point that needs to be dealt with is that of writing the length of ξ as an integral. It turns out that every rectifiable curve ξ admits a reparametrization of time that transforms it into an absolutely continuous curve, without changing the length or the area. This means that we might as well limit ourselves to working with absolutely continuous curves[4]

Let us use "AC" as an abbreviation for "absolutely continuous." If ξ is AC, then the length of ξ is just the integral $\int_a^b \|\dot{\xi}(t)\|\,dt$. If $n=2$, and we write $\xi(t) = (x(t), y(t))$, then the length of ξ is equal to $\int_a^b \sqrt{\dot{x}(t)^2 + y(t)^2}\,dt$.

We are now ready to formulate the isoperimetric problem as follows,

THE ISOPERIMETRIC PROBLEM, VERSION I

Given	a number L such that $L > 0$,
maximize	$\int_0^1 \Big(x(t)\dot{y}(t) - y(t)\dot{x}(t) \Big)\,dt$
subject to	$\int_0^1 \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\,dt = L$,
	$x(1) = x(0)$, $y(1) = y(0)$,
	the curve $(x(\cdot), y(\cdot))$ is simple and AC.

(We do not include the requirement that the curve should be "oriented counterclockwise" because for simple curves ξ that are oriented clockwise the integral of (VI.3) is negative, so these curves are automatically eliminated as possible solutions.)

Alternatively, we can use the equivalence between Problems (IP) and (IP') above to cast our problem as one of minimizing the perimeter given the enclosed area:

[3]Actually, even assigning a precise meaning to (VI.3) requires some work. Indeed, once we have given the precise definition of "rectifiable curve," it will turn out that ξ is rectifiable if and only if it is continuous and of bounded variation. Then the right-hand side of the formula defining $A(\xi)$ makes sense as a Riemann-Stieltjes integral.

[4]In general, $W^{k,p}(\Omega, \mathbb{R}^n)$ denotes the Sobolev space of all $\mathbb{R}^n$-valued functions f on Ω such that f and all the distributional derivatives of f of order $\leq k$ are in L^p. In particular, $W^{1,1}([a,b], \mathbb{R}^n)$ consists of the functions having an integrable distributional derivative, i.e., of the absolutely continuous functions. So we are choosing $W^{1,1}([a,b], \mathbb{R}^n)$ as our space of curves.

THE ISOPERIMETRIC PROBLEM, VERSION II

Given	a number A such that $A > 0$,
Minimize	$\int_0^1 \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\, dt$,
subject to	$\int_0^1 \Big(x(t)\dot{y}(t) - y(t)\dot{x}(t) \Big)\, dt = A$,
	$x(1) = x(0)$, $y(1) = y(0)$,
	the curve $(x(\cdot), y(\cdot))$ is simple and AC.

Versions I and II are "isoperimetric problems" in the precise sense of §II.8. As explained in §II.10, such problems can be transformed into optimal control problems of the standard form, by adding extra state variables for the constraints. If we do that for our two versions, we arrive at

THE ISOPERIMETRIC PROBLEM, VERSION III

Given	a number L such that $L > 0$,
maximize	$\int_0^1 \Big(x(t)v(t) - y(t)u(t) \Big)\, dt$,
subject to	$\dot{x}(t) = u(t)$,
	$\dot{y}(t) = v(t)$,
	$\dot{z}(t) = \sqrt{u(t)^2 + v(t)^2}$,
	$x(1) = x(0)$, $y(1) = y(0)$, $z(0) = 0$, $z(1) = L$,
	the curve $(x(\cdot), y(\cdot))$ is simple,
	the functions $x(\cdot)$, $y(\cdot)$, $z(\cdot)$ are AC.

and

THE ISOPERIMETRIC PROBLEM, VERSION IV

Given	a number L such that $L > 0$,
minimize	$\int_0^1 \sqrt{u(t)^2 + v(t)^2}\, dt$,
subject to	$\dot{x}(t) = u(t)$,
	$\dot{y}(t) = v(t)$,
	$\dot{z}(t) = x(t)v(t) - y(t)u(t)$,
	$x(1) = x(0)$, $y(1) = y(0)$, $z(0) = 0$, $z(1) = A$,
	the curve $(x(\cdot), y(\cdot))$ is simple,
	the functions $x(\cdot)$, $y(\cdot)$, $z(\cdot)$ are AC.

As an alternative to Version III, we could reparametrize the plane curve $t \mapsto (x(t), y(t))$ by arc length. If we do this, then of course we can no longer take our curves to be defined on the interval $[0,1]$, and have to work on $[0,L]$ instead. The function $z(\cdot)$ becomes unnecessary, since now $z(t)$ must be equal to t.

In addition, it si clear that any translate of a curve ξ has the same length and encloses the same area as ξ. We can therefore impose the extra restriction that $x(0) = 0$ and $y(0) = 0$. We end up with

THE ISOPERIMETRIC PROBLEM, VERSION V

Given	a number L such that $L > 0$,
maximize	$\int_0^L \Big(x(t)v(t) - y(t)u(t) \Big)\, dt$,
subject to	$\dot{x}(t) = u(t)$,
	$\dot{y}(t) = v(t)$,
	$u(t)^2 + v(t)^2 = 1$,
	$x(L) = x(0) = 0$, $y(L) = y(0) = 0$,
	the curve $(x(\cdot), y(\cdot))$ is simple and AC.

This is, again, a two-dimensional problem. It is close to belonging to the class of calculus of variations problems, but does not quite qualify, because the curves $(x(\cdot), y(\cdot))$ have to satisfy the additional "velocity constraint" $\dot{x}^2 + \dot{y}^2 = 1$.

VI.2.1 Application of the maximum principle

From now on, L is a fixed positive real number. We let Γ_L be the set of all curves $\gamma : [0,L] \mapsto \mathbb{R}^2$ such that

1. γ is absolutely continuous,
2. $\|\dot{\gamma}(t)\| = 1$ for almost all $t \in [0,L]$,
3. $\gamma(L) = \gamma(0) = (0,0)$,
4. γ is simple (that is, the restriction of γ to the half-open interval $[0,L[$ is one-to-one),

For $\gamma \in \Gamma_L$, we define

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

where we are writing $\gamma(t) = (x(t), y(t))$.

So the isoperimetric problem (Version V) can be restated as follows:

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

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

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

and consider the problem

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

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

Proposition VI.1. *If (IP.V.a) has a solution γ_* such that $\gamma_* \in \Gamma_L$, then γ_* is a solution of (IP.V).* □

VI.2.2 Finding the solutions of (IP.V.a)

To find the solutions of (IP.V.a), we apply the necessary conditions for optimality given by the maximum principle. Suppose that

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

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

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

such that the Hamilton equations

$$\dot{m}_x(t) = -\frac{\partial H}{\partial x}(\gamma_*(t), \mu(t), \dot{\gamma}(t)), \tag{VI.4}$$

$$\dot{m}_y(t) = -\frac{\partial H}{\partial y}(\gamma_*(t), \mu(t), \dot{\gamma}(t)), \tag{VI.5}$$

and the Hamiltonian maximization condition

$$H(\gamma_*(t),\mu(t),\dot\gamma_*(t)) = \max\Big\{H(\gamma_*(t),\mu(t),w) : w\in\mathbb{R}^2, \|w\|\le 1\Big\} \qquad \text{(VI.6)}$$

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

$$\mathbb{R}^6 \ni (x,y,p_x,p_y,u,v)\mapsto H(x,y,p_x,p_y,u,v) \stackrel{\text{def}}{=} p_x u + p_y v + xv - yu\,. \qquad \text{(VI.7)}$$

In view of (XV.6), we can rewrite (XV.3), (XV.4), as

$$\begin{aligned} \dot m_x(t) &= -\dot y(t)\,, && \text{(VI.8)}\\ \dot m_y(t) &= \dot x(t)\,. && \text{(VI.9)} \end{aligned}$$

Define

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

Then

$$\begin{aligned} \dot\alpha(t) &= -2\dot y(t)\,, && \text{(VI.10)}\\ \dot\beta(t) &= 2\dot x(t)\,, && \text{(VI.11)} \end{aligned}$$

and

$$H(\gamma_*(t),\mu(t),u,v) = \alpha(t)u + \beta(t)v\,. \qquad \text{(VI.12)}$$

Therefore, if we define

$$\nu(t) = \sqrt{\alpha(t)^2+\beta(t)^2}\,, \qquad \text{(VI.13)}$$

(XV.5) says that

$$\left.\begin{aligned} \dot x(t) &= \tfrac{\alpha(t)}{\nu(t)}\,,\\ \dot y(t) &= \tfrac{\beta(t)}{\nu(t)}\,, \end{aligned}\right\} \quad \text{if} \quad \nu(t)\neq 0\,. \qquad \text{(VI.14)}$$

If $\nu(t)=0$ then (XV.5) allows the vector $\dot\gamma_*(t)$ to be an arbitrary member of $\{w\in\mathbb{R}^2 : \|w\|\le 1\}$. We can combine both cases into the following restatement of (XV.13) which is also valid when $\nu(t)=0$:

$$\nu(t)\dot x(t) = \alpha(t) \qquad \text{and} \qquad \nu(t)\dot y(t) = \beta(t)\,. \qquad \text{(VI.15)}$$

A simple calculation then shows that

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

Therefore *the function* ν *is constant.* Let $\bar\nu$ be the constant value of $\nu(t)$. If $\bar\nu = 0$ then $\alpha(t)\equiv\beta(t)\equiv 0$, so $\dot\alpha(t)\equiv\dot\beta(t)\equiv 0$, and then (XV.9) and (XV.10) imply that

$\dot{x}(t) \equiv \dot{y}(t) \equiv 0$, so γ_* is a constant curve and $A(\gamma_*) = 0$. It is clear that there do exist curves $\gamma \in \tilde{\Gamma}_L$ for which $A(\gamma) > 0$. Therefore the fact that $A(\gamma_*) = 0$ contradicts the assumption that γ_* is a solution of (IP"). So we can exclude the possibility that $\bar{v} = 0$.

From now on we assume that $\bar{v} > 0$. Then (XV.13) implies that $\dot{\gamma}$ is continuous, and

$$\|\dot{\gamma}(t)\| = 1 \qquad \text{for all } t \in [0,L]\,. \tag{VI.16}$$

Write

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

Then, using (XV.9), (XV.10), and (XV.13) we can conclude that

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

Therefore

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

This implies that

$$\eta(t) = c_1 \cos\left(\frac{2t}{\bar{v}}\right) + c_2 \sin\left(\frac{2t}{\bar{v}}\right)$$

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

$$c_1 = r\cos\left(\frac{2t_0}{\bar{v}}\right)\,, \qquad c_2 = r\sin\left(\frac{2t_0}{\bar{v}}\right)\,,$$

we find that

$$\eta(t) = r\cos\left(\frac{2(t-t_0)}{\bar{v}}\right)\,.$$

Then

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

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

$$\xi(t) = -\sin\left(\frac{2(t-t_0)}{\bar{v}}\right) \qquad \text{and} \qquad \eta(t) = \cos\left(\frac{2(t-t_0)}{\bar{v}}\right)\,. \tag{VI.17}$$

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

$$\xi(t) = -\sin\left(\frac{2\pi k(t-t_0)}{L}\right) \qquad \text{and} \qquad \eta(t) = \cos\left(\frac{2\pi k(t-t_0)}{L}\right)\,. \tag{VI.18}$$

We then get, by integration,

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

where

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

The above formulas define a family $\left\{\gamma^{t_0,k}\right\}_{t_0\in\mathbb{R},k\in\mathbb{Z}}$ of curves belonging to $\tilde{\Gamma}_L$, depending on the real parameter t_0 and the nonzero integer parameter k. If we let $\bar{\gamma}^{t_0,k}$ be the point with coordinates $\frac{-L}{2\pi k}\cos\left(\frac{2\pi k t_0}{L}\right)$, $\frac{L}{2\pi k}\sin\left(\frac{2\pi k t_0}{L}\right)$, then the curves $\gamma^{t_0,k}$ satisfy

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

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

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

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

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

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

Summarizing, we have shown

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

VI.2.3 Existence of area-maximizers

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

Proposition VI.3. $\tilde{\Gamma}_L$ *is a compact subset of* $C^0([0,L],\mathbb{R}^2)$.

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

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

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

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

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

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

We now prove the continuity of A.

Proposition VI.4. *The functional A is continuous on $\tilde{\Gamma}_L$.*

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

$$\lim_{j\to\infty}\Big(A(\gamma_j)-A(\gamma_\infty)\Big)=0\,. \tag{VI.19}$$

Write

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

Then

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

where

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

We must therefore show that

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

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

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

where

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

We will show that

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

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

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

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

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

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

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

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

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

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

It then follows that

$$|\tilde{I}_j| \le \varepsilon \quad \text{whenever} \quad j \ge J\,,$$

completing our proof. □

It then follows from Propositions 3 and 4 that

Proposition VI.5. *The maximization problem (IP”) has a solution.* □

VI.2.4 The final result

By combining Propositions XV.2 and XV.5, we get

Theorem VI.6. *The solutions of the modified isoperimteric problem (IP”) are the circles $\gamma^{t_0,1}$, for $t_0 \in \mathbb{R}$. Moreover, these circles are also the solutions of the original isoperimetric isoperimetric problem (IP’). In particular, the area enclosed by the area-maximizing circles is $\frac{L^2}{4\pi}$.* □

VI.2.5 How to make the solution of the isoperimetric problem rigorous

COMMENT NO. 20. This section will eventually be greatly expanded and possibly moved to a chapter on the issue of mathematical rigor. The expanded section will include short but completely self-contained proofs of the Jordan curve theorem (following the beautiful presentation in Dugundji's book), and of Green's formula for simple rectifiable loops.
At the very end of this draft we include our notes on the isoperimetric problem. What remains to be done is to incorporate them into the text and merge them with what we already have.
We will also include A. Hurwitz's 1902 proof based on Wirtinger's inequality. Furthermore, Wirtinger's inequality itself offers a wonderful example of how to apply optimal control or calculus of variations to prove inequalities. We will include the optimal control proof together with the standard Fourier series proof.

COMMENT NO. 21. As it stands, the section contains half-finished paragraphs, several of which correspond to an earlier approach that we are now planning to drop.

We use $\mathbb{S}^1$ to denote the unit circle $\{(x,y) \in \mathbb{R}^2 : x^2+y^2 = 1\}$. Alternatively, if we think of $\mathbb{R}^2$ as the complex plane $\mathbb{C}$, then $\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}$.

We fix a positive real number L, and let $\kappa_{_L}$ be the map

$$\mathbb{R} \ni \theta \mapsto \kappa_{_L}(\theta) \stackrel{\text{def}}{=} e^{\frac{2\pi i}{L}\theta} \in \mathbb{S}^1 .$$

A *loop* (or *closed curve*) in $\mathbb{R}^2$ is a continuous map $\xi : \mathbb{S}^1 \mapsto \mathbb{R}^2$. Then $C^0(\mathbb{S}^1, \mathbb{R}^2)$ is the set of all loops in $\mathbb{R}^2$. The *carrier*, or *locus*, of the loop ξ is the set $\{\xi(s) : s \in \mathbb{C}^1\}$. Clearly, $|\xi|$ is compact.

If $L > 1$, then $C^0(\mathbb{S}^1, \mathbb{R}^2)$ can be identified in an obvious way with $C^0_{per,L}(\mathbb{R}, \mathbb{R}^2)$, the set of all continuous maps $\gamma : \mathbb{R} \mapsto \mathbb{R}^2$ that are L-periodic, i.e., such that $\gamma(\theta + L) = \gamma(\theta)$ for all $\theta \in \mathbb{R}$. Naturally, the identification is made by assigning to each $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ the map $\xi \circ \kappa_{_L} \in C^0_{per,L}(\mathbb{R}, \mathbb{R}^2)$.

If $\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ is a loop, and $p \in \mathbb{C}\backslash|\xi|$, then the *winding number* of ξ about p in the integer $\nu(\xi, p)$ defined as follows: pick any positive L, and choose a continuous function $f : \mathbb{R} \mapsto \mathbb{C}$ such that $e^{f(t)} = (\xi \circ \kappa_{_L})(t) - p$ for all $t \in \mathbb{R}$. (The existence of f is trivial locally, because the map $\mathbb{C} \ni z \mapsto e^z \in \mathbb{C}\backslash\{0\}$ is surjective and has a continuous right inverse near every $s \in \mathbb{C}\backslash\{0\}$. Given this, the global existence of

f follows easily by patching up local versions.) Then the number $\frac{f(t+L)-f(t)}{2\pi i}$ necessarily has to be an integer, because $e^{f(t+L)} = (\xi \circ \kappa_L)(t+L) - p = (\xi \circ \kappa_L)(t) - p = e^{f(t)}$. Clearly, this integer depends continuously on t. Furthermore, if $f_1, f_2 : \mathbb{R} \mapsto \mathbb{C}$ are any two continuous such that $e^{f_1(t)} = e^{f_2(t)} = (\xi \circ \kappa_L)(t) - p$ for all $t \in \mathbb{R}$ must differ by a constant, since $e^{f_1(t)-f_2(t)} = 1$, so that $f_1(t) - f_2(t) = 2\pi i k(t)$, where $k(t)$ is an integer that depends continuously on t. So $\frac{f(t+L)-f(t)}{2\pi i}$ is in fact independent of t and f, and it is easy to see that it is also independent of L. We define $\nu(\xi, p) \stackrel{\text{def}}{=} \frac{f(t+L)-f(t)}{2\pi i}$ for any $t \in \mathbb{R}$, where L and f are as above.

A *reparametrization* of a loop $\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ is a loop of the form $\xi \circ \psi$, where $\psi : \mathbb{S}^1 \mapsto \mathbb{S}^1$ is an orientation-preserving homeomorphism. ("Orientation-preserving" means, for example, that if $\bar{t}$ is such that $\psi(e^{2\pi i \bar{t}}) = 1$, then the imaginary part of $\psi(e^{2\pi i(\bar{t}+\varepsilon)})$ is positive for all sufficiently small positive ε.) It is clear that reparametrization preserves carriers and winding numbers. Precisely,

(*) *if η is a reparametrization of ξ, then $|\eta| = |\xi|$ and $\nu(\eta, p) = \nu(\xi, p)$ for all $p \in \mathbb{C} \backslash |\xi|$.*

(Proof: the fact that $|\eta| = |\xi|$ is trivial. Let $f : \mathbb{R} \mapsto \mathbb{C}$ be continuous and such that $e^{f(t)} = \eta(e^{2\pi i t})$ for all t. Let $\psi : \mathbb{S}^1 \mapsto \mathbb{S}^1$ be an orientation-preserving homeomorphism such that $\eta = \xi \circ \psi$. Then there exist $a \in \mathbb{R}$ and a homeomorphism $\tilde{\psi} : [0,1] \mapsto [a, a+1]$ such that

$$e^{2\pi i \tilde{\psi}(t)} = \psi(e^{2\pi i t}) \tag{VI.20}$$

for all $t \in [0,1]$, $\tilde{\psi}(0) = a$ and $\tilde{\psi}(1) = a+1$. Extend $\tilde{\psi}$ to the whole real line by defining $\tilde{\psi}(k+t) = k + \psi(t)$ for $t \in [0,1]$ and $k \in \mathbb{Z}$. Let $g = f \circ \tilde{\psi}$. Then (XV.34) holds for all $t \in \mathbb{R}$. Moreover,

$$e^{g(t)} = e^{f(\tilde{\psi}(t))} = \eta(e^{2\pi i \tilde{\psi}(t)}) = \eta\Big(\psi(e^{2\pi i t})\Big) = (\eta \circ \psi)(e^{2\pi i t}) = \xi(e^{2\pi i t}).$$

Therefore $\nu(\xi, p) = \frac{g(1)-g(0)}{2\pi i} = \frac{f(\tilde{\psi}(1))-f(\tilde{\psi}(0))}{2\pi i} = \frac{f(a+1)-f(a)}{2\pi i} = \nu(\eta, p)$.)

A loop $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ is *absolutely continuous* if for some (and hence every) L the function $\gamma = \xi \circ \kappa_L$ is locally absolutely continuous. If ξ is absolutely continuous and $\gamma = \xi \circ \kappa_L$, then we can choose f by letting

$$f(t) = \bar{f} + \int_0^t \frac{\dot{\gamma}(s)}{\gamma(s) - p}\, ds,$$

where $\bar{f} \in \mathbb{C}$ is such that $e^{\bar{f}} = \gamma(0) - p$. Indeed, if we define $g : \mathbb{R} \mapsto \mathbb{C}$ by letting $g(t) = (\gamma(t) - p).e^{-f(t)}$, we have

$$\begin{aligned} \dot{g}(t) &= -(\gamma(t) - p).\dot{f}(t).e^{-f(t)} + \dot{\gamma}(t).e^{-f(t)} \\ &= -(\gamma(t) - p).\frac{\dot{\gamma}(t)}{\gamma(t) - p}.e^{-f(t)} + \dot{\gamma}(t).e^{-f(t)} \\ &= -\dot{\gamma}(t).e^{-f(t)} + \dot{\gamma}(t).e^{-f(t)} = 0. \end{aligned}$$

Therefore $g(t) \equiv$ constant. Since $g(0) = 1$, we have $g(t) \equiv 1$ for all t, so $e^{f(t)} \equiv \gamma(t) - p$. We then get the formula

$$\nu(\xi, p) = \frac{1}{2\pi i} \int_0^L \frac{\dot{\gamma}(s)}{\gamma(s) - p}\, ds \text{ whenever } p \notin |\xi|.$$

A loop $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ is said to be *simple* if it is a one-to-one map, i.e., if $\xi(z_1) \neq \xi(z_2)$ whenever $z_1, z_2 \in \mathbb{S}^1$ and $z_1 \neq z_2$.

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

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

The *vertices* of a triangle

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

The *Jordan curve theorem* says that if ξ is a simple loop then $\mathbb{R}^2 \backslash |\xi|$ has exactly two connected components. In other words, $\mathbb{R}^2 \backslash |\xi|$ is the union of two disjoint connected open subsets Ω_1, Ω_2 of $\mathbb{R}^2$. If $R \in \mathbb{R}$ is such that $R > 0$ and $\|\xi(s)\| \leq R$ for all $s \in \mathbb{S}^!$, then the set $S = \{q \in \mathbb{R}^2 : \|q\| > R\}$ is open, connected, and disjoint from $|\xi|$. Therefore S is contained in one of the components Ω_1, Ω_2. The other component must then be contained in $\{q \in \mathbb{R}^2 : \|q\| \leq R\}$. So one of the components is unbounded and the other one is bounded. We use $\mathrm{Ins}(\xi)$, $\mathrm{Out}(\xi)$ (the "inside" and the "outside" of ξ) to denote, respectively, the bounded and the unbounded connected component of $\mathbb{R}^2\setminus|\xi|$. We refer to $\mathrm{Ins}(\xi)$ as *the region enclosed by the simple loop* ξ. We use $A(\xi)$ to denote the area of $\mathrm{Ins}(\xi)$.

A loop $\xi \in C^0(\mathbb{S}^1,\mathbb{R}^2)$ is said to be *rectifiable* if the map $\xi \circ \kappa_L$ is locally absolutely continuous for some positive L, in which case it follows on easily that $\xi \circ \kappa_L$ is locally absolutely continuous for every positive L.

Lemma VI.7. *Let $\xi : \mathbb{S}^1 \mapsto \mathbb{R}^2 = \mathbb{C}$ be a simple loop. Then the funcion $\mathbb{C}\setminus|\xi| \ni p \mapsto \nu(\xi,p) \in \mathbb{Z}$ is constant on* $\mathrm{Out}(\xi)$ *and on* $\mathrm{Ins}(\xi)$*, and has the value* 0 *on* $\mathrm{Out}(\xi)$ *and the value* 1 *or* -1 *on* $\mathrm{Ins}(\xi)$.

PROOF. The winding number $\nu(\xi,p)$ clearly depends continuously on p as long as p varies on $\mathbb{C}\setminus|\xi|$. (Let $p \in \mathbb{C}\setminus|\xi|$, and let $\{p_j\}_{j=1}^{\infty}$ be a sequence in $\mathbb{C}\setminus|\xi|$ converging to p. Let $f : \mathbb{R} \mapsto \mathbb{C}$ be a continuous map with the property that $e^{f(t)} = \xi(e^{2\pi it}) - p \stackrel{\mathrm{def}}{=} h(t)$ for all $t \in \mathbb{R}$. Let $h_j(t) = \xi(e^{2\pi it}) - p_j$. Then $h_j \to h$ as $j \to \infty$, uniformly on compact sets, so $h_j \to h$ uniformly, since the h_j are 1-periodic. Since $h(t) \neq 0$ for all t, the functions $g_j \stackrel{\mathrm{def}}{=} \frac{h_j}{h}$ converge uniformly to the constant 1 as $j \to \infty$. Let D be the open disc $\{z \in \mathbb{C} : \|z-1\| < 1\}$. Let $\lambda : D \mapsto \mathbb{C}$ be the standard branch of the natural logarithm, so that $\lambda(1+s) = \sum_{k=1}^{\infty}(-1)^{k+1}\frac{s^k}{k}$ whenever $|s| < 1$. Then $\lambda(1) = 0$, λ is continuous, and $e^{\lambda(z)} = z$ whenever $z \in D$. Fix $J \in \mathbb{N}$ such that $g_j(t) \in D$ whenever $t \in \mathbb{R}$ and $j \geq J$. Define

$$f_j(t) = f(t) + \lambda\Big(g_j(t)\Big) \text{ for } t \in \mathbb{R},\ j \geq J.$$

Then

$$e^{f_j(t)} = e^{f(t)}.e^{\lambda(g_j(t))} = h(t).g_j(t) = h_j(t) = \xi(e^{2\pi it}) - p_j.$$

Therefore

$$\begin{aligned}
\nu(\xi,p_j) &= \frac{1}{2\pi i}(f_j(1) - f_j(0)) \\
&= \frac{1}{2\pi i}\Big(f(1) - f(0) + \lambda(g_j(1)) - \lambda(g_j(0))\Big) \\
&= \frac{1}{2\pi i}\Big(f(1) - f(0)\Big) \\
&= \nu(\xi,p)
\end{aligned}$$

whenever $j \geq J$, since $\lambda(g_j(1)) = \lambda(g_j(0))$, because the function g_j is 1-periodic.) Therefore $\mathbb{C}\setminus|\xi| \ni p \mapsto \nu(\xi,p) \in \mathbb{Z}$ is constant on $\mathrm{Out}(\xi)$ and on $\mathrm{Ins}(\xi)$, as stated.

Let $R \in \mathbb{R}$ be such that $R > 0$ and $|\xi(e^{2\pi it})| \leq R$ for all $t \in R$. Let p be such that $|p| > R$. Then the complex number $1 - \frac{\xi(e^{2\pi it})}{p}$ belongs to D for every $t \in \mathbb{R}$, so we can define a continuous function $\sigma : \mathbb{R} \mapsto \mathbb{C}$ by letting

$$\sigma(t) = \lambda\Big(1 - \frac{\xi(e^{2\pi it})}{p}\Big) \text{ for } t \in \mathbb{R}.$$

Let $\bar{f} \in \mathbb{C}$ be such that $e^{\bar{f}} = -p$. Define $f(t) = \bar{f} + \sigma(t)$. Then f is continuous and

$$e^{f(t)} = e^{\bar{f}}.e^{\sigma(t)} = (-p).\Big(1 - \frac{\xi(e^{2\pi it})}{p}\Big) = \xi(e^{2\pi it}) - p\,.$$

Therefore $\nu(\xi,p) = \frac{1}{2\pi i}(f(1) - f(0)) = 0$, since f is clearly 1-periodic. So $\nu(\xi,p) = 0$ for one point p of $\mathrm{Out}(\xi)$, and then $\nu(\xi,p) = 0$ for all the points p of $\mathrm{Out}(\xi)$.

Now choose a point $p \in \mathrm{Ins}(\xi)$. After a translation of the coordinate axes of the plane, we may assume that $p = 0$. Let $\bar{q}$ be a point of $|\xi|$ closest to 0. Then $\bar{q} \neq 0$ and, after a rotation of the coordinate axes, we may assume that $\bar{q}$ is a positive real number $\bar{r}$. Fix a $\bar{t} \in \mathbb{R}$ such that, $\xi(e^{2\pi i\bar{t}}) = \bar{r}$. Fix a number $\bar{\varepsilon}$ such that $0 < \bar{\varepsilon} < \frac{1}{2}$, and two continuous functions $\rho : [\bar{t}-\varepsilon, \bar{t}+\varepsilon] \mapsto]0,\infty[$, $\varphi : [\bar{t}-\varepsilon, \bar{t}+\varepsilon] \mapsto \mathbb{R}$, such that $\xi(e^{2\pi it}) = \rho(t)e^{2\pi i\varphi(t)}$ whenever $t \in [\bar{t}-\varepsilon, \bar{t}+\varepsilon]$, and $\varphi(\bar{t}) = 0$. Then $\rho(t) \geq \bar{r}$ for all $\in [\bar{t}-\varepsilon, \bar{t}+\varepsilon]$.

Then, if $\varepsilon > 0$, it cannot be the case that $\varphi(t) = 0$ for all $t \in [\bar{t}-\varepsilon, \bar{t}+\varepsilon]$. (Indeed, if $\varphi(t) = 0$ whenever $t \in [\bar{t}-\varepsilon, \bar{t}+\varepsilon]$, then $\xi(e^{2\pi it}) = \rho(t) \in \mathbb{R}$ for $t \in [\bar{t}-\varepsilon, \bar{t}+\varepsilon]$. Then the function $[\bar{t}-\varepsilon, \bar{t}+\varepsilon] \ni t \mapsto \xi(e^{2\pi it})$ would be real-valued and would have a local minimum at $\bar{t}$, contradicting the fact that ξ is one-to-one on some neighborhood of $\bar{t}$.) Therefore there exists a sequence $\{\tau_j\}_{j=1}^{\infty}$ in $\mathbb{R}$ such that $\tau_j \to \bar{t}$ as $j \to \infty$ and $\varphi(\tau_j) \neq 0$ for all j. If we let L_j be the closed half-line $\{re^{2\pi i\varphi(\tau_j)} : r \geq 0\}$, then $L_j \cap |\xi|$ is a compact set. Furthermore, $L_j \cap |\xi|$ is nonempty, because $\rho(t_j)e^{2\pi i\varphi(t_j)} \in L_j \cap |\xi|$. So $L_j \cap |\xi|$ has a point q_j closest to 0. Write $q_j = \xi(e^{2\pi it_j})$, $t_j \in [\bar{t}-\frac{1}{2}, \bar{t}+\frac{1}{2}]$. Also, let $q_j = r_je^{2\pi i\varphi(\tau_j)}$, so $0 < r_j \leq \rho(\tau_j)$.

We claim that $\lim_{j\to\infty} t_j = \bar{t}$. Indeed, if this was not so there would exist a subsequence $\{t_{j(\ell)}\}_{\ell=1}^{\infty}$ of $\{t_j\}_{j=1}^{\infty}$ that converges to a $\tilde{t} \in [\bar{t}-\frac{1}{2}, \bar{t}+\frac{1}{2}]\backslash\{\bar{t}\}$. Let $\tilde{q} = \xi(e^{2\pi i\tilde{t}})$. Then $\tilde{q} = \lim_{\ell\to\infty} p_{j(\ell)}$. Since $e^{2\pi i\varphi(\tau_j)} \to 1$ as $j \to \infty$, we conclude that $\tilde{q}$ is a positive real number $\tilde{r}$, and $\tilde{r} = \lim_{\ell\to\infty} r_{j(\ell)}$. Since $0 < r_{j(\ell)} \leq \rho(\tau_{j(\ell)})$, we have $0 < \tilde{r} \leq \rho(\bar{t}) = \bar{r}$. Since $\bar{r}$ is a point of $|\xi$ closest to 0, we must have $\tilde{r} = \bar{r}$, and then $\tilde{q} = \bar{q}$. Since $|\tilde{t} - \bar{t}| \leq \frac{1}{2}$, the fact that $\xi(e^{2\pi i\tilde{t}}) = \xi(e^{2\pi i\bar{t}})$ implies $\tilde{t} = \bar{t}$, which is a contradiction, proving our claim.

By passing to a subsequence we may assume that $|t_j - \bar{t}| < \bar{\varepsilon}$ for all j, and either $t_j < \bar{t}$ for all j or $t_j > \bar{t}$ for all j. Then $q_j = \rho(t_j)e^{2\pi i\varphi(t_j)}$—so $r_j = \rho(t_j)$ and $\varphi(t_j) = \varphi(\tau_j)$—and $re^{2\pi i\varphi(t_j)} \notin |\xi|$ whenever $0 \leq r < r_j$. By passing to a subsequence, we assume that either $\varphi(t_j) < 0$ for all j or $\varphi(t_j) > 0$ for all j.

Let us consider the case when $t_j < \bar{t}$ and $\varphi(t_j) < 0$ for all j. (The other three cases are similar.) Write $\varphi_j = -\varphi(t_j)$, so $\varphi_j > 0$. Fix a number δ such that $0 < \delta < \bar{r} = \mathrm{dist}(0, |\xi|)$. For each j, let S_j be the sector $\{re^{2\pi i\theta} : r \geq 0 \wedge -\varphi_j \leq \theta \leq 0\}$. Let Δ be the closed disc $\{z \in \mathbb{C} : |z| \leq \delta\}$. Let $T_j = S_j \cup \Delta$. Let $\eta_j : [0,3] \mapsto \mathbb{C}$ be the curve given by

$$\eta_j(u) = \begin{cases} \Big((1-u)r_j + u\delta\Big)e^{-2\pi i\varphi_j} & \text{if} \quad 0 \leq u \leq 1\,, \\ \delta e^{-2\pi i((2-u)\varphi_j + 1 - u)} & \text{if} \quad 1 \leq u \leq 2\,, \\ (u-2)\bar{r} + (3-u)\delta & \text{if} \quad 2 \leq u \leq 3\,. \end{cases}$$

Then, as u goes from 0 to 3, $\eta_j(u)$ starts at the point $\xi(e^{2\pi i t_j}) = r_j e^{-2\pi i\varphi_j}$, moves along the straight-line segment M_j^1 from $r_j e^{-2\pi i\varphi_j}$ to $\delta e^{-2\pi i\varphi_j}$, then follows an arc M_j^2 of length $\delta.(2\pi - \varphi(t_j))$ of the circle $\{z : |z| = \delta\}$, clockwise from $\delta e^{-2\pi i\varphi_j}$ to δ, and finally moves along the straight-line segment M_3 from δ to $\bar{r}$, ending at $\bar{r} = \xi(e^{2\pi i\bar{t}})$. It is clear that η_j is a one-to-one map. Furthermore, $\eta_j(u) \neq |\xi|$ whenever $0 < u < 3$.

Define $\hat{\zeta}_j : [\bar{t} - \frac{1}{2}, \bar{t} + \frac{1}{2}] \mapsto \mathbb{C}$ by letting

$$\hat{\zeta}_j(t) = \begin{cases} \xi(e^{2\pi i t}) & \text{if} \quad t \in [\bar{t} - \frac{1}{2}, \bar{t} + \frac{1}{2}] \backslash [\bar{t} - \bar{\varepsilon}, \bar{t} + \bar{\varepsilon}]\,, \\ \eta_j\Big(\frac{3t - 3\bar{t} + 3\bar{\varepsilon}}{2\bar{\varepsilon}}\Big) & \text{if} \quad t \in [\bar{t} - \bar{\varepsilon}, \bar{t} + \bar{\varepsilon}]\,. \end{cases}$$

Then $\hat{\zeta}_j$ is continuous, and $\hat{\zeta}_j(\bar{t} + \frac{1}{2}) = \hat{\zeta}_j(\bar{t} + \frac{1}{2})$, so we can extend $\hat{\zeta}_j$ to a 1-periodic map ζ_j from $\mathbb{R}$ to $\mathbb{C}$. Write $\zeta_j(t) = \omega_j(e^{2\pi i t})$ for $t \in \mathbb{R}$. Then ω is a simple loop.

Let $\Omega_j = \mathrm{Ins}(\xi) \backslash T_j$. We claim that Then Ω is open and bounded.

$$\Omega_j = \mathrm{Ins}(\omega_j)\,. \tag{VI.21}$$

To prove (VI.21, we first observe that Ω_j is clearly open and bounded. Let us show that Ω_j is connected. Let p_0, p_1 be points of Ω_j. Then there exists a continuous path $\alpha : [0,1] \mapsto \mathrm{Ins}(\xi)$ such that $\alpha(0) = p_0$ and $\alpha(1) = p_1$, since $\mathrm{Ins}(\xi)$ is path-connected, because it is open and connected.

□

We use $C^1(\mathbb{S}^1, \mathbb{R}^2)$ to denote the the set of all loops $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ that are of class C^1. Equivalently, a loop $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$ belongs to $C^1(\mathbb{S}^1, \mathbb{R}^2)$ if and only if $\xi \circ \kappa$ is continuously differentiable. Clearly, $C^1(\mathbb{S}^1, \mathbb{R}^2)$ can be identified, via the map $\xi \mapsto \xi \circ \kappa$, with $C^1_{per,2\pi}(\mathbb{R}, \mathbb{R}^2)$, the set of all continuously differentiable 2π-periodic maps $\gamma : \mathbb{R} \mapsto \mathbb{R}^2$. If $\xi \in C^0(\mathbb{S}^1, \mathbb{R}^2)$, we use $\dot{\xi}$ to denote the unique map $\zeta : \mathbb{S}^1 \mapsto \mathbb{R}^2$ such that $\zeta \circ \kappa$ is the derivative of the function $\xi \circ \kappa$.

A loop $\xi \in C^1(\mathbb{S}^1, \mathbb{R}^2)$ is *parametrized by arc-length* (abbr. PAL) if $\|\dot{\xi}(z)\| = 1$ for all $z \in \mathbb{S}^1$.

VI.3 Example III: the catenary

Assume that we are given a flexible rope or chain having length Λ and a constant mass density per unit length. The rope is to hang, its endpoints being held fixed at points A and B. We want to determine the resulting shape, from the fact that *the center of mass should be as low as possible*.

Let P be a vertical plane containing A and B, and let us choose Cartesian coordinates x, y on P so that the x axis is horizontal, the y axis is vertical and points upwards, and A, B have coordinates (a,α), (b,β), respectively. Let us assume that $a < b$. (This amounts to excluding the possibility that A and B may lie on the same vertical line. The solution in this case is easy to find directly.) Let us make the *Ansatz* that the shape of the hanging rope is given by a function $y = y(x)$ belonging to the space $W^{1,1}([a,b],\mathbb{R})$ of all absolutely continuous functions $[a,b] \ni x \to y(x) \in \mathbb{R}$.

Then the coordinates x_c, y_c of the center of mass are

$$\begin{aligned} x_c &= \frac{1}{\Lambda}\int_a^b x\,ds = \frac{1}{\Lambda}\int_a^b x\sqrt{1+y'(x)^2}\,dx\,,\\ y_c &= \frac{1}{\Lambda}\int_a^b y\,ds = \frac{1}{\Lambda}\int_a^b y(x)\sqrt{1+y'(x)^2}\,dx\,, \end{aligned}$$

where ds is the element of arc length. Since Λ is fixed, our goal is to minimize the integral

$$I = \int_a^b y(x)\sqrt{1+y'(x)^2}\,dx\,,$$

in the set $\mathscr{Y}$ of all curves $y(\cdot) \in W^{1,1}([a,b],\mathbb{R})$ that satisfy the constraints

$$\begin{aligned} y(a) &= \alpha\,,\\ y(b) &= \beta\,,\\ \int_a^b \sqrt{1+y'(x)^2}\,dx &= \Lambda\,. \end{aligned}$$

We observe that the distance Δ from A to B is given by

$$\Delta = \sqrt{(b-a)^2+(\beta-\alpha)^2}\,,$$

so the set $\mathscr{Y}$ is nonempty if and only if

$$\Lambda \geq \Delta = \sqrt{(b-a)^2+(\beta-\alpha)^2}\,, \tag{VI.22}$$

Let us introduce a new function s, the arc-length, defined by

$$s(x) = \int_a^x \sqrt{1+y'(r)^2}\,dr\,.$$

Then our curves $y(\cdot) \in \mathscr{Y}$ are in an obvious one-to-one correspondence with the members of the set Ξ of all absolutely continuous curves

$$[a,b] \ni x \to \xi(x) = (y(x), s(x)) \in \mathbb{R}^2$$

in y, s space that satisfy

$$\xi(a) = (\alpha, 0)\,,\ \ \xi(b) = (\beta, \Lambda)\,,$$

as well as

$$\xi'(x) \in U \quad \text{for a.e.} \quad x \in [a,b]\,,$$

where U is the subset of $\mathbb{R}^2$ given by

$$U = \left\{ (u,v) \in \mathbb{R}^2 : v = \sqrt{1+u^2} \right\}.$$

COMMENT NO. 22. Put section number in GNU below

We thus have a calculus of variations problem of the kind considered in §GNU, whose Lagrangian $\mathbb{R}^2 \times U \ni (y,s,u,v) \to L(y,s,u,v)$ is given by

$$L(y,s,u,v) = yv\,.$$

Assume that $y_* \in \mathscr{Y}$ is a solution, and let ξ_* be the corresponding curve in Ξ, so $\xi_*(x) = (y_*(x), s_*(x))$, where $s_*(x) = \int_0^x \sqrt{1+y_*'(r)^2}\,dr$. Let $u_* = y_*'$, $v_* = s_*'$, so $v_*(x) = \sqrt{1+u_*(x)^2}$ for a.e. x.

COMMENT NO. 23. Put conditions name in GNU below

The necessary conditions GNU then assert that there exist an absolutely continuous vector-valued function $[a,b] \ni x \to \pi(x) = (\pi_y(x), \pi_z(x)) \in \mathbb{R}^2$ and a nonnegative constant π_0 such that

$$\pi_y'(x) = \frac{\partial L}{\partial y}(\xi_*(x), \xi_*'(x)) = \pi_0 v_*(x) = \pi_0\sqrt{1+u_*(x)^2}\,, \qquad \text{(VI.23)}$$

$$\pi_s'(x) = \frac{\partial L}{\partial s}(\xi_*(x), \xi_*'(x)) = 0 \qquad \text{(VI.24)}$$

for a.e. $x \in [a,b]$, $(\pi_y(x), \pi_s(x), \pi_0) \neq (0,0,0)$ for all x, and

$$\begin{aligned}
&\pi_y(x)u_*(x) + \pi_s(x)v_*(x) - \pi_0 y_*(x)v_*(x)\\
&\quad = \pi(x)\cdot\xi_*'(x) - \pi_0 L(\xi_*(x), \xi_*'(x))\\
&\quad = \max\left\{ \pi_y(x)u + \pi_s(x)v - \pi_0 L(\xi_*(x), w) : w \in U \right\}\\
&\quad = \max\left\{ \pi_y(x)u + \pi_s(x)v - \pi_0 y_*(x)v : (u,v) \in U \right\}
\end{aligned}$$

for almost all $x \in [a,b]$. It then follows that

$$\begin{aligned}\pi_y(x)u_*(x) + (\pi_s(x) - \pi_0 y_*(x))\sqrt{1+u_*(x)^2} \\ = \max\left\{\pi_y(x)u + (\pi_s(x) - \pi_0 y_*(x))\sqrt{1+u^2} : u \in \mathbb{R}\right\}\end{aligned} \tag{VI.25}$$

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

We now extract from the above conditions more detailed information about the solutions of the optimization problem. To begin with, observe that (VI.24) implies in particular that the number $\pi_s(x)$ is in fact equal to a constant $\bar{\pi}_s$.

Assume first that $\pi_0 = 0$. Then (VI.23) and (VI.24) imply that the vector $\pi(x)$ is equal to a constant $\bar{\pi} = (\bar{\pi}_y, \bar{\pi}_s)$, which is not equal to $(0,0)$. The possibility that $\bar{\pi}_s = 0$ is excluded, because the maximization condition (VI.25) would imply, if $\bar{\pi}_s = 0$, that the function $\mathbb{R} \ni u \to \bar{\pi}_y u$ has a maximum, contradicting the fact that $\bar{\pi}_y \neq 0$. So $\bar{\pi}_s \neq 0$. Then the u-derivative of the right-hand side of (VI.25) is $\bar{\pi}_y + \frac{\bar{\pi}_s u}{\sqrt{1+u^2}}$, which can only vanish if

$$\frac{u}{\sqrt{1+u^2}} = -\frac{\bar{\pi}_y}{\bar{\pi}_s}\,. \tag{VI.26}$$

The function $\mathbb{R} \ni u \mapsto \frac{u}{\sqrt{1+u^2}}$ is strictly increasing, so the solution u of (VI.26) is unique. Hence u_* is a constant. Thus y_* is a linear function.

We now study the case when $\pi_0 \neq 0$. Then we may assume $\pi_0 = 1$. Equation (VI.23) then says that

$$\pi_y'(x) = \sqrt{1+u_*(x)^2} = v_*(x) = s_*'(x)\,.$$

Therefore $\pi_y(x) = \bar{\pi}_y + s_*(x)$ for some constant $\bar{\pi}_y$. Let

$$Y(x) = y_*(x) - \bar{\pi}_s\,.$$

Then (VI.25) says that $u = u_*(x)$ maximizes the function

$$u \to (\bar{\pi}_y + s_*(x))u - Y(x)\sqrt{1+u^2}\,, \tag{VI.27}$$

from which it follows that

$$\frac{u_*(x)}{\sqrt{1+u_*(x)^2}} = -\frac{\bar{\pi}_y + s_*(x)}{Y(x)}\,,$$

as long as $Y(x) \neq 0$. So

$$u_*(x)Y(x) = (\bar{\pi}_y + s_*(x))\sqrt{1+u_*(x)^2}\,,$$

and then

$$Y(x)Y'(x) = S(x)S'(x)\,,$$

where $S(x) = \bar{\pi}_y + s_*(x)$. It then follows that

$$Y(x)^2 = S(x)^2 + c\,, \tag{VI.28}$$

where c is a constant. So we have shown that the function $Y^2 - S^2$ is constant on every subinterval of $[0,1]$ where Y has no zeros.

If $x_0 \in [a,b]$ is such that $Y(x_0) = 0$, then $S(x_0)$ must vanish too, because the function $\mathbb{R} \ni u \to S(x_0)u$ does not have a maximum if $S(x_0) \neq 0$. It follows that $s_*(x_0) = -\bar{\pi}_y$. Since s_* is strictly increasing, the point x_0, if it exists, is unique. Moreover, (VI.28) holds on both sides of x_0, with possibly different constants c. If we then let x go to x_0 from the left and from the right, we see that both constants are equal.

So we have shown that (VI.28) holds throughout $[a,b]$ for some constant c, even if Y has zeros. On the other hand, $Y(x)$ cannot be negative, because if it was then the function (VI.27) would not have a maximum. So $Y(x) \geq 0$ for all x, and then

$$Y(x) = \sqrt{S(x)^2 + c}\,,$$

where, as usual, $\sqrt{r}$ means "the nonnegative square root of r." In addition, the inequality $|S(x)| < Y(x)$ must hold for almost all x, because if $|S(x)| \geq Y(x)$ the function (VI.27) does not have a maximum. Therefore $c > 0$, and we can write $c = \kappa^2$, and then

$$Y(x) = \sqrt{S(x)^2 + \kappa^2}\,, \quad \kappa > 0\,.$$

Define a function $\tau : [a,b] \to \mathbb{R}$ by letting

$$S(x) = \kappa.\sinh(\tau(x))\,.$$

Then

$$\begin{aligned} Y(x) &= \kappa.\cosh(\tau(x))\,, \\ Y'(x) &= \kappa.\tau'(x).\sinh(\tau(x))\,, \\ S'(x) &= \kappa.\tau'(x).\cosh(\tau(x))\,. \end{aligned}$$

The equations $Y' = u_*$ and $S' = \sqrt{1+u_*^2}$ then imply $S' = \sqrt{1+(Y')^2}$, from which it follows that $(S')^2 = 1 + (Y')^2$, and then

$$\kappa^2\tau'(x)^2\cosh^2(\tau(x)) = 1 + \kappa^2\tau'(x)^2\sinh^2(\tau(x))\,. \tag{VI.29}$$

Since $\cosh^2 r = 1 + \sinh^2 r$ for all r, (VI.29) implies

$$\kappa^2 \tau'(x)^2 = 1\,.$$

Moreover, S is strictly increasing, because s_* is, so $\tau'(x) > 0$. This implies that

$$\tau'(x) \equiv \frac{1}{\kappa}\,,$$

so

$$\tau(x) = \frac{x - x_0}{\kappa}$$

for some constant x_0, and then

$$y_*(x) = \gamma + \kappa \cosh\left(\frac{x - x_0}{\kappa}\right), \tag{VI.30}$$

where the constant γ is equal to $\bar{\pi}_y$.

To determine the numbers γ, κ, x_0 in terms of the data a, b, α, β, Λ, we first use (VI.30) to conclude that

$$\begin{aligned}
\alpha &= y_*(a) = \gamma + \kappa \cosh \mathscr{A}\,, \\
\beta &= y_*(b) = \gamma + \kappa \cosh \mathscr{B}\,, \\
\frac{\beta - \alpha}{\kappa} &= \cosh \mathscr{B} - \cosh \mathscr{A} \\
&= 2 \sinh \frac{\mathscr{B} + \mathscr{A}}{2} \sinh \frac{\mathscr{B} - \mathscr{A}}{2} \\
&= 2 \sinh \mathscr{C} \sinh \mathscr{D}\,,
\end{aligned}$$

where

$$\mathscr{A} = \frac{a - x_0}{\kappa}\,, \qquad \mathscr{B} = \frac{b - x_0}{\kappa}\,, \qquad \mathscr{C} = \frac{a + b - 2x_0}{2\kappa}\,, \qquad \mathscr{D} = \frac{b - a}{2\kappa}\,.$$

On the other hand, $y'_*(x) = \sinh \frac{x - x_0}{\kappa}$, so $\sqrt{1 + y'_*(x)^2} = \cosh \frac{x - x_0}{\kappa}$, and then

$$\Lambda = \int_a^b \sqrt{1 + y'_*(x)^2}\,dx = \kappa(\sinh \mathscr{B} - \sinh \dashv)\,,$$

so

$$\begin{aligned}
\frac{\Lambda}{\kappa} &= \sinh \mathscr{B} - \sinh \mathscr{A} \\
&= 2 \sinh \frac{\mathscr{B} - \mathscr{A}}{2} \cosh \frac{\mathscr{B} + \mathscr{A}}{2} \\
&= 2 \sinh \mathscr{D} \cosh \mathscr{C}\,.
\end{aligned}$$

Then

$$\frac{\Lambda^2}{\kappa^2} = 4\sinh^2\mathscr{D}\cosh^2\mathscr{C}\,,$$

and

$$\frac{(\beta-\alpha)^2}{\kappa^2} = 4\sinh^2\mathscr{C}\sinh^2\mathscr{D}\,,$$

so

$$\frac{\Lambda^2-(\beta-\alpha)^2}{\kappa^2} = 4\sinh^2\mathscr{D}\,,$$

and then

$$2\kappa\sinh\mathscr{D} = \sqrt{\Lambda^2-(\beta-\alpha)^2}\,,$$

so

$$\frac{\sinh\mathscr{D}}{\mathscr{D}} = \frac{\sqrt{\Lambda^2-(\beta-\alpha)^2}}{b-a}\,. \qquad \text{(VI.31)}$$

The function $[0,+\infty[\ni u \to \frac{\sinh u}{u} \in \mathbb{R}$ is strictly increasing, and takes the value 1 when $u=0$. If we assume that $\Lambda > \Delta$, then the right-hand side of (VI.31) is >1, so there exists a unique positive solution $\mathscr{D}$ of (VI.31). Then κ is determined by

$$\kappa = \frac{b-a}{2\mathscr{D}}\,. \qquad \text{(VI.32)}$$

In addition,

$$\tanh\mathscr{C} = \frac{\beta-\alpha}{\Lambda}\,, \qquad \text{(VI.33)}$$

and the assumption that $\Lambda > \Delta$ implies that the right-hand side of (VI.33) belongs to the interval $]-1,1[$, so there exists a unique real solution $\mathscr{C}$ of (VI.33). Then x_0 is determined by

$$x_0 = \frac{a+b}{2} - \kappa\mathscr{C} = \frac{a+b}{2} - \frac{(b-a)\mathscr{C}}{2\mathscr{D}}\,. \qquad \text{(VI.34)}$$

Finally, we determine γ by

$$\gamma = \alpha - \kappa\cosh\mathscr{A} = \alpha - \frac{b-a}{2\mathscr{D}}\cosh(\mathscr{C}-\mathscr{D})\,. \qquad \text{(VI.35)}$$

This leads to the formula

$$y_*(x) = \alpha + \frac{b-a}{2\mathscr{D}}\left(\cosh\left(\mathscr{C} + \frac{\mathscr{D}(2x-a-b)}{b-a}\right) - \cosh(\mathscr{C}-\mathscr{D})\right), \qquad \text{(VI.36)}$$

where $\mathscr{C}$ and $\mathscr{D}$ are determined by (VI.31) and (VI.33).

All this was done under the assumption that $\Lambda > \Delta$. Let us now examine the limit of the right-hand side of (VI.36) as $\Lambda \downarrow \Delta$. For this purpose, we rewrite (VI.36) as

$$\begin{aligned} y_*(x) &= \alpha + \frac{b-a}{\mathscr{D}}\left(\sinh\left(\mathscr{C} + \frac{\mathscr{D}(x-b)}{b-a}\right)\sinh\left(\frac{\mathscr{D}(x-a)}{b-a}\right)\right) \\ &= \alpha + (x-a)\frac{\sinh\left(\frac{\mathscr{D}(x-a)}{b-a}\right)}{\frac{\mathscr{D}(x-a)}{b-a}}\sinh\left(\mathscr{C} + \frac{\mathscr{D}(x-b)}{b-a}\right), \end{aligned}$$

and observe that $\mathscr{D} \to 0$ and $\mathscr{C} \to \mathscr{C}_0$ as $\Lambda \downarrow \Delta$, where $\mathscr{C}_0$ is characterized by

$$\tanh \mathscr{C}_0 = \frac{\beta - \alpha}{\Delta}. \tag{VI.37}$$

It follows that, as $\Lambda \downarrow \Delta$, $y_*(x)$ converges to

$$\bar{y}_*(x) = \alpha + (\sinh \mathscr{C}_0)(x-a). \tag{VI.38}$$

Equation (VI.37) easily implies $\sinh \mathscr{C}_0 = \frac{\beta-\alpha}{b-a}$, so the final expression for $\bar{y}_*(x)$ is

$$\bar{y}_*(x) = \alpha + \frac{\beta - \alpha}{b-a}\cdot(x-a), \tag{VI.39}$$

which is, as expected, the unique member of $\mathscr{Y}$ for the limiting case $\Lambda = \Delta$.

COMMENT NO. 24. Must put in remarks on how in the "singular case" $\Lambda = \Delta$ the catenary problem gives rise to an abnormal extremal.

Chapter VII

The control Hamiltonian and the Hamiltonian formulation of relativistic mechanics

In our discussion of Weierstrass' failure to discover the Hamiltonian interpretation of his side conditions and its profound implications, we pointed out that Lagrangians that are positively homogeneous of degree one are particularly unsuited to a Hamiltonian formulation of the necessary conditions for optimality in terms of the classical Hamiltonian, so that for this class of Lagrangians the superiority of the control Hamiltonian is apparent.

Such Lagrangians are in fact quite important in applications. They occur, for example, in relativistic mechanics, where we look at paths in space-time and there is no natural "time function" to parametrize space-like curves.

For example, the Lagrangian for the motion of a charged particle in an electromagnetic field is

$$L = -mc||\dot{q}|| + \frac{\rho}{c}A(q)\cdot\dot{q}\,. \tag{VII.1}$$

Here

1. the variable q takes values in 4-dimensional Minkowski space-time Q,
2. $||\dots||$ is length with respect to the Lorentz-Minkowski metric, so that

$$||v|| = \sqrt{\langle v,v\rangle}\,, \tag{VII.2}$$

where $\langle\cdot,\cdot\rangle$ is the Minkowski inner product, given, with respect to any Lorentz frame, by

$$\langle v,w\rangle = \sum_{i=0}^{3} g_i v_i w_i\,, \text{ with } g_0 = c^2\,,\ g_1 = g_2 = g_3 = -1\,, \tag{VII.3}$$

3. $q \mapsto A(q)$ is the electromagnetic potential, thought of as a field of covectors on Q,

4. m is the rest mass, ρ is the charge, and c is the speed of light,

5. $w \cdot v$ is the ordinary pairing between the covector w and the vector v, given in components by

$$w \cdot v = w_0 v_0 + w_1 v_1 + w_2 v_2 + w_3 v_3\,. \tag{VII.4}$$

6. $L(q,\dot{q})$ is defined for $q \in Q$ and $\dot{q} \in C$, where C is the open cone of forward time-like vectors.

Remark VII.1. Formula (VII.1) can be "derived" (or, more precisely, shown to be mathematically natural) by the following elementary invariance arguments. The action functional $S = \int_a^b L(q(s),\dot{q}(s))ds$ ought to be defined for a curve $s \mapsto q(s)$ independently of the choice of parametrization, which forces L to be positively homogeneous of degree 1 with respect to $\dot{q}$. For a free particle, the Lagrangian should be translation-invariant and therefore must be a function of $\dot{q}$ alone. Lorentz invariance then dictates that it should be a function of $||\dot{q}||$ alone, and homogeneity then forces it to be proportional to $||\dot{q}||$. The choice of a constant of proportionality equal to αmc, for some dimensionless scalar α, is then needed for dimensional reasons.

Finally, the fact that $\alpha = -1$ is necessary to have the appropriate nonrelativistic limit for small velocities. Indeed, let us write $q = (q_0, q_\#)$, relative to some Lorentz frame, where $q_\#$ is a 3-dimensional vector and q_0 is time, and use the frame-dependent physical time $t = q_0$ to parametrize our curves. Then

$$\alpha mc||\dot{q}|| = \alpha mc\sqrt{c^2 - ||\dot{q}_\#||^2}\,, \tag{VII.5}$$

where $||\dot{q}_\#||$ is now the ordinary Euclidean length of the 3-vector $\dot{q}_\#$. Therefore

$$\alpha mc||\dot{q}|| = \alpha mc^2\sqrt{1 - \frac{||\dot{q}_\#||^2}{c^2}}\,, \tag{VII.6}$$

which behaves like $\alpha mc^2 - \frac{\alpha m}{2}||\dot{q}_\#||^2$ as $\dot{q}_\# \to 0$. Modulo an additive constant—which in any case makes no contribution to the Euler-Lagrange equation—this should look like the classical kinetic energy $\frac{m||\dot{q}_\#||^2}{2}$ as $\dot{q}_\# \to 0$, so α must be equal to -1.

As for the other term, given that the electromagnetic potential A is a field of covectors, the simplest way to associate to A a function of q and $\dot{q}$ which is positively homogeneous of degree 1 with respect to $\dot{q}$ is to take it proportional to $A(q)\cdot\dot{q}$, and the choice of $\frac{\rho}{c}$ for the constant is once again natural for dimensional reasons. □

It is well known in physics that there are nontrivial difficulties when one attempts to formulate the relativistic equations of motion in Hamiltonian form. For example, the textbook [65], by H. Goldstein, contains the following remarkable assertion, that we quote adding our own emphasis:

> for Lagrangians homogeneous in the $\dot{q}_i$'s in the first degree the usual Hamiltonian procedure does not work. A number of alternate ways of formulating the Hamiltonian in the "homogeneous problem" have been devised. All take into account explicitly the constraint implied by the vanishing of the Jacobian determinant, but they are all complicated, and no one has been accepted as the standard technique. ***It appears that if at all possible the homogeneous "noninvertible" Lagrangian should be avoided like the plague.***

For a not very convincing example of one of the "alternate ways," the reader may consult Lanczos [86].

On the other hand, a truly succesful approach was proposed by Tulczyjew in 1977 in [150] (cf. also [151]), using symplectic geometry.

From the control perspective, is it easy to see why there appears to be a problem with the classical Hamiltonian formulation, since the transformation (V.15) cannot be inverted. On the other hand, ***the control Hamiltonian*** H ***makes perfect sense, and can be used to write the equations of motion, which turn out to be the usual relativistic equations of motion.***

To see this, write

$$H = p\cdot u - L = p\cdot u + mc||u|| - \frac{\rho}{c}A(q)\cdot u\,, \tag{VII.7}$$

so that $H = \pi\cdot u + mc||u||$, where $\pi = p - \frac{\rho}{c}A(q)$.

Then $\frac{\partial H}{\partial u} = 0$ yields

$$\pi_0 = -mc^3\frac{u_0}{||u||}\,,\quad \pi_i = mc\frac{u_i}{||u||}\ \text{ for }\ i = 1,2,3\,, \tag{VII.8}$$

while the other Hamilton equations say that

$$\frac{dq_i}{ds} = u_i\,, \quad \frac{dp_i}{ds} = \frac{\rho}{c}\sum_j \frac{\partial A_j}{\partial q_i}(q)\, u_j\,. \tag{VII.9}$$

Naturally, the "time" parameter s here is not physical time. Physical time is one of the state variables, namely, the component q_0 of the vector q, and depends on the choice of a Lorentz frame. The quantity s is just a parameter for the curve $q(\cdot)$. Notice that the system of equations (VII.8), (VII.9) is invariant under arbitrary reparametrizations of the curve. Notice also that the 4-momentum covector π satisfies

$$||\pi||^2 = m^2c^2\,, \tag{VII.10}$$

where $||\pi||$ is the norm $\sqrt{\langle \pi, \pi\rangle}$ of $||\pi||$ arising from the inner product

$$\langle p, p'\rangle = \frac{p_0 p_0'}{c^2} - \sum_{i=1}^{3} p_i p_i' \tag{VII.11}$$

of covectors. This in particular implies the famous formula

$$E = mc^2 \tag{VII.12}$$

for the "rest energy," i.e., the 0-th component $E \stackrel{\text{def}}{=} \pi_0$ of π relative to a a frame with respect to which the particle is at rest.

From (VII.9) we get

$$\frac{d\pi_i}{ds} = \frac{\rho}{c}\sum_j \left(\frac{\partial A_j}{\partial q_i}(q) - \frac{\partial A_i}{\partial q_j}(q)\right) u_j\,, \tag{VII.13}$$

i.e.,

$$\frac{d\pi_i}{ds} = \frac{\rho}{c}\sum_j F_{ij}(q) u_j\,, \tag{VII.14}$$

where

$$F_{ij} = \frac{\partial A_j}{\partial q_i} - \frac{\partial A_i}{\partial q_j}\,. \tag{VII.15}$$

If we now choose a Lorentz frame, use the time q_0 as the curve parameter, and write v to denote the vector $(\dot q_1, \dot q_2, \dot q_3)$, we get $\dot q_0 = 1$, and $||\dot q|| = \sqrt{c^2 - ||v||^2}$. (The dots now denote derivatives with respect to q_0.) So

$$\pi_i = \frac{m v_i}{\sqrt{1 - \frac{||v||^2}{c^2}}}\,, \tag{VII.16}$$

showing that π_1, π_2, π_3 are the components of the relativistic momentum of our particle. Equation (VII.13) then gives us the time derivative of the momentum, i.e., the force acting on the particle. Setting $cE_i = F_{0i}$ for $i = 1, 2, 3$, and $B_1 = F_{23}$, $B_2 = F_{31}$, $B_3 = F_{12}$, as is customary, we find

$$\text{Force} = \rho\left(E + \frac{1}{c}(v \times B)\right), \tag{VII.17}$$

which is the usual relativistic law of motion of a charged particle in an electromagnetic field.

Clearly, the above argument involves no *ad hoc* considerations, and is no more than a straightforward appplication of the Hamiltonian formalism, in "control" form.

Chapter VIII

The value function and the comeback of Hamilton's great idea

Recall that, according to J. L. Synge's account of Hamilton's "principal" or "characteristic" function V in §V.7, page 112, one defines $V(A',A)$, for two points A', A, to be the minimum of the values of the integral (V.24) for all curves joining A' to A. For this to make sense, one needs the "rash" assumption that there always is a curve that gives a minimum value to the integral.

In modern terms, and without resorting to this "rash assumption," we can define the *value function* $V : Q \times Q \mapsto \bar{\mathbb{R}}$, for a controlled dynamics $\dot{q} = f(q,u)$, $q \in Q$ and Lagrangian $L(q,u)$, by letting $V(\bar{q},\hat{q})$ be, for a pair of points $\bar{q},\hat{q} \in Q$, the infimum of the integrals $\int_a^b L(q(t),u(t))dt$, taken over all choices of a, b and all solutions $t \mapsto (q(t),u(t))$ of $\dot{q}(t) = f(q(t),u(t))$, $t \in [a,b]$, such that $q(a) = \bar{q}$, $q(b) = \hat{q}$. (Here $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty,+\infty\}$. The infimum is $+\infty$ if there are no trajectories of the system going from $\bar{q}$ to $\hat{q}$.)

If we fix a "target set" $\mathscr{T} \subseteq Q$, then we can also consider the value function $q \mapsto V_{\mathscr{T}}(q)$, defined by letting

$$V_{\mathscr{T}}(q) = \inf\{V(q,q') : q' \in \mathscr{T}\}\,. \tag{VIII.1}$$

It is clear that $V_{\mathscr{T}}$ is very closely related to Hamilton's principal function.

Optimal control theory has fully vindicated Synge's view of the importance of Hamilton's idea of the principal function, by making this function (or, rather, its close relative $V_{\mathscr{T}}$) the central object of the method of "dynamic programming," one of the theory's main tools.

According to this method, one can determine the value function $V_{\mathscr{T}}$ by solving a certain partial differential equation—the "Hamilton-Jacobi-Bellman (HJB) equation"—with appropriate boundary conditions. One can then use the value function to find the optimal controls in "feedback form."

More precisely, the HJB equation says

$$\inf\{\langle \nabla V(q), f(q,u)\rangle + L(q,u) : u \in U\} = 0\,. \tag{VIII.2}$$

and it can be proved, under appropriate technical conditions, that the function $V_{\mathscr{T}}$ satisfies (VIII.2) and is the unique solution that satisfies a certain boundary condition.

However, $V_{\mathscr{T}}$ in general fails to be continuously differentiable, so (VIII.2) has to be interpreted in a weak sense, for example as a "viscosity solution" (cf. [59] for details). On the other hand, at those points q where $V_{\mathscr{T}}$ is differentiable and the infimum of (VIII.2) is a true minimum, attained by a unique value $u(q)$ of of u, this value has a natural interpretation: *$u(q)$ is the initial value of the control that gives rise to an optimal trajectory from q to the target.*

For a simple example, consider the minimum-time optimal control problem in $\mathbb{R}^2$ with equations

$$\dot{x} = \frac{u}{1+y^2}\,,\ \dot{y} = v\,, \tag{VIII.3}$$

with control constraint $|u|+|v| \leq 1$, and target $\{(0,0)\}$. It is easy to see intuitively, and to prove rigorously, that the optimal trajectory from an initial point (x,y) to the origin consists of a vertical segment from (x,y) to $(x,0)$, with $v = \pm 1$, followed by a horizontal segment from $(x,0)$ to the origin, with $u = \pm 1$. The value function V is given by

$$V(x,y) = |x| + |y|\,. \tag{VIII.4}$$

Notice that V is not continuously differentiable. However, V is continuously differentiable on the set $S = \{(x,y) : xy \neq 0\}$, and

$$\nabla V(x,y) = (\mathrm{sgn}(x), \mathrm{sgn}(y)) \ \text{ for } \ (x,y) \in S\,. \tag{VIII.5}$$

The HJB equation says

$$\inf\Big\{\frac{u.\mathrm{sgn}(x)}{1+y^2} + v.\mathrm{sgn}(y) : |u|+|v| \leq 1\Big\} = -1\,, \tag{VIII.6}$$

from which we get $v = -\mathrm{sgn}(y)$ and $u = 0$ for $xy \neq 0$. Notice that at the points of differentiability of V we completely recover the optimal feedback from the function V.

On the other hand, it turns out that *the points of nondifferentiability of V play a crucial role*, even though they form a small set. (For example, they form a set of measure zero, and in fact of Hausdorff codimension 1.) Indeed, it is at these points that the optimal control "turns" and becomes horizontal rather than vertical.

This turns out to be a general feature: the value function is in general not of class C^1, and the points of nondifferentiability play a crucial role in determining the structure of the optimal controls. (For an even simpler example with a clear geometrical interpretation, let Q be the unit sphere in $\mathbb{R}^3$, let the cost be arc-length, and let the target be the South pole. Then $V(q)$ is the distance along the sphere from q to the South pole. Clearly, V is continuously differentiable everywhere except at the South and North poles. The HJB equation says that $||\nabla V|| = 1$. Given any point q which is not one of the poles, the direction u that minimizes the inner product $\langle \nabla V(q), u\rangle$ is exactly the direction of the meridian through q, pointing down, i.e., the direction of the minimizing arc from q to the South pole. The nondifferentiability of V at the North pole is a reflection of the fact that at this point the minimizing arc going to the South pole is not unique.)

Therefore, while Synge was right in claiming that mathematicians should have paid more attention to Hamilton's principal function, it is also true that their reluctance to proceed as if this function had nice regularity properties, was justified. What was truly needed to do justice to Hamilton's idea was theory that would face, rather than avoid, the issue if nonsmoothness, and would make it possible to deal with the value function as a truly nonsmooth object, as modern optimal control does.

VIII.1 The true meaning of the momentum variable

The value function has turned out to have many important properties that we will not enumerate here. In the remainder of this section, we will limit ourselves to describing how the value function provides a nice interpretation of the meaning of the momentum vector p.

In our outline of the evolution leading to the maximum principle, we have seen how the meaning of the "adjoint vector" p—i.e., the "momentum"—has changed. Originally defined by the equation

$$p(t) = \frac{\partial L}{\partial u}(q_*(t), \dot{q}_*(t), t)\,, \tag{VIII.7}$$

(cf. (V.7)), the momentum was thought of as a time-varying vector uniquely associated to a given curve $t \mapsto q_*(t)$. The Weierstrass side condition can then be formulated in terms of this quantity, and says that the function $u \mapsto \langle p(t), u\rangle - L(q_*(t), u)$, with $p(t)$ defined by (VIII.7), i.e., by

$$p(t) = \frac{\partial L}{\partial u}(q_*(t), \dot{q}_*(t)), \tag{VIII.8}$$

is maximized by the choice of $u = \dot{q}_*(t)$. (Recall that in Weierstrass' setting L is supposed not to depend on t),

On the other hand, without requiring *a priori* that $p(t)$ be given by (VIII.7), one could simply say that

(A) *$p(t)$ is such that $u \mapsto \langle p, u\rangle - L(q(t), u, t)$ is maximized by $u = \dot{q}(t)$.*

Since (A) implies (VIII.7) anyhow, it would appear that nothing is gained by substituting (A) for (VIII.7). This appearance, however, is quite deceptive, as was explained in Chapter V. We saw there that *the gain achieved by using (A) instead of (VIII.7) is in fact enormous,* because (A) implies (VIII.7) under the usual conditions of the classical calculus of variations, but not otherwise. Hence, as soon as the "usual conditions" fail to be satisfied, it turns out that (A) is still correct but (VIII.7) is not. (We have seen examples of this in Chapter V. In Example V.5 L fails to be differentiable with respect to u at $u = \dot{q}(t)$, so (VIII.7) does not make sense. In Example V.7 there is an inequality constraint on $\dot{q}(t)$, so (A) does not imply (VIII.7) when the maximum is attained on the boundary of the set of permissible values of $\dot{q}(t)$.)

So (A) is much more fundamental than (VIII.7) as a characterization of "the momentum along a curve." The substitution of (A) for (VIII.7) has a cost, however: while (VIII.7) clearly defines $p(t)$ uniquely, Condition (A) does not, since there may be more than one value of the vector p for which (A) holds. So we are now no longer allowed to talk about ***the*** momentum vector corresponding to a given curve, but only of ***a*** momentum vector.[1] And in fact the maximum principle then says that a necessary condition for optimality of a trajectory is that there exist a momentum function $t \mapsto p(t)$ that satisfies certain properties.

It turns out, however, that there is another identity about p that remains valid under much more general conditions, and provides a much better interpretation of the "true" meaning of the momentum. Precisely, it can be proved under very general conditions that ***the momentum*** p ***is the negative of the gradient of the value function***, as we now explain.

We first illustrate the general situation with a simple example.

[1] with apologies to foreign readers whose native language fails to make a precise distinction between definite and indefinite articles. Here the distinction is crucial.

Example VIII.1. Consider the minimum-time optimal control problem in $\mathbb{R}^2$ discussed earlier, with equations (VIII.3), control constraint $|u|+|v|\leq 1$, and target $\{(0,0)\}$. As explained before, the value function V is given by the formula $V(x,y)=|x|+|y|$, so V is continuously differentiable on the set $S=\{(x,y): xy\neq 0\}$, and (VIII.5) holds.

Pick a point $(\bar{x},\bar{y})\in S$, and let $\xi:[0,T]\mapsto\mathbb{R}^2$ be the optimal trajectory from $(\bar{x},\bar{y})$ to $(0,0)$ (so that $T=|\bar{x}|+|\bar{y}|$). Write $\xi(t)=(x(t),y(t))$, and let $t\mapsto(u^\xi(t),v^\xi(t))$ be the corresponding control. It then follows that $u^\xi(t)=0$ and $v^\xi(t)=-\mathrm{sgn}(\bar{y})$ for $0\leq t\leq|\bar{y}|$, and $v^\xi(t)=0$, $u^\xi(t)=-\mathrm{sgn}(\bar{x})$ for $|\bar{y}|<t\leq|\bar{x}|+|\bar{y}|$. The Hamiltonian is given by

$$H=\frac{p_x u}{1+y^2}+p_y v-p_0\,, \tag{VIII.9}$$

where p_x,p_y are the components of the adjoint covector p.

Using p_x^ξ, p_y^ξ to denote the components of the momentum along ξ, we get the equations

$$\dot{p}_x^\xi=0\,,\;\dot{p}_y^\xi=-\frac{2yp_x^\xi u^\xi}{(1+y^2)^2}\,, \tag{VIII.10}$$

so that $\dot{p}_x^\xi\equiv\dot{p}_y^\xi\equiv 0$, since $y(t)u^\xi(t)\equiv 0$. The maximum value of the Hamiltonian is

$$M(t)=\tilde{M}(t)-p_0\,,\text{ where }\tilde{M}(t)=\max\left(\frac{|p_x^\xi(t)|}{1+y(t)^2},|p_y^\xi(t)|\right). \tag{VIII.11}$$

Clearly, $\tilde{M}(t)\neq 0$ unless both $p_x^\xi(t)$ and $p_y^\xi(t)$ vanish. The possibility that $p_x^\xi(t)=p_y^\xi(t)=0$ is excluded because it implies $p_0=0$, contradicting the nontriviality condition of the maximum principle. So $\tilde{M}(t)>0$, and then $p_0>0$, so we may normalize the adjoint vector and assume that $p_0=1$. To maximize the Hamiltonian we must take

$$u(t)=\mathrm{sgn}(p_x^\xi(t))\,,\;v(t)=0\text{ if }\frac{|p_x^\xi(t)|}{1+y(t)^2}>|p_y^\xi(t)|\,, \tag{VIII.12}$$

and

$$u(t)=0\,,\;v(t)=\mathrm{sgn}(p_y^\xi(t))\text{ if }\frac{|p_x^\xi(t)|}{1+y(t)^2}<|p_y^\xi(t)|\,. \tag{VIII.13}$$

It follows that at time $t=|y|$ the equality $|p_x^\xi(t)|=|p_y^\xi(t)|$ holds, from which it easily follows that $p_x^\xi(t)=\mathrm{sgn}(\bar{x})$ and $p_\xi^y(t)=\mathrm{sgn}(\bar{y})$.

This shows that

$$(p_x^\xi(0),p_y^\xi(0))=-\nabla V(\bar{x},\bar{y})\text{ if }\bar{x}\neq 0\,,\;\bar{y}\neq 0\,. \tag{VIII.14}$$

In other words, *at every point $\bar{q} \in Q$ where the value function is differentiable, if $\xi : [0,T] \mapsto Q$ is an optimal trajectory from $\bar{q}$ to the target, and $t \mapsto p(t)$ is the corresponding momentum vector given by the maximum principle, normalized so that $p_0 = 1$, then the identity*

$$p(0) = -\nabla V(\bar{q}) \tag{VIII.15}$$

holds. □

What has been shown in the above example turns out to be true under fairly general conditions, at points where V is of class C^1. Moreover, the result identifying p with $-\nabla V$ has been generalized to the case when V is Lipschitz but not C^1, using the tools and concepts of nonsmooth analysis (cf. [44, 43, 59, 126] for details).

Remark VIII.2. Often, the maximum principle is written using a Hamiltonian $H = \langle p, f(q,u,t)\rangle + p_0 L(q,u,t)$, and requiring that $u_*(t)$ minimize—rather than maximize—the function $u \mapsto H(q_*(t),u,p(t),t)$. This corresponds to a choice of p equal to the negative of the p we have been using here. When this version of the maximum principle is used, the relationship between p and V is that $p(0) = \nabla V(\bar{q})$. □

Chapter IX

From principle to theorem

IX.1 Principles and versions

Our discussion so far has only dealt with the formal aspect of the necessary conditions for optimality. The results of the previous sections, from the Euler-Lagrange equation to the maximum principle, should be regarded as *principles* rather than theorems. For us, a *principle* is a generator of theorems, a not yet completely precise statement that can be made into a theorem by filling in the technical details and making all the definitions and conditions completely precise. The resulting theorems are *versions* of the principle. Usually, the choice of technical conditions can be made in more than one way, so a "principle" has more than one version. For example, the statement that "if the derivative of a function f vanishes then f is a constant" is a "principle" in our sense. One version is obtained by taking f to be defined on a subinterval I of $\mathbb{R}$, and requiring that $f'(x)$ exist and be equal to 0 for all $x \in I$. Another one may just require f to be continuous on I, and $f'(x)$ to exist and equal zero for all x, except possibly for a finite or countable set. A third one requires f to be absolutely continuous and $f'(x)$ to exist and equal zero almost everywhere. A fourth one requires f to be locally integrable, and f' to vanish in the sense of distributions. In the first three versions, the conclusion is that $f(x) = 0$ for all x, but in the fourth one the conclusion only says that $f(x) = 0$ almost everywhere. One can produce many more versions, e.g. by taking f to be a distribution rather than a function, or by going to higher-dimensional spaces, or to manifolds.

In some cases, a "principle" becomes identified in the minds of mathematicians with its first published rigorous version. This has happened to some extent in the case of the maximum principle, because the book [112], where the result was first presented, already contains a rigorous version. We will show below that this version does not exhaust the full power of the principle, and the work of stating and proving stronger and more general versions is still very much in progress.

In order to turn the principles of the previous sections into real mathematical theorems, we have to be accurate as to the technical assumptions on L and, in the optimal control case, on f and U. We have to be rigorous as to the problem statement and the meaning of the conclusions. This requires, among other things, that the concept of an "arbitrary function" be replaced by a precise specification of the kind of function involved. Even for a simple dynamical law such as $\dot{x} = u$, it would not make sense to solve for x for a function $t \mapsto u(t)$ which is truly "arbitrary" in the sense of 20th century mathematics, i.e., not even required to be measurable. One needs at least to ask that $u(\cdot)$ be integrable in the sense of some reasonable integration theory such as those of Riemann, Lebesgue or Perron.

These considerations would have been alien to 18th century mathematicians, for whom the concept of function had not yet crystalized into a notion susceptible of rigorous treatment. For example, Lagrange wrote (cf. [124], pp. 388-9) that "We define as a *function* of one or several quantities any mathematical expression in which those quantities appear in any manner." But then he went on to say that if $f(x)$ was a function of one variable, then "by the theory of series" we can expand $f(x+i)$, "i being an arbitrary quantity," in a series of the form $f(x) + pi + qi^2 + ri^3 + \ldots\ldots$, in which $p, q, r, \ldots$ are functions of x. One wonders whether Lagrange would have accepted $f(x) = |x|$ as a "function"—as he certainly should have, given his own definition of function—and how he would have written the power series expansion of $f(0+i)$ in this case.

Regarding the necessary conditions for optimality, while the discovery of new and more general formal conditions progressed, rigorous versions of the formal results were derived at various stages of the process, using in each case the mathematical tools available at the time. For example, Weierstrass used the concepts of calculus in a perfectly rigorous and modern way, and proved theorems that pass the most stringent 20th century tests for rigor and precision. But the tools of 20th century measure and integration theory were not yet available to him, so in many cases his theorems make unnecessarily restrictive assumptions, such as that a function f is piecewise continuously differentiable, even though the truly natural assumption would be that f is absolutely continuous.

The story of how rigor emerged from the confusion of 17th and 18th century mathematics is quite interesting in its own right, but we will not tell it here in full detail. We will only give a brief outline, concentrating only on a few points that are either relevant for the final outcome—namely, the maximum principle and its various versions—or particularly remarkable in their own right. We will then give a much more detailed discussion of the work of translating the maximum principle into rigorous mathematics. This is still very much in progress, and we will attempt to convey the flavor of some of the current research on the subject.

IX.2 Euler

It is well known that Euler was able to perform amazing calculations that yielded correct results but by today's standards would be totally unacceptable. (See, for example, the description in Stillwell [122] of how he found the glorious formula $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.) For curve minimization problems, his method was based on discretizing time, making finite difference approximations for all the derivatives involved, obtaining a finite difference equation for the solution of the approximating problem, and then passing to the limit. The integral of a function was approximated by the area under a polygonal curve approximating the function's graph.

More accurately, one should keep in mind that expressions such as "passing to the limit" are from a later period. What Euler actually did was use the language of "infinitely small quantities" and resort very often to "geometric" considerations: he would divide the interval of definition of a function into "infinitely" many subintervals, and draw a picture (which, of course, only showed a subdivision into finitely many intervals).

IX.3 Lagrange

While Euler's work was full of "geometric" arguments, and contained many pictures, Lagrange used no pictures. (Newton's 1687 *Principia* had 250 figures. Lagrange proudly stated in the preface of his 1788 *Mécanique Analytique* that "no diagrams will be found in this work.") He introduced, by analogy with the usual differentials dx, dy, another class of differential-like objects, the "variations"—for which he used notations such as δy—which were thought of as infinitesimally small functions, or infinitesimal functional increments of a function, exactly as the ordinary dx, dy are to be thought of as infinitesimal changes of numerical variables.

So, for example, to solve a "standard" problem (II.5) by Lagrange's method, we would set $\delta I = 0$, and then write

$$0 = \delta I = \delta \int_a^b L\, dt = \int_a^b (\delta L)\, dt = \int_a^b (M\,\delta q + N\,\delta \dot{q}) dt\,, \tag{IX.1}$$

where $M = \frac{\partial L}{\partial q}$, $N = \frac{\partial L}{\partial \dot{q}}$. Using the commutation rule $\delta \dot{q}\, dt = \delta dq = d\delta q$, together with the integration by parts rule $\int_a^b y\,dx = yx|_a^b - \int_a^b x\,dy$, we get

$$\int_a^b N\delta\dot{q}\,dt = \int_a^b N\,d\delta q = N\delta q\Big|_a^b - \int_a^b (\delta q)dN = N\delta q\Big|_a^b - \int_a^b (\delta q)\dot{N}\,dt\,.$$

So $0 = \int_a^b (M - \dot{N})\delta q\,dt + N\delta q\Big|_a^b$. Since δq vanishes at a and b, because our variations have to preserve the endpoint conditions $q(a) = \bar{q}$, $q(b) = \hat{q}$, we get $\int_a^b (M - \dot{N})\delta q\,dt = 0$. Since δq is arbitrary, we end up with $M = \dot{N}$, i.e., with (V.3).

Remark IX.1. The precise meaning of Lagrange's "variations" was not clear to Lagrange or to his contemporaries. When Lagrange wrote there was a lot of confusion about the meaning of "infinitesimal quantities," "fluxions[1]," derivatives, and differentials. It was clear that there were rules for their manipulation of these mysterious entities, and that these rules somehow appeared to lead to consistent results. But the justification of the rules themselves was lacking, as was a truly coherent account of the rules themselves. Mathematicians following these rules had to know when to apply them and when to refrain to use certain arguments that the rules appeared to permit in order to avoid reaching a contradiction. For example, if $y = x^2$, and they wanted to compute dy/dx, they could write

$$dy = (x + dx)^2 - x^2 = 2x\,dx + (dx)^2\,. \tag{IX.2}$$

It was then permissible to divide both sides by dx, presumably by invoking some kind of "division rule," according to which "if $a \cdot c = b \cdot c$ and $c \neq 0$ then $a = b$." This yields

$$dy/dx = 2x + dx\,. \tag{IX.3}$$

They would then have to contradict themselves and arbitrarily set $dx = 0$, thereby obtaining the desired foimula

$$dy/dx = 2x\,. \tag{IX.4}$$

The problem here is, of course, that $dx \neq 0$ is neded to go from (IX.2) to (IX.3), and $dx = 0$ to go from (IX.3) to (IX.4). So in fact the "division rule" would also have to be applicable in some cases when $c = 0$. But naturally, it could not be *always* applicable when $c = 0$, for this would have led to such arguments as "$1 \cdot 0 = 2 \cdot 0$, therefore $1 = 2$." The precise conditions of applicability of the division rule were

[1]This was Newton's term for "velocities," or "celerities," of "decrease or increase" of "fluents." Newton's "fluents," or "flowing quantities," were "quantities which I [i.e., Newton] consider as gradually and indefinitely increasing," and were represented by letters such as v, w, x, y and z, while "the velocities by which every fluent is increased by its generating motion (which I may call *fluxions*, or simply velocities, or celerities), I shall represent by the same letters pointed, thus $\dot{v}$, $\dot{x}$, $\dot{y}$, $\dot{z}$." (Quoted from Cajori [33], p. 194.)

thus clouded in mystery, and the best that could be said was that the rule could be applied as long as it did not lead to a contradiction or an impossibility.

With such methods of reasoning, it is not at all surprising that some mathematicians, in particular Nieuwentijt[2] as early as 1694, and, later, Bishop George Berkeley (1685-1753), the eminent philosopher, criticized the foundations of differential calculus. Berkeley was perfectly justified when he wrote in the *Analyst*, in 1734, that instantaneous rates of change were "neither finite quantities, nor quantities infinitely small, nor yet nothing." They were but "the ghosts of departed quantities. Certainly ... he who can digest a second or third fluxion ... need not, methinks, be squeamish about any points of Divinity."[3]

If such was the situation regarding ordinary differentials, it was even more so for variations, which were treated in a purely formal way.[4]

Both Euler and Lagrange understood the criticisms perfectly well, agreed with the critics, and accepted that the Calculus was unsound, but they believed that it gave correct results anyhow because of cancellation of errors. And their uncanny intuition enabled them to navigate in a sea of conceptual muddle and focus on the obstacles that truly mattered (see, for example, footnote 7 in page 190), without being distracted by those foundational issues that later generations would eventually settle in a completely rigorous way. □

IX.4 The modern justification of Lagrange's derivation; the Lavrentiev phenomenon

Lagrange's derivation of (V.3), properly reformulated in the vector case, is exactly what we would still do today on the formal level to get the basic equations of the calculus of variations. On the other hand, a rigorous mathematical proof given today would assign a more precise meaning to the "variations." We would begin

[2]The Dutch Bernhard Nieuwentijt (1654-1718) was a physician-burgomaster of Purmerend, near Amsterdam.

[3]According to Cajori ([33], p. 219), "Berkeley's object in writing the *Analyst* was to show that the principles of fluxions are no clearer than those of Christianity. He referred to an 'infidel mathematician' (Edmund Halley), of whom the story is told that, when he jested concerning theological questions, he was repulsed by Newton with the remark, 'I have studied those things, you have not.' A friend of Berkeley, when on a bed of sickness, refused spiritual consolation because the great mathematician Halley had convinced him of the inconceivability of the doctrines of Christianity. This induced Berkeley to write the *Analyst*."

[4]Cf. [66], p. 112, for a more detailed discussion of this point, and also for the observation that Euler in 1771 "imagined a family of comparison curves $y = y(x,t)$ and took for his variation δy the quantity $dt(dy/dt)$," evaluated along the curve, and that Lagrange pursued a similar course in his later work.

by specifying a space F of Q-valued functions on $[a,b]$, for example the space $C^1([a,b],Q)$ of C^1 maps from $[a,b]$ to Q. We would then define a "variation" (with fixed endpoints) of an arc q_* to be a curve $[-\bar{\varepsilon},\bar{\varepsilon}] \ni \varepsilon \mapsto q^\varepsilon$—for some $\bar{\varepsilon} > 0$—in the space F, consisting of arcs q^ε such that the endpoint conditions $q^\varepsilon(a) = \bar{q}$, $q^\varepsilon(b) = \hat{q}$ always hold. We would then observe that, if q_* is a solution of our minimization problem, then for very variation $[-\bar{\varepsilon},\bar{\varepsilon}] \ni \varepsilon \mapsto q^\varepsilon$ the function $\varepsilon \mapsto I(q^\varepsilon)$ must have a minimum at $\varepsilon = 0$. Therefore, if $\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} I(q^\varepsilon)$ exists, then it must vanish. Writing $q(t,\varepsilon)$ rather than $q^\varepsilon(t)$, we would then let δq, $\delta \dot{q}$ stand for $\frac{\partial q}{\partial \varepsilon}(t,0)$, $\frac{\partial}{\partial \varepsilon}\frac{\partial q}{\partial t}(t,0)$, respectively. Then the commutation rule expresses the equality of the mixed derivatives:

$$\frac{\partial}{\partial \varepsilon}\frac{\partial q}{\partial t} = \frac{\partial}{\partial t}\frac{\partial q}{\partial \varepsilon}. \tag{IX.5}$$

(So δq is an "infinitesimal variation" of q_*, and $\delta \dot{q}$ can be thought of as the corresponding "infinitesimal variation" of $\dot{q}_*$, or—thanks to the commutation rule—as the time derivative of δq.)

To avoid unnecessary technical complications, we can just limit ourselves to variations of the form $q(t,\varepsilon) = q_*(t) + \varepsilon r(t)$, where $r \in F$ is such that $r(a) = r(b) = 0$. In this case, the commutation rule is valid if, for example, q_* and r are of class C^1.

Then the precise meaning of the formal calculations of (IX.1) is, of course, that we are differentiating the integral with respect to the parameter ε at $\varepsilon = 0$:

$$\begin{aligned} &\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \int_a^b L\Big(q_*(t) + \varepsilon r(t), \dot{q}_*(t) + \varepsilon \dot{r}(t), t\Big)\, dt \\ &= \int_a^b \left[\frac{\partial L}{\partial q}(q_*(t),\dot{q}_*(t),t)\cdot r(t) + \frac{\partial L}{\partial \dot{q}}(q_*(t),\dot{q}_*(t),t)\cdot \dot{r}(t)\right] dt\,. \end{aligned} \tag{IX.6}$$

Clearly, this step needs careful justification.

Once this step is justified, we end up having rigorously proved that $\int_a^b (Mr + N\dot{r}) = 0$. An integration by parts—which also needs to be properly justified—then yields $\int_a^b (M - \dot{N})r = 0$. Finally, since this is true for all r, we should be able to conclude that $M = \dot{N}$, i.e., the that Euler-Lagrange equation holds.

So there are three steps that need rigorous justification, namely,

(i) the differentiation under the integral sign (IX.6),

(ii) the integration by parts,

(iii) the passage from "$\int_a^b F.r = 0$ for all r" to $F \equiv 0$.

The proper justification of (i) and (ii) are delicate problems requiring precise technical hypotheses. The reader should have no difficulty verifying that, under appropriate smoothness conditions on L (for example, if L is of class C^1). Step (i) can be rigorously justified if q_* and r are Lipschitz continuous, but cannot be justified if q_* is only known to be absolutely continuous, even if r is very nice, e.g. a polynomial.

One can also justify (IX.6) for more general Lagrangians, using the Lebesgue dominated convergence theorem. This requires integral bounds. It suffices to assume, for example, that

(H1) $L(q,u,t)$ *is of class* C^1 *with respect to* q *and* u *for each fixed* t,

(H2) $L(q,u,t)$ *is measurable with respect to* t *for each fixed* (q,u),

and

(H3) *there exist* $\alpha > 0$ *and* $\varphi \in L^1$ *such that*

$$\left\|\frac{\partial L}{\partial q}(q,u,t)\right\| + \left\|\frac{\partial L}{\partial u}(q,u,t)\right\| \leq \varphi(t) \text{ for } ||q - q_*(t)|| \leq \alpha\,,\ ||u - \dot{q}_*(t)|| \leq \alpha\,. \tag{IX.7}$$

Notice that these conditions automatically hold if L is C^1 with respect to q,u,t and q_* is Lipschitz, but need not hold, even if L is a polynomial in (q,u,t), when q_* is absolutely continuous but not Lipschitz, since in this case $\dot{q}_*$ could be unbounded and then (IX.7) does not follow.

For example, even if L is very smooth—say a polynomial in all three variables—it can happen that the differentiation under the integral sign *cannot* be justified for a solution $q_*(\cdot)$ of the minimization problem in the class AC of absolutely continuous curves and in fact *this solution does not satisfy the Euler-Lagrange equation*. This is known as the "Lavrentiev phenomenon," discovered in 1926 (cf. [88], [40] §18.5, [5]).

Step (ii)—the integration by parts—also poses a delicate problem. The simplest way to justify it is to assume that L is of class C^2 with respect to all its variables and the curve q_* is of class C^2, in which case the function $t \mapsto \frac{\partial L}{\partial \dot{q}}(q_*(t),\dot{q}_*(t),t)$ is a composite of functions of calss C^1 and hence of class C^1.

Finally, to carry out Step (iii), and infer that $M \equiv \dot{N}$ from the fact that $\int_a^b (M - \dot{N})r\,dt = 0$ "for arbitrary r," one has to be precise about the meaning of "arbitrary

r," and then one has to give a rigorous proof. For example, if "arbitrary r" means "continuously differentiable r vanishing at the boundary of the interval $[a,b]$," then one can use some result such as the "duBois-Raymond Lemma," better known today as the "fundamental lemma of the theory of distributions," according to which, if a locally integrable function $\varphi : [a,b] \mapsto \mathbb{R}$ satisfies $\int_a^b \varphi(t)\psi(t)dt = 0$ for all $\psi \in C^1([a,b],\mathbb{R})$ such that $\psi(a) = \psi(b) = 0$, then $\varphi(t) = 0$ for almost all t.

A more sophisticated analysis, using needle variations, shows that many of the above hypotheses are not needed, as will be shown below. The truly important conditions are, first of all, that L be of class C^1 with respect to q and $||\frac{\partial L}{\partial q}(q,u,t)||$ satisfy an integral bound. Rather surprisingly, *the integral bound for* $\frac{\partial L}{\partial u}(q,u,t)||$ *and even the differentiability of L with respect to u, can be dispensed with altogether,* provided that one reinterprets (V.3) appropriately, as was already explained in Chapter V. This may seem fairly surprising at first, since the elementary derivation of (V.3) that we have suggested appears to require the existence and boundedness of $\frac{\partial L}{\partial u}(q,u,t)$ in a crucial way. (Actually, it may even appear that $\frac{\partial L}{\partial u}$ itself has to be differentiable, since its derivatives are needed in order to be able to write the left-hand side of (V.3).) The true explanation of this strange fact is that Conjecture M2, in Chapter V, is true, and the statement of Conjecture M2 contains no derivatives of L with respect to u.

IX.5 Legendre

Legendre used the same formalism of variations, and derived his second-order necessary condition by looking at the "second variation" $\delta^2 I$. We now present his derivation, mainly to persuade the reader that it is not at all simple, and contains at least one argument which is rather ingenious but somewhat artificial and has a minor[5] flaw.

Let us proceed formally, using Lagrange's δ notation, and compute $\delta^2 I$, at a curve that already satisfies $\delta I = 0$, setting

$$P = \frac{\partial^2 L}{\partial q^2}\,,\ Q = \frac{\partial^2 L}{\partial q \partial \dot{q}}\,,\ R = \frac{\partial^2 L}{\partial \dot{q}^2}\,.$$

We get

$$\delta^2 I = \int_a^b (M\,\delta^2 q + N\,\delta^2 \dot{q} + P(\delta q)^2 + 2Q\,\delta q \delta \dot{q} + R(\delta \dot{q})^2)\,.$$

[5]Relative to the standards of rigor of the time. We are of course ignoring the lack of rigor of the language of infinitesimal quantities, differentials and variations.

Since $M\delta^2 q + N\delta^2\dot{q}$ is the derivative of $N\delta^2 q$ (because $M = \dot{N}$, since $\delta I = 0$), and δq vanishes at the endpoints, we get $\int_a^b (M\,\delta^2 q + N\,\delta^2\dot{q}) = 0$, so

$$\delta^2 I = \int_a^b (P(\delta q)^2 + 2Q\,\delta q\delta\dot{q} + R(\delta\dot{q})^2)\,. \tag{IX.8}$$

Legendre then had to deal with the problem of inferring that $R \geq 0$ from the fact that $\delta^2 I \geq 0$ for "all variations" δq.

Remark IX.2. Today, we would take care of this using "needle variations": take $q(t,\varepsilon) = q(t) + u_\varepsilon$, where u_ε is a continuous function supported in an interval $[t, t+2\varepsilon]$, with value $\sqrt{\varepsilon}$ at $t+\varepsilon$, and linear on $[t, t+\varepsilon]$ and on $[t+\varepsilon, t+2\varepsilon]$. An easy calculation then shows that, as $\varepsilon \downarrow 0$, $\delta^2 I$ behaves like $2R(q(t), \dot{q}(t), t)$. So $\delta^2 I \geq 0$ implies $R \geq 0$.

The point of this argument is that, by allowing u_ε to be supported in an interval of length proportional to ε, we are guaranteeing that $\dot{u}_\varepsilon$ will be much larger than u_ε as $\varepsilon \to 0$, thus making sure that the last term of the right-hand side of (IX.8) dominates the other ones. □

This method was not available to Legendre, so he used another, very ingenious trick. Observe that if α is an arbitrary function, then the integral $\int_a^b (\dot{\alpha}(\delta q)^2 + 2\alpha\,\delta q\,\delta\dot{q})$ is equal to 0, since the integrand is the derivative of $\alpha(\delta q)^2$, and $\delta q = 0$ at the endpoints. So we can rewrite (IX.8) as

$$\delta^2 I = \int_a^b ((P+\dot{\alpha})(\delta q)^2 + 2(Q+\alpha)\,\delta q\delta\dot{q} + R(\delta\dot{q})^2)\,. \tag{IX.9}$$

Now suppose we can choose α to be a solution of the differential equation

$$(P+\dot{\alpha})R = (Q+\alpha)^2\,. \tag{IX.10}$$

Then $\delta^2 I = \int_a^b R^{-1}((Q+\alpha)\delta q + R\delta\dot{q})^2$, which is < 0 if $R < 0$. So R has to be ≥ 0.

This proof is correct except for the fact, pointed out by Lagrange in 1797, that (IX.10) need not have a solution defined on the whole interval $[a,b]$. The reader will recognize (IX.10) as a Riccati differential equation.[6] It is well known that

[6]Here we see the first mention of the Riccati differential equation in relation to problems of the optimal control type. It wasn't the last! Matrix versions of this equation have since the 1960s come to play a central role in systems and control theory. Stationary points of this quadratic differential equation are determined by quadratic matrix equations, which are called Algebraic Riccati Equations. Finding conditions for the existence of real solutions of this algebraic matrix equation is like finding generalizations of the discriminant condition for quadratic polynomials. Matrix Riccati equations yield very effective algorithms for computing optimal (feedback) gains in linear quadratic control problems, in Kalman filtering, and in H^∞ problems.

such equations can have explosions in finite time. Lagrange, who used v to denote $-\alpha$, and x for the independent variable, wrote that "it would be necessary for the correctness of this result that one be able to make sure that the value of v could not become infinite for a value of x lying between the given values of a and b, which will most often be impossible to do because of the difficultly in finding the function $v(x)$."[7]

This, however, is a rather minor flaw, as was shown by Weierstrass in 1879. All one has to do to fix this aspect of Legendre's proof is to observe that given any $t \in [a,b]$ there is a small interval J containing t in its interior and such that a solution of (IX.10) exists on J. Since the restriction to J of a solution of (II.5) is a solution of the minimization problem on J, we can apply Legendre's argument to this restriction, and conclude that $R \geq 0$ on J. Since this is true on a neighborhood of every point of $[a,b]$, Legendre's conclusion follows.

IX.6 The needle variations approach

COMMENT NO. 25. **This section still has to be written.**

IX.7 The classical version of the maximum principle

The first rigorous version of the maximum principle appears in the book [112]. This "classical" version was then improved by other authors, e.g. L. Cesari [40], M. Hestenes [74], E. B. Lee and L. Markus [89], and L. D. Berkovitz [13]. We choose to quote a version appearing in Berkovitz's 1974 book.

We let $f^1,\ldots,f^n$ be the components of f, and write f^0 for L. It is assumed that the f^i, for $i=0,\ldots,m$, are defined on $Q \times U_0 \times [a,b]$, where Q, U_0 are open subsets of $\mathbb{R}^n$, $\mathbb{R}^m$, respectively. Moreover, each function $q \mapsto f^i(q,u,t)$ is required to be of class C^1 with respect to q for each $(u,t) \in U_0 \times [a,b]$, and each map $(u,t) \mapsto f^i(q,u,t)$ has to be Borel measurable for each fixed $q \in Q$. The set U is a subset of U_0. An *admissible control* is a map $[a,b] \mapsto u(t) \in U$ such that for every compact subset K of Q there is an integrable function $t \mapsto \varphi_K(t)$ such that

[7] Lagrange obviously understood that a solution could fail to exist globally on the full interval. Yet, one wonders what differences between his picture of the mathematical universe and our own might lurk behind his reference to "the difficultly in finding the function" rather than to "the possibility that the function might not exist." But what makes Lagrange's criticism of Legendre's argument truly amazing is how, at a time when the calculus was rife with confusion and poorly justified reasoning, Lagrange was able to zero in on a real mathematical difficulty (namely, that a certain initial value problem for a differential equation may fail to have a global solution), while he appeared to have no problem using arguments that he knew were unsound when working with infinitesimal quantities, cf. Remark IX.1 in page 184.

the bound $||f^i(q,u(t),t)|| + ||\frac{\partial f^i}{\partial q}(q,u(t),t)|| \leq \varphi_K(t)$ holds for all $(q,t) \in K \times [a,b]$ and all $i = 0,\ldots,m$. For a general class $\mathscr{U}$ of U-valued functions on $[a,b]$, and $\bar{q},\hat{q} \in Q$, let us use $\mathscr{C}(\mathscr{U},\bar{q},\hat{q})$ to denote the set of all pairs $(q(\cdot),u(\cdot))$ such that $u(\cdot) \in \mathscr{U}$, $q(\cdot)$ is a solution of (XVI.1) (i.e., $q(\cdot)$ is an absolutely continuous curve $[a,b] \mapsto Q$ such that (XVI.1) holds for almost every t), $q(a) = \bar{q}$, and $q(b) = \hat{q}$. Use $\mathscr{U}_{adm}$ to denote the class of all admissible controls. Then the optimization problem is that of minimizing the integral $I = \int_a^b L(q(t),u(t),t)\,dt$ in the class $\mathscr{C}(\mathscr{U}_{adm},\bar{q},\hat{q})$. The conclusion of the theorem is that of (MP), with p_* absolutely continuous, and the adjoint equation and the maximization condition holding almost everywhere.

One could also weaken this, and work with a more restricted class $\mathscr{U}$ of controls. (For example, we may want $\mathscr{U}$ to be the class of all piecewise constant maps from $[a,b]$ to U, or the class of all piecewise continuous maps.) In order to state the result in this case, let us use the notation $u\#_E v$ to denote, if u,v are maps defined on $[a,b]$, and E is a subset of $[a,b]$, the map w on $[a,b]$ defined by "substituting v for u on E," that is, $w(t) = u(t)$ for $t \notin E$, $w(t) = v(t)$ for $t \in E$.

With this terminology, the "classical" result is also valid for the minimization problem in $\mathscr{C}(\mathscr{U},\bar{q},\hat{q})$, where $\mathscr{U}$ is any class of controls such that *(a)* $\mathscr{U} \subseteq \mathscr{U}_{adm}$, *(b)* $\mathscr{U}$ *is closed under interval substitutions (that is, if* $u(\cdot) \in \mathscr{U}$, $v(\cdot) \in \mathscr{U}$, *and* J *is a subinterval of* $[a,b]$, *then* $u\#_J v \in \mathscr{U}$*), and (c) all constants belong to* $\mathscr{U}$. The conclusion is the same as in the classical result.

Remark IX.3. If $\mathscr{U} = \mathscr{U}_{adm}$, then $\mathscr{U}$ satisfies (a) and (b) but not necessarily (c). So, strictly speaking, the "classical" result does not follow from the extension discussed in the previous paragraph. It is easy, however, to find a hypothesis that includes both. What is needed is to substitute for (c) the weaker requirement (c') *for every* $u(\cdot) \in \mathscr{U}$ *and every constant map* $v : [a,b] \mapsto U$ *there exist compact subsets* E_j *of* $[a,b]$ *such that (i)* $u\#_{E_j} v \in \mathscr{U}$ *for all* j, *and (ii)* $\mathrm{meas}([a,b]\backslash E_j) \to 0$ *as* $j \to \infty$.

One can easily verify that (c') always holds when $\mathscr{U} = \mathscr{U}_{adm}$, and it is trivial that (c) implies (c'). □

Remark IX.4. The hypothesis that U is a subset of $\mathbb{R}^m$, often found in early works on the maximum principle, is not needed. One can just take U to be a completely arbitrary separable metric space, and nothing changes.

COMMENT NO. 26. Rewrite the paragraph below.

There is even a more general version where U is an arbitrary set, with no structure whatsoever, but in this case one has to be more careful about the technical hypotheses and the conclusion. Full details will be given in [140, 141, 142]. □

Remark IX.5. Notice that with the above assumptions it need not be true, for a general calculus of variations problem with a Lagrangian $L(q,u,t)$ which is very smooth—e.g. polynomial—in all three variables, that every integrable function $u : [a,b] \mapsto Q = \mathbb{R}^n$ is an admissible control. (Just consider the trivial example of $f(x,u,t) = u^2$, with $Q = U = \mathbb{R}$, $a = 0$, $b = 1$, and take $u(t) = \frac{1}{\sqrt{t}}$.) This makes it at least understandable why the "Lavrentiev phenomenon" mentioned earlier can occur. □

The proof of this first version of the maximum principle is rather long, and we will not present it in full or even sketch it here. But we will mention a couple of ideas that go into this proof and are crucial for the understanding of generalizations to be discussed later.

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

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

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

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

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

$$\Phi_{t_i,t_i+s} \Phi^{u_i}_{t_i+s,t_i}(q) = q + s\Big(f(q,u_i,t_i) - f(q,u_*(t_i),t_i)\Big) + o(s), \tag{IX.11}$$

we see that $\frac{\partial \gamma}{\partial s_i}(0)$ is obtained by solving the variational equation with initial condition $y(t_i) = f(q,u_i,t_i) - f(q,u_*(t_i),t_i)$, and evaluating the solution at time b. The justification of (IX.11) when $f(q,u_i,t)$ and $f(q,u(t),t)$ are jointly continuous in q and t is trivial. In the general case, it is a much more a delicate point, and the conclusion is not valid for all possible choices of the t_i, but only for some special t_i's. However, these t_i's form a set of full measure, so the "almost everywhere" conclusion of the theorem is still valid.

With a lot of work, these basic facts lead to the versions of (MP) described above, which are essentially those of [112] and [13].

IX.8 Recent versions of the maximum principle: the nonsmooth connection

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

IX.9 The nonsmooth analysis approach

One important improvement of the classical version resulted from the use of *nonsmooth analysis* (cf. Clarke [41, 42]). Clarke replaced the continuous differentiability of f and L with respect to q by the weaker requirement that f and L be Lipschitz with respect to q. (Other than that, the hypotheses are exactly as above. Naturally, in all the integral bounds the term $||\frac{\partial f^i}{\partial q}(q,u(t),t)||$ has to be replaced by a Lipschitz constant for $f^i(\cdot,u(t),t)$ in the appropriate set K.)

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

$$\frac{\partial H}{\partial q} = \langle p, \frac{\partial f}{\partial q}\rangle - \frac{\partial L}{\partial q}, \tag{IX.12}$$

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

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

In addition, a second direction developed in which (MP) is formulated not for controlled differential equations $\dot{q} = f(q,u,t)$, but for *differential inclusions* $\dot{q} \in F(q,t)$, where F is a set-valued map (cf. for example [42]).

IX.10 High-order necessary conditions

While the "nonsmooth" generalizations were being developed, other authors pursued a different direction, for very smooth systems. They observed that one could get stronger results by allowing a class of variations richer than that used in the classical proof. One can then obtain "high-order necessary conditions for optimality," in which the inequality $\langle p(t), v\rangle \geq 0$ must hold not only for v of the form $f(q_*(t),u_*(t),t) - f(q_*(t),u,t)$, as in (MP), but also for other vectors resulting from more elaborate variations involving Lie brackets. (We are assuming here for simplicity that $L \equiv 1$, i.e., that we are dealing with minimum-time control.) This is a vast subject, and we refer the reader to [10, 61, 64, 80] for results and surveys of work done in the 1960s, and to [1, 2, 3, 23, 82, 84] for later, more Lie-bracket oriented, work.

IX.11 The "combined systems" problem

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

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

Remark IX.6. The classical calculus of variations is in this respect exceptional and misleading, because its Hamiltonian is $\langle p,u\rangle - L(q,u,t)$, so (MC) implies that $p(t) = \frac{\partial L}{\partial u}(q_*(t),u_*(t),t)$, showing that p is unique, and then the various p_j corresponding to the I_j have to match. So in this case extremality is a local property.

□

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

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

COMMENT NO. 27. JAN: the following section was written before I found the Generalized Differential Quotients and the Path-Integral Generalized Differentials. For a skecth of where I am now, see my French paper. (Item No. 58 of the "papers avaliable in the Web" link of my Web page.

IX.12 A unified nonsmooth approach

Recently one of us (Sussmann [130, 131, 135, 137, 140, 141, 142]) has obtained a general version of (MP) that contains all the above results, applies to some new cases as well, and actually covers the "hybrid" case.

The key point of this work is that the main ideas of the classical proof of [112] apply under much more general conditions, provided only that a suitable theory of "generalized differentials" is used. This theory need not be—and, in fact, *cannot be*—that of classical differentials or that of Clarke's generalized Jacobian. What is truly needed is a definition of "generalized differentials" for which there is a good calculus (and in particular a chain rule) and a nice open mapping theorem.[8]

In [130, 131, 135, 137, 140, 141, 142] it is proposed to use the theories of "semidifferentials," "multidifferentials," and "generalized differential quotients. Basically, a continuous map f is "semidifferentiable" at a point x if it can be approximated to first order near x by a Lipschitz map. (For example, the function $f(x) = |x| + |x|^{3/2}\sin\frac{1}{x}$ is semidifferentiable at 0, because $f(x) = |x| + o(|x|)$. This function does not, however, have a Clarke generalized Jacobian at 0, because it is not Lipschitz near 0, and it is clear that it does not have a classical differential at 0 either.) This theory can also be extended to define "semidifferentiability" of certain set-valued maps, needed to handle "needle variations" when $f(q,u,t)$ is only continuous in q, so that the flow maps $\Phi^{u}_{t',t}$ are set-valued. It turns out that, using semidifferentiable maps, one can prove a general version of (MP) that contains and combines the classical, nonsmooth and Lojasiewicz versions. In [137, 140, 141, 142, 143] we propose two even more general differentiability theories—the "multidifferentials" and the "generalized differential quotients"—which make it possible to extend (MP) farther into a version that also covers the differential inclusions case.

Remark IX.7. The conventional wisdom of the subject holds that differential inclusions are more general than controlled dynamical systems, because a set-valued function usually does not have continuous selections. While this is true, A. Bressan has shown (cf. [21, 22, 24, 25, 27, 28]) that lower semicontinuous set-valued maps admit "conically continuous selections," and that these selections are as good as continuous ones as far as the properties of the corresponding flows are concerned. Adapting Bressan's theory, it is shown in [135, 137, 140] that the general differential inclusion case can be reduced to the case of a controlled dynamical system, and one ends up with a result that covers all cases. □

[8] The classical open mapping theorem says that if a continuous map F from an open subset U of a finite-dimensional real linear space X to another such space Y is such that F is differentiable at a point q and the differential $DF(q)$ is a surjective linear map, then F is *open at* q, in the sense that F maps neighborhoods of q to neighborhoods of $F(q)$. What we need for a proof of the maximum principle under general conditions is a generalization of this to a large enough class of maps, including many maps that are not necessarily differentiable in the ordinary sense, and not Lipschitz or even single-valued (since, for example, flow maps generated by continuous vector fields should be in the class, and these flow maps are in general set-valued). The main point of [130, 131, 135, 137, 140, 141, 142] is to single out a class of maps which is large enough to contain all the maps that are needed for the proof of the maximum principle, and at the same time such that a concept of "generalized differential," with good properties—in particular, a chain rule and an open mapping theorem—exists for that class.

Chapter X

The proof of the maximum principle

X.1 Introduction and overview

The purpose of this chapter is to present a detailed and completely rigorous proof of a very general version of the Maximum Principle, giving a necessary condition for a trajectory-control pair to be a minimizer for an optimal control problem with endpoint constraints.

Following the approach of the book [112], we will "geometrize" the problem of necessary conditions for optimal control, by forgetting about the optimization, and working instead with a ***separation problem***, in which we seek necessary conditions for a set $\mathscr{S}$ of possible endpoint pairs to be separated (in sense to be made precise below, cf. Def. XVIII.18) from the set $\mathbf{EPS}(\Sigma)$ of endpoint pairs arising from a given system Σ at a given.

Furthermore, the result for the separation problem will be proved by using a construction similar to that utilized in [112], based on making "needle variations" and combining them into "packets of needle variations."

Remark X.1. The idea of needle variations goes back to Weierstrass. □

Remark X.2. It would also be possible to use needle variations to give a direct proof of the maximum principle for optimal control, without going through the reduction to the separation problem, but the geometric method presented here is more elegant and leads to more general results. □

The general setting: control systems, optimal control, and separation problems. The general situation to be considered in our various versions of the maximum principle will involve

1. a ***state space*** Ω;
2. a set $\mathscr{P}$ of ***paths*** in Ω or a set $\mathscr{P}^a$ of ***augmented paths*** in Ω;
3. an ***endpoint constraint set*** $\mathscr{S}$;
4. a ***reference path*** ξ_*, or ***reference augmented path*** Ξ_*.

The precise definitions are as follows. Suppose Ω is a set.

1. A ***path*** in Ω is a function $\xi : [a_\xi, b_\xi] \mapsto \Omega$, defined on some closed bounded subinterval domain$(\xi) = [a_\xi, b_\xi]$ of $\mathbb{R}$, known as the ***domain*** of ξ.
2. An ***augmented path*** in Ω is a path in $\mathbb{R} \times \Omega$. If $\Xi : [a,b] \mapsto \mathbb{R} \times \Omega$ is an augmented path in Ω, and we write $\Xi(t) = (\xi^0(t), \xi(t))$ for $t \in [a,b]$, then the map $\xi : [a,b] \mapsto \Omega$ is the ***state space path*** associated to Ξ, and the map $\xi^0 : [a,b] \mapsto \mathbb{R}$ is the ***running cost*** for Ξ.
3. An ***endpoint constraint set*** for Ω is a subset $\mathscr{S}$ of $\Omega \times \Omega$.
4. The ***cost functional*** for a set $\mathscr{P}^a$ of augmented paths in Ω is the function $\mathscr{J}_{\mathscr{P}^a} : \mathscr{P}^a \mapsto \mathbb{R}$ that assigns to each $\Xi = (\xi^0, \xi) \in \mathscr{J}_{\mathscr{P}^a}$ the number $\mathscr{J}_{\mathscr{P}^a}(\Xi)$ given by
$$\mathscr{J}_{\mathscr{P}^a}(\Xi) \stackrel{\text{def}}{=} \xi^0(b_\Xi) - \xi^0(a_\Xi)\,. \tag{X.1}$$
5. The ***boundary*** (or ***endpoint condition***, or ***boundary value***, or ***boundary condition***) of an arc $\xi : [a,b] \mapsto \Omega$ is the ordered pair $\partial\xi \in \Omega \times \Omega$ given by
$$\partial\xi \stackrel{\text{def}}{=} (\xi(a), \xi(b))\,. \tag{X.2}$$
6. If $\mathscr{P}$ is a set of arcs in Ω, we use $\partial\mathscr{P}$ to denote the set $\{\partial\xi : \xi \in \mathscr{P}\}$ of all boundaries of all arcs $\xi \in \mathscr{P}$, and refer to $\partial\mathscr{P}$ as the ***boundary value set*** of $\partial\mathscr{P}$.
7. If $\mathscr{P}^a$ is a set of augmented arcs in Ω, we use $\partial\mathscr{P}^a$ to denote the set $\partial\mathscr{P}$, where $\mathscr{P} = \{\partial\xi : (\xi^0, \xi) \in \mathscr{P}^a\}$ for some ξ^0, so $\partial\mathscr{P}^a$ is the set of all boundaries of all state space paths of the augmented arcs $\Xi \in \mathscr{P}^a$. We refer to $\partial\mathscr{P}^a$ as the ***boundary value set*** of $\partial\mathscr{P}^a$.

8. Suppose that $\mathscr{P}^a$ is a set of augmented paths in Ω, and $\mathscr{S} \subseteq \Omega \times \Omega$ is an endpoint constraint set for Ω. A $(\mathscr{P}^a, \mathscr{S})$***-minimizer*** is an augmented path $\Xi_* = (\xi_*^0, \xi_*) \in \mathscr{P}^a$ such that $\partial \xi_* \in \mathscr{S}$ and

   (MIN) $\mathscr{J}_{\mathscr{P}^a}(\Xi_*) \leq \mathscr{J}_{\mathscr{P}^a}(\Xi)$ *for all* $\Xi = (\xi^0, \xi) \in \mathscr{P}^a$ *such that* $\partial \xi \in \mathscr{S}$,

   i.e., Ξ_* minimizes the cost $\mathscr{J}_{\mathscr{P}^a}(\Xi)$ among all augmented paths $\Xi = (\xi^0, \xi)$ of $\mathscr{P}^a$ that satisfy the boundary condition $\partial \xi \in \mathscr{S}$.

The optimal control problem. We will try to solve the following

OPTIMAL CONTROL PROBLEM (OCP)
Given a set Ω, a set $\mathscr{P}^a$ of augmented paths in Ω, an endpoint constraint set $\mathscr{S} \subseteq \Omega \times \Omega$, and a "reference augmented path" $\Xi_* \in \mathscr{P}^a$, **find**: a necessary condition for Ξ_* to be a $(\mathscr{P}^a, \mathscr{S})$-minimizer.

The separation problem. If A, B are subsets of a set Ω, we say that A is ***separated*** from B at a point $m \in M$ if $A \cap B = \{m\}$.

SEPARATION PROBLEM (SP)
Given: a set Ω, a set $\mathscr{P}$ of paths in Ω, an endpoint constraint set $\mathscr{S} \subseteq \Omega \times \Omega$, and a "reference path" $\xi_* \in \mathscr{P}$, **find**: a necessary condition for the boundary value set $\partial \mathscr{P}$ to be separated from $\mathscr{S}$ at $\partial \xi_*$.

Reduction of the optimal control problem to the separation problem. Suppose we have solved the SP for a certain class $\mathscr{C}$ of 4-tuples $(\Omega, \mathscr{P}, \mathscr{S}, \xi_*)$ of data. We will show how to derive as a corollary of this a solution of the OCP for a corresponding class $\mathscr{C}^a$ of 4-tuples $(\Omega, \mathscr{P}^a, \mathscr{S}, \Xi_*)$.

Fix data $\Omega, \mathscr{P}^a, \mathscr{S}$, and suppose that the augmented path $\Xi_* = (\xi_*^0, \xi_*)$ is a $(\mathscr{P}^a, \mathscr{S})$-minimizer. Let $\Omega^a = \mathbb{R} \times \Omega$. Then $\mathscr{P}^a$ is a set of paths in Ω^a. Write

$$\begin{array}{rclcl} a_* & = & a_{\Xi_*} \,, & \qquad & b_* = b_{\Xi_*} \,, \\ x_* & = & \xi_*(a_*) \,, & & \tilde{x}_* = \xi_*(b_*) \,, \\ x_*^0 & = & \xi_*^0(a_*) \,, & & \tilde{x}_*^0 = \xi_*^0(b_*) \,, \end{array}$$

and define

$$c_* = \tilde{x}^0_* - x^0_* = \mathscr{J}_{\mathscr{P}^a}(\Xi_*)\,.$$

Let $\mathscr{S}^a$ to be the set of all ordered pairs $\Big((x^0,x),(\tilde{x}^0,\tilde{x})\Big) \in \Omega^a \times \Omega^a$ such that

(a) $(x,\tilde{x}) \in \mathscr{S}$,

(b) $\tilde{x}^0 - x^0 \le c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2$,

(c) $x^0 = x^0_*$.

Then it is clear that $\partial\Xi_* \in \mathscr{S}^a$. Furthermore, if $Q \in \partial\mathscr{P}^a \cap \mathscr{S}^a$, then there must exist a $\Xi \in \mathscr{P}^a$ such that $Q = \partial\Xi \in \mathscr{S}^a$. If we write $a = a_\Xi$, $b = b_\Xi$, $\Xi = (\xi^0,\xi)$, $x = \xi(a)$, $\tilde{x} = \xi(b)$, $x^0 = \xi^0(a)$, $\tilde{x}^0 = \xi^0(b)$, $c = \tilde{x}^0 - x^0 = \mathscr{J}_{\mathscr{P}^a}(\Xi)$, then the fact that $\partial\Xi \in \mathscr{S}^a$ implies that $\partial\xi \in \mathscr{S}$ and

$$c \le c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2\,. \tag{X.3}$$

Since Ξ_* is a $(\mathscr{P}^a,\mathscr{S})$-minimizer, and $\partial\xi \in \mathscr{S}$, we have $\mathscr{J}_{\mathscr{P}^a}(\Xi_*) \le \mathscr{J}_{\mathscr{P}^a}(\Xi)$, i.e., $c_* \le c$. Therefore $c_* \le c \le c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2$, from which it follows that $x = x_*$, $\tilde{x} = \tilde{x}_*$, and $c = c_*$. Since $x^0 = x^0_*$, we can conclude that $\tilde{x}^0 = \tilde{x}^0_*$, so that $\partial\Xi = \partial\Xi_*$. Hence $Q = \partial\Xi_*$. So we have shown that $\partial\mathscr{P}^a \cap \mathscr{S}^a = \{\partial\Xi_*\}$. It follows that $\mathscr{P}^a$ and $\mathscr{S}^a$ are separated at $\partial\Xi_*$. If the 4-tuple $(\Omega^a,\mathscr{P}^a,\mathscr{S}^a,\Xi_*)$ belongs to the class $\mathscr{C}$ for which we have a solution of the SP, then we can apply the necessary condition for separation given by this solution, and obtain a necessary condition for minimization for the OCP.

Naturally, the class $\mathscr{C}^a$ of 4-tuples for which this reduction method produces a necessary condition for Ξ_* to be a $(\mathscr{P}^a,\mathscr{S})$-minimizer consists exactly of the 4-tuples $(\Omega,\mathscr{P}^a,\mathscr{S},\Xi_*)$ for which the 4-tuple $(\Omega^a,\mathscr{P}^a,\mathscr{S}^a,\Xi_*)$ constructed above belongs to the class $\mathscr{C}^a$. Any time a specific set of technical conditions characterizing a class $\mathscr{C}$, the above reduction will yield a corresponding class $\mathscr{C}^a$, and then the necessary condition for separation for $\mathscr{C}$ can be translated into a necessary condition for optimality for $\mathscr{C}^a$. This will be carry out in detail later, for each of our necessary conditions for separation.

X.2 Proof of a classical maximum principle

As explained in the previous section, the optimal control problem can be reduced to the separation problem. We will therefore focus our efforts from now on the SP. The first step will be to introduce some notations and review various basic mathematical concepts that will be used in the proof.

X.2.1 Notations, definitions, and background facts

Numbers, intervals, Euclidean spaces, matrices. We use $\mathbb{Z}$, $\mathbb{R}$ to denote, respectively, the set of all integers and the set of all real numbers, and write

$$\begin{array}{lllll} \mathbb{Z}_+ & = & \{n \in \mathbb{Z} : n \geq 0\}\,, & \mathbb{N} & = \{n \in \mathbb{Z} : n \geq 1\}\,, \\ \overline{\mathbb{N}} & = & \mathbb{N} \cup \{+\infty\}\,, & \mathbb{R}_+ & = \{x \in \mathbb{R} : x \geq 0\}\,, \\ \mathbb{R}_- & = & \{x \in \mathbb{R} : x \leq 0\}\,, & \overline{\mathbb{R}} & = \mathbb{R} \cup \{-\infty, +\infty\}\,, \\ \overline{\mathbb{R}}_+ & = & \mathbb{R}_+ \cup \{+\infty\}\,, & \overline{\mathbb{R}}_- & = \mathbb{R}_- \cup \{-\infty\}\,. \end{array}$$

An ***extended real interval*** is a subset I of $\overline{\mathbb{R}}$ such that whenever $a \in I$, $b \in I$, and $a < x < b$, it follows that $x \in I$. We use the notations $[a,b]$, $]a,b[$, $[a,b[$, $]a,b]$ for the closed, open, left-closed right-open and right-closed left-open intervals with endpoints a,b. (The expression (a,b) will always denote the ordered pair whose components are a and b.) A ***real interval*** is an extended real interval I such that $I \subseteq \mathbb{R}$. An extended real interval is ***nontrivial*** if it contains more than one point.

If $m,n \in \mathbb{N}$, we use $\mathbb{R}^{m\times n}$ to denote the space of all real ***matrices***

$$A = \left\{a^i_j\right\}_{1\leq i\leq m, 1\leq i\leq n} = \begin{pmatrix} a^1_1 & a^1_2 & \cdots & a^1_n \\ \vdots & \vdots & \ddots & \vdots \\ a^m_1 & a^m_2 & \cdots & a^m_n \end{pmatrix}$$

with m rows and n columns. The ***transpose*** of a matrix $A \in \mathbb{R}^{m\times n}$ is the matrix $A^\dagger \in \mathbb{R}^{n\times m}$ such that $A^\dagger = \left\{b^k_\ell\right\}_{1\leq k\leq n, 1\leq \ell\leq m}$, where $b^k_\ell = a^\ell_k$.

We use $\mathbb{R}^n$ to denote the space of all real n-dimensional ***column*** vectors

$$x = \left\{x^i\right\}_{1\leq i\leq n} = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{pmatrix} = (x^1, x^2, \cdots, x^n)^\dagger\,,$$

so $\mathbb{R}^n = \mathbb{R}^{n\times 1}$. Also, we use $\mathbb{R}_n$ to denote the space of all real n-dimensional ***covectors***—i.e. row vectors—$p = (p_1, p_2, \ldots, p_n)$, so $\mathbb{R}_n = \mathbb{R}^{1\times n}$.

The ***matrix product*** $A \cdot B$ of two matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{\mu\times\nu}$ is defined if and only if $n = \mu$, and in that case, if $A = \left\{a^i_j\right\}_{1\leq im, 1\leq j\leq n}$ and $B = \left\{b^k_\ell\right\}_{1\leq k\leq n, 1\leq \ell\leq \nu}$, we let $A \cdot B = C$, where $C \in \mathbb{R}^{m\times\nu}$ is given by

$$C = \left\{c^i_\ell\right\}_{1\leq i\leq m, 1\leq \ell\leq \nu}\,, \quad c^i_\ell = \sum_{j=1}^n a^i_j b^j_\ell\,.$$

The matrix product operation is associative: if $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{p\times q}$, then $(A\cdot B)\cdot C = A\cdot(B\cdot C)$.

If $A \in \mathbb{R}^{m\times n}$, then A induces a linear map $\Lambda_A : \mathbb{R}^n \mapsto \mathbb{R}^m$, given by $\Lambda_A(x) = A\cdot x$. It is then easy to see that if $A, B \in \mathbb{R}^{m\times n}$, then $A = B$ if and only if $\Lambda_A = \Lambda_B$, and that if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, then $\Lambda_{A\cdot B} = \Lambda_A \circ \Lambda_B$.

The ***inner product*** of two vectors

$$x = (x^1, x^2, \cdots, x^n)^\dagger \in \mathbb{R}^n\,, \qquad y = (y^1, y^2, \cdots, y^n)^\dagger \in \mathbb{R}^n\,,$$

is the number $\langle x, y\rangle = x^\dagger \cdot y$. (In other words, $\langle x, y\rangle = \sum_{j=1}^n x^j y^j$.) The ***Euclidean norm*** of a vector $x = (x^1, x^2, \cdots, x^n)^\dagger$ is the number $\|x\| = \sqrt{\langle x, x\rangle}$. (That is, $\|x\| = \sqrt{\sum_{j=1}^n (x^j)^2}$.) The inner product of two covectors $p = (p_1, p_2, \cdots, p_n) \in \mathbb{R}_n$, $q = (q_1, q_2, \cdots, q_n) \in \mathbb{R}_n$ is defined in a similar way, to be the number $\langle p, q\rangle = p\cdot q^\dagger$. The Euclidean norm of a covector $p = (p_1, p_2, \ldots, p_n)$ is the number $\|p\| = \sqrt{\langle p, p\rangle}$.

Convex sets, cones. If X is a real linear space, a subset S of X is ***convex*** if $rx + (1-r)x' \in S$ whenever $x \in S$, $x' \in S$, $r \in \mathbb{R}$, and $0 \le r \le 1$.

A ***cone*** in a real linear space X is a nonempty subset C of X such that $r\cdot v \in C$ whenever $v \in C$, $r \in \mathbb{R}$, and $r \ge 0$. It is then clear that *if C is a cone in X then automatically* $0 \in C$. (Indeed, we may pick a $v \in C$, because C is nonempty, and then $0\cdot v \in C$ as well, so $0 \in C$.)

It follows easily from the above definitions that

Fact X.3. *A subset S of a real linear space X is a convex cone if and only if S is nonempty and closed under addition (that is, such that $x + x'$ belongs to S whenever $x, x' \in S$) as well as under multiplication by nonnegative reals (that is, such that $rx \in S$ whenever $x \in S$ and $r \in \mathbb{R}_+$).* □

The following trivial observation, which will play a crucial role later, is an immediate consequence of Fact X.3.

Fact X.4. *Let C be a convex cone in a real linear space X. Then C is a linear subspace of X if and only if $c \in C \Rightarrow -c \in C$ whenever $c \in C$.*

Proof. The "only if" part is evidently true. To prove the "if" statement, assume that C is a convex cone such that $c \in C \Rightarrow -c \in C$ whenever $c \in C$. To prove that C is a linear subspace we have to show that C is closed under the operations of addition and multiplication by real scalars. Fact X.3 tells us that C is closed under the addition and multiplication by nonnegative real scalars. The assumption that

$c \in C \Rightarrow -c \in C$ whenever $c \in C$ then implies that C is closed under multiplication by -1. From this it follows easily that C is closed under multiplication by negative real scalars, completing the proof. □

□

If $\mathscr{C}$ is any set of subsets of a real linear space X, then the intersection

$$\bigcap \mathscr{C} \stackrel{\text{def}}{=} \Big\{ x \in X : x \in C \text{ whenever } C \in \mathscr{C} \Big\}$$

is a subset of X that is contained in all the members of $\mathscr{C}$, and inherits from these sets some important properties. (Notice that $\bigcap \mathscr{C} = X$ if $\mathscr{C}$ is empty.) For example, *if all the members of $\mathscr{C}$ are convex, then $\bigcap \mathscr{C}$ is convex.* Also, *if all the members of $\mathscr{C}$ are cones, then $\bigcap \mathscr{C}$ is a cone*. It follows that *if all the members of $\mathscr{C}$ are convex cones, then $\bigcap \mathscr{C}$ is a convex cone.*

In particular, given a subset S of X, we can let $\mathscr{C}^{conv}(S)$, $\mathscr{C}^{cone}(S)$, $\mathscr{C}^{convcone}(S)$, be, respectively, the set of all cones in X that contain S, the set of all convex subsets of X that contain S, and the set of all convex cones in X that contain S. Then $\bigcap \mathscr{C}^{conv}(S)$, $\bigcap \mathscr{C}^{cone}(S)$, $\bigcap \mathscr{C}^{convcone}(S)$, are, respectively, *the smallest convex subset of X containing S*, *the smallest cone in X containing S*, and *the smallest convex cone in X containing S*.

Metric spaces, neighborhoods, open and closed sets. If X is a set, a ***pseudodistance*** (or ***pseudometric***) on X is a function $d : X \times X \mapsto \mathbb{R}$ such that

(1) $d(x,x') = d(x',x)$ *whenever* $x,x' \in X$,

(2) $d(x,x'') \leq d(x,x') + d(x',x'')$ *whenever* $x,x',x'' \in X$,

(3) $d(x,x) = 0$ *whenever* $x \in X$.

A pseudodistance d on X is a ***distance*** (or ***distance function***, or ***metric***) if it satisfies

(4) *if* $x,x' \in X$ *and* $d(x,x') = 0$ *then* $x = x'$.

A ***pseudometric space*** is a set X endowed with a pseudometric on X. A ***metric space*** is a pseudometric space (X,d) such that d is a metric.

If X is a pseudometric space, we use d_X to denote the pseudodistance function on X. If $\bar{x} \in X$, we write $\mathbb{B}_X(\bar{x},r)$, $\overline{\mathbb{B}}_X(\bar{x},r)$, to denote, respectively, the ***open and closed balls with center $\bar{x}$ and radius*** r, that is, the sets

$$\mathbb{B}_X(\bar{x},r) \stackrel{\text{def}}{=} \{x \in X : d_X(x,\bar{x}) < r\}, \qquad \overline{\mathbb{B}}_X(\bar{x},r) \stackrel{\text{def}}{=} \{x \in X : d_X(x,\bar{x}) \leq r\}.$$

If X is a pseudometric space, $\bar{x} \in X$, and U is a subset U of X we say that U is a ***neighborhood*** of $\bar{x}$—and that $\bar{x}$ is an ***interior point*** of U—if $\mathbb{B}_X(\bar{x}, r) \subseteq U$ for some $r \in \mathbb{R}$ such that $r > 0$. The ***interior of*** U ***in*** X is the set $\mathrm{Int}_X(U)$ of all interior points of U. A subset U of X is ***open*** (or ***open in*** X) if $\mathrm{Int}_X(U) = U$, i.e., if U is a neighbourhood of every point of U. If $S \subseteq X$, then $\mathrm{Int}_X(\mathrm{Int}_X(S)) = \mathrm{Int}_X(S)$, so $\mathrm{Int}_X(S)$ is open in X. A subset C of X is ***closed in*** X if $X \backslash C$ is open in X. The ***closure in*** X of a subset S of X is the smallest closed subset of X that contains S, i.e., the intersection of all closed subsets of X that contain S. We use $\mathrm{Clos}_X(S)$ to denote the closure of S in X. Clearly, then, $\mathrm{Clos}_X(S) = X \backslash \Big(\mathrm{Int}(X \backslash S)\Big)$.

If S is a subset of a pseudometric space X, then S is itself a pseudometric space with the distance function $d_S : S \times S$ given by $d_S(x, x') = d_X(x, x')$ whenever $x, x' \in S$. Clearly, if f d_X is a metric then d_S is a metric as well.

If X and Y are pseudometric spaces, then the Cartesian product $X \times Y$ is a pseudometric space, equipped with the ***product pesudometric*** $d_{X \times Y}$, given by

$$d_{X\times Y}\Big((x,y),(x',y')\Big) = d_X(x,x') + d_Y(y,y') \quad \text{if} \quad x,x' \in X\,,\ y,y' \in Y\,.$$

Limits. A ***sequence*** is a mapping $\mathbf{x}$ defined on $\mathbb{N}$. When dealing with sequences, we will often use the "family of objects" notation, rather than the more common "mapping" notation, and write expressions such as $\mathbf{x} = \{x_j\}_{j\in\mathbb{N}}$, or $\mathbf{x} = \{x_j\}_{j=1}^{\infty}$, in which case it will be understood that "x_j" denotes the value at j of the map $\mathbf{x}$, i.e., the object usually referred to as $\mathbf{x}(j)$. If S is a set, then a sequence $\mathbf{x} = \{x_j\}_{j\in\mathbb{N}}$ is a ***sequence in*** S, or a ***sequence of members of*** S, if $x_j \in S$ for all $j \in \mathbb{N}$.

If $\mathbf{x} = \{x_j\}_{j\in\mathbb{N}}$ is a sequence, and σ is a strictly increasing map from $\mathbb{N}$ to $\mathbb{N}$, then the composite map $\mathbf{x} \circ \sigma$ is a sequence. Any sequence of the form $\mathbf{x} \circ \sigma$, with $\sigma : \mathbb{N} \mapsto \mathbb{N}$ strictly increasing, will be called a ***subsequence*** of $\mathbf{x}$. In that case, we will also write $\{x_{\sigma(j)}\}_{j\in\mathbb{N}}$, or even $\{x_{\sigma_j}\}_{j\in\mathbb{N}}$ since, after all, the map σ is itself a sequence, so we can write it in "family of objects" notation. If J is an infinite subset of $\mathbb{N}$, then there exists a unique strictly increasing map $\sigma : \mathbb{N} \mapsto \mathbb{N}$ such that $\sigma(\mathbb{N}) = J$. We use σ^J to denote this map, so the sequence $\sigma^J = \{\sigma^J_j\}_{j\in\mathbb{N}}$ is "the sequence of members of J, arranged in strictly increasing order." Clearly, if $\sigma : \mathbb{N} \mapsto \mathbb{N}$ is strictly increasing, then $\sigma = \sigma^J$, if we let $J = \sigma(\mathbb{N})$. It follows that a subsequence $\{x_{\sigma_j}\}_{j\in\mathbb{N}}$ of $\mathbf{x} = \{x_j\}_{j\in\mathbb{N}}$ is completely determined by specifying the set $J = \sigma(\mathbb{N})$, and any infinite subset J of $\mathbb{N}$ determines a subsequence.

If X is a pseudometric space, $\{x_j\}_{j\in\mathbb{N}}$ is a sequence in X, and $\bar{x} \in X$, we say that ***the sequence*** $\{x_j\}_{j\in\mathbb{N}}$ ***converges (or "goes") to*** $\bar{x}$, or that ***the*** x_j ***converge (or "go") to*** $\bar{x}$ ***as*** j ***goes to infinity*** (and write "$x_j \to \bar{x}$ as $j \to \infty$" or "$x_j \to \bar{x}$") if for every positive $\varepsilon \in \mathbb{R}$ there exists an $N \in \mathbb{N}$ such that $d_X(x_j,\bar{x}) < \varepsilon$ whenever $j \in \mathbb{N}$ and $j \geq N$. A sequence $\{x_j\}_{j\in\mathbb{N}}$ is ***convergent*** if there exists $\bar{x}$ such that $x_j \to \bar{x}$ as $j \to \infty$.

It is clear that if X is a *metric* space and $\{x_j\}_{j\in\mathbb{N}}$ is a convergent sequence in X, then the limit $\bar{x}$ of $\{x_j\}_{j\in\mathbb{N}}$ is unique. (The converse implication is also true: if X is a pseudometric space such that the limit of every convergent sequence$\{x_j\}_{j\in\mathbb{N}}$ is unique, then X is metric.) In that case, we also write "$\lim_{j\to\infty} x_j = \bar{x}$," or "$\lim_{j\to\infty} x_j = \bar{x}$."

The following two facts are easily verified.

Fact X.5. *If $\{x_j\}_{j\in\mathbb{N}}$ is a sequence in a pseudometric space X, and $\bar{x} \in X$, then $\{x_j\}_{j\in\mathbb{N}}$ converges to $\bar{x}$ if and only if* $\lim_{j\to\infty} d_X(x_j,\bar{x}) = 0$. ☐

Fact X.6. *If $\{x_j\}_{j\in\mathbb{N}}$ is a convergent sequence in a pseudometric space X, then $\{x_j\}_{j\in\mathbb{N}}$ has the* ***Cauchy property****:*

(CP) for every positive real number ε there exists an $N \in \mathbb{N}$ such that $d_X(x_j,x_k) < \varepsilon$ whenever $j,k \in \mathbb{N}$, $j \geq N$, and $k \geq N$. ☐

A sequence that has the Cauchy property is called a ***Cauchy sequence***. A pseudometric space X is ***complete*** if every Cauchy sequence in X is convergent.

It is easy to verify that

Fact X.7. *The closure in X of a subset S of a psedudometric space X is the set of all points $\bar{x} \in X$ such that there is a sequence in S that converges to $\bar{x}$. In particular, a subset S of a pseudometric space X is closed if and only if whenever a sequence in S converges to an $\bar{x} \in X$ it follows that $\bar{x} \in S$.* ☐

Fact X.8. *A closed subset of a complete pseudometric space is a complete pseudometric space.* ☐

If X, Y are pseudometric spaces, S is a subset of X, f is a mapping from S to Y, $\bar{x} \in X$, and $\bar{y} \in Y$, we say that $\bar{y}$ is a ***limit of*** $f(x)$ ***as*** x ***approaches*** $\bar{x}$ (or "as x approaches $\bar{x}$ along S," or "as x approaches $\bar{x}$ via S," or "as x approaches $\bar{x}$ via values in S"), and write "$f(x) \to \bar{y}$ as $x \to_S \bar{x}$") if for every positive $\varepsilon \in \mathbb{R}$ there exists a positive $\delta \in \mathbb{R}$ such that $d_Y(f(x),\bar{y}) < \varepsilon$ whenever $x \in S$ and $d_X(x,\bar{x}) < \delta$. (In all the above notations, expressions such as "along S" or "via values in S" are really unnecessary, because S is determined by f, since $S = \text{domain}(f)$. We will, however, often use such expressions to emphasize the fact that f is only intended to be defined on S.)

Remark X.9. Let $\beta : \bar{\mathbb{N}} \mapsto \mathbb{R}$ be the map given by

$$\beta(n) = 2^{-n} \quad \text{for} \quad n \in \mathbb{N}, \qquad \beta(+\infty) = 0.$$

Define $d_{\bar{\mathbb{N}}}(\mu, \nu) = |\beta(\mu) - \beta(\nu)|$ for $\mu, \nu \in \bar{\mathbb{N}}$. Then $d_{\bar{\mathbb{N}}}$ is a metric on $\bar{\mathbb{N}}$. It is then easy to verify that a sequence $\mathbf{x} = \{x_j\}_{j\in\mathbb{N}}$ of members of a pseudometric space X converges to a point $\bar{x} \in X$ if and only if $\mathbf{x}(j) \to \bar{x}$ as $j \to_{\mathbb{N}} +\infty$. So the notion of limit of a sequence is a special case of the more general concept of limit of a function along a set. □

If Y is metric, and $\bar{x} \in \mathrm{Clos}_X(S)$, then the limit $\bar{y}$ of $f(x)$ as $x \to_S \bar{x}$ is unique. (Indeed, if $f(x) \to \bar{y}_1$ as $x \to_S \bar{x}$ and $f(x) \to \bar{y}_2$ as $x \to_S \bar{x}$, then, given $\varepsilon > 0$, we can find a positive $\delta \in \mathbb{R}$ such that $d_Y(f(x), \bar{y}_1) < \varepsilon$ and $d_Y(f(x), \bar{y}_2) < \varepsilon$ whenever $x \in S$ and $d_X(x, \bar{x}) < \delta$. Since $\bar{x} \in \mathrm{Clos}_X(S)$, we can pick $x \in X$ such that $d_X(x, \bar{x}) < \delta$. Then $d_Y(\bar{y}_1, \bar{y}_2) < 2\varepsilon$. Since ε is arbitrary, $d_Y(\bar{y}_1, \bar{y}_2) = 0$. Since Y is metric, we can conclude that $\bar{y}_1 = \bar{y}_2$.) In that case we can talk about *the* limit of $f(x)$ as x approaches $\bar{x}$, and use $\lim_{x\to\bar{x}} f(x)$ (or $\lim_{x\to\bar{x}, x\in S} f(x)$) to denote this limit.

Notice that, with our definitions, if $\bar{x} \in S$ and Y is metric, then $\lim_{x\to\bar{x}} f(x) = \bar{y}$ implies that $f(\bar{x}) = \bar{y}$, i.e., that f is continuous at $\bar{x}$ along S (cf. page 208). If we wish to have $f(x)$ "go to $\bar{x}$" without allowing x to take the value $\bar{x}$, we will take the limit $\lim_{x\to\bar{x}, x\in S\setminus\{\bar{x}\}} f(x)$, for which we will also use the notation $\lim_{x\to\bar{x}, x\in S, x\neq\bar{x}} f(x)$.

It is easy to verify that

Fact X.10. *If X, Y are pseudometric spaces, $\bar{x} \in X$, $\bar{y} \in Y$, $S \subseteq X$, and $f : S \mapsto Y$, then $f(x) \to \bar{y}$ as $x \to \bar{x}$ if and only if whenever $\{x_j\}_{j\in\mathbb{N}}$ is a sequence in S that converges to $\bar{x}$ it follows that the sequence $\{f(x_j)\}_{j\in\mathbb{N}}$ converges to $\bar{y}$.* □

The contraction mapping theorem. Let X be a metric space. A mapping $f : X \mapsto X$ is said to be a ***contraction*** if there exists a real number c such that $0 \leq c < 1$ having the property that $d_X(f(x), f(x')) \leq c d_X(x, x')$ whenever $x, x' \in X$.

A ***fixed point*** of a mapping $f : X \mapsto X$ is a point $x \in X$ such that $f(x) = x$.

The following well-known theorem will play a crucial role, so we include its statement and proof.

Theorem X.11. *Assume that X is a complete metric space and $f : X \mapsto X$ is a contraction. Then f has a unique fixed point.*

Proof. Uniqueness is trivial, because if x and x' are fixed points then

$$d_X(x,x') = d_X(f(x), f(x')) \leq c d_X(x,x')\,,$$

which is only possible if $d_X(x,x')$, i.e., if $x = x'$.

To prove existence, we pick a point $x_1 \in X$, and define $x_2 = f(x_1)$, $x_3 = f(x_2)$, and, recursively, $x_k = f(x_{k-1})$ if $k \in \mathbb{N}$, $k > 1$. Then

$$d_X(x_{k+1},x_k) = d_X(f(x_k), f(x_{k-1})) \leq c d_X(x_k,x_{k-1}) \qquad \text{if} \quad k > 1\,,$$

from which it follows by a trivial induction that

$$d_X(x_{k+1},x_k) \leq c^{k-1} d_X(x_2,x_1) \qquad \text{whenever} \quad k \geq 1\,.$$

Then the triangle inequality implies that

$$\begin{aligned}
d_X(x_k,x_{k+j}) &\leq d_X(x_k,x_{k+1}) + d_X(x_{k+1},x_{k+2}) + \cdots + d_X(x_{k+j-1},x_{k+j}) \\
&\leq \Big(c^{k-1} + c^k + \cdots + c^{k+j-1}\Big) d_X(x_2,x_1) \\
&= c^{k-1}\Big(1 + c + \cdots + c^j\Big) d_X(x_2,x_1) \\
&\leq c^{k-1}\Big(\sum_{\ell=0}^{\infty} c^\ell\Big) d_X(x_2,x_1) \\
&\leq \frac{c^{k-1}}{1-c} d_X(x_2,x_1)\,,
\end{aligned}$$

whenever $k, j \in \mathbb{N}$.

Given a positive ε, we can pick $k \in \mathbb{N}$ such that $\frac{c^{k-1}}{1-c} d_X(x_2,x_1) \leq \varepsilon$, and then $d_X(x_\ell,x_s) \leq \varepsilon$ whenever $\ell \geq k$ and $s \geq k$. So the sequence $\{x_k\}_{k\in\mathbb{N}}$ satisfies the Cauchy condition. Hence $\{x_k\}_{k\in\mathbb{N}}$ converges to a point $\bar{x} \in X$. If $k \in \mathbb{N}$, then $d_X(f(\bar{x}),x_{k+1}) \leq c d_X(\bar{x},x_k)$. So

$$d_X(f(\bar{x}),\bar{x}) \leq d_X(f(\bar{x}),x_{k+1}) + d_X(x_{k+1},\bar{x}) \leq c d_X(\bar{x},x_k) + d_X(x_{k+1},\bar{x})\,.$$

Since $c d_X(\bar{x},x_k) + d_X(x_{k+1},\bar{x}) \to 0$ as $k \to \infty$ (because $\{x_k\}_{k\in\mathbb{N}}$ converges to $\bar{x}$), we have $d_X(f(\bar{x}),\bar{x}) = 0$. Hence $f(\bar{x}) = \bar{x}$, so . Hence $\bar{x}$ is a fixed point of f. □

□

Continuous maps. Suppose that X, Y are pseudometric spaces and $f : X \mapsto Y$ is a mapping. If $\bar{x} \in X$, we say that f is ***continuous at*** $\bar{x}$ if, whenever $\{x_j\}_{j\in\mathbb{N}}$ is a sequence in X that converges to $\bar{x}$ it follows that the sequence $\{f(x_j)\}_{j\in\mathbb{N}}$ converges to $f(\bar{x})$. We say that f is ***continuous*** if f is continuous at $\bar{x}$ for every $\bar{x} \in X$. If Y is metric, then it is clear that f is continuous at $\bar{x}$ if and only if $\lim_{x\to\bar{x}} f(x) = f(\bar{x})$.

It is well known, and easy to prove, that

Fact X.12. *If X and Y are pseudometric spaces and $f : X \mapsto Y$ is a mapping, then the following are equivalent:*

1. *f is continuous.*
2. *For every open subset U of Y, the set $f^{-1}(U) = \{x \in X : f(x) \in U\}$ is open in X.*
3. *For every closed subset C of Y, the set $f^{-1}(C) = \{x \in X : f(x) \in C\}$ is closed in X.* □

Compact sets. If X is a metric space, a subset K of X is said to be ***compact*** if every sequence in K has a subsequence that converges to a member of K. The following facts are well known.

Fact X.13. *If X is a metric space, then a subset K of S is compact if and only if whenever $\mathscr{U}$ is a set of open subsets of X such that $K \subseteq \bigcup \mathscr{U}$ it follows that there exists a finite subset $\mathscr{U}_0$ of $\mathscr{U}$ such that $K \subseteq \bigcup \mathscr{U}_0$.* □

Fact X.14. *If X is a finite-dimensional real normed space, then a subset K of S is compact if and only if it is closed and bounded. In particular, a nonempty real interval I is compact if and only if $I = [a,b]$ for some $a,b \in \mathbb{R}$ such that $a \leq b$.* □

Fact X.15. *If X, Y are metric spaces, X is compact, and $f : X \mapsto Y$ is continuous, then $f(X)$ is a compact subset of Y.* □

Fact X.16. *If X is a compact metric space, and $f : X \mapsto \mathbb{R}$ is continuous, then f has a maximum and a minimum on X. (That is, there exist $x_-, x_+ \in X$ such that $f(x_-) \leq f(x) \leq f(x_+)$ for all $x \in X$.)* □

Normed spaces. Let X be a real linear space. A ***norm*** on X is a function $X \ni x \mapsto \|x\| \in \mathbb{R}$ such that

1. $\|x\| \geq 0$ for all x,
2. $\|x\| = 0$ if and only if $x = 0$,
3. $\|rx\| = |r|\,\|x\|$ whenever $x \in X$ and $r \in \mathbb{R}$,
4. $\|x + x'\| \leq \|x\| + \|x'\|$ whenever x and x' belong to X.

It is clear that, if X is a real linear space endowed with a norm $\|\cdot\| : X \mapsto \mathbb{R}$, then the function $X \times X \ni (x, x') \mapsto d_X(x, x') \stackrel{\text{def}}{=} \|x - x'\| \in \mathbb{R}$ is a metric on X.

A ***real normed space*** is a real linear space endowed with a norm. Clearly, if X is a real normed space then X is a metric space if endowed with the metric d_X defined above.

A ***real Banach space*** is a real normed space such that the metric space (X, d_X) is complete.

If X, Y are real linear spaces, $L : X \mapsto Y$ is a linear map, and $x \in X$, we will use $L \cdot x$ and Lx as alternative notations for $L(x)$.

If X, Y are real normed spaces, a linear map $L : X \mapsto Y$ is ***bounded*** if there exists a real number C such that

$$\|L \cdot x\| \leq C\|x\| \quad \text{for all} \quad x \in X\,. \tag{X.4}$$

The infimum of the set of all Cs for which (X.4) holds is the ***norm*** (or ***operator norm***) of L. We use $\|L\|$, or $\|L\|_{op}$, to denote this norm. It is then easy to see that

$$\|L\|_{op} = \sup\Big\{\, \|L \cdot x\| : x \in X\,,\ \|x\| \leq 1 \,\Big\}\,. \tag{X.5}$$

Furthermore, the following are well known and easy to prove.

Fact X.17. *If X and Y are real normed spaces then*

1. *A linear map $L : X \mapsto Y$ is bounded if and only if it is continuous.*
2. *The set* $\mathrm{BLin}(X, Y)$ *of all bounded linear maps from X to Y, endowed with the operator norm, is a normed space.*
3. $\mathrm{BLin}(X, Y)$ *is a Banach space if Y is a Banach space.* □

Fact X.18. *Every finite-dimensional real normed space is complete, i.e., a Banach space.* □

Fact X.19. *If X and Y are real normed spaces and X is finite-dimensional then every linear map $L : X \mapsto Y$ is bounded.* □

The spaces $\mathbb{R}^n$, $\mathbb{R}_n$ are real normed spaces (and hence Banach spaces, by Fact X.18) with the Euclidean norm defined above. We write $\mathbb{B}^n(x,r)$, $\overline{\mathbb{B}}^n(x,r)$, rather than $\mathbb{B}_{\mathbb{R}^n}(x,r)$, $\overline{\mathbb{B}}_{\mathbb{R}^n}(x,r)$, to denote the open and closed balls in these spaces.

The matrix space $\mathbb{R}^{m\times n}$ is identified with the space $\mathrm{Lin}(\mathbb{R}^n,\mathbb{R}^m)$ of all linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$ via the map $\mathbb{R}^{m\times n} \ni A \mapsto \Lambda_A \in \mathrm{Lin}(\mathbb{R}^n,\mathbb{R}^m)$ defined above. Furthermore, $\mathrm{Lin}(\mathbb{R}^n,\mathbb{R}^m)$ coincides with $\mathrm{BLin}(\mathbb{R}^n,\mathbb{R}^m)$ by Fact X.19. Hence we can endowed $\mathbb{R}^{m\times n}$ with the operator norm, by defining $\|A\| = \|\Lambda_A\|_{op}$ (that is, $\|A\| = \max\{\|A\cdot x\| : x \in \mathbb{R}^n, \|x\| \leq 1\}$. Clearly, then, $\mathbb{R}^{m\times n}$, endowed with the operator norm, is a Banach space.

The Brouwer fixed point theorem. The following well-known and very important result, known as the ***Brouwer fixed point theorem***, will play a crucial role in our proof of the maximum principle.

COMMENT NO. 28. We must put a reference for the Brouwer fixed point theorem

Theorem X.20. *Let $n \in \mathbb{N}$, and assume that $\bar{x} \in \mathbb{R}^n$, $r \in \mathbb{R}$, and $r > 0$. Let f be a continuous map from $\bar{\mathbb{B}}^n(\bar{x},r)$ to $\bar{\mathbb{B}}^n(\bar{x},r)$. Then f has a fixed point, i.e., there exists $x \in \bar{\mathbb{B}}^n(\bar{x},r)$ such that $f(x) = x$.* □

Linear functionals, hyperplanes, duals. If X is a real linear space, then

- A linear map from X to $\mathbb{R}$ is a ***linear functional*** on X.
- A ***hyperplane*** in X is a subset H of X such that $H = \{x \in X : \lambda(x) = \alpha\}$ for some real number α and some linear map $\lambda : X \mapsto \mathbb{R}$ (called a ***defining linear functional*** for H) which does not vanish identically.

It is clear that if H is a hyperplane in X then any two defining linear functionals λ, λ', for H are scalar multiples of each other. (That is, there exists $r \in \mathbb{R}\backslash\{0\}$ such that $\lambda' = r\lambda$.)

The set $X^{\dagger,alg}$ of all linear functionals on X is the ***algebraic dual*** of X. Clearly, $X^{\dagger,alg}$ is a real linear space with the obvious definitions of addition and multiplication by scalars.

If X is normed, then we know from Fact X.17 that a linear functional $\lambda : X \mapsto \mathbb{R}$ is continuous if and only if it is bounded. We use $X^\dagger$ to denote the space of all bounded linear functionals on X, and refer to $X^\dagger$ as the ***dual*** of X. Since $\mathbb{R}$ is a Banach space, it is clear that $X^\dagger$ is a Banach space. If X is finite-dimensional then $X^{\dagger,alg} = X^\dagger$.

It is well known that

Fact X.21. *If X is a real normed space, λ belongs to $X^\dagger$, $\lambda \neq 0$, $\alpha \in \mathbb{R}$, and $H = \{x \in X : \lambda(x) = \alpha\}$, then H is closed if and only if λ is bounded.* □

Spaces of continuous functions. If X, Y are metric spaces, we use $C^0(X,Y)$ to denote the space of all continuous maps from X to Y. Given a compact subset K of X, we define the ***K-pseudodistance*** $\delta_{X,Y,K}(f,g)$ between two members f,g of $C^0(X,Y)$ to be the number

$$\delta_K(f,g) = \max\{d_X(f(x,),g(x)) : x \in K\},$$

which exists because the function $K \ni x \mapsto d_X(f(x,),g(x)) \in \mathbb{R}$ is continuous and the set K is compact. It is then clear that $\delta_{X,Y,K}$ is a pseudodistance. Furthermore, if X is itself compact then $\delta_{X,Y,X}$ is a distance, so $C^0(X,Y)$ is a metric space.

It is easy to see that *if X is compact and Y is complete then the metric space $C^0(X,Y)$, with the metric $\delta_{X,Y,X}$, is complete.*

If X is compact and Y is a normed space then the function

$$C^0(X,Y) \ni f \mapsto \|f\|_{sup} \stackrel{\text{def}}{=} \max\{\|f(x)\| : x \in K\}$$

is a norm, whose associated metric is precisely $\delta_{X,Y,X}$. Hence $C^0(X,Y)$ is a normed space. Clearly, *if Y is a Banach space and X is a compact metric space then $C^0(X,Y)$ is a Banach space.*

Vector fields. A ***vector field*** on an open subset Ω of $\mathbb{R}^n$ is a mapping $X : \Omega \mapsto \mathbb{R}^n$. Naturally, we can regard a vector field $X : \Omega \mapsto \mathbb{R}^n$ as a column $X = (X^1,\ldots,X^n)^\dagger$ of real-valued functions. Such a vector field acts as a ***differential operator*** $\varphi \mapsto X\varphi$ on functions: if $\varphi : \Omega \mapsto \mathbb{R}$ is a function, and x is a point of Ω where φ is differentiable, then $X\varphi(x)$ is the number

Hamiltonian vector fields. Assume that $n \in \mathbb{N}$ and Ω is an open subset of $\mathbb{R}^n$. Let $h : \Omega \times \mathbb{R}_n \mapsto \mathbb{R}$ be a function of class C^1. The ***Hamilton vector field***

corresponding to h is the vector field $\vec{h} : \Omega \times \mathbb{R}_n \mapsto \mathbb{R}^n \times \mathbb{R}_n$ defined by

$$\vec{h} = \sum_{k=1}^{n} \left(\frac{\partial h}{\partial p_k} \frac{\partial}{\partial x^k} - \frac{\partial h}{\partial x^k} \frac{\partial}{\partial p_k} \right),$$

where $\frac{\partial}{\partial x^k}$ is the constant vector field on $\mathbb{R}_n \mapsto \mathbb{R}^n \times \mathbb{R}_n$ whose $\mathbb{R}^n$ component is $(\delta_k^1, \ldots, \delta_k^n)^\dagger$, and whose $\mathbb{R}_n$ component is 0, and $\frac{\partial}{\partial p_k}$ is the constant vector field on $\mathbb{R}_n \mapsto \mathbb{R}^n \times \mathbb{R}_n$ whose whose $\mathbb{R}^n$ component is 0 and whose $\mathbb{R}_n$ component is $(\delta_1^k, \ldots, \delta_n^k)^\dagger$, and $\delta_i^j = 1$ if $i = j$, $\delta_i^j = 0$ if $i \neq j$.

Linear differential equations. We want to study the linear differential equation

$$\dot{x} = A(t)x, \tag{X.6}$$

under the assumption that

(X.A2.1) *$n \in \mathbb{N}$, $I = [a,b]$ is a compact real interval, and $A \in C^{0,pc}(I, \mathbb{R}^{n \times n})$.*

Definition X.22. *Assume that (X.A2.1) holds. A* ***vector solution*** *of (X.6) on I is a continuous map $\xi : I \mapsto \mathbb{R}^n$ such that $\dot{\xi}(t) = A(t)\xi(t)$ for all $t \in I$ that are points of continuity of A. A* ***matrix solution*** *of (X.6) on I is a continuous map $\Xi : I \mapsto \mathbb{R}^{n \times n}$ such that $\dot{\Xi}(t) = A(t)\Xi(t)$ for all $t \in I$ that are points of continuity of A.* □

The following existence and uniqueness theorem is well known.

Proposition X.23. *Assume that (X.A2.1) holds. Let $\bar{t} \in I$, $\bar{x} \in \mathbb{R}^n$. Then there exists one and only one vector solution ξ of (X.6) such that $\xi(\bar{t}) = \bar{x}$.*

Proof. Fix $\bar{t}$. Endow the space $\mathbb{R}^{n \times n}$ with the operator norm. Let

$$k = \sup\{\|A(t)\| : t \in I\}.$$

Let $\mathscr{V}$ be the space $C^0(I, \mathbb{R}^n)$ of all continuous maps from I to $\mathbb{R}^n$, endowed with the norm $\|\cdot\|_{\mathscr{V}}$ given by $\|\xi\|_{\mathscr{V}} = \max\{e^{-k|t-\bar{t}|}\|\xi(t)\| : t \in I\}$. Then $\mathscr{V}$ is a Banach space. Define a map $\Phi : \mathscr{V} \mapsto \mathscr{V}$ by letting

$$\Phi(\xi)(t) = \bar{x} + \int_{\bar{t}}^{t} A(s)\xi(s)\, ds.$$

It is then clear that ξ is a vector solution of (X.6) such that $\xi(\bar{t}) = \bar{x}$ if and only if $\Phi(\xi) = \xi$.

If $\xi, \xi' \in \mathscr{V}$, and we write $\zeta = \Phi(\xi)$, $\zeta' = \Phi(\xi')$, we have

$$\begin{aligned}
e^{-k|t-\bar{t}|}\|\zeta(t) - \zeta'(t)\| &= e^{-k|t-\bar{t}|}\left\|\int_{\bar{t}}^{t} A(s)(\xi(s) - \xi'(s))\,ds\right\| \\
&\leq ke^{-k|t-\bar{t}|}\int_{\min(\bar{t},t)}^{\max(\bar{t},t)} \|\xi(s) - \xi'(s)\|\,ds \\
&\leq ke^{-k|t-\bar{t}|}\int_{\min(\bar{t},t)}^{\max(\bar{t},t)} e^{k|s-\bar{t}|}e^{-k|s-\bar{t}|}\|\xi(s) - \xi'(s)\|\,ds \\
&\leq ke^{-k|t-\bar{t}|}\|\xi - \xi'\|_{\mathscr{V}}\int_{\min(\bar{t},t)}^{\max(\bar{t},t)} e^{k|s-\bar{t}|}\,ds \\
&= ke^{-k|t-\bar{t}|}\|\xi - \xi'\|_{\mathscr{V}}\left(\frac{1}{k}\left(e^{k|t-\bar{t}|} - 1\right)\right) \\
&= \|\xi - \xi'\|_{\mathscr{V}}\left(1 - e^{-k|t-\bar{t}|}\right) \\
&\leq \|\xi - \xi'\|_{\mathscr{V}}\left(1 - e^{-k(b-a)}\right) \\
&= \kappa\|\xi - \xi'\|_{\mathscr{V}},
\end{aligned}$$

where $\kappa = 1 - e^{-k(b-a)}$.

It follows that $\|\zeta - \zeta'\|_{\mathscr{V}} \leq \kappa\|\xi - \xi'\|_{\mathscr{V}}$. Hence Φ is a map from $\mathscr{V}$ to $\mathscr{V}$ that satisfies $\|\Phi(\xi) - \Phi(\xi')\|_{\mathscr{V}} \leq \kappa\|\xi - \xi'\|_{\mathscr{V}}$ whenever $\xi, \xi' \in \mathscr{V}$. Since $\kappa < 1$, the contraction mapping principle implies that there exists a unique $\xi \in \mathscr{V}$ such that $\Phi(\xi) = \xi$. Hence our statement is proved. ☐

The following result can be proved by an argument similar to the proof of Proposition X.23, using instead of $\mathscr{V}$ an appropriate space of matrix-valued functions. Alternatively, if $e_1, \ldots, e_n$ is the canonical basis of $\mathbb{R}^n$ (that is, $e_i = (\delta_i^1, \delta_i^2, \ldots, \delta_i^n)$, where $\delta_i^j = 1$ if $i = j$ and $\delta_i^j = 0$ if $i \neq j$), and $\xi_1, \ldots, \xi_n$ are the unique vector solutions of (X.6) such that $\xi_j(\bar{t}) = \bar{Z}e_j$ for $j = 1, \ldots, n$, we can define $\Xi(t) \in \mathbb{R}^{n\times n}$ by letting $\Xi(t)$ be the matrix with columns $\xi_1, \ldots, \xi_n$. It is then easy to verify that Ξ is the unique matrix solution of (X.6) such that $\Xi(\bar{t}) = \bar{Z}$.)

Proposition X.24. *Assume that (X.A2.1) holds. Let $\bar{t} \in I$, $\bar{Z} \in \mathbb{R}^{n\times n}$. Then there exists one and only one matrix solution Ξ of (X.6) such that $\Xi(\bar{t}) = \bar{Z}$.* ☐

If $s, t \in I$, then there exists a unique matrix solution Ξ of (X.6) such that $\Xi(s) = \mathrm{id}_{\mathbb{R}^n}$. (Recall that $\mathrm{id}_{\mathbb{R}^n}$ is the identity map of $\mathbb{R}^n$.) The value of this solution is a function of t, which also depends on s as a parameter, so we will use $\Xi^A(t,s)$ to denote this value. Then Ξ^A is a function from $I \times I$ to $\mathbb{R}^{n\times n}$, uniquely characterized by the integral equation

$$\Xi^A(t,s) = \mathrm{id}_{\mathbb{R}^n} + \int_s^t A(r)\Xi^A(r,s)\,dr\,. \tag{X.7}$$

It is clear that Ξ^A satisfies

$$\Xi^A(t,t) = \mathrm{id}_{\mathbb{R}^n} \qquad \text{whenever} \quad t \in I\,. \tag{X.8}$$

If we fix two members t_2, t_3 of I, then both functions $I \ni t \mapsto \Xi^A(t,t_2)\Xi^A(t_2,t_3)$ and $I \ni t \mapsto \Xi^A(t,t_3)$ are matrix solutions of (X.6) which take the value $\Xi^A(t_2,t_3)$ when $t = t_2$. Hence these two solutions coincide, from which it follows that

$$\Xi^A(t_1,t_2)\Xi^A(t_2,t_3) = \Xi^A(t_1,t_3) \qquad \text{whenever} \quad t_1,t_2,t_3 \in I\,. \tag{X.9}$$

Equations (X.8) and (X.9) are the ***flow identities***. They imply, in particular, that

$$\Xi^A(t,s) \text{ is invertible for every } t,s \tag{X.10}$$

(because of the identity $\Xi^A(t,s)\Xi^A(s,t) = \mathrm{id}_{\mathbb{R}^n}$) and that

$$\Xi^A(t,s) = \Xi^A(t,\bar t)\Xi^A(s,\bar t)^{-1} \qquad \text{whenever} \quad t,s,\bar t \in I \tag{X.11}$$

(because $\Xi^A(t,\bar t) = \Xi^A(t,s)\Xi^A(s,\bar t)$). In particular, (X.11) implies that

$$\text{the map } I \times I \ni (t,s) \mapsto \Xi^A(t,s) \in \mathbb{R}^{n\times n} \text{ is continuous}\,, \tag{X.12}$$

because $\Xi^A(t,\bar t)$ and $\Xi^A(s,\bar t)$ are continuous functions of t,s.

Definition X.25. Assume that conditions (28.1-2-3-4) hold. Then the map $\Xi^A \in C^0(I \times I, \mathbb{R}^{n\times n})$ is the ***fundamental solution*** of (X.6). □

Naturally, using the fundamental solution Ξ^A it is easy to write explicit formulas for all the vector and matrix solutions of (X.6):

(1) *The vector solution ξ of (X.6) that satisfies $\xi(\bar t) = \bar x \in \mathbb{R}^n$ is given by* $\boxed{\xi(t) = \Xi^A(t,\bar t)\bar x}$.

(2) *The matrix solution Ξ of (X.6) that satisfies $\Xi(\bar t) = \bar Z \in \mathbb{R}^{n\times n}$ is given by* $\boxed{\Xi(t) = \Xi^A(t,\bar t)\bar Z}$.

We now study how the fundamental solution Ξ^A depends on A. First of all, if A is fixed, we let $k_A = \max\{\|A(t)\| : t \in I\}$, $\kappa_A = e^{k_A(b-a)}$, and observe that the integral equation (X.7) implies the bound

$$\|\Xi^A(t,s)\| \leq 1 + k_A \int_{\min(t,s)}^{\max(t,s)} \|\Xi^A(r,s)\|\, dr,$$

and then Gronwall's inequality yields the estimate

$$\|\Xi^A(t,s)\| \leq e^{k_A|t-s|}, \tag{X.13}$$

so in particular $\|\Xi^A(t,s)\| \leq \kappa_A$.

Now suppose we are given two piecewise continuous functions $A, A' : I \mapsto \mathbb{R}^{n\times n}$. Define

$$\theta_{A,A'} = \int_a^b \|A(t) - A'(t)\|\, dt. \tag{X.14}$$

Then, if $t, s \in I$, we have

$$\begin{aligned}
&\Xi^A(t,s) - \Xi^{A'}(t,s)\\
= &\int_s^t \Big(A(r)\Xi^A(r,s) - A'(r)\Xi^{A'}(r,s)\Big)\, dr\\
= &\int_s^t \Big(A(r) - A'(r)\Big)\Xi^A(r,s)\, dr + \int_s^t A'(r)\Big(\Xi^A(r,s) - \Xi^{A'}(r,s)\Big)\, dr
\end{aligned}$$

so

$$\|\Xi^A(t,s) - \Xi^{A'}(t,s)\| \quad \leq \quad \kappa_A \theta_{A,A'} + k_{A'} \int_{\min(t,s)}^{\max(t,s)} \Big\|\Xi^A(r,s) - \Xi^{A'}(r,s)\Big\|\, dr,$$

and then Gronwall's inequality implies

$$\|\Xi^A(t,s) - \Xi^{A'}(t,s)\| \leq \kappa_A \kappa_{A'} \theta_{A,A'}. \tag{X.15}$$

The following result then follows immediately:

Proposition X.26. *Let $n \in \mathbb{N}$, $a, b \in \mathbb{R}$, $a \leq b$. Assume that S is a subset of the space $C^{0,pc}([a,b], \mathbb{R}^{n\times n})$ such that there exists a constant $k \in \mathbb{R}$ with the property that $\|A(t)\| \leq k$ whenever $A \in S$, $t \in [a,b]$. Then the fundamental solution map $S \ni A \mapsto \Xi^A \in C^0([a,b]\times[a,b], \mathbb{R}^{n\times n})$ is Lipschitz, if S is endowed with the metric d_S given by $d_S(A,A') = \theta_{A,A'}$ (cf. (X.14)). That is, there exists a constant $\bar{\kappa}$ such that $\|\Xi^A - \Xi^{A'}\|_{sup} \leq \bar{\kappa}\theta_{A,A'}$ whenever $A, A' \in S$.*

Proof. It suffices to take $\bar{\kappa} = e^{2k(b-a)}$ and use (X.15). □

The Hahn-Banach theorem. THIS SECTION HAS TO BE WRITTEN

X.2.2 The classical differential and its properties

In this subsection we assume that

(X.A3.1) *X is a normed space and S is a subset of X*,

(X.A3.2) *Y is a normed space*,

(X.A3.3) $s_* \in S$,

(X.A3.4) *L is a bounded linear map from X to Y*,

(X.A3.5) *$F : S$ı---> Y is a ppd map.*

Definition X.27. Assume that (X.A3.1) holds. The ***Bouligand tangent cone*** of S at s_* is the set $T^B_{s_*}S$ of all limits $v = \lim_{j\to\infty} \frac{s_j - s_*}{h_j}$, for all sequences $\{s_j\}_{j\in\mathbb{N}}$, $\{h_j\}_{j\in\mathbb{N}}$, such that (i) $s_j \in S$ and $h_j > 0$ for all j, and (ii) $h_j \to 0$ as $j \to \infty$. ☐

Definition X.28. Assume that (X.A3.1-2-3-4-5) hold. We say that L is a ***Fréchet differential***—or, simply, a ***differential—of** F **at** s_** if

(X.28.a) *$F(s)$ is defined for all points s that belong to some relative neighborhood $S \cap \mathbb{B}_X(s_*, \delta) = \{s \in S : \|s - s_*\| < \delta\}$ of s_* in S,*

and

(X.28.b) *the identity*

$$\lim_{s\to s_*, s\in S, s\neq s_*} \frac{\|F(s) - F(s_*) - L\cdot(s - s_*)\|}{\|s - s_*\|} = 0 \qquad \text{(X.16)}$$

holds.

We say that F is ***differentiable at*** s_* if there exists a bounded linear map $L : X \mapsto Y$ such that L is a differential of F at s_*. ☐

Clearly, if F is differentiable at s_*, then the differential L of F at s_* is uniquely determined on the closed linear subspace

$$\Sigma(S, s_*) = \mathrm{Clos}_X\left(\mathrm{lin\,span} T^B_{s_*}S\right).$$

(Indeed, if L_1, L_2 are two differentials of F at s_*, and $v \in T^B_{s_*}S$, then we can find sequences $\{s_j\}_{j=1}^\infty$ in S, $\{h_j\}_{j=1}^\infty$ in $]0,+\infty[$, such that $v = \lim_{j\to\infty} \frac{s_j - s_*}{h_j}$. Then, if $i = 1,2$, and j is large enough, we have $L_i \cdot (s_j - s_*) = F(s_j) - F(s_*) + o(\|s_j - s_*\|)$, and $s_j - s_* = h_j v + o(h_j)$, so $L_i \cdot (s_j - s_*) = F(s_j) - F(s_*) + o(h_j)$, and then

$$L_i \cdot \left(\frac{s_j - s_*}{h_j}\right) = \frac{F(s_j) - F(s_*)}{h_j} + o(1) \text{ as } j \to \infty,$$

from which it follows that

$$L_i \cdot v = \lim_{j\to\infty} \left(\frac{F(s_j) - F(s_*)}{h_j}\right).$$

Hence $L_1 \cdot v = L_2 \cdot v$. So L_1 and L_2 agree on $T^B_{S_*}S$ and then they agree on $\Sigma(S, s_*)$, since they are linear and continuous.)

We use $\mathscr{D}F(s_*)$ to denote the set of all differentials of F at s_*. Then $\mathscr{D}F(s_*)$ consists of a single linear map if and only if F is differentiable at s_* and $\Sigma(S, s_*) = X$. In that case, the unique member of $\mathscr{D}F(s_*)$ will be referred to as ***the differential of** F **at** s_**, and will be represented with the notation $DF(s_*)$.

Remark X.29. More generally, if $S \subseteq S' \subseteq X$, $s_* \in S$, and $F : S' \mapsto\!\!\!\dashrightarrow Y$, we may want to consider the "differentials of F at s_* in the direction of S." We will not, however, introduce a separate definition because, if we follow the terminology introduced earlier, differentiating F at s_* in the direction of S just amounts to differentiating at s_* the restriction $F\lceil S$ of F to S. □

The chain rule.

Proposition X.30. *Let X_1, X_2, X_3 be normed real linear spaces, let S_1, S_2 be subsets of X_1, X_2, let F_1, F_2 be ppd maps from S_1 to X_2 and from S_2 to X_3, respectively, such that $F_1(S_1) \subseteq S_2$. Let $s_* \in S_1$ be such that F_1 is differentiable at s_*, so in particular $\sigma_* \stackrel{\text{def}}{=} F_1(s_*)$ exists and belongs to S_2. Assume that F_2 is differentiable at σ_*. Let $L_1 \in \mathscr{D}F_1(s_*)$, $L_2 \in \mathscr{D}F_2(\sigma_*)$. Let $F = F_2 \circ F_1$, $L = L_2 \circ L_1$. Then $L \in \mathscr{D}F(s_*)$.*

Proof. We assume, as we clearly may without loss of generality, that $s_* = 0$, $\sigma_* = 0$. Fix positive numbers $r_1^\#$, $\bar{r}_2^\#$, such that F_i is everywhere defined on $U_i^\# \cap S_i$ for $i = 1,2$, where $U_i^\# = \mathbb{B}_{X_i}(0, r_i^\#)$. Since $\|s\|^{-1}\Big(F_1(s) - L_1(s)\Big)$ goes to zero as $s \to 0$ via values in $U_1^\# \cap S_1$, we may pick a positive $\tilde{r}_1$ such that $\tilde{r}_1 \le r_1^\#$ and $\|F_1(s) -$

$L_1(s)\| \leq \|s\|$ whenever $s \in S_1$ and $\|s\| \leq \tilde{r}_1$. Let $\kappa = 1 + \|L_1\|$. Then $\|F_1(s)\| \leq \kappa\|s\|$ whenever $s \in S_1$ and $\|s\| \leq \tilde{r}_1$. Pick a positive r_1 such that $r_1 \leq \tilde{r}_1$ and $\kappa m r_1 \leq r_2^{\#}$. Then $F_1(s) \in U_2^{\#} \cap S_2$ whenever $s \in S_1$ and $\|s\| \leq r_1$. Define $r_2 = r_2^{\#}$, and let $U_i = \mathbb{B}_{X_i}(0, r_i)$ for $i = 1,2$. Then, for $i = 1,2$, F_i is everywhere defined on $U_i \cap S_i$, and in addition $F_1(U_1 \cap S_1) \subseteq U_2 \cap S_2$.

For $i = 1,2$, $0 < r \leq r_i$, let

$$\theta_i(r) = \sup\left\{ \frac{\|F_i(s) - L_i \cdot s\|}{\|s\|} : s \in S_i, 0 < \|s\| \leq r \right\},$$

so that $\theta_i(r) \to 0$ as $r \downarrow 0$. Then, if $s \in U_1 \cap S_1$, and we write $\sigma = F_1(s)$, we have

$$\begin{aligned}
\|F(s) - L(s)\| &= \|F_2(F_1(s)) - L_2(L_1(s))\| \\
&\leq \|F_2(F_1(s)) - L_2(F_1(s))\| + \|L_2(F_1(s)) - L_2(L_1(s))\| \\
&\leq \theta_2(\|\sigma\|) + \|L_2\|\,\|F_1(s) - L_1(s)\| \\
&\leq \theta_2(\|\sigma\|)\|\sigma\| + \|L_2\|\theta_1(\|s\|)\|s\|\,.
\end{aligned}$$

On the other hand,

$$\|\sigma\| = \|F_1(s)\| \leq \|F_1(s) - L_1(s)\| + \|L_1(s)\| \leq \theta_1(\|s\|)\,\|s\| + \|L_1\|\,\|s\|\,.$$

Therefore $\|F(s) - L(s)\| \leq \Theta(\|s\|)\,\|s\|$, where

$$\Theta(r) \stackrel{\text{def}}{=} \theta_2\Big(r(\theta_1(r) + \|L_1\|)\Big)\Big(\theta_1(r) + \|L_1\|\Big) + \|L_2\|\theta_1(r)\,.$$

Clearly, $\lim_{r\downarrow 0} \Theta(r) = 0$, so $\|s\|^{-1}\|F(s) - L(s)\|$ goes to 0 as $s \to 0$ via values in S_1, and our proof is complete. □

Open mapping theorems for classical differentials. If X, Y are metric spaces, $F : X \mapsto Y$ is a map, and $x \in X$, we say that F is ***open at*** x if F maps neighborhoods of x to neighborhoods of $F(x)$ (that is, whenever U is a subset of X such that $x \in \mathrm{Int}_X U$ it follows that $F(x) \in \mathrm{Int}_X F(U)$).

Definition X.31. Let X, Y be normed real linear spaces, let $\bar{x}$, v be points of X, Y, respectively, let S be a subset of X such that $\bar{x} \in S$, and let $F : S \mapsto Y$ be a map. We say that F is ***directionally open at*** $\bar{x}$ ***along*** S ***in the direction of*** v if there exists a convex cone D in Y such that

(X.31.1) $v \in \mathrm{Int}(D)$,

(X.31.2) *for every positive number α there exists a positive number $\beta = \beta(\alpha)$ such that*

(X.31.2.i) *the image under F of the set $\{x \in S : \|x - \bar{x}\| \leq \alpha\}$ contains the set $\{F(\bar{x}) + y : y \in D, \|y\| \leq \beta\}$.*

If the function $\alpha \mapsto \beta(\alpha)$ can be chosen so that there exist positive numbers α^*, μ such that $\beta(\alpha) = \mu\alpha$ for $0 < \alpha \leq \alpha^*$, then F is said to be ***directionally open with linear rate at $\bar{x}$ along S in the direction of v.*** □

Theorem X.32. (The directional open mapping theorem for the classical differential). *Let X, Y be normed real linear spaces such that Y is finite-dimensional, let $\bar{x} \in X$, $v \in Y$, let C be a convex cone in X, and let*

$$S = \bar{x} + C = \{x \in X : x - \bar{x} \in C\}.$$

Let U be a neighborhood of $\bar{x}$ in X, let $F : U \cap S \mapsto Y$ be a continuous map, and let $L : X \mapsto Y$ be a bounded linear map such that $L \in \mathscr{D}F(\bar{s})$. Assume that $v \in \mathrm{Int}(L \cdot C)$. Then F is directionally open with linear rate at $\bar{x}$ along S in the direction of v.

Proof. We assume, without loss of generality, that $\bar{x} = 0$ and $F(\bar{x}) = 0$. We let $m = \dim Y$. If $w \in Y$ and $\rho > 0$, we use $\bar{\mathbb{B}}_Y(w,\rho)$ to denote the closed ball $\{y \in Y : \|y - w\| \leq \rho\}$. We let $(\tilde{e}_1, \dots, \tilde{e}_m)$ be a basis of Y, write $\tilde{e}_0 = -(\tilde{e}_1 + \dots + \tilde{e}_m)$, and pick an r such that $r > 0$ and $v + r\tilde{e}_j \in L \cdot C$ for every $j \in \{0, \dots, m\}$. We then let $e_j = v + r\tilde{e}_j$, and observe that the vectors $e_0, \dots, e_m$ are affinely independent, i.e., that whenever a linear combination $a_0e_0 + a_1e_1 + \dots + a_me_m$ with $a_0 + a_1 + \dots + a_m = 0$ is equal to zero, it follows that $a_0 = a_1 = \dots = a_m = 0$. (Indeed, if $a_0e_0 + a_1e_1 + \dots + a_me_m = 0$ and $a_0 + a_1 + \dots + a_m = 0$, we get $(a_0 + a_1 + \dots + a_m)v + ra_0\tilde{e}_0 + ra_1\tilde{e}_1 + \dots + ra_m\tilde{e}_m = 0$. Then $a_0\tilde{e}_0 + a_1\tilde{e}_1 + \dots + a_m\tilde{e}_m = 0$, so $(a_1 - a_0)\tilde{e}_1 + \dots + (a_m - a_0)\tilde{e}_m = 0$, which implies $a_1 - a_0 = a_2 - a_0 = \dots = a_m - a_0 = 0$, so $a_0 = a_1 = a_2 = \dots = a_m$, and then $0 = a_0 + a_1 + \dots + a_m = (m+1)a_0$, so $a_0 = a_1 = a_2 = \dots = a_m = 0$.) We then remark that every member y of Y can be expressed in a unique way as an affine combination (that is, a linear combination whose coefficients add up to 1) $a_0(y)e_0 + a_1(y)e_1 + \dots + a_m(y)e_m$ of the e_j. (Indeed, we can find $b_1, \dots, b_m$ such that $r^{-1}(y - v) = b_1\tilde{e}_1 + \dots + b_m\tilde{e}_m$, since the $\tilde{e}_j$ form a basis of Y. Then we can let $a_0 = (m+1)^{-1}(1 - b_1 - \dots - b_m)$ and $a_j = a_0 + b_j$ for $j = 1, \dots, m$. It is then easy to verify that $y = a_0e_0 + a_1e_1 + \dots + a_me_m$ and $a_0 + a_1 + \dots + a_m = 1$. Uniqueness follows from the fact that $e_0, \dots, e_m$ are affinely independent.) It is clear that the functions $Y \ni y \mapsto a_j(y) \in \mathbb{R}$ are continuous.

Since $v = (m+1)^{-1}(e_0 + \ldots + e_m)$, we have $a_j(v) = (m+1)^{-1}$ for $j = 0, \ldots, m$. Therefore we can fix a ρ such that $\rho > 0$ and $a_j(y) > 0$ for all $j \in \{0, \ldots, m\}$ and all $y \in \bar{\mathbb{B}}_Y(v, \rho)$. Let $W = \bar{\mathbb{B}}_Y(v, \rho)$. Then $W \subseteq L \cdot C$, because if $y \in \bar{\mathbb{B}}_Y(v, \rho)$ then $y = \sum_{j=0}^m a_j(y) e_j$, the e_j are in $L \cdot C$, $\sum_{j=0}^m a_j(y) = 1$, and $a_j(y) > 0$, so y is a convex combination of the e_j.

Since $e_j \in L \cdot C$, we can pick $c_j \in C$ such that $L \cdot c_j = e_j$. Let $M : Y \mapsto X$ be the map that sends $y \in Y$ to the affine combination $M \cdot y \stackrel{\text{def}}{=} \sum_{j=0}^m a_j(y) c_j$. Then M is a continuous (and in fact affine, but not necessarily linear) map, and $L \circ M = \text{identity}_Y$, because $L(M \cdot y) = \sum_{j=0}^m a_j(y) L \cdot c_j = \sum_{j=0}^m a_j(y) e_j = y$.

Let $\kappa = \sup\{\|M \cdot y\| : y \in W\}$. For $s > 0$, let

$$\theta(s) = \sup\left\{\frac{\|F(x) - L \cdot x\|}{\|x\|} : x \in C, \|x\| \leq s\right\},$$

so $\theta(s) \to 0$ as $s \downarrow 0$.

Now fix a ρ^* such that (i) $0 < \rho^* < \rho$, and (ii) if $v \neq 0$ then $\rho^* < \|v\|$. Then define $W^* = \bar{\mathbb{B}}_Y(v, \rho^*)$. Also, let δ^* be such that $\delta^* > 0$ and $\bar{\mathbb{B}}_X(0, \delta^*) \subseteq U$. Then pick a τ^* such that $0 < \tau^*$, $\kappa\tau^* \leq \delta^*$, and $\kappa\theta(\tau^*\kappa) \leq \rho - \rho^*$. Let $\alpha^* = \kappa\tau^*$, so $0 < \alpha^* \leq \delta^*$. For each positive number α, define $V_\alpha = \bar{\mathbb{B}}_X(0, \tilde{\alpha})$ and $\tau_\alpha = \frac{\tilde{\alpha}}{\kappa}$, where $\tilde{\alpha} = \min(\alpha, \alpha^*)$. Then $V_\alpha \subseteq U$, and

$$tM \cdot y \in V_\alpha \cap C \quad \text{whenever} \quad \alpha > 0, y \in W, 0 \leq t \leq \tau^*.$$

We then define

$$G_t(y) = \frac{1}{t}\Big(F(tM \cdot y)\Big) \quad \text{for} \quad y \in W,\ 0 < t \leq \tau^*.$$

Clearly, G_t is a continuous map from W to Y, if $t \in]0, \tau^*]$, and

$$\begin{aligned}
\|G_t(y) - y\| &= \frac{1}{t}\Big(\|F(tM \cdot y) - ty\|\Big) \\
&\leq \frac{1}{t}\Big(\|F(tM \cdot y) - \|t(L \circ M) \cdot y)\|\Big) \\
&= \frac{1}{t}\Big(\|F(tM \cdot y) - \|L(tM \cdot y)\|\Big) \\
&\leq \frac{1}{t}\Big(\theta(t\kappa)t\kappa\Big) \\
&= \kappa\theta(t\kappa).
\end{aligned}$$

If $y \in W^*$ and $0 < t \leq \tau^*$, we can consider the map $\mu_{t,y} : W \mapsto Y$ given by

$$\mu_{t,y}(z) = z - G_t(z) + y.$$

Then

$$\|\mu_{t,y}(z) - v\| = \|z - G_t(z) + y - v\| \leq \|z - G_t(z)\| + \|y - v\| \leq \kappa\theta(t\kappa) + \rho^* \leq \rho\,,$$

since $\kappa\theta(\tau^*\kappa) \leq \rho - \rho^*$. Therefore $\mu_{t,y}$ is a continuous map from W to W. By the Brouwer fixed point theorem, there exists, for each $y \in W^*$ and each $t \in]0,t^*]$, a $z = z(t,y) \in W$ such that $\mu_{t,y}(z) = z$. Then $z - G_t(z) + y = z$, so $G_t(z) = y$. But then $F(tM \cdot (z(t,y))) = ty$. Since this is true for every $y \in W^*$ and every $t \in]0,\tau^*]$, we have shown that

$$tW^* \subseteq F(tM(W)) \text{ whenever } 0 < t \leq \tau^*\,.$$

Furthermore, the inclusion of the previous line also holds for $t = 0$, since $F(0) = 0$. It follows that

$$\bigcup_{0 \leq t \leq \tau_\alpha} tW^* \subseteq F(V_\alpha \cap C) \text{ whenever } \alpha > 0\,.$$

Let D be the smallest convex cone containing W^*. Then v is an interior point of D. Furthermore, D is closed. (Indeed, if $v = 0$ then $D = Y$, so D is closed. Assume that $v \neq 0$. If $\{y_j\}_{j=1}^{\infty}$ is a sequence of members of D that converges to a limit y, then we can write $y_j = t_j w_j$, $t_j \in [0,\infty[$, $w_j \in W^*$. After passing to a subsequence, if necessary, we may assume that $w = \lim_{j\to\infty} w_j$ exists and belongs to W^*. Since $v \neq 0$, the inequality $\rho^* < \|v\|$ implies that $0 \notin W^*$, so w and the w_j are all $\neq 0$. Then $t_j = \frac{\|y_j\|}{\|w_j\|}$, and $\lim_{j\to\infty} t_j = t = \frac{\|y\|}{\|w\|}$. It follows that $y = tw$, so $y \in D$.)

Now fix a constant ν such that $\mathbb{B}_Y(0,1) \cap D \subseteq \bigcup_{0 \leq t \leq \nu} tW^*$. (The existence of ν is trivial. Indeed, if $v = 0$ we can pick $\nu = \frac{1}{\rho^*}$. If $v \neq 0$, then every $w \in W^*$ satisfies $\|w - v\| \geq \rho^*$, so $\|v\| \leq \rho^* + \|w\|$, and then $\|w\| \geq \|v\| - \rho^*$. Since $\rho^* < \|v\|$, we can take ν to be any nmuber such that $\nu \geq (\|v\| - \rho^*)^{-1}$, and conclude that if $y \in \mathbb{B}_Y(0,1) \cap D$, and we write $y = tw$, with $w \in W^*$, then $t = \frac{\|y\|}{\|w\|} \leq \frac{1}{\|v\| - \rho^*} \leq \nu$.) It follows that $\mathbb{B}_Y(0,s) \cap D \subseteq \bigcup_{0 \leq t \leq \nu s} tW^*$ for every positive s.

Now, if α is an arbitrary positive number, and we let $\beta(\alpha) = \nu^{-1}\tau_\alpha$, we have

$$\mathbb{B}_Y(0,\beta(\alpha)) \cap D \subseteq \bigcup_{0 \leq t \leq \nu\beta(\alpha)} tW^* = \bigcup_{0 \leq t \leq \tau_\alpha} tW^* \subseteq F(V_\alpha \cap C)\,.$$

The proof of Theorem X.32 is thus complete. $\square$

Theorem X.33. (The open mapping theorem for the classical differential). *Let X, Y be normed real linear spaces such that Y is finite-dimensional, let $\bar{x} \in X$, let U be a neighborhood of $\bar{x}$ in X, and let $F : U \mapsto Y$ be a continuous map. Assume that F is differentiable at $\bar{x}$, and let $L = DF(\bar{x})$. Assume that L is surjective. Then F is open at $\bar{x}$.*

Proof. It suffices to apply Theorem X.32 with $v = 0$, $C = X$. $\square$

Approximating cones. Assume that Y is a normed real linear space, S is a subset of Y, $s_* \in S$, and C is a convex cone in Y. We say that C is an ***approximating cone to S at*** s_* if there exist a normed space X, a neighborhood U of 0 in X, a convex cone D in X, a continuous map $F : U \cap D \mapsto Y$, and a bounded linear map $L \in \mathrm{Lin}(X,Y)$, such that $F(0) = s_*$, $F(U \cap D) \subseteq S$, $LD = C$, and $L \in \mathscr{D}F(0)$.

We say that C is a ***limiting approximating cone to S at*** s_* if there exists an increasing sequence $\{C_j\}_{j=1}^{\infty}$ of convex cones in Y such that $C = \mathrm{Clos} \bigcup_{j=1}^{\infty} C_j$ and each C_j is an approximating cone to S at s_*.

Transversality of cones. Suppose X is a real linear space. We say that two convex cones C^1, C^2 in X are ***transversal***, and write

$$C^1 \overline{\pitchfork} C^2 \ ,$$

if $C^1 - C^2 = X$. We say that C^1 and C^2 in X are ***strongly transversal***, and write

$$C^1 \overline{\pitchfork\!\!\!\!\pitchfork} C^2 \ ,$$

if $C^1 \overline{\pitchfork} C^2$ and $C^1 \cap C^2 \neq \{0\}$.

The transversal intersection theorem for the classical differential. Let us assume that Y is a normed real linear space, S^1 and S^2 are subsets of Y, and $s_* \in S^1 \cap S^2$. We say that S^1 and S^2 are ***locally separated at*** s_* if every neighborhood U of s_* contains a point s of $S^1 \cap S^2$ such that $s \neq s_*$.

We recall that if Y is a normed real linear space, then $Y^\dagger$ denotes the dual space of Y.

Definition X.34. Let Y be a normed real linear space. Let C be a convex cone in Y. The ***polar*** of C is the set

$$C^\dagger \stackrel{\mathrm{def}}{=} \{\lambda \in Y^\dagger : \lambda(c) \leq 0 \ \text{ for all } \ c \in C\} \, . \qquad \square$$

Definition X.35. *Let Y be a normed real linear space, and let C^1, C^2 be two convex cones in Y. We say that the cones C^1 and C^2 are* ***linearly separated*** *if* $(C^1)^\dagger \cap (-C_2)^\dagger \neq \{0\}$. $\square$

In other words,

Two convex cones C^1 and C^2 are linearly separated if and only if there exists a nontrivial continuous linear functional $\lambda : Y \mapsto \mathbb{R}$ such that $\langle \lambda, v \rangle \leq 0$ for all $v \in C^1$ and $\langle \lambda, v \rangle \geq 0$ for all $v \in C^2$.

Proposition X.36. *Let C^1, C^2 be convex cones in a finite-dimensional real linear space Y. Then the following three conditions are equivalent:*

(X.36.1) $C^1 \overline{\pitchfork} C^2$*;*
(X.36.2) C^1 *and* C^2 *are not linearly separated;*
(X.36.3) either
(X.36.3.a) C^1 *are* C^2 *are linear subspaces of* Y *such that* $Y = C^1 \oplus C^2$*,*
or
(X.36.3.b) $C^1 \overline{\overline{\pitchfork}} C^2$.

Proof. The equivalence of (X.36.1) and ((X.36.2) is very easy to prove: if (X.36.2) does not hold, then there exists a $\lambda \in Y^\dagger \backslash \{0\}$ such that $\langle \lambda, c \rangle \leq 0$ for all $c \in C^1$ and $\langle \lambda, c \rangle \geq 0$ for all $c \in C^2$, so that $\langle \lambda, c \rangle \leq 0$ for all $c \in C^1 - C^2$ and then $C^1 - C^2$ cannot be the whole space because λ is nontrivial, so (X.36.1) does not hold; conversely, if (X.36.1) does not hold, then the convex cone $C = C^1 - C^2$ is not the whole space, and then the closure $\bar{C} \stackrel{\text{def}}{=} \text{Clos}_Y(C)$ is not the whole space either, because Y is finite-dimensional; therefore the Hahn-Banach theorem implies the existence of a nontrivial linear functional $\lambda \in Y^\dagger$ which is nonpositive on $C^1 - C^2$; then λ is nonpositive on C^1 and nonnegative on C^2, so $\lambda \in (C^1)^\dagger \cap (-C^2)^\dagger$ and $\lambda \neq 0$, showing that (X.36.2) does not hold.

The implication (X.36.3)⇒(X.36.2) is clearly trivial. To prove that (X.36.2)⇒(X.36.3), it suffices to show that if C^1 and C^2 are transversal but not strongly transversal then (a) holds. So assume that $C^1 \overline{\pitchfork} C^2$ but C^1 and C^2 are not strongly transversal. It then follows that $C^1 \cap C^2 = \{0\}$. We will show that C^1 and C^2 are linear subspaces. This will then imply—since $C^1 - C^2 = Y$ and $C^1 \cap C^2 = \{0\}$—that $C^1 + C^2 = Y$ and $C^1 \cap C^2 = \{0\}$, so $C^1 \oplus C^2 = Y$, establishing (X.36.3.a).

To prove that C^1 and C^2 are linear subspaces, we will just do it for C^1, the other case being similar. Since C^1 is a convex cone, it suffices to show (cf. Fact X.4) that

$$c \in C^1 \Longrightarrow -c \in C^1 \qquad \text{whenever} \quad c \in C\,.$$

Pick $c \in C^1$. Using the fact that $C^1 \overline{\pitchfork} C^2$, i.e., that $C^1 - C^2 = X$, express $-c$ as a difference $-c = c^1 - c^2$, $c^i \in C^i$. Then $c^1 + c = c^2$. But $c^1 + c \in C^1$ and $c^2 \in C^2$.

Since $C^1 \cap C^2 = \{0\}$, we conclude that $c^1 + c = 0$. So $-c = c^1$, showing that $-c \in C^1$, and completing our proof. $\square$

Theorem X.37. (The transversal intersection theorem for classical differentials). *Assume that Y is a finite-dimensional real linear space, S^1 and S^2 are subsets of Y, $s_* \in S^1 \cap S^2$, and C^1, C^2 are limiting approximating cones to S^1, S^2 at s_*. Assume that the cones C^1 and C^2 are strongly transversal. Then the sets S^1 and S^2 are not locally separated at s_*.*

Proof. Let us write $C^1 = \mathrm{Clos}\,\bigcup_{j=1}^{\infty} C_j^1$, $C^2 = \mathrm{Clos}\,\bigcup_{j=1}^{\infty} C_j^2$, where each C_j^i is an approximating cone to S^i at s_* and $C_j^i \subseteq C_{j+1}^i$ for all i, j. Fix a basis $(e_1, \dots, e_m)$ of Y, and then let $e_0 = -(e_1 + \dots + e_m)$. Using the fact that $C_1 \mathrel{\overline{\pitchfork}} C_2$, find e_k^i, for $i = 1, 2$, $k = 0, \dots, m$, such that $e_k^i \in C^i$ and $e_k^1 - e_k^2 = e_k$. Then find sequences $\{e_{k,j}^i\}_{j=1}^{\infty}$, for each i and each k, such that $e_{k,j}^i \in C_j^i$ for every i, j, k, and $\lim_{j \to \infty} e_{k,j}^i = e_k^i$. Let $e_{k,j} = e_{k,j}^1 - e_{k,j}^2$. Then $\lim_{j \to \infty} e_{k,j} = e_k$ for $k = 0, \dots, m$. Since 0 is an interior point of the convex hull of $e_0, \dots, e_m$, it is clear that there exists a j^* such 0 is an interior point of the convex hull Σ_j of $e_{0,j}, \dots, e_{m,j}$ if $j \geq j^*$. Since $\Sigma_j \subseteq C_j^1 - C_j^2$, we see that $C_j^1 \mathrel{\overline{\pitchfork}} C_j^2$ if $j \geq j^*$.

We now claim that

(&) *There exists a $j^\#$ such that $j^\# \geq j^*$ and $C_j^1 \mathrel{\overline{\pitchfork\!\!\!|}} C_j^2$ whenever $j \geq j^\#$.*

If (&) was not true, there would exist an infinite subset J of $\{j \in \mathbb{N} : j \geq j^*\}$ such that C_j^1 is not strongly transversal to C_j^2 for $j \in J$. In view of Proposition X.36, since $C_j^1 \mathrel{\overline{\pitchfork}} C_j^2$ whenever $j \in J$, we can conclude that for every $j \in J$ the cones C_j^1 and C_j^2 are in fact linear subspaces and $C_j^1 \oplus C_j^2 = Y$. The fact that the sequences $\{C_j^1\}_{j=1}^{\infty}$, $\{C_j^2\}_{j=1}^{\infty}$, are increasing implies that $\bigcup_{j=1}^{\infty} C_j^1 = \bigcup_{j \in J} C_j^1$ and $\bigcup_{j=1}^{\infty} C_j^2 = \bigcup_{j \in J} C_j^2$. Furthermore, since for $i = 1, 2$ the families $\{C_j^i\}_{j \in J}$ are increasing sequences of linear subspaces, and Y is finite-dimensional, there exists a $j^\dagger$ such that $\{C_j^i\} = C_{j^\dagger}^i$ whenever $j \in J$, $j \geq j^\dagger$, $i = 1, 2$. But then

$$\bigcup_{j=1}^{\infty} C_j^1 = \bigcup_{j \in J} C_j^1 = C_{j^\dagger}^1 \qquad \text{and} \qquad \bigcup_{j=1}^{\infty} C_j^2 = \bigcup_{j \in J} C_j^2 = C_{j^\dagger}^2 .$$

It follows that $\bigcup_{j=1}^{\infty} C_j^1$ and $\bigcup_{j=1}^{\infty} C_j^2$ are linear subspaces of Y. In particular, these two sets are closed, and then $C^1 = \bigcup_{j \in J} C_j^1 = C_{j^\dagger}^1$ and $C^2 = \bigcup_{j \in J} C_j^1 = C_{j^\dagger}^2$. But then $C^1_{j^\dagger}\overline{\pitchfork}C^2_{j^\dagger}$, since $C^1\overline{\pitchfork}C^2$. This contradicts the fact that $C^1_{j^\dagger} \oplus C^2_{j^\dagger} = Y$. This contradiction proves (&).

If we pick any j such that $j \geq j^\#$, and substitute C^1_j, C^2_j for C^1, C^2, then C^1, C^2 are approximating cones to S^1, S^2 at s_*, and in addition $C^1\overline{\pitchfork}C^2$. In other words, we have shown that we may assume, without loss of generality, that

(&&) *C^1 and C^2 are approximating cones to S^1 and S^2 at s_*.*

From now on, we assume that (&&) holds, and also that $s_* = 0$.

Fix a neighborhood V of 0 in Y. For each $i \in \{1,2\}$, choose a normed space X^i, a convex cone D^i in X^i, a neighborhood U^i of 0 in X^i, a continuous map $F^i : U^i \cap D^i \mapsto Y$, and a continuous linear map $L^i : X^i \mapsto Y$, such that $C^i = L \cdot D^i$, $F(0) = 0$, $L \in \mathscr{D}F^i(0)$, and $F^i(U^i \cap D^i) \subseteq S^i$. After shrinking the U^i, if necessary, we may assume that $F^i(U^i \cap D^i) \subseteq V$ for $i = 1,2$.

Using the fact that $C^1\overline{\pitchfork}C^2$, choose a linear functional $\mu : Y \mapsto \mathbb{R}$ and a vector $\bar{y} \in C^1 \cap C^2$ such that $\mu(\bar{y}) > 0$. Let $\alpha = \mu(\bar{y})$. Write $\bar{y} = L^1\bar{v}^1 = L^2\bar{v}^2$, with $\bar{v}^1 \in D^1$, $\bar{v}^2 \in D^2$. Let $U = U^1 \times U^2$, $D = D^1 \times D^2$. Define a map $F : U \cap D \mapsto Y \times \mathbb{R}$ by letting

$$F(x^1,x^2) = (F^1(x^1) - F^2(x^2), \mu(F^1(x^1))) \text{ .if } x^1 \in U^1 \cap D^1\,,\ x^2 \in U^2 \cap D^2\,.$$

Then F is clearly continuous. A simple calculation shows that F is differentiable at $(0,0)$, and a differential of F at $(0,0)$ is the linear map $L : X^1 \times X^2 \mapsto Y \times \mathbb{R}$ given by

$$L \cdot (v^1,v^2) = (L^1 \cdot v^1 - L^2 \cdot v^2, \mu(L^1 \cdot v^1))$$

for $(v^1,v^2) \in X^1 \times X^2$. Given any $k \in \{0,\ldots,m\}$, the fact that $C^1\overline{\pitchfork}C^2$ implies that we can find $(\tilde{v}^1_k, \tilde{v}^2_k) \in D$ such that $L^1 \cdot \tilde{v}^1_k - L^2 \cdot \tilde{v}^2_k = e_k$. Then $L \cdot (\tilde{v}^1_k, \tilde{v}^2_k) = (e_k, r_k)$ for some $r_k \in \mathbb{R}$. If $r \geq r$, then we can find a pair $(v^1_k(r), v^2_k(r)) \in D$ such that $L \cdot (v^1_k(r), v^2_k(r)) = (e_k, r)$ by letting

$$v^1_k(r) = \tilde{v}^1_k + \beta\bar{v}_1\,, \qquad v^2_k = \tilde{v}^2_k + \beta\bar{v}_2\,, \qquad \beta = \frac{r - r_k}{\alpha}\,.$$

Let $\doteq \max(r_0,\ldots,r_m)$. Then we have shown that the vector (e_k, r) belongs to $L \cdot D$ if $k \in \{0,\ldots.m\}$ and $r \geq \bar{r}$. Let Σ be the convex hull of the set $\{e_0,\ldots,e_m\}$. Then $\Sigma \times [\bar{r},\infty[$ is a subset of $L \cdot D$. In particular, this implies that the vector $(0,\bar{r}+1)$ belongs to the interior $\mathrm{Int}(L \cdot D)$ of $L \cdot D$. Therefore $(0,1) \in \mathrm{Int}(L \cdot D)$ as well, and then Theorem X.32 implies that there exists a positive number ρ such that $(0,r) \in F(U \cap D)$ whenever $0 \leq r \leq \rho$. Pick any r such that $0 < r \leq \rho$.

Then $(0,r) \in F(U \cap D)$, so there exist points $x^1 \in U^1 \cap D^1$, $x^2 \in U^2 \cap D^2$, such that $F^1(x^1) - F^2(x^2) = 0$ and $\mu(F^1(x^1)) = r$. If we let $y = F^1(x^1) = F^2(x^2)$, then $y \in S^1 \cap S^2$. Since $\mu(y) = r \neq 0$, y cannot be the origin of Y. Since y belongs to$F^1(U^1 \cap D^1) \subseteq V$, y satisfies $y \in V$. Since V is arbitrary, we have shown that $S^1 \cap S^2$ contains nonzero points arbitrarily close to the origin, which is the desired conclusion. □

The linear separation theorem for the classical differential.

Theorem X.38. (The linear separation theorem for the classical differential). *Let Y be a finite-dimensional real normed space. Let S^1, S^2 be subsets of Y, let $s_* \in S^1 \cap S^2$, and let C^1, C^2 be limiting approximating cones to S^1, S^2 at s_*. Assume that the sets S^1, S^2 are locally separated at s_*. Then either*

(a) C^1 and C^2 are linear subspaces and $Y = C^1 \oplus C^2$

or

(b) C^1 and C^2 are linearly separated.

Proof. Assume that both (a) and (b) are false. Then Proposition X.36 implies that either (a) holds or $C^1 \overline{\pitchfork} C^2$. But we are assuming that (a) is false. So $C^1 \overline{\pitchfork} C^2$. Then Theorem X.38 tells us that S^1 and S^2 are not locally separated at s_*, and we have reached a contradiction. □

Compatible sets of variational vectors. If V is a subset of a real linear space Y, we use $convcone_Y(V)$ to denote the smallest convex cone in Y that contains all the members of V.

Definition X.39. Let Y be a normed space, let S be a subset of Y, let $s_* \in S$, and let $v \in Y$. We say that v is a ***variational vector*** for S at s_* if there exist a positive number ε and a continuous map $\varphi : [0,\varepsilon] \mapsto S$ such that $\lim_{t\downarrow 0}\left(\frac{\varphi(t)-s_*-tv}{t}\right) = 0$. □

We use $VV_{s_*}S$ to denote the set of all variational vectors for S at s_*. It is easy to see that *if $s_* \in S$ then the set $VV_{s_*}S$ is always a cone in Y.*

Definition X.40. Let Y be a normed space, let S be a subset of Y, let $s_* \in S$, and let $\mathscr{V} \subseteq Y$. We say that $\mathscr{V}$ is a ***compatible set of variational vectors for S at s_**** if for every finite subset $\mathscr{V}_0$ of $\mathscr{V}$ the convex cone $convcone_Y(\mathscr{V}_0)$ is an approximating cone for S at s_*. □

It is easy to see that

Fact X.41. *If $\mathscr{V}$ is a compatible set of variational vectors for S at s_* then every $v \in \mathscr{V}$ is a variational vector for S at s_*.* □

The converse, however, is clearly not true in general. Indeed, if $\mathscr{V}$ is a set of variational vectors for S at s_* then it does not follow that every convex combination of members of $\mathscr{V}$ is in $VV_{s_*}(S)$. (For example, if S is the union of the x axis and the y axis in $\mathbb{R}^2$, and $\mathscr{V}$ consists of the vectors $v = (1,0)$ and $w = (0,1)$, then $\mathscr{V} \subseteq VV_{s_*}(S)$, but the vector $\frac{1}{2}(v+w)$ is not in $VV_{s_*}(S)$.) On the other hand, for $\mathscr{V}$ to be compatible it is necessary that every convex combination of members of $\mathscr{V}$ be in $VV_{s_*}(S)$.

Proposition X.42. *Let Y be a finite-dimensional real normed space, let S be a subset of Y, and let $s_* \in S$. Let $\mathscr{V}$ be a compatible set of variational vectors for S at s_*. Then the cone* $\mathrm{Clos}\Big(convcone_Y(\mathscr{V})\Big)$ *is a limiting approximating cone for S at s_*.*

Proof. Let $C = \mathrm{Clos}\Big(convcone_Y(\mathscr{V})\Big)$. Let $\mathbb{S}_Y = \{y \in Y : \|y\| = 1\}$, so $\mathbb{S}_Y$ is a compact subset of Y. Let $K = C \cap \mathbb{S}_Y$, so K is compact as well.

For each $j \in \mathbb{N}$, let F_j be a finite subset of K such that the balls $\mathbb{B}_Y(x, 2^{-j-1})$, $x \in F_j$, cover K. Then pick, for each $x \in F_j$, a vector $v_{j,x} \in convcone_Y(\mathscr{V})$ such that $\|v_{j,x} - x\| \leq 2^{-j-1}$. Find, for each j and each x, a finite subset $E(j,x)$ of $\mathscr{V}$ such that $v_{j,x}$ is a linear combination of the members of $E(j,x)$ with positive coefficients. Let $E_j = \bigcup_{k=1}^{j} \bigcup_{x \in F_k} E(k,x)$. Let $C_j = convcone_Y(E_j)$. Then the C_j form an increasing sequence of convex cones. Clearly, $C_j \subseteq C$ for all j. Furthermore, if $x \in C$ and $j \in \mathbb{N}$, we can find $x' \in F_j$ such that $\|x - x'\| \leq 2^{-j-1}$, and then $\|x - v_{j,x'}\| \leq 2^{-j}$ and $v_{j,x'} \in C_j$. Hence $C = \mathrm{Clos} \bigcup_{j=1}^{\infty} C_j$. Finally, each C_j is an approximating cone to S at s_*, because C_j is the smallest convex cone in Y containing the finite subset E_j of $\mathscr{V}$. Hence $\mathrm{Clos}\Big(convcone_Y(\mathscr{V})\Big)$ is a limiting approximating cone for S at s_*. □

□

X.2.3 Classical control systems

Piecewise continuous controls. We recall that the concept of a "U-control" was introduced in Definition XVIII.10 in page 468.

If U is a metric space and $I = [a,b]$ is a nonempty compact interval, then a map $\eta : I \mapsto M$ is ***piecewise continuous*** if there exists a finite subset F of the interior of I such that η is continuous at every point t of $I \backslash F$, and has left and right limits $\eta(t-)$, $\eta(t+)$, at every point $t \in F$.

We use $C^{0,pc}(I,U)$ to denote the set of all piecewise continuous U-controls defined on I.

The following trivial observation will be important later.

The closure of the graph $\mathrm{Gr}(\eta) = \{(\eta(t),t) : t \in I\}$ *of a piecewise continuous U-control* $\eta : I \mapsto U$ *is a compact subset of* $U \times \mathbb{R}$.

.

Classical control systems.

Definition X.43. Let n be a positive integer. A ***classical n-dimensional control system*** is a 5-tuple $\Sigma = (\Omega, U, I, f, \mathscr{U})$ such that

(X.43-1) Ω *is an open subset of* $\mathbb{R}^n$,
(X.43-2) U *is a set,*
(X.43-3) I *is a nonempty subinterval of* $\mathbb{R}$,
(X.43-4) f *is a map from* $\Omega \times U \times I$ *to* $\mathbb{R}^n$,
(X.43-3) $\mathscr{U}$ *is a set of U-controls such that* $\mathrm{domain}(\eta) \subseteq I$ *for every* $\eta \in \mathscr{U}$.
$\square$

Definition X.44. If $n \in \mathbb{N}$ and $\Sigma = (\Omega, U, I, f, \mathscr{U})$ is a classical n-dimensional control system, we say that Σ is

- ***of class*** C^k—for a $k \in \mathbb{Z}_+$—if the map $\Omega \ni x \mapsto f(x,u,t) \in X$ is of class C^k for every $(u,t) \in U \times I$;
- ***of class*** $C^{0,0}$ if U is a metric space and f is a continuous map from $\Omega \times U \times I$ to $\mathbb{R}^n$;
- ***of class*** $C^{0,0,pc}$ if it is of class $C^{0,0}$ and in addition every $\eta \in \mathscr{U}$ is piecewise continuous;
- ***of class*** $C^{0,0,0}$ if it is of class $C^{0,0}$ and in addition every $\eta \in \mathscr{U}$ is continuous;

- ***of class*** $C^{1,0}$ if it is of class C^1 and of class $C^{0,0}$, and the map

$$\Omega \times U \times I \ni (x,u,t) \mapsto Df(x,u,t) \in \mathbb{R}^{n\times n}$$

  is continuous, where

  (i) $\mathbb{R}^{n\times n}$ is the space of $n \times n$ real matrices, identified in the standard way with the space of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^n$,

  (ii) if $x_* \in \Omega$, $u_* \in U$, and $t_* \in I$, then $Df(x_*,u_*,t_*)$ is the differential of f at (x_*,u_*,t_*) with respect to x (that is, the matrix $L \in \mathbb{R}^{n\times n}$ having the property that $f(x_*+h,u_*,t_*) - f(x_*,u_*,t_*) - L\cdot h = o(\|h\|)$ as $h \to 0$);

  .

- ***of class*** $C^{1,0,pc}$ if it is of class $C^{1,0}$ and $C^{0,0,pc}$;
- ***of class*** $C^{1,0,0}$ if it is of class $C^{1,0}$ and $C^{0,0,0}$. □

Trajectories. We introduced the concept of " Σ-trajectory" in Definition XVIII.12 (page 469) for general control systems Σ. For classical n-dimensional control systems Σ of class $C^{0,0,pc}$, the definition of " Σ-trajectory" can be reformulated without using the notions of local absolute continuity and equality almost everywhere, thanks to the following trivial fact:

Fact X.45. *If ξ, ζ are $\mathbb{R}^n$-valued functions on a compact interval $[a,b]$ such that ξ is continuous and ζ is piecewise continuous, then the following two conditions are equivalent:*

(X.45.1) the derivative $\dot\xi(t)$ of ξ at t exists and is equal to $\zeta(t)$ for all but finitely many $t \in [a,b]$;

(X.45.2) ξ is absolutely continunous and $\dot\xi(t) = \zeta(t)$ for almost every $t \in [a,b]$. □

Definition X.46. Assume that n is a nonegative integer, $\Sigma = (\Omega, U, I, f, \mathscr{U})$ is a classical n-dimensional control system of class $C^{0,0,pc}$, and $\eta \in \mathscr{U}$.

- A ***trajectory*** for η is a continuous map $\xi : I_\xi \mapsto \Omega$, defined on a nonempty subinterval I_ξ of domain(η), such that the derivative $\dot\xi(t)$ exists and is equal to $f(\xi(t),\eta(t),t)$ for all $t \in I_\xi$ that are points of continuity of η.

- A ***full trajectory*** for η is a trajectory for η whose domain is the entire interval domain(η).
- A ***trajectory-control pair*** for Σ is a pair $\gamma = (\xi,\eta)$ such that $\eta \in \mathscr{U}$ and ξ is a full trajectory for η.
- We use $TCP(\Sigma)$ to denote the set of all trajectory-control pairs (ξ,η) for the system Σ. $\square$

Remark X.47. Assume that $\Sigma = (\Omega,U,I,f,\mathscr{U})$ is a classical n-dimensional control system of class $C^{0,0,pc}$, and $\eta \in \mathscr{U}$. If $\xi : I_\xi \mapsto \Omega$ is a continuous map, then the map $I_\xi \ni t \mapsto f(\xi(t),\eta(t),t)$ is piecewise continuous. If ξ satisfies the integral equation

$$\xi(t) - \xi(s) = \int_s^t f(\xi(r),\eta(r),r)dr \quad \text{whenever} \quad s,t \in I_\xi\,, \tag{X.17}$$

(where the right-hand side is understood as a Riemann integral) then $\dot\xi(t)$ exists and equals $f(\xi(t),\eta(t),t)$ at every $t \in I_\xi$ which is a point of continuity of η. Hence ξ is a trajectory for η. Conversely, if ξ is a trajectory for η, then ξ satisfies (X.17). (Indeed, if we fix s, and define $\zeta(t) = \xi(s) + \int_s^t f(\xi(r),\eta(r),r)dr$ for $t \in I_\xi$, then the function $\xi - \zeta$ is continuous on I_ξ, and its derivative exists and is equal to $\dot\xi(t) - f(\xi(t),\eta(t),t)$—i.e., to 0—at every $t \in I_\xi$ which is a point of continuity of η. Hence $\xi - \zeta$ is constant on every subinterval of I_ξ where η is continuous. Since $\xi - \zeta$ is continuous, and η is piecewise continuous, $\xi - \zeta$ is in fact constant on I_ξ, so $\xi \equiv \zeta$, because $\xi(s) = \zeta(s)$.)

It follows that *a continuous map $\xi : I_\xi \mapsto \Omega$ defined on a nonempty subinterval I_ξ of* domain(η) *is a trajectory for the control $\eta \in \mathscr{U}$ if and only if ξ satisfies equation (X.17).* $\square$

Remark X.48. If $\Sigma = (\Omega,U,I,f,\mathscr{U})$ is a classical n-dimensional control system of class $C^{0,0}$, then for every compact subset K of $\Omega \times U \times I$ there exists a constant $c^0(K)$ such that $\|f(x,u,t)\| \le c^0(K)$ whenever $(x,u,t) \in K$. If in addition f is of class $C^{1,0}$, then there exists a constant $c^1(K)$ such that $\|Df(x,u,t)\| \le c^1(K)$ whenever (x,u,t) belongs to K. $\square$

Proposition X.49. *(Local existence of trajectories.) Let $\Sigma = (\Omega,U,I,f,\mathscr{U})$ be a classical n-dimensional control system which is of class $C^{1,0,pc}$. Then for every $\eta \in \mathscr{U}$ and every point $(\bar x,\bar t) \in \Omega \times$* domain$(\eta)$ *there exist a positive number ε and a trajectory ξ for η such that $\xi(\bar t) = \bar x$ and ξ is defined on the interval $J_{\eta,\bar t,\varepsilon}$, where*

$$J_{\eta,\bar t,\varepsilon} \stackrel{\text{def}}{=} \text{domain}(\eta) \cap [\bar t - \varepsilon, \bar t + \varepsilon]\,,$$

Proof. We fix η and $\bar{t}$, and write J_ε for $J_{\eta,\bar{t},\bar{\varepsilon}}$ is compact. Pick positive numbers $\bar{\delta}$, $\bar{\varepsilon}$ such that the ball $B \stackrel{\text{def}}{=} \bar{\mathbb{B}}^n(\bar{x},\bar{\delta})$ is contained in Ω and the set $J_{\bar{\varepsilon}}$ is compact. (This is possible, since $\bar{t} \in \text{domain}(\eta)$.) Let D be the closure of the set $\{\eta(t) : t \in J_{\bar{\varepsilon}}\}$. Then D is a compact subset of U. Let $K = B \times D \times J_{\bar{\varepsilon}}$. Let $c^0(K)$, $c^1(K)$ be the constants defined in Remark X.48.

For each ε such that $0 < \varepsilon \leq \bar{\varepsilon}$, let M_ε be the space of all continuous functions $\xi : J_\varepsilon \mapsto B$ such that $\xi(\bar{t}) = \bar{x}$. Define a map Φ_ε on M_ε as follows: if $\xi \in M_\varepsilon$, then $\Phi_\varepsilon(\xi)$ is the $\mathbb{R}^n$-valued function on J_ε given by

$$\Phi_\varepsilon(\xi)(t) = \bar{x} + \int_{\bar{t}}^{t} f(\xi(s),\eta(s),s)\,ds \qquad \text{for} \qquad t \in J_\varepsilon\,.$$

Clearly, if $\xi \in M_\varepsilon$ then $\Phi_\varepsilon(\xi)$ is a continuous function such that $\Phi_\varepsilon(\xi)(\bar{t}) = \bar{x}$. Furthermore, the bound $\|\Phi_\varepsilon(\xi)(t) - \bar{x}\| \leq \varepsilon c^0(K)$ clearly holds, implying that Φ_ε takes values in M_ε as long as $\varepsilon c^0(K) \leq \bar{\delta}$.

Define the distance $d(\xi,\zeta)$ between two members ξ,ζ of M_ε by letting

$$d(\xi,\zeta) = \max\{\|\xi(t) - \zeta(t)\| : t \in J_\varepsilon\}\,.$$

It is then clear that M_ε is a complete metric space.

If ξ,ζ belong to M_ε, and $t \in J_\varepsilon$, then

$$\Phi_\varepsilon(\xi)(t) - \Phi_\varepsilon(\zeta)(t) = \int_{\bar{t}}^{t} \Big(f(\xi(s),\eta(s),s) - f(\zeta(s),\eta(s),s) \Big)\,ds$$

so

$$\|\Phi_\varepsilon(\xi)(t) - \Phi_\varepsilon(\zeta)(t)\| \leq \varepsilon c^1(K) d(\xi,\zeta)\,,$$

since

$$\begin{aligned} &f(\xi(s),\eta(s),s) - f(\zeta(s),\eta(s),s) = \\ &\qquad \Big(\int_0^1 Df(\zeta(s) + h(\xi(s) - \zeta(s)),\eta(s),s)\,dh \Big) \cdot \Big(\xi(s) - \zeta(s) \Big)\,. \end{aligned}$$

Hence $d(\Phi_\varepsilon\xi,\Phi_\varepsilon\zeta) \leq \varepsilon c^1(K) d(\xi,\zeta)$ for all $\xi,\zeta \in M_\varepsilon$.

It follows that, if we choose ε such that $\varepsilon c^0(K) \leq \bar{\delta}$ and $\varepsilon c^1(K) < 1$, then Φ_ε is a contraction map from the complete metric space M_ε to itself. Hence Φ_ε has a fixed point by Theorem X.11. It is then clear that this fixed point is a trajectory of $\eta_{\bar{t},\varepsilon}$. □

□

Remark X.50. The above result is also valid if Σ is just of class $C^{0,0,pc}$. In that case, the proof using the contraction mapping principle does not work, and must be replaced by a more sophisticated argument, using a different fixed point theorem. Instead of M_ε, we work in the space N_ε of all Lipschitz-continuous maps $\xi : J_\varepsilon \mapsto B$ such that $\xi(\bar{t}) = \bar{x}$ and $\|\xi(t) - \xi(s)\| \leq c^0(K)|t-s|$ for all $s,t \in J_\varepsilon$. Then N_ε is a compact convex subset of the Banach space $C^0(J_\varepsilon, \mathbb{R}^)$ of all continuous functions from J_ε to $\mathbb{R}^n$. The map Φ_ε constructed above maps N_ε to N_ε if $\varepsilon c^0(K) \leq \bar{\bar{\delta}}$. Furthermore, Φ_ε is continuous. Hence Φ_ε has a fixed point by Schauder's fixed point theorem. □

In the following statements, we write

$$s \wedge t = \min(s,t)\,, \qquad s \vee t = \max(s,t)\,.$$

Proposition X.51. *(Gronwall's inequality.) Assume that α and β are nonnegative continuous functions on an interval $[a,b]$, c is a nonnegative constant, and $s \in [a,b]$ is such that*

$$\alpha(t) \leq c + \int_{s\wedge t}^{s\vee t} \beta(r)\alpha(r)\,dr \text{ for all } t \in [a,b]\,.$$

Then

$$\alpha(t) \leq c e^{\int_{s\wedge t}^{s\vee t} \beta(r)dr} \text{ for all } t \in [a,b]\,.$$

Proof. Define

$$\psi(t) = \Big(c + \int_{s\wedge t}^{s\vee t} \beta(r)\alpha(r)\,dr\Big) e^{-\int_{s\wedge t}^{s\vee t} \beta(r)dr}\,.$$

Then, if $t \geq s$, we have

$$\psi(t) = \Big(c + \int_s^t \beta(r)\alpha(r)\,dr\Big) e^{-\int_s^t \beta(r)dr}\,,$$

so

$$\begin{aligned}
\dot{\psi}(t) &= \beta(t)\alpha(t)e^{-\int_s^t \beta(r)dr} - \beta(t)\Big(c + \int_s^t \beta(r)\alpha(r)\,dr\Big)e^{-\int_s^t \beta(r)dr} \\
&= \beta(t)e^{-\int_s^t \beta(r)dr}\Big(\alpha(t) - c - \int_s^t \beta(r)\alpha(r)\,dr\Big) \\
&\leq 0\,.
\end{aligned}$$

If $t \leq s$, then

$$\psi(t) = \Big(c + \int_t^s \beta(r)\alpha(r)\,dr\Big) e^{-\int_t^s \beta(r)dr}\,,$$

so

$$\begin{aligned}\dot{\psi}(t) &= -\beta(t)\alpha(t)e^{-\int_t^s \beta(r)dr} + \beta(t)\Big(c + \int_t^s \beta(r)\alpha(r)\,dr\Big)e^{-\int_t^s \beta(r)dr} \\ &= \beta(t)e^{-\int_s^t \beta(r)dr}\Big(-\alpha(t) + c + \int_t^s \beta(r)\alpha(r)\,dr\Big) \\ &\geq 0\,.\end{aligned}$$

It follows that $\psi(t) \leq \psi(s)$ for all t. Since $\psi(s) = c$, we have

$$\alpha(t) \leq c + \int_{s\wedge t}^{s\vee t} \beta(r)\alpha(r)\,dr \leq ce^{\int_{s\wedge t}^{s\vee t}\beta(r)dr} \quad \text{for all } t \in [a,b]\,,$$

completing our proof. ☐

Continuity and convergence of controls. We would like to define what it means for a parametrized family of controls $\{\eta_p\}_{p\in P}$—where P is a metric space—to "depend continuously on p." As usual, it suffices to define what it means for a sequence $\{\eta_j\}_{j\in\mathbb{N}}$ of controls to converge to a control η.

Recall that if U is a metric space, $a,b \in \mathbb{R}$, and $a \leq b$, then $C^{0,pc}([a,b],U)$ is the set of all piecewise continuous maps $\eta : [a,b] \mapsto U$.

If U is a metric space and X is a normed space, we use $C^0(U,X)$ to denote the set of all continuous maps $\varphi : U \mapsto X$, and refer to the members of $C^0(U,X)$ as ***X-valued test functions on U***.

WARNING: The "φ-distance" d_φ really is a ***pseudodistance*** on $C^{0,pc}([a,b],U)$, rather than a distance, because it is not necessarily true that $d_\varphi(\eta_1,\eta_2) = 0$ implies $\eta_1 = \eta_2$.

If we are given two controls η_1,η_2 belonging to $C^{0,pc}([a,b],U)$, and a test function $\varphi \in C^0(U\times[a,b],X)$, we can define the "$\varphi$-distance" $d_\varphi(\eta_1,\eta_2)$ to be the number

$$d_\varphi(\eta_1,\eta_2) \stackrel{\text{def}}{=} \sup\left\{\left\|\int_a^t \Big(\varphi(\eta_1(s),s) - \varphi(\eta_2(s),s)\Big)\,ds\right\| : a \leq t \leq b\right\}.$$

Definition X.52. Assume that U is a metric space, $a,b \in \mathbb{R}$, $a \leq b$, $\{\eta_j\}_{j\in J}$ is a sequence in $C^{0,pc}([a,b],U)$, and $\eta \in C^{0,pc}([a,b],U)$. We say that the sequence $\{\eta_j\}_{j\in J}$ ***c-converges*** to η (and write $\eta_j \to_c \eta$, or $\eta = c - \lim_{j\to\infty, j\in J}\eta_j$) if

(X.52.1) there exists a compact subset U_0 of U such that $\eta_j([a,b]) \subseteq U_0$ for all $j \in J$, and

(X.52.2) $\lim_{j\to\infty, j\in J} d_\varphi(\eta_j,\eta) = 0$ for every continuous real-valued function $\varphi \in C^0(U\times[a,b],\mathbb{R})$. $\square$

Definition X.53. Assume that U and P are metric spaces, $a,b\in\mathbb{R}$, $a\le b$, and $\{\eta_p\}_{p\in P}$ is a family of members of $C^{0,pc}([a,b],U)$. We say that the family $\{\eta_p\}_{p\in P}$ ***depends c-continuously on the parameter*** p if whenever $\{p_j\}_{j\in\mathbb{N}}$ is a sequence of points of P that converges to a $p\in P$ it follows that $\eta_{p_j}\to_c \eta_p$. $\square$

Remark X.54. If a sequence $\{\eta_j\}_{j\in J}$ in $C^{0,pc}([a,b],U)$ c-converges to a limit $\eta\in C^{0,pc}([a,b],U)$, then it is clear that $d_\varphi(\eta_j,\eta)\to 0$ as $j\to_J\infty$ not only for all real-valued test functions $\varphi\in C^0(U\times[a,b],\mathbb{R})$, but also for all vector-valued test functions $\varphi\in C^0(U\times[a,b],\mathbb{R}^m)$, for every m, since in that case φ has real-valued components $\varphi^1,\dots,\varphi^m$, so $d_\varphi(\eta_j,\eta)\to 0$ if and only if $d_{\varphi^k}(\eta_j,\eta)\to 0$ for $k=1,\dots,m$. $\square$

If $n\in\mathbb{N}$ and $\Sigma=(\Omega,U,I,f,\mathscr{U})$ is an n-dimensional classical control system of class $C^{0,0}$, $a,b\in I$, $a\le b$, and $\xi:[a,b]\mapsto\Omega$ is continuous, then we can define an $\mathbb{R}^n$-valued test function $\varphi_{f,\xi}:U\times[a,b]\mapsto\mathbb{R}^n$ by letting

$$\varphi_{f,\xi}(u,t)=f(\xi(t),u,t)\ \text{for}\ u\in U\,,\ t\in[a,b]\,.$$

Proposition X.55. *Assume that $\Sigma=(\Omega,U,I,f,\mathscr{U})$ is a classical n-dimensional control system of class $C^{1,0,pc}$. Let $a,b\in I$ be such that $a\le b$. Let Q be a compact convex subset of Ω, and let U_0 be a compact subset of U. Let $K=Q\times U_0\times[a,b]$, so K is a compact subset of $\Omega\times U\times I$. Let $c^0(K)$, $c^1(K)$ be the constants defined in Remark X.48. Assume that η_1,η_2 are members of $\mathscr{U}$ that are defined on $[a,b]$, and ξ_1,ξ_2 are Q-valued full trajectories for η_1,η_2. Then the estimate*

$$\|\xi_1(t)-\xi_2(t)\|\le \tag{X.18}$$
$$e^{|t-s|c^1(K)}\left(\|\xi_1(s)-\xi_2(s)\|+2\min\left(d_{\varphi_{f,\xi_1}}(\eta_1,\eta_2),d_{\varphi_{f,\xi_2}}(\eta_1,\eta_2)\right)\right)$$

holds for all $s,t\in[a,b]$.

Proof. We write

$$\begin{aligned}\xi_1(t)-\xi_2(t) &= \xi_1(s)-\xi_2(s)+\int_s^t \Big(f(\xi_1(r),\eta_1(r),r)-f(\xi_2(r),\eta_2(r),r)\Big)\,dr\\ &= \xi_1(s)-\xi_2(s)+\int_s^t \Big(f(\xi_1(r),\eta_1(r),r)-f(\xi_1(r),\eta_2(r),r)\Big)\,dr\\ &\qquad +\int_s^t \Big(f(\xi_1(r),\eta_2(r),r)-f(\xi_2(r),\eta_2(r),r)\Big)\,dr\,.\end{aligned}$$

Clearly,

$$\Big\|\int_s^t \Big(f(\xi_1(r),\eta_1(r),r)-f(\xi_1(r),\eta_2(r),r)\Big)\,dr\Big\| \le 2d_{\varphi_{f,\xi_1}}(\eta_1,\eta_2),$$

and

$$\Big\|f(\xi_1(r),\eta_2(r),r)-f(\xi_2(r),\eta_2(r),r)\Big\| \le c^1(K)\|\xi_1(r)-\xi_2(r)\|\,,$$

because

$$\begin{aligned}&f(\xi_1(r),\eta_2(r),r)-f(\xi_2(r),\eta_2(r),r)=\\ &\qquad\Big(\int_0^1 Df(\xi_2(r)+h(\xi_1(r)-\xi_2(r)),\eta_2(r),r)\,dh\Big)\cdot\Big(\xi_1(r)-\xi_2(r)\Big)\,.\end{aligned}$$

Hence

$$\begin{aligned}\Big\|\xi_1(t)-\xi_2(t)\Big\| &\le \|\xi_1(s)-\xi_2(s)\|+2d_{\varphi_{f,\xi_1}}(\eta_1,\eta_2)\\ &\qquad +c^1(K)\int_{s\wedge t}^{s\vee t}\|\xi_1(r)-\xi_2(r)\|\,dr\,.\end{aligned}\tag{X.19}$$

It then follows from Proposition X.51 and (X.19) that

$$\|\xi_1(t)-\xi_2(t)\| \le e^{|t-s|c^1(K)}\Big(\|\xi_1(s)-\xi_2(s)\|+2d_{\varphi_{f,\xi_1}}(\eta_1,\eta_2)\Big)\,.\tag{X.20}$$

Interchanging the pairs (ξ_1,η_1) and (ξ_2,η_2) yields a similar inequality with $d_{\varphi_{f,\xi_1}}$ replaced by $d_{\varphi_{f,\xi_2}}$. Inequality (X.18) then follows. □

□

Proposition X.56. *(Uniqueness of solutions of the Cauchy problem.) Assume that $\Sigma=(\Omega,U,I,f,\mathscr{U})$ is a classical n-dimensional control system of class $C^{1,0,pc}$. Let $a,b\in I$ be such that $a\le b$, and let η be such that $\eta\in\mathscr{U}$ and* domain$(\eta)=[a,b]$. *Let ξ_1,ξ_2 be full trajectories for η such that $\xi_1(a)=\xi_2(a)$. Then $\xi_1\equiv\xi_2$.*

Proof. Let G be the set of all $t \in [a,b]$ such that $\xi_1(s) = \xi_2(s)$ for all s such that $a \le s \le t$. Then it is clear that G is a closed interval of the form $[a,c]$, with $a \le c \le b$. We prove that $c = b$. Suppose that $c < b$. Use Proposition X.55, taking $s = c$, letting Q be some closed ball centered at $\xi_1(c)$ such that $Q \subseteq \Omega$ and both $\xi_1(t)$ and $\xi_2(t)$ belong to Q for all t in some interval $[c, c+\varepsilon]$, and taking η_1, η_2 to be equal to each other and equal to the restriction of η to $[c, c+\varepsilon]$. Then Proposition X.55 implies that $\|\xi_1(t) - \xi_2(t)\| = 0$ for $t \in [c, c+\varepsilon]$. Hence $[c, c+\varepsilon] \subseteq G$, contradicting the definition of c. □

□

Proposition X.57. *(Continuous dependence of trajectories on the control and initial condition.) Assume that $\Sigma = (\Omega, U, I, f, \mathscr{U})$ is a classical n-dimensional control system of class $C^{1,0,pc}$. Let $a, b \in I$ be such that $a \le b$. Let U_0 be a compact subset of U. Let $\eta_* \in \mathscr{U}$ be such that* domain$(\eta_*) = [a,b]$ *and $\eta_*([a,b]) \subseteq U_0$, and let $\xi_* : [a,b] \mapsto \Omega$ be a full trajectory for η_*. Then for every positive number ε there exist positive numbers δ, μ, such that*

(1) $\bar{\mathbb{B}}^n(\xi_*(a), \delta) \subseteq \Omega$,

(2) if $\bar{x} \in \bar{\mathbb{B}}^n(\xi_*(a), \delta)$, and $\eta \in \mathscr{U}$ is such that domain$(\eta) = [a,b]$, $\eta([a,b]) \subseteq U_0$, *and $d_{\varphi_{f,\xi_*}}(\eta, \eta_*) \le \mu$, then*

(2.a) there exists a full trajectory ξ for η such that $\xi(a) = \bar{x}$,

(2.b) the trajectory $\xi = \xi_{\bar{x},\eta}$ described in (2.a) (which is clearly unique because of Proposition X.56) satisfies $\|\xi_{\bar{x},\eta}(t) - \xi_*(t)\| \le \varepsilon$ for all $t \in [a,b]$.

Proof. If we let $Q_0 = \{\xi_*(t) : a \le t \le b\}$, then Q_0 is a compact subset of Ω. Choose a positive number $\bar{\delta}$ such that the compact set

$$Q = \{x \in \mathbb{R}^n : \|x - \xi_*(t)\| \le \bar{\delta} \text{ for some } t \in [a,b]\}$$

is contained in Ω. Let $K = Q \times U_0 \times [a,b]$. Let $c^0(K)$, $c^1(K)$ be the constants defined in Remark X.48.

Let $\tilde{\varepsilon} = \min(\varepsilon, \bar{\delta})$. Let δ, μ be such that $0 < \delta \le \bar{\delta}$ and $\mu > 0$. Let $\bar{x} \in \Omega$ be such that $\|\bar{x} - \xi_*(a)\| \le \delta$, and let η be a U_0-valued piecewise continuous control defined on $[a,b]$ such that $d_{f,\xi_*}(\eta, \eta_*) \le \mu$. For each $\beta \in [a,b]$, let η^β be the restriction of η to the interval $[a,\beta]$. Let $S(\bar{x}, \eta)$ be the set of all $\beta \in [a,b]$ such that there exists a trajectory $\xi : [a,\beta] \mapsto \Omega$ for η^β such that $\xi(a) = \bar{x}$ and $\|\xi(t) - \xi_*(t)\| \le \tilde{\varepsilon}$ for all $t \in [a,\beta]$. Let $\bar{\beta} = \bar{\beta}(\bar{x}, \eta)$ be the supremum of all the members of $S(\bar{x}, \eta)$. Then

$\bar\beta \in S(\bar x,\eta)$. (Indeed, suppose that $\bar\beta \notin S(\bar x,\eta)$. Then there exists an increasing sequence $\{\beta_j\}_{j\in\mathbb{N}}$ of members of $S(\bar x,\eta)$ such that $\beta_j \uparrow \bar\beta$ as $j\to\infty$. For each j, we can pick a trajectory $\xi_j : [a,\beta_j] \mapsto \Omega$ such that $\xi_j(\bar t) = \bar x$ and $\|\xi_j(t) - \xi_*(t)\| \le \tilde\varepsilon$ for all $t \in [a,\beta_j]$. The uniqueness property (Proposition X.56) implies that $\xi_{j+1} \equiv \xi_j$ on $[a,\beta_j]$. Hence there exists a continuous curve $\xi : [a,\bar\beta[\mapsto \Omega$ such that $\xi \equiv \xi_j$ on $[a,\beta_j]$. Then $\dot\xi(t) = f(\xi(t),\eta(t),t)$ for all except finitely many t's. Therefore $\|\dot\xi(t)\| \le c^0(K)$ for all except finitely many t's, and then the inequality $\|\xi(t) - \xi(s)\| \le c^0(K)|t-s|$ holds whenever $t,s \in [a,\bar\beta[$. It follows that the limit $\hat x = \lim_{t\uparrow\bar\beta}\xi(t)$ exists. We can then extend ξ to the closed interval $[a,\bar\beta]$ by letting $\xi(\bar\beta) = \hat x$, and it is clear that the extended ξ is a trajectory for $\eta^{\bar\beta}$ such that $\xi(\bar t) = \bar x$ and $\|\xi(t) - \xi_*(t)\| \le \tilde\varepsilon$ for all $t \in [a,\bar\beta]$. Hence $\bar\beta \in S(\bar x,\eta)$.)

We now show that $\bar\beta(\bar x,\eta) = b$ if δ, μ are sufficiently small. To see this, we let $\xi_{\bar x,\eta}$ be the trajectory ξ for $\eta^{\bar\beta(\bar x,\eta)}$ such that $\xi(\bar t) = \bar x$. (This trajectory exists and satisfies $\|\xi(t) - \xi_*(t)\| \le \tilde\varepsilon$ for all $t \in [a,\bar\beta(\bar x,\eta)]$ because $\bar\beta(\bar x,\eta) \in S(\bar x,\eta)$, and is unique by Proposition X.56.) Then Proposition X.55 (applied with $s = a$) yields the bound

$$\|\xi_{\bar x,\eta}(t) - \xi_*(t)\| \le e^{(b-a)c^1(K)}\Big(\delta + 2d_{f,\xi_*}(\eta,\eta_*)\Big) \quad \text{for all} \quad t \in [a,\bar\beta(\bar x,\eta)]\,.$$

Now choose

$$\delta = \mu = \frac{1}{4}e^{-(b-a)c^1(K)}\tilde\varepsilon\,. \tag{X.21}$$

Then

$$\|\xi_{\bar x,\eta}(t) - \xi_*(t)\| \le \frac{3\tilde\varepsilon}{4} \quad \text{for all} \quad t \in [a,\bar\beta(\bar x,\eta)] \tag{X.22}$$

provided that

$$\eta \in C^{0,pc}([a,b],U_0)\,, \qquad \|\bar x - \xi_*(a)\| \le \delta\,, \qquad \text{and} qquad d_{f,\xi_*}(\eta,\eta_*) \le \mu\,. \tag{X.23}$$

It follows that (Indeed, suppose that (X.23) holds and $\bar\beta(\bar x,\eta) < b$. Let $\bar\beta = \bar\beta(\bar x,\eta)$. Then (X.22) implies that $\|\xi_{\bar x,\eta}(\bar\beta) - \xi_*(\bar\beta)\| \le \frac{3\tilde\varepsilon}{4}$. But then $\|\xi_{\bar x,\eta}(\bar\beta) - \xi_*(\bar\beta)\| < \tilde\varepsilon$. Proposition X.49 implies that there exists a positive ρ such that $\bar\beta + \rho \le b$, having the property that on the interval $[\bar\beta,\bar\beta+\rho]$ there is a trajectory ξ for the control $\eta\lceil[\bar\beta,\bar\beta+\rho]$ such that $\xi(\bar\beta) = \xi_{\bar x,\eta}(\bar\beta)$. Since $\|\xi_{\bar x,\eta}(\bar\beta) - \xi_*(\bar\beta)\| < \tilde\varepsilon$ we may assume, after making ρ smaller, if necessary, that $\|\xi_{\bar x,\eta}(t) - \xi_*(t)\| < \tilde\varepsilon$ for all $t \in [\bar\beta,\bar\beta+\rho]$. If we then define a map $\zeta : [a,\bar\beta+\rho] \mapsto \Omega$ by letting $\zeta(t) = \xi_{\bar x,\eta}(t)$ for $t \in [a,\bar\beta]$ and $\zeta(t) = \xi(t)$ for $t \in [\bar\beta,\bar\beta+\rho]$, we see that ζ is a trajectory for the control $\eta^{\bar\beta+\rho}$, $\zeta(\bar t) = \bar x$, and $\|\zeta(t) - \xi_*(t)\| \le \tilde\varepsilon$ for all $t \in [\bar\beta,\bar\beta+\rho]$. Therefore $\bar\beta + \rho \in S(\bar x,\eta)$, contradicting the definition of $\bar\beta$.)

It is then clear that, if we choose δ and μ such that (X.21) holds, then all our desired conclusions are satisfied. $\square$

$\square$

Corollary X.58. *Assume that $\Sigma = (\Omega, U, I, f, \mathscr{U})$ is a classical n-dimensional control system of class $C^{1,0,pc}$. Let $a, b \in I$ be such that $a \leq b$. Let $\{\eta_j\}_{j\in J}$ be a sequence of members of $\mathscr{U}$ that are defined on $[a,b]$ and c-converge to an $\eta_* \in \mathscr{U}$ as $j \to_J \infty$. Let $\xi_* : [a,b] \mapsto \Omega$ be a full trajectory of η_*. Let $\{x_j\}_{j\in J}$ be a sequence of members of Ω such that $\lim_{j\to\infty, j\in J} x_j = \xi_*(a)$. Then there exists a $j^\# \in \mathbb{N}$ such that*

(1) if $j \in J$ and $j \geq j^\#$ then there exists a full trajectory $\xi_j : [a,b] \mapsto \Omega$ for η_j such that $\xi_j(a) = x_j$,

(2) the ξ_j described in (1), which are clearly unique because of Proposition X.56, converge uniformly to ξ_* as $j \to_J \infty$.

Proof. Since the η_j c-converge to η_*, we can pick a compact subset U_0 of U such that η_* and all the η_j take values in U_0. Proposition X.57 then implies the existence, for each $k \in \mathbb{N}$, of positive numbers δ_k, μ_k such that $\bar{\mathbb{B}}^n(\xi_*(a), \delta_k) \subseteq \Omega$, having the property that whenever $\bar{x} \in \bar{\mathbb{B}}^n(\xi_*(a), \delta_k)$, $\eta \in \mathscr{U}$, domain$(\eta) = [a,b]$, and $d_{\varphi_{f,\xi_*}}(\eta, \eta_*) \leq \mu_k$, it follows that there exists a full trajectory $\xi = \xi_{\bar{x},\eta}$ for η such that $\xi(a) = \bar{x}$ and $\|\xi(t) - \xi_*(t)\| \leq 2^{-k}$ for all $t \in [a,b]$. Clearly, we may assume that the sequence $\{\delta_k\}_{k\in\mathbb{N}}$ is decreasing. Let $j^\#$ be such that $\|x_j - \xi_*(a)\| \leq \delta_1$ and $d_{\varphi_{f,\xi_*}}(\eta_j, \eta_*)\| \leq \mu_1$ whenever $j \in J$ and $j \geq j^\#$. Let $\xi_j = \xi_{x_j,\eta_j}$ for $j \in J$, $j \geq j^\#$. Then ξ_j is a full trajectory for η_j and $\xi_j(a) = x_j$.

Given k, we can pick a $j(k) \in J$ having the property that $\|x_j - \xi_*(a)\| \leq \delta_k$ and $d_{\varphi_{f,\xi_*}}(\eta_j, \eta_*) \leq \mu_k$ whenever $j \in J$ and $j \geq j(k)$. Then $\|\xi_j(t) - \xi_*(t)\| \leq 2^{-k}$ whenever $j \in J$, $j \geq j(k)$ and $t \in [a,b]$. Therefore $\xi_j \to \xi_*$ uniformly as $j \to_J \infty$. $\square$

$\square$

Corollary X.59. *Assume that $\Sigma = (\Omega, U, I, f, \mathscr{U})$ is a classical n-dimensional control system of class $C^{1,0,pc}$. Let $a, b \in I$ be such that $a \leq b$. Let P be a metric space, and let $\{\eta_p\}_{p\in P}$ be a family of members of $\mathscr{U}$ that are defined on $[a,b]$ and depend c-continuously on p. Let*

$$\mathscr{O} = \left\{(\bar{x}, p) : (\exists \xi \in C^0([a,b], \Omega))((\xi, \eta_p) \in TCP(\Sigma) \wedge \xi(a) = \bar{x})\right\}.$$

Then

(1) *$\mathscr{O}$ is an open subset of $\Omega \times P$,*

(2) *if we let $\xi_{\bar{x},p}$ be, for $(\bar{x},p) \in \mathscr{O}$, the member ξ of $C^0([a,b],\Omega)$ such that $(\xi,\eta_p) \in TCP(\Sigma)$ and $\xi(a) = \bar{x}$ (which is unique because of Proposition X.56), then the map*

$$\mathscr{O} \ni (\bar{x},p) \mapsto \xi_{\bar{x},p} \in C^0([a,b],\Omega)$$

is continuous. (That is, whenever $\{(\bar{x}_j,p_j)\}_{j\in J}$ is a sequence in $\mathscr{O}$ that converges to a limit $(\bar{x},p) \in \mathscr{O}$, it follows that $\xi_{\bar{x}_j,p_j} \to_J \xi_{\bar{x},p}$ uniformly.)

Proof. Let $\tilde{\mathscr{O}} = (\Omega \times P)\backslash\mathscr{O}$. We prove that $\tilde{\mathscr{O}}$ is a closed subset of $\Omega \times P$. Suppose that $\{(\bar{x}_j,p_j)\}_{j\in J}$ is a sequence in $\tilde{\mathscr{O}}$ that converges to a limit $(\bar{x},p) \in \Omega \times P$. We show that $(\bar{x},p) \in \tilde{\mathscr{O}}$ by contradiction. Suppose that $(\bar{x},p) \in \mathscr{O}$. Let ξ_* be the full trajectory for η_p such that $\xi_*(a) = \bar{x}$. Then Corollary X.58 implies that if $j \in J$ is large enough then there exists a full trajectory ξ_j for η_{p_j} such that $\xi_j(a) = \bar{x}_j$. But then $(\bar{x}_j,p_g) \in \mathscr{O}$ for large j, and we have reached a contradiction. This proves that $\mathscr{O}$ is open. The second statement then follows immediately from Corollary X.58. □

□

The variational equation. In this subsection we will assume that

(X.2.3.1) $\Sigma = (\Omega,U,I,f,\mathscr{U})$ *is a classical n-dimensional control system of class $C^{1,0,pc}$;*

(X.2.3.2) *a,b are such that $a,b \in I$ and $a \le b$;*

(X.2.3.3) $\gamma_* = (\xi_*,\eta_*)$ *is a trajectory-control pair of Σ such that* $\text{domain}(\eta) = [a,b]$.

Definition X.60. Assume that conditions (X.2.3.1-2-3) hold. The ordinary differential equation

$$\dot{y} = Df(\xi_*(t),\eta_*(t),t)\cdot y \tag{X.24}$$

is called the ***variational equation for Σ along the trajectory-control pair γ_****. □

It is clear that the variational equation is of the form (X.6)—with

$$A(t) = Df(\xi_*(t),\eta_*(t),t) \tag{X.25}$$

and that the matrix-valued function A satisfies $A \in C^{0,pc}([a,b],\mathrm{Lin}(X,X))$, because Df is continuous on $\Omega \times U \times [a,b]$, ξ_* is continuous, and the control η_* is piecewise continuous.

Definition X.61. Assume that conditions (X.2.3.1-2-3) hold. Then

- *the matrix-valued function $A : [a,b] \mapsto \mathbb{R}^{n\times n}$ defined by (X.25) is known as the* ***variational generator for Σ along γ_****;
- *the fundamental solution of (X.24) is called the* ***variational propagator for Σ along γ_*.*** □

We will use $\mathscr{A}^{\Sigma,\gamma_*}$, Π^{Σ,γ_*} to denote, respectively, the variational generator and the variational propagator for Σ along γ_*, so $\boxed{\mathscr{A}^{\Sigma,\gamma_*}(t) \stackrel{\text{def}}{=} Df(\xi_*(t),\eta_*(t),t)}$ and $\boxed{\Pi^{\Sigma,\gamma_*} \stackrel{\text{def}}{=} \Xi^{\mathscr{A}^{\Sigma,\gamma_*}}}$.

With the above notations, the variational equation (X.24) becomes

$$\dot{y} = \mathscr{A}^{\Sigma,\gamma_*}(t)\cdot y\,. \tag{X.26}$$

The vector solution of (X.26) with an initial condition $y(\bar{t}) = \bar{y}$ is then given by

$$y(t) = \Pi^{\Sigma,\gamma_*}(t,\bar{t})\bar{y}\,.$$

The infinitesimal reachable cone and the infinitesimal endpoint value cone. We recall that in Definition XVIII.17 (page 471) we introduced the ***endpoint value set*** $\mathbf{EPS}(\Sigma)$ for a control system Σ, i.e., the set of all endpoint values $\partial\xi$, for all pairs $(\xi,\eta) \in TCP(\Sigma)$. We will now define a terminal point analogue, the "reachable set" $\mathbf{RS}(\Sigma,a,x_*)$, and then introduce, under the assumption that conditions (X.2.3.1-2-3) hold, two sets of vectors—the "infinitesimal reachable cone" $IRC(\Sigma,\gamma_*)$, and the "infinitesimal endpoint value cone" $IEVC(\Sigma,\gamma_*)$—that are meant to be first-order approximations to $\mathbf{RS}(\Sigma,a,x_*)$ and $\mathbf{EPS}(\Sigma)$ at the points $\xi(b)$, $\partial\xi$, of $\mathbb{R}^n$, $\mathbb{R}^n\times\mathbb{R}^n$, respectively. The main technical point of the proof of the Maximum Principle will then be, precisely, to show that these "infinitesimal" cones are indeed limiting approximating cones to the corresponding sets.

Definition X.62. For a control system Σ, a time a, and a point x_* belonging to the state space of Σ, we use $\mathbf{RS}(\Sigma,a,x_*)$ to denote the set of all points $\xi(b)$, for all pairs (b,ξ) such that $b \geq a$, $(\xi,\eta) \in TCP(\Sigma)$ for some $\eta \in \mathscr{U}$ having the property that $\text{domain}(\eta) = [a,b]$, and $\xi(a) = x_*$. The set $\mathbf{RS}(\Sigma,a,x_*)$ is the ***Σ-reachable set starting at x_* at time a.*** □

First, we define vectors v^{Σ,u,t,γ_*}, $w^{\Sigma,u,t,\gamma_*} \in \mathbb{R}^n$, W^{Σ,u,t,γ_*}, $W^{\Sigma,\gamma_*}(h) \in \mathbb{R}^n \times \mathbb{R}^n$, for $a \leq t < b$, $u \in U$, $h \in \mathbb{R}^n$, by letting ,

$$v^{\Sigma,u,t,\gamma_*} \stackrel{\text{def}}{=} f(\xi_*(t),u,t) - f(\xi_*(t),\eta_*(t+),t)\,, \tag{X.27}$$

$$w^{\Sigma,u,t,\gamma_*} \stackrel{\text{def}}{=} \Pi^{\Sigma,\gamma_*}(b,t)\cdot v^{\Sigma,u,t,\gamma_*}\,, \tag{X.28}$$

$$W^{\Sigma,u,t,\gamma_*} \stackrel{\text{def}}{=} (0,w^{\Sigma,u,t,\gamma_*})\,, \tag{X.29}$$

$$W^{\Sigma,\gamma_*}(h) \stackrel{\text{def}}{=} \Big(h,\Pi^{\Sigma,\gamma_*}(b,a)\cdot h\Big)\,, \tag{X.30}$$

and observe that W^{Σ,γ_*} is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n \times \mathbb{R}^n$.

Definition X.63. Assume that conditions (X.2.3.1-2-3) hold.

1. The ***infinitesimal reachable cone*** for (Σ,γ_*) is the smallest closed convex cone in $\mathbb{R}^n$ that contains all the vectors w^{Σ,u,t,γ_*}, for all $(u,t) \in U \times [a,b[$.

2. The ***infinitesimal two-endpoint value cone*** for (Σ,γ_*) is the smallest closed convex cone in $\mathbb{R}^n \times \mathbb{R}^n$ that contains all the vectors W^{Σ,u,t,γ_*}, for all $(u,t) \in U \times [a,b[$, and all the $W^{\Sigma,t,\gamma_*}(h)$, for all $h \in \mathbb{R}^n$.

We use $IRC(\Sigma,\gamma_*)$ and $ITEVC(\Sigma,\gamma_*)$, to denote, respectively, the infinitesimal reachable cone and the infinitesimal two-endpoint value cone for (Σ,γ_*). ☐

Remark X.64. The name commonly used in the literature for what we have called "infinitesimal reachable cone" is "Pontryagin cone." ☐

Packets of needle variations. Throughout this subsection, it will be assumed that

(X.2.3.1) *$n \in \mathbb{N}$, and $\Sigma = (\Omega,U,I,f,\mathscr{U})$ is a classical n-dimensional control system of class $C^{1,0,pc}$;*

(X.2.3.2) *$a \in I$, $b \in I$, and $a \leq b$;*

(X.2.3.3) *$\gamma_* = (\xi_*,\eta_*)$ (the "reference trajectory-control pair") is a member of $TCP(\Sigma)$;*

(X.2.3.4) $\text{domain}(\eta_*) = [a,b]$.

We recall that if $\mu \in \mathbb{N}$ then $\mathbb{R}_\mu$ denotes the set of all real μ-dimensional *row* vectors $\vec{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_\mu)$. We use $\mathbb{R}_{+,\mu}$ to denote the nonnegative orthant of $\mathbb{R}_\mu$, i.e., the set of all μ-tuples $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_\mu) \in \mathbb{R}_\mu$ such that $\varepsilon_j \geq 0$ for $j = 1, \dots, \mu$. For $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_\mu) \in \mathbb{R}_{+,\mu}$, we write

$$|\vec{\varepsilon}| \stackrel{\text{def}}{=} \varepsilon_1 + \cdots + \varepsilon_\mu \,.$$

If $r > 0$, we use $\mathscr{S}_\mu(r)$ to denote the μ-dimensional simplex

$$\mathscr{S}_\mu(r) \stackrel{\text{def}}{=} \{\vec{\varepsilon} \in \mathbb{R}_{+,\mu} : |\vec{\varepsilon}| \leq r\} \,.$$

Definition X.65. Assume that conditions (X.2.3.1-2-3) hold.

1. An ***infinitesimal needle variation packet for*** Σ ***at*** γ_* is a finite sequence

$$\mathbf{V} = ((u_1, t_1), (u_2, t_2), \dots, (u_m, t_m)) \tag{X.31}$$

   of members of $U \times \mathbb{R}$, such that $a \leq t_1 \leq t_2 \leq \cdots \leq t_m < b$.

2. The ***infinitesimal two-endpoint value map associated to*** Σ***,*** γ_****, and the infinitesimal needle variation packet (X.31)*** is the linear map $ITEVM_{\Sigma,,\gamma_*,\mathbf{V}} : \mathbb{R}^n \times \mathbb{R}_m \mapsto \mathbb{R}^n \times \mathbb{R}^n$ given by

$$ITEVM_{\Sigma,\gamma_*,\mathbf{V}}(h,\vec{\varepsilon}) = W^{\Sigma,\gamma^*}(h) + \sum_{j=1}^{m} \varepsilon_j W^{\Sigma,u_j,t_j,\gamma_*} \text{ if } h \in \mathbb{R}^n\,, \vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_m). \ \square$$

Definition X.66. Assume that conditions (X.2.3.1-2-3) hold, $\mu \in \mathbb{N}$, and $\gamma_* = (\xi_*, \eta_*) \in TCP(\Sigma)$. A ***$\mu$-parameter equal-time variation*** of γ_* is a family

$$\mathscr{V} = \{\gamma^{\vec{\varepsilon}}\}_{\vec{\varepsilon} \in \mathscr{S}_\mu(r)}$$

of members of $TCP(\Sigma)$, defined for some positive real number r, such that, if $\gamma^{\vec{\varepsilon}} = (\xi^{\vec{\varepsilon}}, \eta^{\vec{\varepsilon}})$, then

1. each $\eta^{\vec{\varepsilon}}$ belongs to $\mathscr{U} \cap C^{0,pc}([a,b],U)$,
2. the controls $\eta^{\vec{\varepsilon}}$ depend c-continuously on the parameter $\vec{\varepsilon}$,
3. the trajectories $\xi^{\vec{\varepsilon}}$ depend continuously on $\vec{\varepsilon}$, in the sense that if $\vec{\varepsilon}^j \to \vec{\varepsilon}$ then $\xi^{\vec{\varepsilon}^j} \to \xi^{\vec{\varepsilon}}$ uniformly on $[a,b]$,
4. $\gamma^0 = \gamma_*$. $\square$

Lemma X.67. *Assume that conditions (X.2.3.1-2-3) hold, and $\mu \in \mathbb{N}$. Let $\mathscr{F} = \{\eta^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_\mu(r)}$ be a family of controls belonging to $\mathscr{U} \cap C^{0,pc}([a,b],U)$ and such that $\eta^0 = \eta_*$ and the $\eta^{\vec{\varepsilon}}$ depend c-continuously on $\vec{\varepsilon}$. Let $\mathbf{h} = \{h^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_\mu(r)}$ be a family of vectors in $\mathbb{R}^n$ such that $h^0 = 0$ and the $h^{\vec{\varepsilon}}$ depend continuously on $\vec{\varepsilon}$. Then there exists an $\bar{r}$ such that $0 < \bar{r} \le r$ having the property that*

(#) for every $\vec{\varepsilon} \in \mathscr{S}_\mu(\bar{r})$ there exists a unique full trajectory $\xi^{\vec{\varepsilon}} : [a,b] \mapsto \Omega$ for $\eta^{\vec{\varepsilon}}$ such that $\xi^{\vec{\varepsilon}}(a) = \xi_*(a) + h^{\vec{\varepsilon}}$.

Proof. Corollary X.59 implies that $\xi^{\vec{\varepsilon}}$ exists if $|\vec{\varepsilon}|$ is sufficiently small. Proposition X.56 implies that is unique. Corollary X.59 then implies that the $\xi^{\vec{\varepsilon}}$ depend continuously on $\vec{\varepsilon}$. □

Corollary X.68. *Assume that conditions (X.2.3.1-2-3) hold, and $\mu \in \mathbb{N}$. Let $\mathscr{F} = \{\eta^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_\mu(r)}$, $\mathbf{h} = \{h^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_\mu(r)}$ be as in the statement of Lemma X.67. Then there exist an $\bar{r}$ such that $0 < \bar{r} \le r$ and a μ-parameter equal-time variation $\mathscr{V} = \{\gamma^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_\mu(r)}$ of γ_* such that each $\gamma^{\vec{\varepsilon}}$ is of the form $(\xi^{\vec{\varepsilon}}, \eta^{\vec{\varepsilon}})$ for some $\xi^{\vec{\varepsilon}}$, and $\xi^{\vec{\varepsilon}}(a) = \xi_*(a) + h^{\vec{\varepsilon}}$.*
□

Definition X.69. Assume that conditions (X.2.3.1-2-3) hold, $\mu \in \mathbb{N}$, and $\mathscr{V} = \{\gamma^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_m(h)}$ is a μ-parameter equal-time variation of γ_*. For each $\vec{\varepsilon} \in \mathbb{R}_{+,m}$ sufficiently close to 0, let $\gamma^{\vec{\varepsilon}} = (\xi^{\vec{\varepsilon}}, \eta^{\vec{\varepsilon}})$. Let

$$\mathbf{EVM}_{\Sigma,,\gamma_*,\mathscr{V}}(\vec{\varepsilon}) \stackrel{\text{def}}{=} \partial\xi^{\vec{\varepsilon}}.$$

Then $\mathbf{EVM}_{\Sigma,,\gamma_*,\mathscr{V}}$—which is defined on some relative neighborhood of 0 in $\mathbb{R}_{+,\mu}$, and takes values in $\mathbb{R}^n \times \mathbb{R}^n$—is the ***endpoint value map associated to*** $\mathscr{V}$ ***at*** γ_*. □

Suppose that $\mathscr{V} = \{\gamma^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_\mu(r)}$ is a μ-parameter equal-time variation of the trajectory-control pair $\gamma_* = (\xi_*, \eta_*)$. Let $\gamma^{\vec{\varepsilon}} = (\xi^{\vec{\varepsilon}}, \eta^{\vec{\varepsilon}})$. Then each $\eta^{\vec{\varepsilon}}$ can be integrated against a test function $\varphi \in C^0(U \times [a,b], \mathbb{R})$. This gives rise to a function $\theta_{\varphi,\mathscr{V}} : \mathscr{S}_\mu(r) \mapsto \mathbb{R}$, given by

$$\theta_{\varphi,\mathscr{V}}(\vec{\varepsilon}) = \int_a^b \varphi(\eta^{\vec{\varepsilon}}(t), t)\, dt\,.$$

Definition X.70. Assume that conditions (X.2.3.1-2-3) hold, and suppose that $\mathbf{V}$ is an infinitesimal needle variation packet for Σ at γ_* given by (X.31). A ***needle variation packet associated to*** $\mathbf{V}$ is an $m+1$-parameter equal-time variation $\mathscr{V} = \{\gamma^{\vec{\varepsilon}}\}_{\vec{\varepsilon}\in\mathscr{S}_m(h)}$ of γ_* such that, if $\gamma^{\vec{\varepsilon}} = (\xi^{\vec{\varepsilon}}, \eta^{\vec{\varepsilon}})$, then

(1) for every test function φ belonging to $C^0(U \times [a,b], \mathbb{R})$, the function $\theta_{\varphi,\mathscr{V}}$ is differentiable at $\vec{\varepsilon} = 0$ along $\mathscr{S}_{m+1}(r)$, and its differential $D\theta_{\varphi,\mathscr{V}}(0)$ is the linear map from $\mathbb{R}_{m+1}$ to $\mathbb{R}$ given by

$$D\theta_{\varphi,\mathscr{V}}(0)(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_m) = \sum_{j=1}^{m} \varepsilon_j \Big(\varphi(u_j, t_j) - \varphi(\eta_*(t_j+), t_j) \Big) ; \qquad \text{(X.32)}$$

(2) the map $\mathscr{S}_{m+1}(r) \ni \vec{\varepsilon} \mapsto \xi^{\vec{\varepsilon}}(a)$ is differentiable at $\vec{\varepsilon} = 0$ along $\mathscr{S}_{m+1}(r)$, and its differential is the linear map $\mathbb{R}_{m+1} \ni (\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_m) \mapsto \varepsilon_0 h \in \mathbb{R}^n$;

(3) There exists a constant κ such that for every $\vec{\varepsilon} \in \mathscr{S}_{m+1}(r)$ there is a subset $I(\vec{\varepsilon})$ of $[a,b]$ which is a finite union of intervals the sum of whose measures is $\leq \kappa|\vec{\varepsilon}|$ and is such that $\eta^{\vec{\varepsilon}}(t) = \eta_*(t)$ whenever $t \in [a,b] \backslash I(\vec{\varepsilon})$. $\square$

Proposition X.71. *Assume that conditions (X.2.3.1-2-3) hold, and suppose that* $\mathbf{V}$ *is an infinitesimal needle variation packet for* Σ *at* γ_**, given by (X.31). Let* $\mathscr{V} = \{\eta^{\vec{\varepsilon}}\}_{\vec{\varepsilon} \in \mathscr{S}_m(h)}$ *be a needle variation packet in* $\mathscr{U}$ *associated to* $\mathbf{V}$*. Then the endpoint map* $\mathscr{E}_{\Sigma,,\gamma_*,\mathscr{V}}$ *is differentiable at* 0 *along* $\mathscr{S}_m(h)$*, and*

$$D\mathscr{E}_{\Sigma,,\gamma_*,\mathscr{V}}(0) = \mathscr{I}\mathscr{E}_{\Sigma,,\gamma_*,\mathbf{V}} .$$

Proof. We pick $\bar{h}$ such that (#) holds. The fact that the family $\mathscr{V}$ depends continuously on $\vec{\varepsilon}$ implies that there exists a compact subset U_0 of U such that $\eta^{\vec{\varepsilon}}([a,b]) \subseteq U_0$ for all $\vec{\varepsilon} \in \mathscr{S}_m(\bar{h})$. (Indeed, take U_0 to be the closure of W, where W is the union of the sets $\eta^{\vec{\varepsilon}}([a,b])$, for all $\vec{\varepsilon} \in \mathscr{S}_m(\bar{h})$. We show that U_0 is compact. Suppose $\mathbf{u} = \{u_j\}_{j=1}^{\infty}$ is a sequence of members of U_0. Then for each u_j there exists a $\tilde{u}_j \in W$ such that $d_U(u_j, \tilde{u}_j) \leq 2^{-j}$. Since $\tilde{u}_j \in W$, we can write $\tilde{u}_j = \eta^{\vec{\varepsilon}_j}(t_j)$ for some $\vec{\varepsilon}_j \in \mathscr{S}_m(\bar{h})$, $t_j \in [a,b]$. Since both $\mathscr{S}_m(\bar{h})$ and $[a,b]$ are compact, we can find an infinite subset J of $\mathbb{N}$ such that the limits $\vec{\varepsilon}^{\#} = \lim_{j\to\infty, j\in J} \vec{\varepsilon}_j$ and $t^{\#} = \lim_{j\to\infty, j\in J} t_j$ exist and belong to $\mathscr{S}_m(\bar{h})$, $[a,b]$, respectively. The sequence of controls $\{\eta^{\vec{\varepsilon}_j}\}_{j\in J}$ c-converges to $\eta^{\vec{\varepsilon}^{\#}}$. Therefore there exists a compact subset Z of U such that $\eta^{\vec{\varepsilon}_j}([a,b]) \subseteq K$ for all $j \in J$. Since $\eta^{\vec{\varepsilon}_j}(t_j) \in K$, we can find an infinite subset J' of J such that the limit $u^{\#} = \lim_{j\to\infty, j\in J'} \eta^{\vec{\varepsilon}_j}(t_j)$ exists. Then $u^{\#} = \lim_{j\to\infty, j\in J'} \tilde{u}_j$. Furthermore, $u^{\#} \in U_0$, because the $\tilde{u}_j$ belong to W. Finally, $u^{\#} = \lim_{j\to\infty, j\in J'} u_j$ as well. So we have extracted a subsequence of $\mathbf{u}$ that converges to a point of U_0. This proves the desired compactness.)

For $\vec{\varepsilon} \in \mathscr{S}_m(\bar{h})$, we write $\zeta^{\vec{\varepsilon}}(t) = \xi^{\vec{\varepsilon}}(t) - \xi_*(t)$. Proposition X.57 implies that the curves $\xi^{\vec{\varepsilon}}$ converge uniformly to ξ_* as $\vec{\varepsilon} \to 0$, so $\zeta^{\vec{\varepsilon}} \to 0$ uniformly as $\vec{\varepsilon} \to 0$.

Hence we may assume, after making $\bar{h}$ smaller, if necessary, that there exists a positive number $\bar{\delta}$ such that the ball $\{x \in \mathbb{R}^n : \|x - \xi_*(t)\| \leq \delta\}$ is contained in Ω for every $t \in [a,b]$, and $\|\zeta^{\vec{\varepsilon}}(t)\| \leq \bar{\delta}$ whenever $\vec{\varepsilon} \in \mathscr{S}_m(\bar{h})$ and $t \in [a,b]$. We let $\Omega_0 = \{x \in \Omega : (\exists t \in [a,b])(\|x - \xi_*(t)\| \leq \bar{\delta}\}$. Then Ω_0 is compact. We then define

$$K = \Omega_0 \times U_0 \times [a,b],$$

and let $c^0(K)$, $c^1(K)$ be the constants introduced in Remark X.48.

Clearly,

$$\begin{aligned}
\dot{\zeta}^{\vec{\varepsilon}}(t) &= f(\xi^{\vec{\varepsilon}}(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi_*(t), \eta_*(t), t) \\
&= f(\xi^{\vec{\varepsilon}}(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi^{\vec{\varepsilon}}(t), \eta_*(t), t) + f(\xi^{\vec{\varepsilon}}(t), \eta_*(t), t) - f(\xi_*(t), \eta_*(t), t) \\
&= f(\xi^{\vec{\varepsilon}}(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi^{\vec{\varepsilon}}(t), \eta_*(t), t) + \mathscr{A}^{\vec{\varepsilon}}(t) \cdot \zeta^{\vec{\varepsilon}}(t) \\
&= E_1^{\vec{\varepsilon}}(t) + E_2^{\vec{\varepsilon}}(t) + + \mathscr{A}^{\vec{\varepsilon}}(t) \cdot \zeta^{\vec{\varepsilon}}(t),
\end{aligned}$$

where

$$\begin{aligned}
\mathscr{A}^{\vec{\varepsilon}}(t) &= \int_0^1 Df(\xi_*(t) + \alpha\zeta^{\vec{\varepsilon}}(t), \eta_*(t), t)\, d\alpha, \\
E_1^{\vec{\varepsilon}}(t) &= f(\xi_*(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi_*(t), \eta_*(t), t) \\
E_2^{\vec{\varepsilon}}(t) &= \Big(f(\xi^{\vec{\varepsilon}}(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi^{\vec{\varepsilon}}(t), \eta_*(t), t) \Big) \\
&\quad - \Big(f(\xi_*(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi_*(t), \eta_*(t), t) \Big).
\end{aligned}$$

Let $\Pi^{\vec{\varepsilon}}$ be the fundamental matrix solution of the equation $\dot{x} = \mathscr{A}^{\vec{\varepsilon}}(t) \cdot x$. Then

$$\zeta^{\vec{\varepsilon}}(t) = \int_a^t \Pi^{\vec{\varepsilon}}(t,s) \cdot \Big(E_1^{\vec{\varepsilon}}(s) + E_2^{\vec{\varepsilon}}(s) \Big)\, ds. \tag{X.33}$$

(Indeed, if we let $\rho^{\vec{\varepsilon}}(t)$ denote the right-hand side of (X.33), then

$$\begin{aligned}
\dot{\rho}^{\vec{\varepsilon}}(t) &= E_1^{\vec{\varepsilon}}(t) + E_2^{\vec{\varepsilon}}(t) + \int_a^t \frac{\partial \Pi^{\vec{\varepsilon}}}{\partial t}(t,s) \Big(E_1^{\vec{\varepsilon}}(s) + E_2^{\vec{\varepsilon}}(s) \Big)\, ds \\
&= E_1^{\vec{\varepsilon}}(t) + E_2^{\vec{\varepsilon}}(t) + \mathscr{A}^{\vec{\varepsilon}}(t) \int_a^t \Pi^{\vec{\varepsilon}}(t,s) \Big(E_1^{\vec{\varepsilon}}(s) + E_2^{\vec{\varepsilon}}(s) \Big)\, ds \\
&= E_1^{\vec{\varepsilon}}(t) + E_2^{\vec{\varepsilon}}(t) + \mathscr{A}^{\vec{\varepsilon}}(t) \rho^{\vec{\varepsilon}}(t).
\end{aligned}$$

So $\rho^{\vec{\varepsilon}}$ satisfies $\dot{\rho}^{\vec{\varepsilon}} = E_1^{\vec{\varepsilon}} + E_2^{\vec{\varepsilon}} + \mathscr{A}^{\vec{\varepsilon}} \rho^{\vec{\varepsilon}}$, while we have shown that $\dot{\zeta}^{\vec{\varepsilon}} = E_1^{\vec{\varepsilon}} + E_2^{\vec{\varepsilon}} + \mathscr{A}^{\vec{\varepsilon}} \zeta^{\vec{\varepsilon}}$. Since $\rho^{\vec{\varepsilon}}(a) = 0$ and $\zeta^{\vec{\varepsilon}}(a) = 0$, we have $\rho^{\vec{\varepsilon}} \equiv \zeta^{\vec{\varepsilon}}$, as stated.)

It is clear that $\mathscr{A}^0$ is the variational generator $\mathscr{A}^{f,\gamma_*}$. Therefore $\mathscr{A}^{\vec{\varepsilon}}$ converges uniformly to $\mathscr{A}^{f,\gamma_*}$, and then Proposition X.26 implies that the fundamental solutions $\Pi^{\vec{\varepsilon}}$ converge uniformly to Π^{f,γ_*}. Therefore, if we define

$$\sigma_1(\vec{\varepsilon}) = \sup\left\{ \|\Pi^{\vec{\varepsilon}}(t,s) - \Pi^{f,\gamma_*}(t,s)\| : t,s \in [a,b] \right\},$$

we have $\lim_{\vec{\varepsilon}\to 0} \sigma_1(\vec{\varepsilon}) = 0$.

Let $I(\vec{\varepsilon})$ be a subset of $[a,b]$ which is a finite union of intervals the sum of whose measures is $\leq \kappa|\vec{\varepsilon}|$ and is such that $\eta^{\vec{\varepsilon}}(t) = \eta_*(t)$ whenever $t \in [a,b]\backslash I(\vec{\varepsilon})$. The vector $E_1^{\vec{\varepsilon}}(t)$ and $E_2^{\vec{\varepsilon}}(t)$ clearly vanish whenever $t \neq I(\vec{\varepsilon})$, and are bounded in norm by $2c^0(K)$ and $4c^0(K)$, resepctively, for every t. Hence

$$\int_a^b \Big(\|E_1^{\vec{\varepsilon}}(t)\| + \|E_2^{\vec{\varepsilon}}(t)\| \Big)\, dt \leq 6c^0(K)\kappa|\vec{\varepsilon}|\,. \tag{X.34}$$

Therefore, if we write

$$\zeta^{\vec{\varepsilon}}(b) = \int_a^b \Pi^{f,\gamma_*}(b,s)\cdot \Big(E_1^{\vec{\varepsilon}}(s) + E_2^{\vec{\varepsilon}}(s) \Big)\, ds + R_1(\vec{\varepsilon})\,, \tag{X.35}$$

so that

$$R_1(\vec{\varepsilon}) = \int_a^b \Big(\Pi^{\vec{\varepsilon}}(b,s) - \Pi^{f,\gamma_*}(b,s) \Big) \cdot \Big(E_1^{\vec{\varepsilon}}(s) + E_2^{\vec{\varepsilon}}(s) \Big)\, ds\,,$$

we have

$$\|R_1(\vec{\varepsilon})\| \leq 6c^0(K)\kappa\sigma_1(\vec{\varepsilon})|\vec{\varepsilon}|\,. \tag{X.36}$$

If we let

$$\tilde{\kappa} = \sup\left\{ \|\Pi^{f,\gamma_*}(t,s)\| : t,s \in [a,b] \right\},$$

then (X.33) and (X.34) imply the bound

$$\|\zeta^{\vec{\varepsilon}}(t)\| \leq 6c^0(K)\kappa\tilde{\kappa}|\vec{\varepsilon}| \quad \text{for} \quad t \in [a,b]\,.$$

Therefore

$$\begin{aligned}
&\|E_2^{\vec{\varepsilon}}(t)\| \\
&= \ \left\| \Big(f(\xi^{\vec{\varepsilon}}(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi_*(t), \eta^{\vec{\varepsilon}}(t), t) \Big) - \Big(f(\xi^{\vec{\varepsilon}}(t), \eta_*(t), t) - f(\xi_*(t), \eta_*(t), t) \Big) \right\| \\
&\leq \ \left\| f(\xi^{\vec{\varepsilon}}(t), \eta^{\vec{\varepsilon}}(t), t) - f(\xi_*(t), \eta^{\vec{\varepsilon}}(t), t) \right\| + \left\| f(\xi^{\vec{\varepsilon}}(t), \eta_*(t), t) - f(\xi_*(t), \eta_*(t), t) \right\| \\
&\leq \ 2c^1(K)\|\xi^{\vec{\varepsilon}}(t) - \xi_*(t)\| \\
&\leq \ 12c^0(K)c^1(K)\kappa\tilde{\kappa}|\vec{\varepsilon}|\,.
\end{aligned}$$

Hence
$$\int_a^b \|E_2^{\vec{\varepsilon}}(t)\|\,dt \le 12c^0(K)c^1(K)\kappa^2\tilde{\kappa}|\vec{\varepsilon}|^2\,.$$
Therefore, if we write
$$\zeta^{\vec{\varepsilon}}(b) = \int_a^b \Pi^{f,\gamma_*}(b,s)\cdot E_1^{\vec{\varepsilon}}(s)\,ds + R_1(\vec{\varepsilon}) + R_2(\vec{\varepsilon})\,, \tag{X.37}$$
so that
$$R_2(\vec{\varepsilon}) = \int_a^b \Pi^{f,\gamma_*}(b,s)\cdot E_2^{\vec{\varepsilon}}(s)\,ds\,,$$
we have
$$\|R_2(\vec{\varepsilon})\| \le 12c^0(K)c^1(K)\kappa^2\tilde{\kappa}^2|\vec{\varepsilon}|^2\,. \tag{X.38}$$
If we combine (X.39), (X.36), and (X.38), and observe that $\zeta^{\vec{\varepsilon}}(b) = \mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}(\vec{\varepsilon}) - \mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}(0)$, we find that
$$\mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}(\vec{\varepsilon}) = \mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}(0) + \int_a^b \Pi^{f,\gamma_*}(b,s)\cdot E_1^{\vec{\varepsilon}}(s)\,ds + o(|\vec{\varepsilon}|) \quad \text{as} \quad \vec{\varepsilon}\to 0\,. \tag{X.39}$$
Hence, to prove that $\mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}(\vec{\varepsilon})$ is differentiable at $\vec{\varepsilon}=0$ with a certain differential Λ, if suffices to show that the map $\vec{\varepsilon}\mapsto M(\vec{\varepsilon}) \stackrel{\text{def}}{=} \int_a^b \Pi^{f,\gamma_*}(b,s)\cdot E_1^{\vec{\varepsilon}}(s)\,ds$ is differentiable at $\vec{\varepsilon}=0$ with differential Λ.

Now, let $\varphi(u,t) = \Pi^{f,\gamma_*}(b,t)f(\xi_*(t),u,t)$. Then φ is a continous vector-valued function on $U\times[a,b]$. Furthermore, it is clear that
$$M(\vec{\varepsilon}) = \theta_{\varphi,\mathscr{V}}(\vec{\varepsilon}) - \theta_{\varphi,\mathscr{V}}(0)\,.$$
It follows that M is differentiable at $\vec{\varepsilon}=0$, and $DM(0)$ is the map
$$(\varepsilon_1,\ldots,\varepsilon_m)\mapsto \sum_{j=1}^m \varepsilon_j\Big(\varphi(u_j,t_j) - \varphi(\eta_*(t_j+),t_j)\Big)\,,$$
i.e., the map
$$(\varepsilon_1,\ldots,\varepsilon_m)\mapsto \sum_{j=1}^m \varepsilon_j\Pi^{f,\gamma_*}(b,t_j)\Big(f(\xi_*(t_j),u_j,t) - f(\xi_*(t_j),\eta_*(t_j+),t)\Big)\,.$$
In other words, $DM(0)$ is the map $(\varepsilon_1,\ldots,\varepsilon_m)\mapsto \sum_{j=1}^m \varepsilon_j w^{u_j,t_j,\gamma_*}$, i.e., the map $\mathscr{I}\mathscr{E}_{\Sigma,\gamma_*,\mathbf{V}}$. Hence $\mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}$ is differentiable at $\vec{\varepsilon}=0$ and $D\mathscr{E}_{\Sigma,\gamma_*,\mathscr{V}}(0) = \mathscr{I}\mathscr{E}_{\Sigma,\gamma_*,\mathbf{V}}$, completing our proof. □

Existence of packets of needle variations. In this subsection, we assume that

(X.2.3.1) $\Sigma = (\Omega, U, I, f, \mathscr{U})$ *is a control system in* $\mathbb{R}^n$ *of class* $C^{0,0,pc}$;

(X.2.3.2) $a \in I$, $b \in I$, *and* $a \le b$;

(X.2.3.3) $\gamma_* = (\xi_*, \eta_*)$ *(the "reference trajectory-control pair") is a member of* $TCP(\Sigma)$;

(X.2.3.4) $\mathrm{domain}(\eta_*) = [a,b]$.

Definition X.72. *Let U be a set, and let $\mathscr{U}$ be a set of U-valued controls. We say that $\mathscr{U}$ is* ***closed under interval intertwining*** *if, whenever J is an interval, J' is a subinterval of J, and η_1, η_2 are two members of $\mathscr{U}$ defined on J, it follows that $\eta_3 \in \mathscr{U}$, where $\eta_3 : J \mapsto U$ is the U-control defined by*

$$\eta_3(t) = \begin{cases} \eta_1(t) & \text{if} \quad t \in J' \\ \eta_2(t) & \text{if} \quad t \in J \backslash J' \end{cases}.$$

Proposition X.73. *Assume that conditions (X.2.3.1-2-3) hold and suppose in addition that $\mathscr{U}$ contains all constant U-controls defined on $[a,b]$ and is closed under interval intertwining. Let* $\mathbf{V}$ *be an infinitesimal needle variation packet for Σ at γ_*. Then there exists a needle variation packet in $\mathscr{U}$ associated to* $\mathbf{V}$.

Proof. Let

$$\mathbf{V} = \big((u_1,t_1), (u_2,t_2), \ldots, (u_m,t_m)\big), \qquad a \le t_1 \le t_2 \le \ldots \le t_m < b. \tag{X.40}$$

We define functions $\tau_j^- : \mathbb{R}_{+,m} \mapsto \mathbb{R}$, $\tau_j^+ : \mathbb{R}_{+,m} \mapsto \mathbb{R}$, for $j = 1, \ldots, m$, by letting $\tau_1^-(\vec{\varepsilon}) = t_1$, $\tau_1^+(\vec{\varepsilon}) = t_1 + \varepsilon_1$, if $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)$, and then

$$\tau_j^-(\vec{\varepsilon}) = \max(\tau_{j-1}^+(\vec{\varepsilon}), t_j), \qquad \tau_j^+(\vec{\varepsilon}) = \tau_j^-(\vec{\varepsilon}) + \varepsilon_j \qquad \text{for} \quad j = 2, \ldots, m.$$

We let $I_j(\vec{\varepsilon}) = [\tau_j^-(\vec{\varepsilon}), \tau_j^+(\vec{\varepsilon})[$, and write $\mathscr{I}(\vec{\varepsilon}) = \bigcup_{j=1}^m I_j(\vec{\varepsilon})$. It is then clear that

(a) the τ_j^- and τ_j^+ are continuous functions;

(b) $t_j \le \tau_j^-(\vec{\varepsilon})$ for all $\vec{\varepsilon}, j$,

(c) $\tau_j^-(0) = \tau_j^+(0) = t_j$ for all j;

(d) the intervals $I_j(\vec{\varepsilon})$, for $j = 1, \ldots, m$, are pairwise disjoint, and ordered from left to right (that is, $t < t'$ whenever $t \in I_{j-1}(\vec{\varepsilon}), t' \in I_j(\vec{\varepsilon})$, $j = 2, \ldots, m$);

(e) $I_j(\vec{\varepsilon})$ has length ε_j for each $\vec{\varepsilon}, j$.

Furthermore, for each j the $I_j(\vec{\varepsilon})$ converge to $\{t_j\}$ as $\vec{\varepsilon} \to 0$, in the sense that the endpoints $\tau_j^-(\vec{\varepsilon})$ and $\tau_j^+(\vec{\varepsilon})$ go to t_j.

We pick h such that $\tau_m^+(\vec{\varepsilon}) \le b$ whenever $|\vec{\varepsilon}| \le h$. (This is possible because $\tau_m^+(\vec{\varepsilon}) = t_m < b$.) Then $\mathscr{I}(\vec{\varepsilon}) \subseteq [a,b]$ whenever $|\vec{\varepsilon}| \le h$.

We then construct a control variation $\mathscr{V} = \{\eta^{\vec{\varepsilon}}\}_{\vec{\varepsilon} \in \mathscr{S}_m(h)}$, by defining

$$\eta^{\vec{\varepsilon}}(t) = \begin{cases} \eta_*(t) & \text{if} \quad t \in [a,b] \backslash \mathscr{I}(\vec{\varepsilon})\,, \\ u_j & \text{if} \quad t \in I_j(\vec{\varepsilon})\,. \end{cases} \tag{X.41}$$

It clearly follows from our assumptions on $\mathscr{U}$ that the $\eta^{\vec{\varepsilon}}$ belong to $\mathscr{U}$. Also, it is easy to see that the $\eta^{\vec{\varepsilon}}$ depend c-continuously on the parameter $\vec{\varepsilon}$ and $\eta^0 = \eta^*$. So $\mathscr{V}$ is an m-parameter equal-time variation of η_* in $\mathscr{U}$. We now prove that $\mathscr{V}$ is in fact a needle variation packet in $\mathscr{U}$ associated to $\mathbf{V}$. For this purpose, we verify the two conditions of Definition X.70. Condition (2) is trivial (with $\kappa = 1$), since we can take $I(\vec{\varepsilon})$ to be $\mathscr{I}(\vec{\varepsilon})$. To verify Condition (1), we pick a test function $\varphi \in C^0(U \times [a,b], \mathbb{R})$ and compute

$$\begin{aligned} \theta_{\varphi,\mathscr{V}}(\vec{\varepsilon}) - \theta_{\varphi,\mathscr{V}}(0) &= \int_a^b \Big(\varphi(\eta^{\varepsilon}(t),t) - \varphi(\eta^0(t),t)\Big)\,dt \\ &= \sum_{j=1}^m \int_{\tau_j^-(\vec{\varepsilon})}^{\tau_j^+(\vec{\varepsilon})} \Big(\varphi(u_j,t) - \varphi(\eta_*(t),t)\Big)\,dt \\ &= \sum_{j=1}^m \varepsilon_j\Big(\varphi(u_j,t_j) - \varphi(\eta_*(t_j+),t_j))\Big) + E(\vec{\varepsilon})\,, \end{aligned}$$

where the error $E(\vec{\varepsilon})$ is given by

$$E(\vec{\varepsilon}) = \sum_{j=1}^m \int_{\tau_j^-(\vec{\varepsilon})}^{\tau_j^+(\vec{\varepsilon})} \left(\Big(\varphi(u_j,t) - \varphi(u_j,t_j)\Big) - \Big(\varphi(\eta_*(t),t) - \varphi(\eta_*(t_j+),t_j)\Big)\right) dt\,.$$

Since $\lim_{t\to t_j}\Big((\varphi(u_j,t) - \varphi(u_j,t_j)\Big) = \lim_{t\to t_j+}\Big((\varphi(\eta_*(t),t) - \varphi(\eta_*(t_j+),t_j)\Big) = 0$, the error term satisfies $|E(\vec{\varepsilon})| = o(|\vec{\varepsilon}|)$ as $\vec{\varepsilon} \to 0$. Hence $\theta_{\varphi,\mathscr{V}}$ is differentiable at 0 along $\mathscr{S}_m(h)$, and the differential $D\theta_{\varphi,\mathscr{V}}(0)$ is given by (X.32). This completes the proof that Condition (1) holds. □

Definition X.74. *If U is a metric space, $I = [a,b]$ is a compact intevral, and $\eta_* \in C^0([a,b],U)$, an L^1* ***neighborhood of*** η_* ***in*** $C^0([a,b],U)$ *is a set $\mathscr{N}$ such that*

$$\mathscr{N} = \Big\{\eta \in C^0([a,b],U) : \int_a^b d_U(\eta(t),\eta_*(t)\,dt \le \delta\Big\}$$

for some positive number δ. □

Proposition X.75. *Assume that conditions (X.2.3.1-2-3) hold and suppose in addition that U is path-connected, η_* is continuous, and $\mathscr{U}$ contains an L^1 neighborhood of η_* in $C^0([a,b],U)$. Let $\mathbf{V}$ be an infinitesimal needle variation packet for Σ at γ_*. Then there exists a needle variation packet in $\mathscr{U}$ associated to $\mathbf{V}$.*

Proof. Let $\mathbf{V}$ be given by (X.40). For each $j \in \{1,\ldots,m\}$, choose a continuous path $\gamma_j : [0,1] \mapsto U$ such that $\gamma_j(0) = \eta_*(t_j)$ and $\gamma_j(1) = u_j$. We let

$$U_0 = \eta_*([a,b]) \cup \{u_1,\ldots,u_m\} \cup \bigcup_{j=1}^{m} \gamma_j([0,1])\,,$$

so U_0 is a compact subset of U. We let

$$\hat{\kappa} = \max\{d_U(u,v) : u \in U_0, v \in U_0\}\,.$$

For $\vec{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_m)$, we define $\tilde{\varepsilon}_j = \varepsilon_j + 4\varepsilon_j^2$. We then define a pair of functions $\tau_j^- : \mathbb{R}_{+,m} \mapsto \mathbb{R}$, $\tau_j^+ : \mathbb{R}_{+,m} \mapsto \mathbb{R}$, for each $j \in \{1,\ldots,m\}$, by letting $\tau_1^-(\vec{\varepsilon}) = t_1$, $\tau_1^+(\vec{\varepsilon}) = t_1 + \tilde{\varepsilon}_1$, and then writing

$$\tau_j^-(\vec{\varepsilon}) = \max(\tau_{j-1}^+(\vec{\varepsilon}), t_j)\,, \qquad \tau_j^+(\vec{\varepsilon}) = \tau_j^-(\vec{\varepsilon}) + \tilde{\varepsilon}_j \qquad \text{for} \quad j = 2,\ldots,m\,.$$

We let

$$\begin{aligned}
I_j^1(\vec{\varepsilon}) &= [\tau_j^-(\vec{\varepsilon}), \tau_j^-(\vec{\varepsilon}) + \varepsilon_j^2\,[\,,\\
I_j^2(\vec{\varepsilon}) &= [\tau_j^-(\vec{\varepsilon}) + \varepsilon_j^2, \tau_j^-(\vec{\varepsilon}) + 2\varepsilon_j^2\,[\,,\\
I_j(\vec{\varepsilon})^3 &= [\tau_j^-(\vec{\varepsilon}) + 2\varepsilon_j^2, \tau_j^-(\vec{\varepsilon}) + 2\varepsilon_j^2 + \varepsilon_j\,[\,,\\
I_j^4(\vec{\varepsilon}) &= [\tau_j^-(\vec{\varepsilon}) + 2\varepsilon_j^2 + \varepsilon_j, \tau_j^-(\vec{\varepsilon}) + 3\varepsilon_j + \varepsilon_j,[\,,\\
I_j^5(\vec{\varepsilon}) &= [\tau_j^-(\vec{\varepsilon}) + 3\varepsilon_j + \varepsilon_j, \tau_j^-(\vec{\varepsilon}) + 4\varepsilon_j + \varepsilon_j\,[= [\tau_j^-(\vec{\varepsilon}) + 3\varepsilon_j + \varepsilon_j, \tau_j^+(\vec{\varepsilon})\,[\,,\\
I_j(\vec{\varepsilon}) &= \bigcup_{i=1}^{5} I_j^i(\vec{\varepsilon}) = [\tau_j^-(\vec{\varepsilon}), \tau_j^+(\vec{\varepsilon})\,[\,,\\
\mathscr{I}(\vec{\varepsilon}) &= \bigcup_{j=1}^{m} I_j(\vec{\varepsilon})\,.
\end{aligned}$$

It is then clear that

(a) the τ_j^- and τ_j^+ are continuous functions;

(b) $t_j \leq \tau_j^-(\vec{\varepsilon})$ for all $\vec{\varepsilon}, j$,

(c) $\tau_j^-(0) = \tau_j^+(0) = t_j$ for all j;

(d) the intervals $I_j(\vec{\varepsilon})$, for $j = 1, \ldots, m$, are pairwise disjoint, and ordered from left to right (that is, $t < t'$ whenever $t \in I_{j-1}(\vec{\varepsilon}), t' \in I_j(\vec{\varepsilon})$, $j = 2, \ldots, m$);

(e) $I_j(\vec{\varepsilon})$ has length $\tilde{\varepsilon}_j$ for each $\vec{\varepsilon}, j$.

Furthermore, for each j the $I_j(\vec{\varepsilon})$ converge to $\{t_j\}$ as $\vec{\varepsilon} \to 0$, in the sense that the endpoints $\tau_j^-(\vec{\varepsilon})$ and $\tau_j^+(\vec{\varepsilon})$ go to t_j.

We pick h_0 such that $\tau_m^+(\vec{\varepsilon}) \le b$ whenever $|\vec{\varepsilon}| \le h_0$. (This is possible because $\tau_m^+(\vec{\varepsilon}) = t_m < b$.) Then $\mathscr{I}(\vec{\varepsilon}) \subseteq [a,b]$ whenever $|\vec{\varepsilon}| \le h_0$.

We then construct a family $\mathscr{V}_0 = \{\eta^{\vec{\varepsilon}}\}_{\vec{\varepsilon} \in \mathscr{S}_m(h_0)}$ of U-controls by defining

$$\eta^{\vec{\varepsilon}}(t) = \begin{cases} \eta_*(t) & \text{if} \quad t \in [a,b] \backslash \mathscr{I}(\vec{\varepsilon}), \\ \eta_*\Big(\tau_j^-(\vec{\varepsilon}) - \frac{(t-\tau_j^-(\vec{\varepsilon}))(\tau_j^-(\vec{\varepsilon})-t_j)}{\varepsilon_j^2}\Big) & \text{if} \quad t \in I_j^1(\vec{\varepsilon}), \\ \gamma_j\Big(\frac{t-\tau_j^-(\vec{\varepsilon})-\varepsilon_j^2}{\varepsilon_j^2}\Big) & \text{if} \quad t \in I_j^2(\vec{\varepsilon}), \\ u_j & \text{if} \quad t \in I_j^3(\vec{\varepsilon}), \\ \gamma_j\Big(\frac{\tau_j^-(\vec{\varepsilon})+3\varepsilon_j^2+\varepsilon_j-t}{\varepsilon_j^2}\Big) & \text{if} \quad t \in I_j^4(\vec{\varepsilon}), \\ \eta_*\Big(t_j + \frac{(t-\tau_j^-(\vec{\varepsilon})-3\varepsilon_j^2-\varepsilon_j)(\tau_j^+(\vec{\varepsilon})-t_j)}{\varepsilon_j^2}\Big) & \text{if} \quad t \in I_j^5(\vec{\varepsilon}). \end{cases} \tag{X.42}$$

It is then easy to verify that all the controls η^ε are continuous U-valued functions. Furthermore, if $\vec{\varepsilon} \in \mathscr{S}_m(h_0)$, then

$$\begin{aligned} \int_a^b d_U(\eta^{\vec{\varepsilon}}(t), \eta_*(t))dt &= \sum_{j=1}^m \int_{I_j(\vec{\varepsilon})} d_U(\eta^{\vec{\varepsilon}}(t), \eta_*(t))dt \\ &\le \hat{\kappa} \sum_{j=1}^m (\varepsilon_j + 4\varepsilon_j^2) \\ &\le \hat{\kappa}(|\vec{\varepsilon}| + 4|\vec{\varepsilon}|^2). \end{aligned}$$

Pick a positive number δ such that

$$\Big\{\eta \in C^0([a,b],U) : \int_a^b d_U(\eta(t), \eta_*(t)\,dt \le \delta\Big\} \subseteq \mathscr{U},$$

and then choose h such that $0 < h \le h_0$ and $\hat{\kappa}(h + 4h^2) \le \delta$, and define

$$\mathscr{V} = \{\eta^{\vec{\varepsilon}}\}_{\vec{\varepsilon} \in \mathscr{S}_m(h)}.$$

It is then clear that the $\eta^{\vec{\varepsilon}}$ belong to $\mathscr{U}$ if $\vec{\varepsilon} \in \mathscr{S}_m(h)$. Also, it is easy to see that the $\eta^{\vec{\varepsilon}}$ depend c-continuously on the parameter $\vec{\varepsilon}$ and $\eta^0 = \eta^*$. So $\mathscr{V}$ is an m-parameter equal-time variation of η_* in $\mathscr{U}$. We now prove that $\mathscr{V}$ is in fact a needle variation packet in $\mathscr{U}$ associated to $\mathbf{V}$. For this purpose, we verify the two conditions of Definition X.70. Condition (2) is trivial (with $\kappa = 1+4h$), since we can take $I(\vec{\varepsilon})$ to be $\mathscr{I}(\vec{\varepsilon})$.

To verify Condition (1), we pick a test function $\varphi \in C^0(U \times [a,b], \mathbb{R})$, define

$$\kappa_\varphi = \max\Big\{|\varphi(u,t)| : u \in U_0, t \in [a,b]\Big\},$$

and compute

$$\begin{aligned}
\theta_{\varphi,\mathscr{V}}(\vec{\varepsilon}) - \theta_{\varphi,\mathscr{V}}(0) &= \int_a^b \Big(\varphi(\eta^\varepsilon(t),t) - \varphi(\eta^0(t),t)\Big)\,dt \\
&= \sum_{j=1}^m \sum_{k=1}^5 \int_{I_j(\vec{\varepsilon})} \Big(\varphi(\eta^\varepsilon(t),t) - \varphi(\eta^0(t),t)\Big)\,dt \\
&= \sum_{j=1}^m \varepsilon_j \Big(\varphi(u_j,t_j) - \varphi(\eta_*(t_j),t_j))\Big) + E(\vec{\varepsilon}),
\end{aligned}$$

where the error $E(\vec{\varepsilon})$ is given by

$$\begin{aligned}
E(\vec{\varepsilon}) &= E_1(\vec{\varepsilon}) + E_2(\vec{\varepsilon}), \\
E_1(\vec{\varepsilon}) &= \sum_{j=1}^m \int_{I_j^3(\vec{\varepsilon})} \Bigg(\Big(\varphi(u_j,t) - \varphi(u_j,t_j)\Big) - \Big(\varphi(\eta_*(t),t) - \varphi(\eta_*(t_j),t_j)\Big)\Bigg)\,dt, \\
E_2(\vec{\varepsilon}) &= \sum_{j=1}^m \sum_{k\in\{1,2,4,5\}} \int_{I_j^k(\vec{\varepsilon})} \Big(\varphi(\eta^\varepsilon(t),t) - \varphi(\eta^0(t),t)\Big)\,dt.
\end{aligned}$$

Since $\lim_{t\to t_j}\Big((\varphi(u_j,t) - \varphi(u_j,t_j)\Big) = \lim_{t\to t_j}\Big((\varphi(\eta_*(t),t) - \varphi(\eta_*(t_j),t_j)\Big) = 0$, the first error term satisfies $|E_1(\vec{\varepsilon})| = o(|\vec{\varepsilon}|)$ as $\vec{\varepsilon} \to 0$. Since the length of $I_j^k(\vec{\varepsilon})$ is ε_j^2 whenever $j = 1,\dots,m$ and $k \in \{1,2,4,5\}$, the second error term satisfies $|E_2(\vec{\varepsilon})| \le 8\kappa_\varphi|\vec{\varepsilon}|^2$, so $|E_2(\vec{\varepsilon})| = o(|\vec{\varepsilon}|)$ as well. Hence $|E(\vec{\varepsilon})| = o(|\vec{\varepsilon}|)$.

Hence $\theta_{\varphi,\mathscr{V}}$ is differentiable at 0 along $\mathscr{S}_m(h)$, and the differential $D\theta_{\varphi,\mathscr{V}}(0)$ is given by (X.32). This completes the proof that Condition (1) holds. □

□

X.3 The main theorem

Theorem X.76. ***(The fixed time separation theorem.)*** *Assume that*

(X.76.1) $n \in \mathbb{N}$ *and* $\Sigma = (\Omega, U, I, f, \mathscr{U})$ *is a classical n-dimensional control system of class* $C^{1,0,pc}$*;*

(X.76.2) a, b *are real numbers such that* $a < b$*;*

(X.76.3) $\gamma_* = (\xi_*, \eta_*) \in TCP(\Sigma)$ *and* $\text{domain}(\eta_*) = [a,b]$*;*

(X.76.4) $\mathscr{S}$ *is a subset of* $\Omega \times \Omega$ *such that* $\partial\xi_* \in \mathscr{S}$*;*

(X.76.5) $\mathscr{C}$ *is a limiting approximating cone to* $\mathscr{S}$ *at* $\partial\xi_*$*;*

(X.76.6) *the sets* $\mathscr{R}^\Sigma$ *and* $\mathscr{S}$ *are locally separated at* $\partial\xi_*$*;*

(X.76.7) *either*

(X.76.7.a) $\mathscr{U}$ *contains all constant* U*-controls defined on* $[a,b]$ *and is closed under interval intertwining,*

or

(X.76.7.b) U *is path-connected,* η_* *is continuous, and* $\mathscr{U}$ *contains an* L^1 *neighborhood of* η_* *in* $C^0([a,b],U)$.

Define the ***Hamiltonian***

$$\Omega \times \mathbb{R}_n \times U \times [a,b] \ni (x,p,u,t) \mapsto H(x,p,u,t) \in \mathbb{R}$$

by letting

$$H(x,p,u,t) \stackrel{\text{def}}{=} p \cdot f(x,u,t)\,.$$

Then there exists a π *such that*

(X.76.a) $\pi : [a,b] \mapsto \mathbb{R}_n$ *is a continuous map;*

(X.76.b) *(the "adjoint equation")*

$$\dot{\pi}(t) = -\pi(t) \cdot \frac{\partial f}{\partial x}(\xi_*(t), \eta_*(t), t) \tag{X.43}$$

at all points of continuity t *of* η_**;*

(X.76.c) *(the "Hamiltonian maximization condition")*

$$H(\xi_*(t), \pi(t), \eta_*(t), t) = \max\Big\{H(\xi_*(t), \pi(t), u, t) : u \in U\Big\} \tag{X.44}$$

at all points of continuity t *of* η_**;*

(X.76.d) *(the "transversality condition")*

$$(\pi(a), -\pi(b)) \in \mathscr{C}^\dagger\,; \tag{X.45}$$

(X.76.d) (the "nontriviality condition") $\pi(t) \neq 0$ for some t (which implies that $\pi(t) \neq 0$ for all t).

REMARK: The adjoint equation (X.43) can also be written as

$$\dot{\pi}(t) = -\frac{\partial H}{\partial x}(\xi_*(t), \pi(t), \eta_*(t), t)\,. \tag{X.46}$$

The fact that the dynamical equation $\dot{x} = f(x,u,t)$ holds when we substitute $\xi_*(t)$ for x and $\eta_*(t)$ for u is equivalent to the condition that

$$\dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\xi_*(t), \pi(t), \eta_*(t), t)\,. \tag{X.47}$$

The pair of equations (X.47), (X.46) is equivalent to the single equation

$$\dot{\Theta}(t) = \vec{H}_{\eta_*(t),t}(\Theta(t))\,, \tag{X.48}$$

evolving in $\Omega \times \mathbb{R}_n$ (i.e., (x,p) space), where

$$\Theta(t) = (\xi_*(t), \pi(t))\,, \qquad H_{u,t}(x,p) \stackrel{\text{def}}{=} H(x,p,u,t)\,.$$

THE INTRINSIC FORM OF THE ADJOINT EQUATION:

The adjoint equation (X.43) is equivalent to the following statement:

For every vector field X of class C^1 on Ω, the equality

$$\frac{d}{dt}\Big(\pi(t) \cdot X(\xi_*(t))\Big) = \pi(t) \cdot [f_{\eta_*(t),t}, X](\xi_*(t))$$

holds, where

$$f_{u,t}(x) \stackrel{\text{def}}{=} f(x,u,t)\,.$$

ASSUMPTIONS FOR THEOREM OC-FT:

(A1) Ω is an open subset of $\mathbb{R}^n$.

(A2) U is a metric space.

(A3) $a, b \in \mathbb{R}$ and $a < b$.

(A4) $\Omega \times U \times [a,b] \ni (x,u,t) \mapsto f(x,u,t) \in \mathbb{R}^n$ and $\Omega \times U \times [a,b] \ni (x,u,t) \mapsto L(x,u,t) \in \mathbb{R}$ are continuous maps such that

(A4.a) the maps $\Omega \ni x \mapsto f(x,u,t) \in \mathbb{R}^n$ and $\Omega \ni x \mapsto L(x,u,t) \in \mathbb{R}$ are of class C^1 for each fixed (u,t),

(A4.b) the maps $\Omega \times U \times [a,b] \ni (x,u,t) \mapsto \frac{\partial f}{\partial x}(x,u,t) \in \mathbb{R}^{n\times n}$ and $\Omega \times U \times [a,b] \ni (x,u,t) \mapsto \frac{\partial L}{\partial x}(x,u,t) \in\in \mathbb{R}_n$ are continuous.

(A5) $\mathscr{U} = C^{0,pc}([a,b],U)$.

(A6) S is a subset of Ω.

(A7) $(\xi_*, \eta_*) \in TCP(\Omega, U, \mathscr{U}, f)$ and $\xi_*(b) \in S$.

(A8) C is a tangent cone of S at $\xi_*(b)$.

(A9) $J(\xi_*, \eta_*) \le J(\xi, \eta)$ for all $(\xi, \eta) \in TCP(\Omega, U, \mathscr{U}, f)$ such that $\xi(b) \in S$, where

$$J(\xi,\eta) \stackrel{\text{def}}{=} \int_a^b L(\xi(t), \eta(t), t)\, dt\,.$$

THEOREM OC-FT: ***(Optimal control theorem, fixed time.)*** *Assume that (A1,2,3,4,5,6,7,8,9) hold. Define the* ***Hamiltonian***

$$\Omega \times \mathbb{R}_n \times U \times \mathbb{R} \times [a,b] \ni (x,p,u,p_0,t) \mapsto H(x,p,u,p_0,t) \in \mathbb{R}$$

by letting

$$H(x,p,u,p_0,t) \stackrel{\text{def}}{=} p \cdot f(x,u,t) - p_0(x,u,t)\,.$$

Then there exists a pair (π, π_0) such that

(#) $\pi : [a,b] \mapsto \mathbb{R}_n$ *is a continuous map and* $\pi_0 \in \mathbb{R}$*;*

(AE) (the "adjoint equation")

$$\dot{\pi}(t) = -\pi(t) \cdot \frac{\partial f}{\partial x}(\xi_*(t), \eta_*(t), t) + \pi_0 \frac{\partial L}{\partial x}(\xi_*(t), \eta_*(t), t) \qquad \text{(X.49)}$$

at all points of continuity t of η_*;

(HM) (the "Hamiltonian maximization condition")

$$H(\xi_*(t), \pi(t), \eta_*(t), \pi_0, t) = \max\Big\{H(\xi_*(t), \pi(t), u, \pi_0, t) : u \in U\}$$

at all points of continuity t of η_*;

(TC) (the "transversality condition")

$$\pi(b) \in -C^{\dagger};$$

(NT) (the "nontriviality condition") $(\pi(t), \pi_0) \neq (0,0)$ *for some t (which implies that* $(\pi(t), \pi_0) \neq (0,0)$ *for all t);*

(NN) (the "nonnegativtiy condition") $\pi_0 \geq 0$.

REMARK: The adjoint equation (X.49) can also be written as

$$\dot{\pi}(t) = -\frac{\partial H}{\partial x}(\xi_*(t), \pi(t), \eta_*(t), \pi_0, t)\,. \tag{X.50}$$

The fact that dynamical equation $\dot{x} = f(x,u,t)$ holds when we substitute $\xi_*(t)$ for x and $\eta_*(t)$ for u is equivalent to the condition that

$$\dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\xi_*(t), \pi(t), \eta_*(t), \pi_0, t)\,. \tag{X.51}$$

The pair of equations (X.51), (X.50) is equivalent to the single equation

$$\dot{\Theta}(t) = \vec{H}_{\eta_*(t),\pi_0,t}(\Theta(t))\,, \tag{X.52}$$

evolving in $\Omega \times \mathbb{R}_n$, i.e., (x,p) space (the cotangent bundle of Ω), where

$$\Theta(t) = (\xi_*(t), \pi(t))\,,$$

and

$$H_{u,p_0,t}(x,p) \stackrel{\text{def}}{=} H(x,p,u,p_0,t)\,.$$

THE INTRINSIC FORM OF THE ADJOINT EQUATION:

The adjoint equation (X.49) is equivalent to the following statement:

For every vector field X of class C^1 on Ω, the equality

$$\frac{d}{dt}\Big(\pi(t)\cdot X(\xi_*(t))\Big) = \pi(t)\cdot[f_{\eta_*(t),t}, X](\xi_*(t)) + \pi_0 X L_{\eta_*(t),t}(\xi_*(t))$$

holds, where

$$f_{u,t}(x) \stackrel{\text{def}}{=} f(x,u,t) \quad \text{and} \quad L_{u,t}(x) \stackrel{\text{def}}{=} L(x,u,t)\,.$$

ASSUMPTIONS FOR THEOREM OC-VT:

(Vaca.1) Ω is an open subset of $\mathbb{R}^n$.

(Vaca.2) U is a metric space.

(Vaca.3) $T_* \in \mathbb{R}$ and $T_* > 0$.

(Vaca.4) $\Omega \times U \ni (x,u) \mapsto f(x,u) \in \mathbb{R}^n$ and $\Omega \times U \ni (x,u) \mapsto L(x,u) \in \mathbb{R}$ are continuous maps such that

(Vaca.4.a) the maps $\Omega \ni x \mapsto f(x,u) \in \mathbb{R}^n$ and $\Omega \ni x \mapsto L(x,u) \in \mathbb{R}$ are of class C^1 for each fixed u,

(Vaca.4.b) the maps $\Omega \times U \ni (x,u) \mapsto \frac{\partial f}{\partial x}(x,u) \in \mathbb{R}^{n\times n}$ and $\Omega \times U \ni (x,u,t) \mapsto \frac{\partial L}{\partial x}(x,u) \in\in \mathbb{R}_n$ are continuous.

(Vaca.5) $\alpha > 0$ and $\mathscr{U} = \bigcup_{T_*-\alpha \le T \le T_*+\alpha} C^{0,pc}([0,T],U)$.

(Vaca.6) S is a subset of Ω.

(Vaca.7) $(\xi_*,\eta_*) \in TCP(\Omega,U,\mathscr{U},f)$, $\eta_* \in C^{0,pc}([0,T_*],U)$, and $\xi_*(T_*) \in S$.

(Vaca.8) C is a tangent cone to S at $\xi_*(T_*)$.

(Vaca.9) $J(\xi_*,\eta_*) \le J(\xi,\eta)$ for all $(\xi,\eta) \in TCP(\Omega,U,\mathscr{U},f)$ such that $\xi(b) \in S$, where

$$J(\xi,\eta) \stackrel{\text{def}}{=} \int_0^T L(\xi(t),\eta(t),t)\,dt \text{ if } \eta \in C^{0,pc}([0,T],U)\,..$$

THEOREM OC-VT: ***(Optimal control theorem, variable time.)*** *Assume that (Vaca.1,2,3,4,5,6,7,8,9) hold. Define the* ***Hamiltonian*** $\Omega \times \mathbb{R}_n \times U \times \mathbb{R} \ni (x,p,u,p_0,t) \mapsto H(x,p,u,p_0) \in \mathbb{R}$ *by letting*

$$H(x,p,u,p_0) \stackrel{\text{def}}{=} p\cdot f(x,u) - p_0(x,u)\,.$$

Then there exists a pair (π,π_0) *such that*

(#) $\pi : [a,b] \mapsto \mathbb{R}_n$ *is a continuous map and* $\pi_0 \in \mathbb{R}$*;*

(AE) (the "adjoint equation")

$$\dot{\pi}(t) = -\pi(t)\cdot\frac{\partial f}{\partial x}(\xi_*(t),\eta_*(t)) + \pi_0\frac{\partial L}{\partial x}(\xi_*(t),\eta_*(t)) \qquad \text{(X.53)}$$

at all points of continuity t *of* η_**;*

(HM) *(the "Hamiltonian maximization condition")*

$$H(\xi_*(t),\pi(t),\eta_*(t),\pi_0) = \max\Big\{H(\xi_*(t),\pi(t),u,\pi_0) : u \in U\Big\}$$

at all points of continuity t of η_**;*

(ZH) *(the "zero Hamiltonian condition")*

$$H(\xi_*(t),\pi(t),\eta_*(t),\pi_0) = 0$$

at all points of continuity t of η_**;*

(TC) *(the "transversality condition")*

$$\pi(b) \in -C^\dagger\,;$$

(NT) *(the "nontriviality condition")* $(\pi(t),\pi_0) \neq (0,0)$ *for some t (which implies that* $(\pi(t),\pi_0) \neq (0,0)$ *for all t);*

(NN) *(the "nonnegativtiy condition")* $\pi_0 \geq 0$.

ASSUMPTIONS FOR THEOREM SEP-VT:

(Ox.1) Ω is an open subset of $\mathbb{R}^n$.

(Ox.2) U is a metric space.

(Ox.3) $T_* \in \mathbb{R}$ and $T_* > 0$.

(Ox.4) $\Omega \times U \ni (x,u) \mapsto f(x,u) \in \mathbb{R}^n$ is a continuous map such that

(Ox.4.a) the map $\Omega \ni x \mapsto f(x,u) \in \mathbb{R}^n$ is of class C^1 for each fixed u,

(Ox.4.b) the map $\Omega \times U \ni (x,u) \mapsto \frac{\partial f}{\partial x}(x,u) \in \mathbb{R}^{n\times n}$ is continuous.

(Ox.5) $\alpha > 0$ and $\mathscr{U} = \bigcup_{T_*-\alpha\leq T\leq T_*+\alpha} C^{0,pc}([0,T],U)$.

(Ox.6) S is a subset of Ω.

(Ox.7) $\eta_* \in C^{0,pc}([0,T_*],U)$, $(\xi_*,\eta_*) \in TCP(\Omega,U,\mathscr{U},f)$, and $\xi_*(T_*) \in S$.

(S8) C is a tangent cone of S at $\xi_*(b)$.

(S9) $\mathscr{R}_{\Omega,U,\mathscr{U},f}(\xi_*(0)) \cap S = \{\xi_*(T)$, where

$$\mathscr{R}_{\Omega,U,\mathscr{U},f}(\xi_*(0)) \stackrel{\text{def}}{=} \Big\{ \xi(T) : \eta \in C^{0,pc}([0,T],U), (\xi,\eta) \in TCP(\Omega,U,\mathscr{U},f) \Big\} .$$

THEOREM SEP-VT: ***(Separation theorem, variable time.)*** *Assume that (Ox.1,2,3,4,5,6,7,8,9) hold. Define the* ***Hamiltonian***

$$\Omega \times \mathbb{R}_n \times U \times \mathbb{R} \times [a,b] \ni (x,p,u,p_0,t) \mapsto H(x,p,u) \in \mathbb{R}$$

by letting

$$H(x,p,u) \stackrel{\text{def}}{=} p \cdot f(x,u) .$$

Then there exists a π such that

(#) $\pi : [a,b] \mapsto \mathbb{R}_n$ *is a continuous map;*

(AE) (the "adjoint equation")

$$\dot{\pi}(t) = -\pi(t) \cdot \frac{\partial f}{\partial x}(\xi_*(t), \eta_*(t)) \tag{X.54}$$

at all points of continuity t of η_*;

(HM) (the "Hamiltonian maximization condition")

$$H(\xi_*(t), \pi(t), \eta_*(t)) = \max \Big\{ H(\xi_*(t), \pi(t), u) : u \in U \}$$

at all points of continuity t of η_*;

(ZH) (the "zero Hamiltonian condition")

$$H(\xi_*(t), \pi(t), \eta_*(t)) = 0$$

at all points of continuity t of η_*;

(TC) (the "transversality condition")

$$\pi(b) \in -C^\dagger ;$$

(NT) (the "nontriviality condition") $\pi(t) \neq 0$ *for some t (which implies that* $\pi(t) \neq 0$ *for all t).*

$$\boxed{\begin{array}{ccc} \text{Theorem SEP} - \text{FT} & \Longrightarrow & \text{Theorem SEP} - \text{VT} \\ \Downarrow & & \Downarrow \\ \text{Theorem OC} - \text{FT} & \Longrightarrow & \text{Theorem OC} - \text{VT} \end{array}}$$

Chapter XI

The geometric connection

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

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

A similar question could, of course, be asked of "geometry." Exactly what is geometry about? Acording to E. T. Bell, in [9], p. 301,

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

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

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

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

- ***first***, by making it clear that "lines" refers not only to "straight lines" but to more general curves (exactly as "surfaces" does not mean "planes");

- ***second***, by incorporating into the definition the revolutionary 19th century discovery that in mathematics there is no such thing as "space" but, rather, there is a multiplicity of "spaces," of which the space of classical Euclidean geometry is one but not the only one;

- ***third***, by taking into account the modern understanding that the concept of a truly general "curve" or "surface" or "solid" is much too broad, and what geometry really studies is some special classes of "curves, surfaces and solids,"

- ***fourth***, by getting rid of the limitation to three dimensions implicit in the reference to "curves, surfaces and solids," and the statement that these objects live in "space."

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

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

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

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

Even though it is far from clear what it is exactly that makes an object "geometric," mathematicians feel and have always felt that they can tell one when they see it. For example, if you were to define an ellipse to be a curve that, relative to some system of Cartesian coordinates x, y in the plane has an equation of the form

$ax^2 + by^2 = 1$, with $a < 0$ and $b > 0$, then this would not be regarded[1] as "geometric." If, on the other hand, you take the very same notion, and characterize it instead as "a curve which is the locus of all the points for which the sum of the distances to two fixed given points is constant," then the "geometricity" of the ellipse becomes apparent.

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

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

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

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

From all this we have to conclude that "geometricity" is in the eyes of the beholder, and the notion of a "geometric object" is undergoing perpetual change. What makes an object "geometric" is that ***it admits a simple characterization***, with

---

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

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

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

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

XI.1 Birds, bees, and the birth of geometry and optimal control

Homo sapiens is not the first entity that has engaged in problem-solving and arrived at "geometric" curves and surfaces as solutions of design problems. The animal kingdom already exhibits many examples of optimal control problems and their geometric solutions[4]. For example, J. L. Coolidge, in his 1940 book [49] (page 1), asked

> What is the earliest example of puposive geometrical construction?

and argued that

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

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

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

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

COMMENT NO. 29. NEED LOTS OF REFERENCES HERE!!!!!

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

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

That is, the bees want their cells to be identical regular polygons ("for bees would have none of figures which are not uniform") and they want them to tile the plane. This leaves them with only three options, namely, triangles, squares and hexagons,

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

[6]Quoted from Coolidge [49], p. 1

and the bees cleverly choose the hexagons because they maximize area ("hold more honey"). Pappus is not entirely precise here. It is not true that any hexagon will have a larger area than any triangle or square. What is true is that, for regular polygons with n sides ***and a given perimeter*** P, the enclosed area grows with n. In particular, a hexagon of a given perimeter will "hold more honey" than a triangle or square of the same perimeter. But, if we assume that our superintelligent bees start from the natural observation that "we are M bees and we have an area A that we can tile with our cells," tben it is quite obvious that, whether they use hexagons or squares or triangles the area of each cell will be $\frac{A}{M}$. This means, of course, that if they had used a triangular tiling then the perimeter of the cells would be larger than that of the hexagonal cells. It is unclear to us exactly what process of reasoning may have led Pappus' wise bees to focus on area maximization ***subject to a fixed perimeter***.

On the other hand,

COMMENT NO. 30. Must complete this section

XI.2 Line segments

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

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

The fascinating story of the "ritual origins of geometry" is told in a series of papers (cf. [119, 120, 121]) by A. Seidenberg, and was incorporated by B. L. van der Waerden in his 1983 study [152] of geometry and algebra in ancient civilizations.

The important point for us is that

COMMENT NO. 31. Must complete this section

Curve optimization problems had been studied and solved at least since the Greeks. The oldest of all is of course the one of determining the shortest path joining two points, whose solution—which must have been well known since very ancient times—is a straight-line segment.

XI.3 The parallel postulate

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

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

Euclid himself appears to have felt that this postulate was less evident than the others, as shown by the fact that he avoids using it for as long as possible. And after Euclid, according to R. Bonola ([19], p. 2),

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

In the 17th century, the Fifth Postulate attracted the interest of many mathematicians, several of whom (for example, P. A. Cataldi in 1603 and 1604, G. A. Borelli in 1658, G. Vitale in 1680, and J. Wallis in 1693, cf. [38, 39], [20],citevitale, [154]) published works entirely devoted to the subject.

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

XI.4 Circles and the isoperimetric problem

The oldest curve optimization problem after the one one about segments is the *isoperimetric problem*, also known as *Dido's problem*, inspired by the mythical story told by Virgil (70-19 B.C) in the *Aeneid* about the foundation of Carthage (*c.* 850 B.C.): the question is (I) *to find the plane curve of a given length that encloses the largest possible area*. (It is quite easy to see that this is equivalent to the problem of (II) *to determine the shortest of all plane curves that enclose a given area*.) The solution was known by the Greeks to be the circle, although it took until the 19th century for this to be proved in a way that meets our contemporary standards of rigor. Some philosophers were convinced that *this had to mean something* in the grand design of the universe. For example, Aristotle believed that "all motion is either rectilinear or circular or a combination of the two, because these are the only 'simple movements' " ([155], p. 4). He wrote (*ibid.*) that:

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

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

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

XI.5 Shortest paths in geometrical optics

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

As we have already pointed out in §V.1, Hero of Alexandria showed that when a light ray emitted by an object is reflected by a mirror it follows a path from the object to the eye which is the shortest of all possible such paths. In Hero's setting, which involved a single medium and therefore a constant speed of light, "shortest" was equivalent to "fastest." This was no longer the case in the work of Fermat (1601-1665), who formulated the general principle that *light rays follow the fastest—i.e., minimum-time—paths*, and showed how this explained not only Hero's observation about reflection but also Snell' law of refraction. (See Scholium

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

VII, in page 135; we shall soon see that Fermat's principle played a crucial role in Johann Bernoulli's solution of the brachistochrone problem, cf. Chapter XIV below, especially §XIV.4 in page 300.)

These observations were no longer speculative guesses about the distant heavens, but empirically verifiable assertions about the world around us, which in fact made nontrivial testable predictions. (For example, as we indicated §Scholium VII, Descartes had an alternative explanation for Snell' law, which required that light travel *faster* in a denser medium, whereas Fermat's principle yielded the correct law of refraction only if light was assumed to travel more slowly in a denser medium. Huygens (1629-1695) subsequently provided a justification of Fermat's principle by deriving it from wave optics.)

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

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

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

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

or

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

Under the second interpretation, Fermat's principle would still say that light follows a shortest path—i.e., a "geodesic"—but in a different, non-Euclidean geometry. With hindsight, we now know that Interpretation 2 is the most fruitful. For example, G. Lochak writes in [94], p. 57–58 [our translation]:

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

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

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

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

COMMENT NO. 32. HAVE TO CHECK LEIBNIZ' ORIGINAL LATIN TO MAKE SURE.

XI.6 Johann Bernoulli misses an opportunity to solve the problem of Euclid's fifth postulate

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

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

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

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

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

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

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

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

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

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

Johann Bernoulli had essentially all the tools he needed to solve the problem. Why didn't he? What he lacked was the understanding of the *geometric* significance of the cycloids and half circles that he had found as solutions of his mini-

[8] Since one usually requires that the state space of a control system be an open subset of $\mathbb{R}^n$ or a smooth manifold, it is more natural to regard the brachistochrone and related problems as defined in the *whole plane* rather than in a half-plane. This is why we are using the absolute value of y. We will pursue this point in XIV.10, where we will study the "reflected brachistochrnoe problem."

mization problems. He certainly saw these curves as "geometric," but failed to see that they could serve as the "lines" of a different geometry.

XI.7 The missing link: curve minimization as a rich source of new geometries

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

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

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

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

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

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

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

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

XI.8 Covariance and invariance, from Lagrange to Riemann

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

Lagrange allowed the configuration of a collection of N particles to be represented in terms of completely general systems of $3N$ coordinates. This was a decisive step in the process that would free geometry from the constraint of having to deal with 3-dimensional Euclidean space. "To appreciate how all was new in this work [the *Mécanique Analytique*], let us point out that, when Lagrange described a motion in terms of a system of Cartesian coordinates in space, this was such a nontrivial step that he devoted two pages to trying to persuade the reader" (Lochak, [94], p. 49).

He then showed that the Euler-Lagrange system (V.3) has the same form in "curvilinear" or "generalized" coordinates, so that (V.3) is invariant under *arbitrary*, nonlinear coordinate changes. This was the first example of what would later become the cornerstone of differential geometry, namely, the principle of "coordinate democracy," according to which all coordinate systems, "rectangular" as well as "curvilinear," have to be treated equally.

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

> It is perhaps one of the principal advantages of our method that it expresses the equations of every problem in the most simple form relative to each set of variables and that it enables us to see beforehand which variables one should use in order to facilitate the integration as much as possible. [Quoted from [155], p. 33.]

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

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

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

The functions g_{ij} define the metric relative to a particular coordinate system $(x^1,\ldots,x^n)$. If we change to another coordinate system $\tilde{x}$, related to x by $\tilde{x} = \Phi(x)$, $x = \Psi(\tilde{x})$, then the coefficients $\tilde{g}_{\alpha\beta}$ expressing the metric in the new system are related to the g_{ij} by

$$\tilde{g}_{\alpha\beta}(\tilde{x}) = \sum_{ij} g_{ij}(x) \cdot \frac{\partial x^i}{\partial \tilde{x}^\alpha} \cdot \frac{\partial x^j}{\partial \tilde{x}^\beta}\,. \tag{XI.1}$$

The metric itself should be thought of as a "coordinate free object," represented in each coordinate system by an array $\{g_{ij}\}$ of functions. These arrays are related

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

by the transformation (XI.1). More generally, *covariant quantities are "objects" that for each coordinate system are represented by an array of functions, in such a way that when one changes coordinates these arrays transform according to some specified law.* There are different types of such quantities, depending on the transformation law, and among them there is a class of quantities called *tensors*.

Remark XI.1. In modern differential geometry, one seeks to define covariant objects intrinsically, rather than as "arrays of functions that transform in a certain way." This is a much better and more powerful language, exactly as, say, the language of linear maps on abstract vector spaces is superior to the language of matrices. But to do justice to this idea one needs much more algebra, so we will keep our discussion simple by sticking with the old-fashioned way of talking about covariance. □

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

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

$$\tilde{R}^{\alpha}_{\beta\gamma\delta}(\tilde{x}) = \sum_{ijk\ell} R^{i}_{jk\ell}(x) \cdot \frac{\partial \tilde{x}^{\alpha}}{\partial x^{i}} \cdot \frac{\partial x^{j}}{\partial \tilde{x}^{\beta}} \cdot \frac{\partial x^{k}}{\partial \tilde{x}^{\gamma}} \cdot \frac{\partial x^{\ell}}{\partial \tilde{x}^{\delta}} . \tag{XI.2}$$

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

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

$$R^{i}_{jk\ell} = \frac{\partial \Gamma^{i}_{j\ell}}{\partial x^{k}} - \frac{\partial \Gamma^{i}_{jk}}{\partial x^{\ell}} + \sum_{s} \left(\Gamma^{i}_{sk} \Gamma^{s}_{j\ell} - \Gamma^{i}_{s\ell} \Gamma^{s}_{jk} \right) , \tag{XI.3}$$

where the Γ_{ij}^k are the so-called *Christoffel symbols*:

$$\Gamma_{jk}^i = \frac{1}{2}\sum_\ell g^{i\ell}\left(\frac{\partial g_{\ell k}}{\partial x^j} + \frac{\partial g_{\ell j}}{\partial x^k} - \frac{\partial g_{jk}}{\partial x^\ell}\right). \tag{XI.4}$$

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

XI.9 Invariants and Klein's Erlangen program

In 1872, F. Klein formulated his *Erlangen program*, whose main idea is that each geometry is characterized by the *group of transformations* which preserve the relevant properties of the geometry. Associated to such a group there is usually a collection of *invariants* and a class of *covariant objects*. For example, for Euclidean space the transformation group is the orthogonal group, an obvious and important invariant is the distance between two points, and the components of a vector are covariant quantities. For the study of general differentiable manifolds, the natural "group" is really a *pseudogroup*, namely, the set of diffeomorphisms, and differential geometry has produced a huge zoo of covariant objects: tensors of various kinds (an in particular vector fields and differential forms), jet spaces, connections, etc. For control systems, many of the properties of interest (e.g. the structure of minimum time trajectories, feedback stabilizability, feedback linearizability) are invariant under this general pseudogroup, so one should try to characterize them in a coordinate-free way, using invariants. A *complete set of invariants* for a class of properties is a set of invariants that complete determines these properties. For example, if we are interested in the solvability of a system of linear equations $Ax = b$, a complete set of invariants is provided by a pair of integers, namely, the rank of A and the rank of the "augmented matrix" $[A;b]$. □

Chapter XII

Beyond the maximum principle, geometrically

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

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

Part of the problem here is to find the right language for talking about these

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

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

XII.1 Covariance and invariance in control theory: the Lie bracket

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

To see this, let us recall that the main aspect that makes optimal control stand out and be different from the calculus of variations is the presence of a nontrivial dynamics, given by Equation (XVI.1). In the special but very important case of an *autonomous system*, i.e., when f does not depend on t, the dynamics (XVI.1) is really a family $\mathbf{f} = \{f(\cdot,u)\}_{u\in U}$ of vector fields, parametrized by $u \in U$. (From now on, we write $f_u(q)$ rather than $f(q,u)$, to emphasize that f_u should be regarded as a vector field.)

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

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

$$[f,g] = \frac{\partial g}{\partial x}.f - \frac{\partial f}{\partial x}.g\,. \qquad \text{(XII.1)}$$

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

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

THE FORMAL DEFINITION OF THE LIE BRACKET

Let us regard $\mathbb{R}^n$ as a space of *column* vectors, so the members of $\mathbb{R}^n$ are n-dimensional columns (or, equivalently, n by 1 matrices) $x = \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}$ whose entries $x^1, x^2, \ldots, x^n$ are real numbers. If Ω is an open subset of $\mathbb{R}^n$, and f is a vector field on Ω (that is, a map from Ω to $\mathbb{R}^n$), we regard f as a column of functions, so that $f(x) = \begin{bmatrix} f^1(x) \\ \vdots \\ f^n(x) \end{bmatrix}$ for $x \in \Omega$. Also, we write D_i for the partial differential operator $\frac{\partial}{\partial x^i}$. If $f : \Omega \mapsto \mathbb{R}^n$ is of class C^1, we use Df to denote the Jacobian matrix of f, so Df is a square matrix of functions, and $Df(x) = \begin{bmatrix} D_1 f^1(x) & \ldots & D_n f^1(x) \\ \vdots & \vdots & \vdots \\ D_1 f^n(x) & \ldots & D_n f^n(x) \end{bmatrix}$ for $x \in \Omega$.

If $f, g : \Omega \mapsto \mathbb{R}^n$ are vector fields of class C^1, the *Lie bracket* $[f, g]$ of f and g is another vector field (of class C^0) defined by

$$[f,g](x) = Dg(x) \cdot f(x) - Df(x) \cdot g(x) \quad \text{for} \quad x \in \Omega\,.$$

If we write $h = [f, g]$, and $h(x) = \begin{bmatrix} h^1(x) \\ \vdots \\ h^n(x) \end{bmatrix}$, then the components h^j of h are given, for $j = 1, \ldots, n$, by

$$h^j(x) = \sum_{i=1}^{n} \Big((D_i g^j)(x) f^i(x) - (D_i f^j)(x) g^i(x) \Big) \quad \text{for} \quad x \in \Omega\,.$$

COMMENT NO. 33. Here we must insert a box with the following simple example showing how the Lie bracket shows up. If you have m dollars in your bank account (that is, you have $m(t)$ dollars at time t, earning interest continuously at a rate ρ), and in addition at time t you spend $s(t)$ dollars per unit time (where $s(t)$ can be negative, in which case you are earning money), then m evolves according to

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

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

COMMENT NO. 34. At this point we should also include Vel Jurdjevic's famous pedagogical example: the operations of jumping into a swimming pool and filling the pool with water do not commute.

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

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

where

$$\delta(t) = qe^{tf}e^{tg}e^{-tf}e^{-tg}\,. \tag{XII.3}$$

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

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

But certain things computable in terms of the f_u are independent of the choice of coordinates. For example, the fact that two vector fields f, g commute (i.e., $[f,g] = 0$), is intrinsic. Moreover, this property has a very direct control-theoretic meaning: $[f,g] \equiv 0$ if and only if, for all control functions $u(\cdot)$, $v(\cdot)$, and all $\bar{q}$, the solution of

$$\dot{q}(t) = u(t)f(q(t)) + v(t)g(q(t)), \quad q(0) = \bar{q}, \tag{XII.4}$$

is always given, for small t, by

$$q(t) = \bar{q}e^{U(t)f}e^{V(t)g}, \tag{XII.5}$$

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

It then turns out that

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

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

Remark XII.1. Actually, a stronger result is true, but we need more terminology to state it. For a system $\{f_u\}_{u \in U}$ of real-analytic vector fields, let L be the Lie algebra of vector fields generated by the f_u. Then *the Lie bracket relations at a point q form a complete set of invariants for all properties of the system that are preserved by local diffeomorphisms of the maximal integral manifold S of L through q.*

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

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

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

XII.2 Control, Lie brackets, zoology and medieval bestiaries

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

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

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

$$\dot{x}_1 = x_2 + 2x_2x_3\,, \qquad \dot{x}_2 = x_3 - x_1x_2 + x_2^3\,, \tag{XII.6}$$
$$\dot{x}_3 = u + x_2^2 + x_1x_3 - x_1^2x_2 + 2x_1x_2^3\,,$$

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

$$\dot{x}_1 = u_1\,,\ \dot{x}_2 = u_2\,,\ \dot{x}_3 = u_1x_2 - u_2x_1\,, \tag{XII.7}$$

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

So ***Lie brackets help us organize our knowledge of nonlinear systems in terms of their deep structural properties.*** This way of classifying systems is as superior to the naïve approach based on a superficial counting of the number of nonlinearities as the animal taxonomy of modern zoology is to the classifications found in medieval bestiaries. Exactly as a dolphin is structurally closer to a zebra than to a shark—in spite of the very visible ways, such as shape and habitat, in which a dolphin resembles a shark more than a zebra—System (XII.6) is much closer to a linear system than System (XII.7). Lie brackets enable us to discover this, and then one can find abundant confirmation by looking at the properties of both systems. (For example, (XII.6) is globally stabilizable by a smooth state feedback, but (XII.7) is not even locally stabilizable near any point.)

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

XII.3 Using Lie brackets to understand optimal trajectories

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

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

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

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

For example, the "singular" control of Example XII.2 is closely related to a property of the Lie brackets of the vector fields f and g. Precisely, *the singular trajectory exists and is contained in the x axis because $[f,g]$ vanishes on the x axis.*

Since the computation of Example XII.2, as presented, does not explicitly display any Lie brackets, we will now make the Lie brackets stand out by looking, more generally, at systems of the form $\dot{x} = f(x) + ug(x)$, with $|u| \leq 1$, without specializing to a particular pair of vector fields f, g.

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

XII.4 Implementing Lagrange's covariance idea in optimal control

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

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

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

An intrinsic version, treating all "generalized components" (sometimes known as "momentum functions") equally, can be formulated in a number of ways, e.g. using Poisson brackets. For a general version of the maximum principle on manifolds, using generalized components as above (and also for an alternative version involving connections along curves), we refer the reader to [134]. The crucial point of the version given in [134] is—considering, for simplicity, the case of the minimum time problem for a control system of the form $\dot{q} = f(q,u)$—that the adjoint equation is rephrased as the following statement (the "intrinsic adjoint equation"):

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

The point of using (IAE) is that it treats the inner product of with X—i.e., "the X component of p"—equally for all vector fields X, thus enabling us to follow Lagrange's prescription (cf. Page 276) that one should express "the equations of every problem in the most simple form relative to each set of variables and ... enables us to see beforehand which variables one should use in order to facilitate the integration as much as possible," by working in each case with the functions of the form $\langle p, X\rangle$ that are most suited to the problem.

XII.5 Why the language of Lie brackets is natural for optimal control

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

The reader should think of the following analogy. When one tries to solve a simple 2×2 system of linear equations such as $2x + 3y = 5$, $6x - 2y = 4$, one may fail to notice that the number $2.(-2) - 3.6$ is important. If, however, one "does it with letters" (as a Calculus student would say), i.e., sets out to solve a ***general*** system $a_{11}x + a_{12}y = b_1$, $a_{21}x + a_{22}y = b_2$, then one discovers immediately that the number $a_{11}a_{22} - a_{12}a_{21}$ determines the properties of the solutions, so it becomes natural to gives this number a name such as "determinant," and to look for a generalization of the concept to higher dimensions. Similarly, when one applies the maximum principle to study a particular problem such as the one of Example XII.2, one is in fact computing a Lie bracket, although this fact may go unnoticed. But when one "does it with letters," e.g. by working with a general problem of the form $\dot{x} = f(x) + ug(x)$, leaving f and g unspecified, then the Lie bracket $[f, g]$ stands out immediately as the object that "determines" what happens, exactly as the expression $a_{11}a_{22} - a_{12}a_{21}$ did in the linear algebra example.

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

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

Chapter XIII

Some open problems

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

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

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

or, perhaps, even more strongly, that

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

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

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

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

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

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

For more details on this problem, the reader may consult [128, 129]. □

Open Problem 2.. Let $f_1,\ldots,f_m$ be C^∞ vector fields on an open subset Q of $\mathbb{R}^n$. Let Λ be the Lie algebra of vector fields generated by $f_1,\ldots,f_m$. Suppose $\{g(q):g\in\Lambda(q)\}=\mathbb{R}^n$ for every $q\in Q$. Consider the minimum time problem for the control system $\dot{q} = u_1f_1(q) + \ldots + u_mf_m(q)$, with $u = (u_1,\ldots,u_m) \in U$, where U is the closed unit ball in $\mathbb{R}^m$. Assume $f_1(q),\ldots,f_m(q)$ are linearly independent at each point.

Are all minimum time trajectories smooth?

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

For more details on this problem and a long list of references, the reader is referred to [92]. □

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

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

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

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

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

For a general discussion of infinite chattering, we refer the reader to the book [159]. Some partial results on the unicycle with angular acceleration control are given in [139]. □

Chapter XIV

The brachistochrone, 300 years ago and today

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

XIV.1 Johann Bernoulli's challenge

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

[1] Jakob's method was quite different fromm Johann's, and at first sight seemed clumsier, but in the long run it has turned out to be more akin to the mainstream ideas of the calculus of variations, Hamilton-Jacobi theory, and dynamic programming, and is therefore widely considered to be of great historical importance in the development of optimal control, as shown in our earlier quotation from Stillwell. It will not, however, be discussed here, due to lack of space. Goldstine's book [66] give an excellent account.

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

COMMENT NO. 35. IN THE EARLIER VERSION WE HAD "DIVULGE." I LIKE "DISCLOSE" BETTER. IT'S CERTAINLY BETTER ENGLISH.

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

COMMENT NO. 36. PUT PICTURE OF PAGE WITH JOHANN'S LATIN TEXT AROUND HERE

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

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

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

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

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

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

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

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

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

XIV.2 Why was the brachistochrone so important in 1696-97?

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

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

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

COMMENT NO. 37. CHECK TO SEE IF THESE WERE JB'S EXACT WORDS

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

In the case of the brachistochrone, ***there is no way to guess***[6]. The problem was posed and solved at a time when the Calculus was being invented, and cannot

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

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

be solved without using Calculus. So it provides an impressive illustration of the power of the Calculus to solve problems that cannot be solved in any other way, and to discover truths about the world that can be stated and understood in "well known" terms[7], but can only be arrived at by using this new mathematical tool.

XIV.3 Formulation of the brachistochrone problem

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

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

(ii) $\frac{1}{2}(|\dot{f}_1(t)|^2 + |\dot{f}_2(t)|^2) = E + gf_2(t)$ for almost all $t \in [0,T]$,

where g is the gravitational constant.

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

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

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

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

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

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

$$H_+(\eta) = \{(x,y) : y \geq -\eta\}, \quad \text{where} \quad \eta = \frac{E}{g}, \tag{XIV.1}$$

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

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

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

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

One fact which is truly obvious is that *the solution cannot always be a straight line*, a possibility that Johann Bernoulli rightly warns against. For example, consider the extreme case when $E=0$ and $a_1=a_2=b_2=0\neq b_1$. We will show that in this situation there is a feasible path that goes from A to B in finite time. Since the straight-line segment S from A to B is horizontal, (ii) implies that the speed of motion along S vanishes, and then it will follow that S cannot be an optimal path, because the motion along S takes infinite time. To construct a feasible path from A to B we follow a circle C with center on the x axis. The circle can be parametrized by x, taking values in the interval $[0,b_1]$. The speed v of motion along C will be proportional to $\sqrt{y}$, and hence bounded away from 0 as long as x is bounded away from 0 and b_1. Near A, we can also use y as a parameter, and then $y \downarrow 0$ as $x \downarrow 0$, and v behaves like $\frac{dy}{dt}$ and also like $\sqrt{y}$, so $\frac{dt}{dy} \sim \frac{1}{\sqrt{y}}$, showing that t stays finite as x and y approach 0, since the function $y \mapsto \frac{1}{\sqrt{y}}$ is integrable near $y=0$. A similar argument shows that t stays finite near B. So C joins A to B in finite time. □

XIV.4 Johann Bernoulli's solution

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

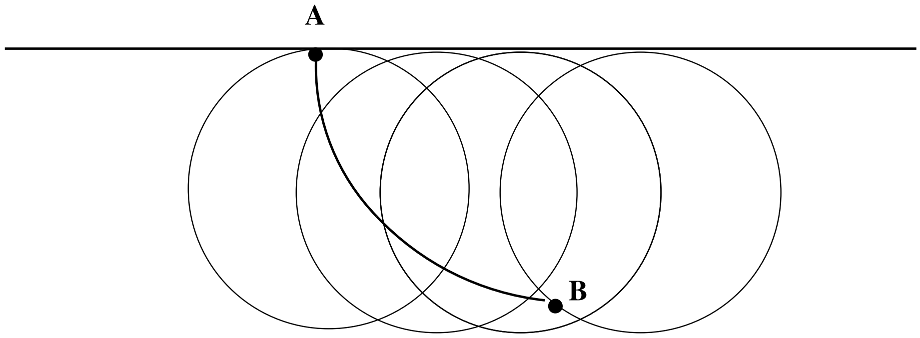


FIGURE 9. The cycloid

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

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

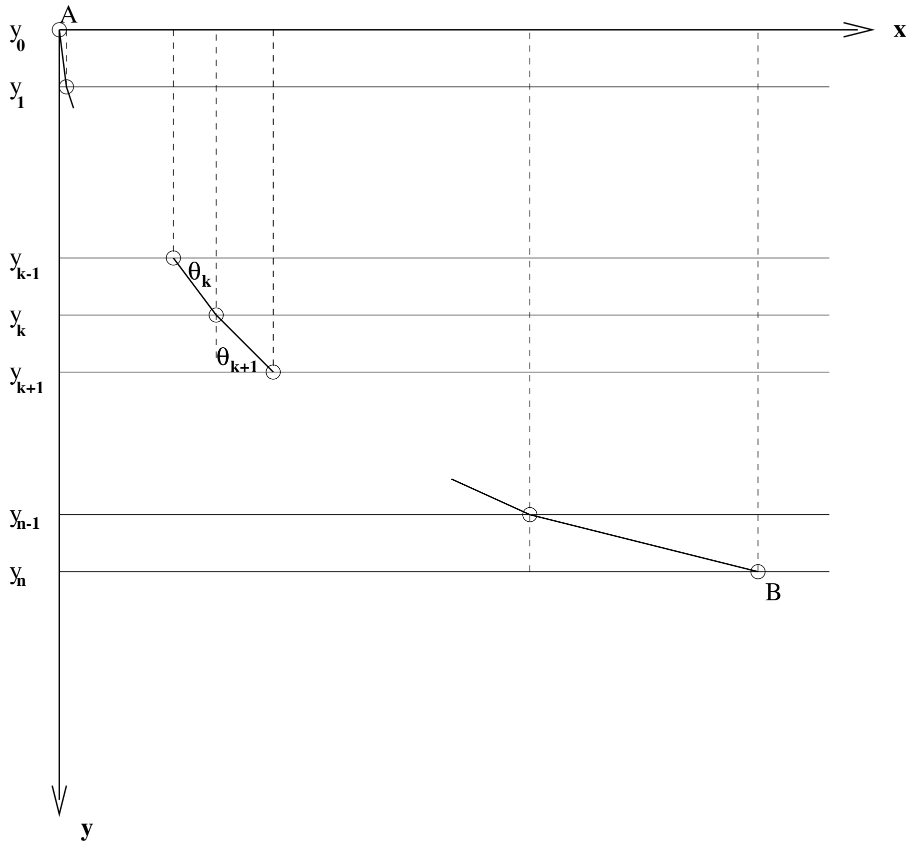


FIGURE 10. Bernouli"s discretization

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

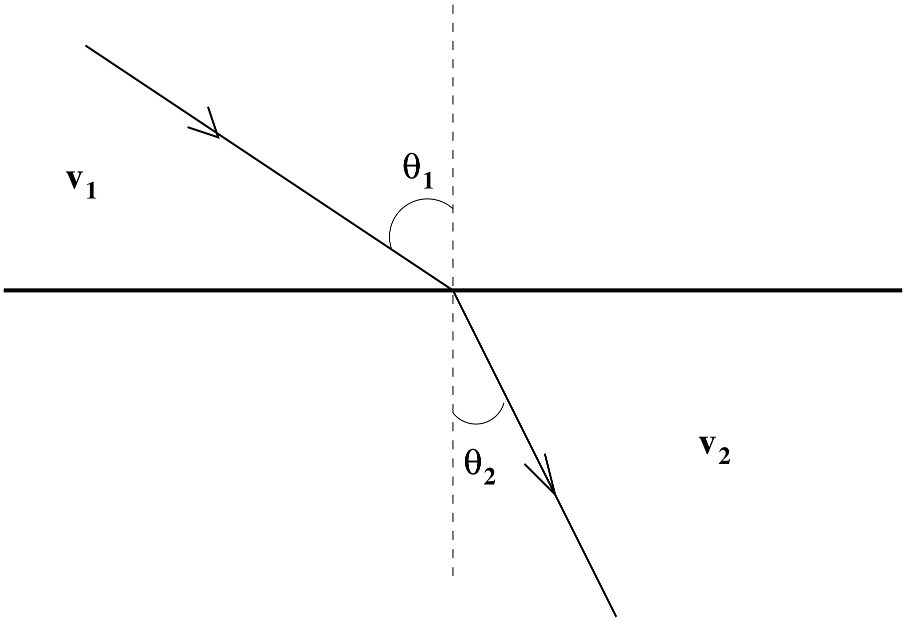


FIGURE 11. Snell's law

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

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

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

Johann Bernoulli used this law to conclude that the quantity $\frac{\sin\theta_k}{\sqrt{y_{k+1}}}$ will be a constant, since in each strip S_k the speed of our light ray is $\sqrt{y_{k+1}}$. Passing to the limit as $\delta \downarrow 0$, we conclude that the sine of the angle θ between the tangent to the brachistochrone and the vertical axis must be proportional to $\sqrt{y}$. Since $\sin\theta =$

$\frac{dx}{\sqrt{dx^2+dy^2}}$, we find that $\frac{dx^2}{dx^2+dy^2} = Ky$, where K is a constant. Then $\frac{dx^2+dy^2}{dx^2} = \frac{1}{Ky}$, i.e., $1+y'(x)^2 = \frac{C}{y}$, where $C = \frac{1}{K}$. So the curve described by expressing the y-coordinate of the brachistochrone as a function of its x-coordinate will satisfy the differential equation

$$y'(x) = \sqrt{\frac{C-y(x)}{y(x)}}\,, \tag{XIV.2}$$

with C a constant. The curves parametrized by

$$x(\varphi) = x_0 + \frac{C}{2}(\varphi - \sin\varphi)\,, \quad y(\varphi) = \frac{C}{2}(1-\cos\varphi)\,, \quad 0 \le \varphi \le 2\pi\,, \tag{XIV.3}$$

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

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

(The parameter count for (*) is obviously right: (XIV.3) describes a 2-parameter family of curves; requiring that the curve pass through A fixes one of the parameters, and then asking that it pass through B fixes the other. Proving (*) requires a bit more work, but is still quite easy.)

XIV.5 Remarks on Johann Bernoulli's argument

The argument that we have presented is the one of Johann Bernoulli, and Equation (XIV.2) is the one that he wrote in his paper, followed by the statement "from which I conclude that the *Brachistochrone* is the ordinary *Cycloid*." (Johann Bernoulli actually wrote $dy = dx\sqrt{\frac{x}{a-x}}$, but he was using x for the vertical coordinate and y for the horizontal one. Cf. [124], p. 394.) In present day notation, it is customary to adopt the convention that the symbol $\sqrt{r}$ stands for the *nonnegative* square root of r, but it is obvious that Johann Bernoulli did not have this in mind. What he meant was, clearly, what we would write as

$$y'(x) = \pm\sqrt{\frac{C-y(x)}{y(x)}}\,, \tag{XIV.4}$$

or, equivalently,

$$y(x)(1+y'(x)^2) = \text{constant}\,. \tag{XIV.5}$$

In particular, the solution curves should be allowed to have a negative slope, contrary to what a "modern" interpretation of (XIV.2) might suggest. But y' should stay continuous, so that a switching from a $+$ to a $-$ solution of (XIV.4) is not permitted.

Even with the more accurate rewriting (XIV.5), the differential equation derived by Johann Bernoulli also has spurious solutions, not given by (XIV.3). Indeed, for any $\bar{y} > 0$, the constant function $y(x) = \bar{y}$ is a solution, corresponding to $C = \bar{y}$. More generally, one can take an ordinary cycloid given by (XIV.3), follow it up to $\varphi = \pi$—so that $dy/dx = 0$—then follow the constant solution $y(x) = C$ for an arbitrary time T, and then continue with a cycloid given by (XIV.3). Such paths are, indeed, compatible with Huygens' law of refraction.

Remark XIV.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. $\square$

XIV.6 The calculus of variations approach

The spurious solutions, and all the other problems, such as the apparent arbitrariness of the requirement that y' be continuous, can be eliminated in a number of ways. For example, one can prove directly that the spurious trajectories are not optimal. Or one can use, as an alternative to Johann Bernoulli's method, the calculus of variations approach, based on writing the Euler-Lagrange equation (V.3).

It is easy to see that the brachistochrone problem can be put in the "standard" form (II.5), provided we ***postulate***[8] that it suffices to consider curves in the x, y plane that are graphs of functions $y = y(x)$ defined on the interval $[a_1, b_1]$. Then Constraint (ii) can be written (taking $E = 0$ and $2g = 1$ as before) as $dx^2 + dy^2 = y\,dt^2$, which gives $dt = \frac{\sqrt{dx^2+dy^2}}{\sqrt{y}} = L(y, \dot{y})dx$, where $L(y,u) = \frac{\sqrt{1+u^2}}{\sqrt{y}}$, and we are using x rather than t for the time variable, and writing $\dot{y}$ for dy/dx. So Johann Bernoulli's problem becomes that of minimizing the integral $\int_{a_1}^{b_1} L(y(x), \dot{y}(x))dx$ subject to $y(a_1) = a_2$ and $y(b_1) = b_2$.

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

This gives the equation

$$1+y'(x)^2+2y(x)y''(x)=0\,, \tag{XIV.6}$$

which is *stronger* than (XIV.5), since (XIV.5) is equivalent to $y'+y'^3+2yy'y''=0$, i.e., to $y'(1+y'^2+2yy'')=0$, whose solutions are those of (XIV.6) plus the spurious solutions found earlier.

It is easy to verify that the solutions of the Euler-Lagrange equation (XIV.6) are *exactly* the curves given by (XIV.3), without any extra spurious solutions, showing that, for the brachistochrone problem, ***the Euler-Lagrange method gives better results than Johann Bernoulli's approach.***

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

XIV.7 The brachistochrone as an optimal control problem

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

$$\dot{x}=u\sqrt{|y|}\,,\quad \dot{y}=v\sqrt{|y|}\,, \tag{XIV.7}$$

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

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

$$u=\frac{p_1}{|p|}\,,\quad v=\frac{p_2}{|p|}\,, \tag{XIV.8}$$

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

$$\dot{p}_1=0\,,\quad \dot{p}_2=-\alpha\,\frac{p_1u+p_2v}{2\sqrt{\alpha y}}=-\frac{\alpha|p|}{2\sqrt{\alpha y}}\,. \tag{XIV.9}$$

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

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

$$y'(x)=\frac{dy}{dx}=\frac{\dot{y}}{\dot{x}}=\frac{v}{u}=\frac{p_2}{p_1}\,, \tag{XIV.10}$$

we have

$$1 + y'(x)^2 = \frac{|p|^2}{p_1^2} \tag{XIV.11}$$

and

$$y''(x) = \frac{1}{p_1} \cdot \frac{dp_2}{dx} = \frac{\dot{p}_2}{p_1 \dot{x}}. \tag{XIV.12}$$

But (XIV.7) and (XIV.8) imply that

$$\dot{x} = \frac{p_1 \sqrt{\alpha y}}{|p|}, \tag{XIV.13}$$

and then (XIV.9) and (XIV.12) yield

$$y''(x) = -\frac{|p|^2}{2yp_1^2}. \tag{XIV.14}$$

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

$$1 + y'^2 + 2yy'' = 0, \tag{XIV.15}$$

which is exactly Equation (XIV.6). As we showed before, this leads to the cycloids, with no "spurious solutions." Notice that this argument does not involve any discretization or any use of Snell' law of refraction.

Notice also that in our control argument *we have not postulated that the solution curves could be represented as graphs of functions $y(x)$. We have proved it!* (Remember that in the calculus of variations case this was an extra assumption.)

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

XIV.8 Optimality proofs and the existence problem

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

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

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

One possibility (cf. Bliss [17]) is to use Hamilton-Jacobi theory. Another, much simpler approach, is to prove first the *existence* of an optimal trajectory. Indeed, once this is known for all $A, B \in H_+$, then it is clear that any optimal trajectory $q_* : [0,T] \mapsto H_+$ going from A to B must be such that its restriction $\xi\lceil[\alpha,\beta]$ to every subinterval $[\alpha,\beta]$ of $[0,T]$ is also optimal for the problem $P(\alpha,\beta)$ with endpoints $q_*(\alpha)$, $q_*(\beta)$. (This is a rather obvious point, known in optimal control theory as the "principle of optimality," and noticed in the case of the brachistochrone by Jakob Bernoulli.) If one assumes that $q_*(\alpha)$ and $q_*(\beta)$ belong to $\overset{o}{H}_+$ for $0 < \alpha < \beta < T$, then the necessary conditions for optimality give a unique candidate for the solution of $P(\alpha,\beta)$, namely, the unique curve in the family given by (XIV.3) that goes through $q_*(\alpha)$ and $q_*(\beta)$. So ξ is such that $\xi\lceil[\alpha,\beta]$ is given by (XIV.3) whenever $0 < \alpha < \beta < T$, and this easily implies that ξ itself is a cycloid given by (XIV.3). To complete the argument, one has to exclude the possibility of a solution of the problem with endpoints A, B which intersects the x axis at some other point. This can be done by a simple qualitative argument that we omit. (The key lemma is this: suppose L is a horizontal line in the upper half-plane H_+. Let P, Q be distinct points of L. Then the straight-line segment joining P to Q is faster than any trajectory from P to Q that lies entirely below L.)

So everything hinges upon the existence result. It turns out that this is not at all trivial if the brachistochrone problem is reformulated in terms of the classical calculus of variations, because the Lagrangian $L(y,u) = \frac{\sqrt{1+u^2}}{\sqrt{y}}$ has a singularity at $y = 0$.[9] (Existence theorems for classical calculus of variations problems that include the brachistochrone can be found in Cesari [40], §14.3 and §14.4. But they are not at all simple.)

In the control theory setting, however, the existence question is easily settled by a completely trivial application of Ascoli's theorem: given $A, B \in H_+$, let T be the infimum of the times of all trajectories going from A to B; let $\{\xi_j\}$ be a sequence of trajectories from A to B, defined on intervals $[0,T_j]$, such that $T_j \downarrow T$. Using Ascoli's theorem, extract a subsequence that converges uniformly to a curve $\xi : [0,T] \mapsto H_+$. In principle, ξ is just a trajectory of the "convexified" system,

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

in which the evolution equations are those of (XIV.7) but the control constraint is $u^2+v^2 \le 1$ 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 $u(t)^2+v(t)^2 \le 1$ was strict on a set of positive measure, then it would be possible to find an even faster trajectory from A to B by just reparametrizing ξ. So we see, once again, that *in the control setting a question that is hard from the point of view of the calculus of variations can become easy or even trivial.*

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

XIV.9 Hestenes' "variation of the brachistochrone problem"

In 1956, M. Hestenes wrote (in [73], p. 64):

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

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

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

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

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

The Hamiltonian is given by

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

The adjoint equation becomes

$$\dot{\pi}_x = 0 \qquad \dot{\pi}_z = -\pi_x \qquad \dot{\pi}_y = 0 \qquad \dot{\pi}_w = -\pi_y\,.$$

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

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

where ω_0 and ω_1 are constant two-dimensional vectors.

The Hamiltonian maximization condition then says that the vector

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

is given by

$$\mathbf{u}(t) = \frac{\omega(t)}{\|\omega(t)\|} . \tag{XIV.16}$$

The Hamiltonian must vanish along our trajectory. Therefore, $\pi(t)$ can never vanish, for if $\pi(t) = 0$ it would follow that $\pi_0 = 0$, contradicting the nontriviality condition. Hence $\omega(t)$ cannot vanish identically. So ω is a linear function which is not identically equal to zero. Therefore there is at most one value of t such that $\omega(t) = 0$. It follows that the function $\mathbf{u}$ of (XIV.16) is well defined except possibly at one point $\bar{t}$. Moreover, if such a $\bar{t}$ exists, then $\omega(t) = (t - \bar{t})\omega_1$, so

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

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

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

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

XIV.10 The reflected brachistochrone problem

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

Let us pursue Hestenes' idea that "it is interesting to speculate on what type of brachistochrone problem the Bernoulli brothers might have formulated had they lived in modern times." An obvious question for the brothers to ask would have been that of finding out what happens if we work in the whole plane. This is in fact a much more natural mathematical setting for the brachistochrone problem. We shall refer to this new question as the *reflected brachistochrone problem.* (It is because we were thinking all along of this extension that we wrote $\sqrt{|y|}$ rather that $\sqrt{y}$ in (XIV.7)).

Solving this more general problem amounts to finding the light rays when the medium is *the whole plane*, and the speed of light is $\sqrt{|y|}$. Notice that this problem is "completely controllable," in the sense that any two points A, B of $\mathbb{R}^2$, even if they lie on opposite sides of the x axis, can be joined by a feasible path. The right-hand side of (XIV.7) vanishes along the x axis, but this does not prevent the existence of feasible paths crossing the x-axis, because the function $\sqrt{|y|}$ is not Lipschitz near the x axis. (If the function was Lipschitz, then by the usual uniqueness theorem of ordinary differential equations, every solution going through a point on the x axis would have to be a constant curve.) However the same non-Lipschitz feature that makes the system controllable also renders the maximum principle inapplicable, in its classical and nonsmooth versions, including the Lojasiewicz version, since all these require a Lipschitz reference vector field.

Suppose we want to find the light rays for a medium which is the whole plane, with "speed of light" equal to $\sqrt{|y|}$. Can the necessary conditions help us find the solution? Suppose, for example, that we want to find an optimal trajectory from A to B, where A lies in the upper half-plane and B is in the lower half-plane. Then one can show, first of all, that an optimal trajectory ξ exists, using Ascoli's theorem as before. Next, using the usual necessary conditions for optimality, e.g. the Euler-Lagrange equation or the classical version of the maximum principle, one shows that any portion of an optimal curve which is entirely contained in the closed upper half plane or in the closed lower half plane is a cycloid given by (XIV.3), or a reflection or such a cycloid with respect to the x axis. Next, one sees that ξ cannot traverse the x axis more than once. (This requires an elementary qualitative lemma that we leave as exercise.) So we know that ξ consists of a cycloid Γ_1 going from A to a point X in the x axis, followed by a second cycloid Γ_2 going from X to B. It remains to find X.

It turns out that ***the version of [135] applies***, since this result does not require Lipschitz continuity—or even continuity—of the right-hand side, and works as long as the reference trajectory arises from a semidifferentiable flow. The result is that ***the cycloids*** Γ_1, Γ_2 ***must be such that the quantity***

$$C = |y(x)|(1 + y'(x)^2)\,. \tag{XIV.17}$$

which is constant along each Γ_i***, actually has the same value for both curves***.

This extra condition determines X. We refer the reader to [135] for the details.

COMMENT NO. 38. JAN: Actually, I am not sure that "the details" are done in my Kobe CDC paper. (I'd have to check. I cannot do it right now.) And I am not sure that, even if the details turn out to be done there, it is a good idea to "refer the reader." Maybe we should put in the full computation. What do you think? If you are in favor, I volunteer to do it. I probably have the whole thing on line somewhere, so it should not be too much work, if I can find it.

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

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

COMMENT NO. 39. JAN: wouldn't it be nice to put a figure here of the two cycloids, displaying the similarity of this situation with that of Snell's law, with the segments replaced by cycloids?

XIV.11 A modest claim

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

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

COMMENT NO. 40. JAN: I think this is a very important point, and the case we are making seems extremely convincing to me, but I am of course far from a neutral objective observer. Please think it over. If you agree that is very convincing, then maybe you can find a better way to say what I am trying to say, or you can add some more stuff on this?

XIV.12 Hamilton's principal function for the brachistochrone problem, and the differential-geometric connection

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

XIV.13 What is Hamilton's principal function for the brachistochrone?

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

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

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

COMMENT NO. 41. WE NEED ONE MORE SUBSECTION HERE. Topic: the Hestenes problem and the brachistochrone are closely related. In both cases we are moving in $\mathbb{R}^2$ and controlling acceleration. In the BP we control acceleration by requiring that acceleration is that of gravity plus something perpendicular to the velocity (D'Alembert's virtual force principle). In Hestenes acceleration=gravity plus thrust. So in both cases the problem should have $\mathbb{R}^4$ as its state space. Then how come the BP is formulated in $\mathbb{R}^2$? It should not be!!! And how come for the BP we specify initial and terminal positions, but for Hestenes we specify initial and terminal positions AND VELOCITIES? Isn't that an anomaly? Then we should give a first hint of the answer that is coming later: the BP is also a 4-dimensional problem, really, but it is anomalous in two ways: first, because there is a nontrivial conserved quantity—the energy—so the state space is foliated by 3-dimensional manifolds such that every trajectory is contained in one of these manifolds. Second, because for the BP the controls are unbounded, even within a single leaf existence of an optimal control joining any two points does not follow. These two anomalies conspire to make the BP look two-dimensional and to conceal its true 4-dimensional nature. It should be obvious from the above that I think I know what I want to say, but if you want to give try your hand and write something, go ahead.

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

COMMENT NO. 42. Maybe we should make a complete study of the general acceleration control problem in $\mathbb{R}^2$ of which both Hestenes and BP are particular cases? Compute the Lie algebra in general and show how it degenerates for the BP? What do you say?

COMMENT NO. 43. One the previous GNUs are taken care of, I'd like to go back to the beginning of the book and put in a few remarks on how the BP is a "degenerate problem" and how this fact has obscured its true meaning and how it fits within a general theory.

XIV.14 Lie brackets, integrals of motion, and the differential-geometric perspective

A much better way to pose the problem is to write down the equations of motion that truly correspond to Johann Bernoulli's problem as he stated it in June 1696. One should then let the mathematical analysis lead us to the discovery of energy conservation and the simplification resulting from it. This will lead us naturally to a four-dimensional problem whose analysis is best carried out from the differential-geometric perspective, by means of Lie-algebraic calculations.

The true equations of motion for the brachistochrone are

$$\dot{x} = v, \quad \dot{y} = w\,, \ \ \dot{v} = uw\,, \quad \dot{w} = -uv - g\,. \tag{XIV.18}$$

Here x and y are the horizontal and vertical coordinates of our moving point, and v, w are the components of the velocity vector. The requirement that the point is freely falling "under the influence of its own weight" means that the force effectively acting on the point should be equal to a vector proportional to $[0,-g]$ plus a "virtual force" that does no work, i.e., is perpendicular to the velocity. Equation (XIV.18) captures these requirements, by introducing a virtual force vector of the form $[uw,-uv]$, where u is an arbitrary "control," taking values in $\mathbb{R}$.

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

$$\dot{q} = F(q) + uG(q)\,, \tag{XIV.19}$$

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

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

XIV.15 The diamond Lie algebra

Having expressed our equations of motion in the form (XIV.19), the first thing that one should do, from the differential-geometric perspective is to compute the iterated Lie brackets of F and G, and determine the ***Lie algebra sturcture of our problem.*** It turns out that the four vector fields F, G, $[F,G]$ and $[F,[F,G]]$ are linearly independent over $\mathbb{R}$ (that is, there exists no nontrivial linear combination of them with constant coefficients that vanishes identically), and in addition the relations

$$[G,[F,G]] = F\,, \quad [F,[F,[F,G]]] = [G,[F,[F,G]]] = 0 \tag{XIV.20}$$

hold. It then follows easily that, if L denotes the Lie algebra of vector fields generated by F and G, then L is four-dimensional. As an abstract Lie algebra, L is isomorphic to the *diamond Lie algebra* G_4, a solvable Lie algebra that plays an important role in representation theory (cf. [14], p. 59). This Lie algebra has a basis H,P,Q,E, with the commutation relations $[H,P] = Q$, $[H,Q] = -P$, $[P,Q] = E$, $[H,E] = [P,E] = [Q,E] = 0$. (To construct an isomorphism Φ from G_4 to L, define $\Phi(P) = [F,G]$, $\Phi(Q) = F$, $\Phi(H) = G$, $\Phi(E) = -[F,[F,G]]$. A concrete realization of G_4 as an algebra of differential operators—acting of smooth functions on the real line—can be defined by "quantizing the harmonic oscillator," i.e., by taking $P = \frac{d}{dx}$, $Q =$ multiplication by x, $E =$ identity, $H = \frac{1}{2}(-P^2 + Q^2) = \frac{1}{2}\Big(-\frac{d^2}{dx^2} + x^2\Big)$, so H is the Hermite operator.

XIV.16 Energy conservation

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

$$G = \psi_1 F + \psi_2 [F,G] + \psi_3 [F,[F,G]] \quad \text{and} \quad [G,[F,G]] = F\,, \tag{XIV.21}$$

where the ψ_i are smooth functions of q, given by

$$\psi_1 = \frac{v}{g}\,, \quad \psi_2 = -\frac{w}{g}\,, \quad \psi_3 = -\frac{v^2 + w^2}{2g^2}\,. \tag{XIV.22}$$

From these facts one can easily show that every iterated Lie bracket of F and G is a linear combination of F, $[F,G]$, and $[F,[F,G]]$ with smooth coefficients. Since $F(q)$, $[F,G](q)$, and $[F,[F,G]](q)$ are linearly independent at each point $q \in \mathbb{R}^4$, we can conclude that the *distribution* Λ generated by L is 3-*dimensional* at each point. (By definition, $\Lambda(q) \stackrel{\text{def}}{=} \{X(q) : X \in L\}$.) This means—using, for example, Frobenius' theorem on the existence of integral manifolds—that, at least locally, there is a nontrivial integral of motion, i.e., a function with nonzero gradient which is constant along all integral curves of F and G, and then also along all solutions of (XIV.19). This integral of motion can be easily computed and turns out to be the energy E, given by

$$E(x,y,v,w) = \frac{v^2+w^2}{2} + gy\,. \tag{XIV.23}$$

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

XIV.17 Naïve application of the maximum principle

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

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

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

XIV.18 Application of the maximum principle on manifolds

Using the integral of motion E, we can regard $\mathbb{R}^4$ as foliated by the 3-dimensional level hypersurfaces of E, and conclude that every trajectory is contained in one of the leaves $S_{\bar{E}}$. If $\bar{E} \in \mathbb{R}$, then the maximum principle on manifolds (cf. Chapter XII and [134]) can be applied to the problem restricted to $S_{\bar{E}}$, and the conclusion turns out to be exactly the same as that for the unrestricted problem, except that now the nontriviality condition says that $p_*(t)$ cannot be orthogonal to $S_{\bar{E}}$ at $q_*(t)$. (In differential-geometric terms, $p_*(t)$ is really a *covector* on $S_{\bar{E}}$ at $q_*(t)$, and has to be nontrivial as such. Equivalently, if one insists on regarding $p_*(t)$ as a vector in $\mathbb{R}^4$, then it should not be orthogonal to the tangent space to $S_{\bar{E}}$ at $q_*(t)$.) Naturally, one could also avoid using the maximum principle on manifolds by using the fact that $S_{\bar{E}}$ is identified with $\mathbb{R}^3$ in obvious way, since it is the graph of a smooth function $y(x,v,w)$. This would amount to eliminating y from the equations by substituting for it its expression in terms of the other variables, for each fixed $\bar{E}$.

The Hamiltonian is $H(q,u,p,p_0) = \varphi(p,q)\rangle + u\psi(q,p) - p_0$, where the functions φ, ψ are defined by $\varphi(p,q) = \langle p, F(q)\rangle$ and $\psi(q,p) = \langle p, G(q)\rangle$. Since u is completely unrestricted, H cannot have a maximum as a function of u unless $\psi(q,p) = 0$. So ψ must vanish along every lifted extremal. Differentiation of $\psi = 0$ yields $\rho = 0$, where $\rho(q,p) = \langle p, [F,G](q)\rangle$. A further differentiation—together with the fact that $[G,[F,G]] = F$—yields $\theta + u\varphi = 0$, where $\theta(q,p) = \langle p, [F,[F,G]](q)\rangle$. This gives $u = -\varphi^{-1}\theta$, i.e.,

$$u = -\frac{\langle p, [F,[F,G]](q)\rangle}{\langle p, F(q)\rangle}. \tag{XIV.24}$$

This, together with the equalities $\psi = \rho = 0$, determine u as a function of q, i.e., provide an optimal control in the form of a feedback law $u = u(q)$. In fact, (XIV.21) implies that $\psi = \psi_1\varphi + \psi_2\rho + \psi_3\theta$ along our trajectory. Since $\psi = \rho = 0$, and $\theta = -u\varphi$, we get (using the fact that $\varphi \neq 0$, because p is nontrivial as a linear functional on the span of F, $[F,G]$ and $[F,[F,G]]$, so that $(\varphi,\rho,\theta) \neq (0,0,0)$, while on the other hand $\psi = \rho = 0$, so that if $\varphi = 0$ implies $\theta = 0$) $u = \psi_3^{-1}\psi_1$, i.e.,

$$u = \frac{2vg}{v^2 + w^2}. \tag{XIV.25}$$

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

$$\Sigma = \{(x,y,v,w) : v^2 + w^2 = 0\}. \tag{XIV.26}$$

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

If we define

$$Z(q) = F(q) - \psi_1(q)G(q), \tag{XIV.27}$$

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

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

Remark XIV.2. The preceding calculations provide another illustration of the implementation of ***Lagrange's idea of choosing the best possible coordinates for a particular problem***. Notice how everything is made much simpler by the fact that, using the intrinsic version of the adjoint equation, we can do all our computations directly for the switching function $\langle p, G\rangle$ and other functions of the same type, without ever writing the components of p. □

XIV.19 The complete synthesis and Hamilton's principal function for the 4-dimensional brachistochrone problem

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

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

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

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

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

The solution of this apparent paradox is as follows. Suppose we fix an energy level $\bar{E}$. Given any A, the set $C(A,\bar{E})$ of those a's such that $(A,a) \in S_{\bar{E}}$ is either empty or a circle, except for the degenerate case when $(A,0) \in S_{\bar{E}}$, in which case

$C(A,\bar{E})$ is a point. If $C(A,\bar{E})$ is a circle, then through each point of $\{A\}\times C(A,\bar{E})$ there passes an integral curve of Z. The set of all points lying on such curves is 2-dimensional, and intersects the 1-dimensional set $\{B\}\times C(B,\bar{E})$ at just one point. This means that there is a unique pair $(\bar{a},\bar{b})$ such that $(A,\bar{a})$ and $(B,\bar{b})$ lie on the same integral curve of Z. The portion of that curve joining $(A,\bar{a})$ to $(B,\bar{b})$ is, of course, our famous cycloid.

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

XIV.20 Deriving Snell' law from the maximum principle

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

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

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

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

Once again, it turns out that our recent version of (MP) given in [135] applies to this problem as well. The reason is that in [135] all that is needed is that the reference trajectory q_* arise from a flow $\{\Phi_{t',t}\}$ consisting of semidifferentiable maps, and there is no need for this flow to be generated by from a continuous vector field. One can formulate our refraction problem in terms of the dynamical law $\dot{q} = f(q,u,t)$, where the control u is now a pair (u_-,u_+) of unit vectors, and $f(q,u,t) = c_-u_-$ for $q \in H_-$, $f(q,u,t) = c_+u_+$ for $q \in H_+$. If $\bar{q} \in H_-$, $\hat{q} \in H_+$, it is easily seen, by applying the classical version of (MP), that an optimal trajectory from $\bar{q}$ to $\hat{q}$ has to consist of a segment going from $\bar{q}$ to some point P in H, followed by another segment from P to $\hat{q}$. The question is to find P. It is easy to verify that, if $q_* : [a,b] \mapsto \mathbb{R}^n$ is a trajectory of this type, then q_* is in fact a trajectory of a constant control $u = (u_-,u_+)$, such that both u_- and u_+ point to the same side of H. Although the vector field $q \mapsto f(q,u)$ is discontinuous, one can easily see that the corresponding flow maps $\Phi_{t,a}$ are continuously differentiable (and *a fortiori* semidifferentiable) at $\bar{q}$, for all t except the one such that $\Phi_{t,a}(\bar{q}) \in H$. This suffices to make the maximum principle of [135] applicable, and derive Snell' law.

Chapter XV

Rigorous solution of the isoperimetric problem

XV.1 Example II: the isoperimetric problem

.

XV.1.1 Problem statement

The standard statement of the isoperimetric problem is as follows:

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

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

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

The solution presented here is not completely rigorous, because

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

ξ is the line integral

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

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

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

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

XV.1.2 Problem reformulations

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

1. γ is absolutely continuous;
2. $\gamma(1) = \gamma(0)$;
3. γ has length L, that is,
$$\int_0^1 \|\dot{\gamma}(t)\| = \int_0^1 \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2}\,dt = L;$$
4. γ is simple (that is, the restriction of γ to the half-open interval $[0,1[$ is one-to-one);
5. γ is oriented counterclockwise.

The precise meaning of the words "oriented counterclockwise" is as follows: if $\gamma : [0,1] \mapsto \mathbb{R}^2$ is continuous, simple, and such that $\gamma(0) = \gamma(1)$, then the Jordan Curve Theorem (cf. Theorem XV.25 below) implies that the complement of the set $|\gamma| = \{\gamma(t) : t \in [0,1]\}$ has two connected components, one bounded and one unbounded; let $\mathrm{Ins}(\gamma)$ denote the bounded connected component of $\mathbb{R}^2 \backslash |\gamma|$; then the winding number $\nu(\xi, p)$ is either equal to 1 for all $p \in \mathrm{Ins}(\gamma)$ or equal to -1 for all $p \in \mathrm{Ins}(\gamma)$, by Theorem XV.26. (The definitions of "connected component" and "winding number" are given later, in §XV.12 and §XV.13.)

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

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

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

(This is "Green's formula," proved below, cf. Theorem XV.30.)

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

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

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

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

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

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

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

and then

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

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

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

Therefore

$$\left\| \frac{d}{ds}(\xi_\gamma(s)) \right\| = 1 \text{ for almost all } s \in [0,L]. \tag{XV.2}$$

The above argument is of course nonrigorous. For a rigorous proof of (XV.27), see Proposition XV.9 below.

Clearly,

So our problem can be restated as follows:

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

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

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

and consider the problem

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

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

Proposition XV.1. *If (IP") has a solution γ_* such that $\gamma_* \in \Gamma_L$, then γ_* is a solution of (IP').* ☐

XV.1.3 Finding the solutions of (IP")

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

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

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

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

such that the Hamilton equations

$$\begin{aligned} \dot{m}_x(t) &= -\frac{\partial H}{\partial x}(\gamma_*(t), \mu(t), \dot{\gamma}(t)), && \text{(XV.3)} \\ \dot{m}_y(t) &= -\frac{\partial H}{\partial y}(\gamma_*(t), \mu(t), \dot{\gamma}(t)), && \text{(XV.4)} \end{aligned}$$

and the Hamiltonian maximization condition

$$H(\gamma_*(t), \mu(t), \dot{\gamma}_*(t)) = \max\Big\{H(\gamma_*(t), \mu(t), w) : w \in \mathbb{R}^2, \|w\| \le 1\Big\} \qquad \text{(XV.5)}$$

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

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

In view of (XV.6), we can rewrite (XV.3), (XV.4), as

$$\begin{aligned} \dot{m}_x(t) &= -\dot{y}(t), && \text{(XV.7)} \\ \dot{m}_y(t) &= \dot{x}(t). && \text{(XV.8)} \end{aligned}$$

Define

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

Then

$$\begin{aligned} \dot{\alpha}(t) &= -2\dot{y}(t), && \text{(XV.9)} \\ \dot{\beta}(t) &= 2\dot{x}(t), && \text{(XV.10)} \end{aligned}$$

and

$$H(\gamma_*(t), \mu(t), u, v) = \alpha(t)u + \beta(t)v. \qquad \text{(XV.11)}$$

Therefore, if we define

$$\nu(t) = \sqrt{\alpha(t)^2 + \beta(t)^2}, \tag{XV.12}$$

(XV.5) says that

$$\left.\begin{aligned} \dot{x}(t) &= \tfrac{\alpha(t)}{\nu(t)}, \\ \dot{y}(t) &= \tfrac{\beta(t)}{\nu(t)}, \end{aligned}\right\} \quad \text{if} \quad \nu(t) \neq 0. \tag{XV.13}$$

If $\nu(t) = 0$ then (XV.5) allows the vector $\dot{\gamma}_*(t)$ to be an arbitrary member of $\{w \in \mathbb{R}^2 : \|w\| \leq 1\}$. We can combine both cases into the following restatement of (XV.13) which is also valid when $\nu(t) = 0$:

$$\nu(t)\dot{x}(t) = \alpha(t) \qquad \text{and} \qquad \nu(t)\dot{y}(t) = \beta(t). \tag{XV.14}$$

A simple calculation then shows that

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

Therefore *the function ν is constant.* Let $\bar{\nu}$ be the constant value of $\nu(t)$. If $\bar{\nu} = 0$ then $\alpha(t) \equiv \beta(t) \equiv 0$, so $\dot{\alpha}(t) \equiv \dot{\beta}(t) \equiv 0$, and then (XV.9) and (XV.10) imply that $\dot{x}(t) \equiv \dot{y}(t) \equiv 0$, so γ_* is a constant curve and $A(\gamma_*) = 0$. It is clear that there do exist curves $\gamma \in \tilde{\Gamma}_L$ for which $A(\gamma) > 0$. Therefore the fact that $A(\gamma_*) = 0$ contradicts the assumption that γ_* is a solution of (IP”). So we can exclude the possibility that $\bar{\nu} = 0$.

From now on we assume that $\bar{\nu} > 0$. Then (XV.13) implies that $\dot{\gamma}$ is continuous, and

$$\|\dot{\gamma}(t)\| = 1 \qquad \text{for all } t \in [0, L]. \tag{XV.15}$$

Write

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

Then, using (XV.9), (XV.10), and (XV.13) we can conclude that

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

Therefore

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

This implies that

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

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

$$c_1 = r\cos\left(\frac{2t_0}{\bar{v}}\right), \qquad c_2 = r\sin\left(\frac{2t_0}{\bar{v}}\right),$$

we find that

$$\eta(t) = r\cos\left(\frac{2(t-t_0)}{\bar{v}}\right).$$

Then

$$\dot{\xi}(t) = \frac{\bar{v}\eta(t)}{2} = -r\sin\left(\frac{2(t-t_0)}{\bar{v}}\right).$$

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

$$\xi(t) = -\sin\left(\frac{2(t-t_0)}{\bar{v}}\right) \qquad \text{and} \qquad \eta(t) = \cos\left(\frac{2(t-t_0)}{\bar{v}}\right). \tag{XV.16}$$

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

$$\xi(t) = -\sin\left(\frac{2\pi k(t-t_0)}{L}\right) \qquad \text{and} \qquad \eta(t) = \cos\left(\frac{2\pi k(t-t_0)}{L}\right). \tag{XV.17}$$

We then get, by integration,

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

where

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

The above formulas define a family $\left\{\gamma^{t_0,k}\right\}_{t_0\in\mathbb{R},k\in\mathbb{Z}}$ of curves belonging to $\tilde{\Gamma}_L$, depending on the real parameter t_0 and the nonzero integer parameter k. If we let $\bar{\gamma}^{t_0,k}$ be the point with coordinates $\frac{-L}{2\pi k}\cos\left(\frac{2\pi k t_0}{L}\right)$, $\frac{L}{2\pi k}\sin\left(\frac{2\pi k t_0}{L}\right)$, then the curves $\gamma^{t_0,k}$ satisfy

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

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

$|\gamma^{t_0,k}|$ $|k|$ *times.* Moreover, *the orientation derived from this map is counterclockwise if* $k > 0$ *and clockwise if* $k < 0$.

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

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

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

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

Summarizing, we have shown

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

XV.2 Existence of area-maximizers

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

Proposition XV.3. $\tilde{\Gamma}_L$ *is a compact subset of* $C^0([0,L],\mathbb{R}^2)$.

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

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

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

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

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

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

We now prove the continuity of A.

Proposition XV.4. *The functional A is continuous on $\tilde{\Gamma}_L$.*

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

$$\lim_{j\to\infty} \Big(A(\gamma_j) - A(\gamma_\infty)\Big) = 0\,. \qquad \text{(XV.18)}$$

Write

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

Then

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

where

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

We must therefore show that

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

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

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

where

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

We will show that

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

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

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

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

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

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

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

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

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

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

It then follows that

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

completing our proof. □

It then follows from Propositions 3 and 4 that

Proposition XV.5. *The maximization problem (IP") has a solution.* □

XV.3 The final result

By combining Propositions XV.2 and XV.5, we get

Theorem XV.6. *The solutions of the modified isoperimetric problem (IP") are the circles $\gamma^{t_0,1}$, for $t_0 \in \mathbb{R}$. Moreover, these circles are also the solutions of the original isoperimetric isoperimetric problem (IP'). In particular, the area enclosed by the area-maximizing circles is $\frac{L^2}{4\pi}$.* □

XV.4 Polygonal loops

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

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

The *vertices* of a triangle

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

XV.5 Rectifiable curves

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

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

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

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

XV.6 Absolutely continuous curves

Proposition XV.7. *Let $\gamma : [a,b] \mapsto \mathbb{R}^n$ be an absolutely continuous map. Then γ is rectifiable and* $\mathrm{Le}(\gamma) = \int_a^b \|\dot\gamma(t)\|\, dt$.

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

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

Hence

$$\mathrm{Le}(\gamma) \le \int_a^b \|\dot\gamma(t)\|\, dt , \tag{XV.19}$$

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

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

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

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

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

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

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

so

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

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

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

from which it follows that

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

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

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

Therefore

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

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

$$\begin{aligned} I_j^1 &= \int_{t_{j-1}}^{t_j} \|\dot\gamma(t)\|\,dt \\ &\le \lambda\mu_j + \mu_j\|\dot\gamma(\hat t)\| \\ &\le \|U_j\| + 2\int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt + 2\lambda\mu_j\,. \end{aligned} \qquad \text{(XV.20)}$$

This inequality was proved assuming that $K \cap [t_{j-1}, t_j] \neq \emptyset$, but is obviously true as well if the set $K \cap [t_{j-1}, t_j]$ is empty, because in that case $\int_{[t_{j-1},t_j]\cap E} \|\dot\gamma(t)\|\,dt = I_j^1$. Therefore (XV.20) holds for all j. Hence

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

If we add over all j, we find

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

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

$$\int_a^b \|\dot\gamma(t)\|\,dt \leq \mathrm{Le}(\gamma)\,. \tag{XV.21}$$

In view of (XV.19) and (XV.21), our conclusion is proved. □

XV.7 Parametrization by arc length

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

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

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

Proposition XV.8. *Suppose that $a,b \in \mathbb{R}$, $a \leq b$, and $\xi : [a,b] \mapsto \mathbb{R}^n$ is a rectifiable curve, and $L = \mathrm{Le}(\xi)$. Then the arc length s_ξ is a continuous monotonically nondecreasing function from $[a,b]$ to $\mathbb{R}$ that satisfies $s_\xi(a) = 0$ and $s_\xi(b) = L$.*

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

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

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

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

We will prove that $\lambda_+ = \lambda_-$. Let $\alpha = \lambda_+ - \lambda_-$. Fix a positive number ε, and then choose a positive δ such that $\|\xi(\tau) - \xi(t)\| \leq \varepsilon$ whenever $\tau \in [a,b]$ and

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

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

$$\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) \geq s_\xi(t_{j+1}) - \varepsilon \geq \lambda_+ - \varepsilon\,. \tag{XV.22}$$

Indeed, the second inequality of (XV.22) follows from the definition of λ_+, the monotonicity of s_ξ, and the fact that $t_{j+1} > t$. If $\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) < s_\xi(t_{j+1}) - \varepsilon$, then there would exist a number ρ and a partition $\hat{\mathbf{u}} = (u_0, \ldots, u_M)$ of $[a, t_{j+1}]$ such that

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

Then $\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{u}}) > \varepsilon + \mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}})$. If we let $\mathbf{u}' = (u_0, \ldots, u_M, t_{j+2}, \ldots, t_N)$, then $\mathbf{u}'$ is a partition of $[a,t']$ and $\mathrm{Le}(\xi_{t'}, \mathbf{u}') > \varepsilon + \mathrm{Le}(\xi_{t'}, \mathbf{t}') \geq s_\xi(t')$. The inequality $\mathrm{Le}(\xi_{t'}, \mathbf{u}') > s_\xi(t')$ clearly contradicts the definition of $s_\xi(t')$, so the first inequality of (XV.22) is proved.

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

$$\mathrm{Le}(\xi_{t_{j-1}}, \tilde{\mathbf{t}}) \leq s_\xi(t_{j-1}) \leq \lambda_-\,. \tag{XV.23}$$

It follows from (XV.22) and (XV.23) that

$$\mathrm{Le}(\xi_{t_{j+1}}, \hat{\mathbf{t}}) - \mathrm{Le}(\xi_{t_{j-1}}, \tilde{\mathbf{t}}) \geq \lambda_+ - \varepsilon - \lambda_- = \alpha - \varepsilon\,. \tag{XV.24}$$

But

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

and then (XV.24) implies that $\alpha \leq 3\varepsilon$.

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

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

$$s_\xi(t) + \|\xi(t') - \xi(t)\| \leq s_\xi(t') \quad \text{whenever} \quad a \leq t < t' \leq b\,. \tag{XV.25}$$

(Proof: the inequality is trivial if $t = t'$; assume that $t < t'$; if $\mathbf{t} = (t_0, \ldots, t_N)$ is a partition of $[a,t]$, then $\mathbf{t}' = (t_0, \ldots, t_N, t')$ is a partition of $[a,t']$, and $\mathrm{Le}(\xi_t, \mathbf{t}) + \|\xi(t') - \xi(t)\| = \mathrm{Le}(\xi_{t'}, \mathbf{t}') \leq s_\xi(t')$. Taking the supremum over $\mathbf{t}$, we find that (XV.25) holds.) It follows from (XV.25) that if $a \leq t < t' \leq b$ and $s_\xi(t) = s_\xi(t')$ then $\xi(t) = \xi(t')$. Therefore there exists a unique map $\gamma_\xi : [0,L] \mapsto \mathbb{R}^n$ having the property that $\gamma_\xi(s) = \xi(t)$ whenever $s = s_\xi(t)$. If $0 \leq s < s' \leq L$, then we can pick $t, t' \in [a,b]$ such that $s = s_\xi(t)$ and $s' = s_\xi(t')$. Then $t < t'$, and (XV.25) implies that $\|\gamma_\xi(s') - \gamma_\xi(s)\| \leq s' - s$.

Therefore

$$\|\gamma_\xi(s') - \gamma_\xi(s)\| \leq |s' - s| \quad \text{whenever} \quad s, s' \in [0,L]\,. \tag{XV.26}$$

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

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

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

and then

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

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

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

Therefore

$$\left\|\frac{d}{ds}(\xi_\gamma(s))\right\| = 1 \text{ for almost all } s \in [0,L]\,. \tag{XV.27}$$

The above argument is of course nonrigorous. For a rigorous proof of (XV.27), see Proposition XV.9 below.

Proposition XV.9. *Let $\xi : [a,b] \mapsto \mathbb{R}^n$ be a rectifiable map of length L, and let γ_ξ be the reparametrization of ξ by arc-length. Then γ_ξ is Lipschitz and $\|\dot{\gamma}_\xi(s)\| = 1$ for almost all $s \in [0,L]$.*

PROOF. Inequality (XV.26), which was rigorously proved in page 338, shows that γ_ξ is Lipschitz with constant 1. Hence $\|\dot{\gamma}_\xi(s)\| \leq 1$ for almost all $s \in [0,L]$.

Let us show that $\mathrm{Le}(\gamma_\xi) = L$. Let $\mathbf{t} = (t_0, \ldots, t_N)$ be a partition of $[a,b]$. Let $\mathbf{s} = (s_0, \ldots, s_N)$, where $s_j = s_\xi(t_j)$. Then $\mathbf{s}$ is a partition of $[0,L]$, except for the fact that the s_j need not be all different. Let $\hat{\mathbf{s}}$ be the partition obtained by eliminating from $\mathbf{s}$ every s_j that is equal to some previous s_k. It is then clear that

$$\mathrm{Le}(\gamma_\xi) \geq \mathrm{Le}(\gamma_\xi, \hat{\mathbf{s}}) = \sum_{j=1}^{N} \|\gamma_\xi(s_j) - \gamma_\xi(s_{j-1})\| = \sum_{j=1}^{N} \|\xi(t_j) - \xi(s_{j-1})\| .$$

Hence $\mathrm{Le}(\gamma_\xi) \geq \mathrm{Le}(\xi, \mathbf{t})$ for an arbitrary partition $\mathbf{t}$ of $[a,b]$. Therefore $\mathrm{Le}(\gamma_\xi) \geq \mathrm{Le}(\xi) = L$. On the other hand, Proposition (XV.7) tells us that $\mathrm{Le}(\gamma_\xi) = \int_0^L \|\dot{\gamma}_\xi(s)\|\,ds$, so $\mathrm{Le}(\gamma_\xi) \leq L$, because $\|\dot{\gamma}_\xi(s)\| \leq 1$ for almost all $s \in [0,L]$. Therefore $\mathrm{Le}(\gamma_\xi) = L$. Since $\int_0^L \|\dot{\gamma}_\xi(s)\|\,ds = L$, and $\|\dot{\gamma}_\xi(s)\| \leq 1$ for almost all $s \in [0,L]$, it must be the case that $\|\dot{\gamma}_\xi(s)\| = 1$ for almost all $s \in [0,L]$, and our proof is complete. $\square$

XV.8 Changing variables

Suppose $a < b$, $c < d$, $f : [c,d] \mapsto \mathbb{R}^n$ is an integrable vector-valued function, and $\sigma : [a,b] \mapsto [c,d]$ is an absolutely continuous, monotonically nondecreasing function such that $\sigma(a) = c$ and $\sigma(b) = d$. Can we "change variables" in the integral $\int_c^d f(s)\,ds$ by writing $s = \sigma(t)$ and $ds = \dot{\sigma}(t)\,dt$, and conclude that the formula

$$\int_c^d f(s)\,ds = \int_a^b \dot{\sigma}(t) \cdot f(\sigma(t))\,dt \qquad \text{(XV.28)}$$

holds?

Notice, first of all, that if f is an "integrable function" in the usual sense (that is, an equivalence class of "true functions" from $[c,d]$ to $\mathbb{R}^n$ modulo equality almost everywhere, such that every member of the class is Lebesgue-integrable), then $f \circ \sigma$ is not a well defined "integrable function" on $[a,b]$, because if two functions f_1, f_2

on $[c,d]$ coincide on the complement of a null set E, and $g_1 = f_1 \circ \sigma$, $g_2 = f_2 \circ \sigma$, then all we can conclude is that g_1 and g_2 coincide on the complement of $\sigma^{-1}(E)$, and $\sigma^{-1}(E)$ need not be a null set. It turns out, however that this is not a serious difficulty because, as we will show, if E is a null set then $\dot\sigma(t) = 0$ for almost all $t \in \sigma^{-1}(E)$, so the function $\dot\sigma \cdot (f \circ \sigma)$ is well defined modulo a.e. equality. Furthermore, $\dot\sigma \cdot (f \circ \sigma)$ is integrable, because $\dot\sigma$ is integrable and $f \circ \sigma$ is bounded. So it is at least plausible that (XV.28) might be true. The following proposition says that this is indeed so. For the reasons explained above, we state it first for "true functions," where a *true function* on $[a,b]$ is a function on $[a,b]$, rather than an equivalence class modulo a.e. equality.

Proposition XV.10. *Let $n \in \mathbb{N}$, $a,b,c,d \in \mathbb{R}$ be such that $a < b$ and $c < d$. Let $f : [c,d] \mapsto \mathbb{R}^n$ be a Lebesgue integrable true function. Assume that $\sigma : [a,b] \mapsto \mathbb{R}$ is an absolutely continuous, monotonically nondecreasing map such that $\sigma(a) = c$ and $\sigma(b) = d$. Then the true function*

$$[a,b] \ni t \mapsto \dot\sigma(t) \cdot f(\sigma(t))$$

is integrable, and (XV.28) holds.

PROOF. If I is a closed interval $[\gamma,\delta]$ (where $c \le \gamma \le \delta \le b$), then $\sigma^{-1}(I)$ is the interval $J = [\alpha,\beta]$, where $\alpha = \min\{s \in [a,b] : \sigma(s) = \gamma\}$ and $\beta = \max\{s \in [a,b] : \sigma(s) = \delta\}$. Therefore

$$\mathrm{meas}(I) = \delta - \gamma = \sigma(\beta) - \sigma(\alpha) = \int_\alpha^\beta \dot\sigma(t)\,dt = \int_{\sigma^{-1}(I)} \dot\sigma(t)\,dt\,.$$

A similar argument applies if I is half-open half-closed, or if I is open. (For example, if $I =]\gamma,\delta]$, then β is defined as above, but α and J must be taken to be $\max\{s \in [a,b] : \sigma(s) = \gamma\}$, and $]\alpha,\beta]$, respectively.) So $\mathrm{meas}(I) = \int_{\sigma^{-1}(I)} \dot\sigma(t)\,dt$ for all intervals I such that $I \subseteq [c,d]$. Using the formula $\sigma^{-1}(E \cup F) = \sigma^{-1}(E) \cup \sigma^{-1}(F)$, it follows that $\mathrm{meas}(E) = \int_{\sigma^{-1}(E)} \dot\sigma(t)\,dt$ for every subset of $[c,d]$ which is a finite union of intervals, since every such subset is in fact a finite union of pairwise disjoint intervals.

Let $\mathscr{E}$ be the class of all Borel measurable subsets E of $[c,d]$ such that $\mathrm{meas}(E) = \int_{\sigma^{-1}(E)} \dot\sigma(t)\,dt$, and let $\mathscr{I}$ be the class of subsets of $[c,d]$ that are finite unions of intervals. Then $\mathscr{I} \subseteq \mathscr{E}$. Furthermore, it is easy to see that $\mathscr{E}$ is a monotone class of subsets of $[c,d]$, that is, (i) if $\{E_j\}_{j=1}^\infty$ is a sequence of members of $\mathscr{E}$ such that $E_j \subseteq E_{j+1}$ for every j, and $E = \bigcup_{j=1}^\infty E_j$, then $E \in \mathscr{E}$, and (ii) if $\{E_j\}_{j=1}^\infty$ is a sequence of members of $\mathscr{E}$ such that $E_{j+1} \subseteq E_j$ for every j, and $E = \bigcap_{j=1}^\infty E_j$, then

$E \in \mathscr{E}$. (Indeed, if $E \in \mathscr{E}$ and $F = [c,d] \backslash E$, then

$$\begin{aligned}\int_{\sigma^{-1}(F)} \dot{\sigma} &= \int_a^b \dot{\sigma} - \int_{\sigma^{-1}(E)} \dot{\sigma} = \sigma(b) - \sigma(a) - \text{meas}(E) \\ &= d - c - \text{meas}(E) = \text{meas}([c,d]) - \text{meas}(E) = \text{meas}(F)\,,\end{aligned}$$

so $F \in \mathscr{E}$. This shows that $\mathscr{E}$ is closed under complementation, from which it follows that if we prove (i) then (ii) is true as well. To prove (i), we assume that $E = \bigcup_{j=1}^{\infty} E_j$, $E_j \in \mathscr{E}$, and $\{E_j\}_{j=1}^{\infty}$ is increasing. Then

$$\begin{aligned}\text{meas}(E) &= \lim_{j\to\infty} \text{meas}(E_j) = \lim_{j\to\infty} \int_{\sigma^{-1}(E_j)} \dot{\sigma} = \lim_{j\to\infty} \int_a^b \dot{\sigma} \cdot \chi_{\sigma^{-1}(E_j)} \\ &= \int_a^b \dot{\sigma} \cdot \chi_{\sigma^{-1}(E)} = \int_{\sigma^{-1}(E)} \dot{\sigma}\,,\end{aligned}$$

by the Lebesgue monotone convergence theorem. So $E \in \mathscr{E}$.)

It follows from the monotone convergence theorem (cf. Halmos, vaca, Theorem vaca) that $\mathscr{E}$ is the set of all Borel subsets of $[c,d]$. So we have proved that $\text{meas}(E) = \int_{\sigma^{-1}(E)} \dot{\sigma}(t)\,dt$ for all Borel subsets of $[c,d]$.

Now[1] suppose that E is a Lebesgue measurable subset of $[c,d]$. Then there exist Borel measurable subsets E_1, E_2 of $[c,d]$ such that $E_1 \subseteq E \subseteq E_2$ and $\text{meas}(E_2 \backslash E_1) = 0$. Let $F_1 = \sigma^{-1}(E_1)$, $F = \sigma^{-1}(E)$, $F_2 = \sigma^{-1}(E_2)$. Then F_1 and F_2 are Borel measurable, $F_1 \subseteq F \subseteq F_2$, and $\int_{F_2 \backslash F_1} \dot{\sigma} = 0$. Since $\dot{\sigma}$ is nonnegative and measurable, there exists a Lebesgue null subset N of $F_2 \backslash F_1$ such that $\dot{\sigma} = 0$ on $(F_2 \backslash F_1) \backslash N$. It follows that the functions $\dot{\sigma} \cdot \chi_F$ and $\dot{\sigma} \cdot \chi_{F_2}$ coincide on the set $[a,b] \backslash N$. Since N is a null set, and $\dot{\sigma} \cdot \chi_{F_2}$ is integrable, we conclude that $\dot{\sigma} \cdot \chi_F$ is integrable as well, and

$$\int_a^b \dot{\sigma} \cdot \chi_F = \int_a^b \dot{\sigma} \cdot \chi_{F_2} = \int_{F_2} \dot{\sigma} = \text{meas}(E_2) = \text{meas}(E)\,.$$

Since $\chi_F = \chi_E \circ \sigma$, we have shown that $\dot{\sigma} \cdot (\chi_E \circ \sigma)$ is integrable and

$$\int_a^b \dot{\sigma} \cdot (\chi_E \circ \sigma) = \int_c^d \chi_E$$

whenever E is a Lebesgue measurable subset of $[c,d]$. In other words, we have proved that $\dot{\sigma} \cdot (f \circ \sigma)$ is integrable and (XV.28) holds whenever f is the indicator function of a Lebesgue measurable subset of $[c,d]$.

[1] If E is Borel measurable then $\sigma^{-1}(E)$ is also Borel measurable, because σ is continuous. But if E is Lebesgue measurable then it does not follow that $\sigma^{-1}(E)$ is Lebesgue measurable as well.

It then follows easily that $\dot{\sigma}\cdot(f\circ\sigma)$ is integrable and (XV.28) holds whenever f is a simple function (i.e., a finite linear combination of indicator functions of a Lebesgue measurable subsets of $[c,d]$). Then $\dot{\sigma}\cdot(f\circ\sigma)$ is integrable and (XV.28) holds whenever f is a bounded Lebesgue measurable real-valued function on $[c,d]$, since every such function is a uniform limit of simple functions. If f is nonnegative and measurable, then $f=\lim f_j$, where $f_j(s)=f(s)$ when $f(s)\le j$ and $f_j(s)=j$ when $f(s)>j$. Therefore $f\circ\sigma=\lim f_j\circ\sigma$, so $\dot{\sigma}\cdot f\circ\sigma$ is measurable and (XV.28) holds. In particular, if f is integrable (i.e., if $\int_a^b f<\infty$) then $\dot{\sigma}\cdot f\circ\sigma$ is integrable as well. Now, if $f:[c,d]\mapsto\mathbb{R}$ is integrable, then we can write $f=f_+-f_-$, where both f_+ and f_- are integrable and nonnegative, and conclude that $\dot{\sigma}\cdot(f\circ\sigma)$ is integrable and (XV.28) holds. Finally, if f is vector-valued, then the desired conclusion follows by applying what we have shown to each of the components of f. □

Now suppose that

(A1) Q is an open subset of $\mathbb{R}^n$,

(A2) $Q\times\mathbb{R}^n\ni(q,u)\mapsto L(q,u)\in\mathbb{R}$ is a continuous function, thought of as a time-independent Lagrangian,

(A3) L is positively homogeneous of degree 1 with respect to the velocity, that is, $L(q,ru)=rL(q,u)$ whenever $q\in Q$, $u\in\mathbb{R}^n$ and $r\ge 0$,

(A4) $\xi:[c,d]\mapsto Q$ is an absolutely continuous curve in Q,

(A5) $\sigma:[a,b]\mapsto\mathbb{R}$ is absolutely continuous, monotonically nondecreasing, and such that $\sigma(a)=c$ and $\sigma(b)=d$,

(A6) $\gamma=\xi\circ\sigma$.

Proposition XV.11. *Assume that conditions (A1)$\cdots$(A6) above hold. Then γ is absolutely continuous,*

$$\frac{d\gamma}{dt}(t)=\dot{\sigma}(t)\cdot\frac{d\xi}{ds}(\sigma(t))\quad\text{for almost all}\quad t\in[a,b]\,, \tag{XV.29}$$

and

$$\int_c^d L\Big(\xi(s),\frac{d\xi}{ds}(s)\Big)\,ds=\int_a^b L\Big(\gamma(t),\frac{d\gamma}{dt}(t)\Big)\,dt\,. \tag{XV.30}$$

PROOF. Let $f(s)=\frac{d\xi}{ds}(s)$, $g(t)=\dot{\sigma}(t)\cdot f(\sigma(t))$. Then Proposition XV.10 implies that g is integrable. Furthermore, we can pick an arbitrary $t\in[a,b]$ and apply

Proposition XV.10 to the interval $[a,t]$, with σ replaced by its restriction to $[a,t]$, and conclude that

$$\gamma(t) - \gamma(a) = \xi(\sigma(t)) - \xi(c) = \int_c^{\sigma(t)} f(s)\,ds = \int_a^t g(\tau)\,d\tau\,.$$

Since g is integrable and t is arbitrary, γ is absolutely continuous and (XV.29) holds.

Now let $h(s) = L\Big(\xi(s), \frac{d\xi}{ds}(s)\Big)$. Then Proposition XV.10 implies that

$$\begin{aligned}
\int_c^d h(s)\,ds &= \int_a^b \dot{\sigma}(t)\cdot h(\sigma(t)) \\
&= \int_a^b \dot{\sigma}(t)\cdot L\Big(\xi(\sigma(t)), \frac{d\xi}{ds}(\sigma(t))\Big)\,dt \\
&= \int_a^b L\Big(\gamma(t), \dot{\sigma}(t)\cdot\frac{d\xi}{ds}(\sigma(t))\Big)\,dt \\
&= \int_a^b L\Big(\gamma(t), \frac{d\gamma}{dt}(t)\Big)\,dt\,,
\end{aligned}$$

where in the last two steps we have used the homogeneity of L and (XV.29).

Therefore (XV.30) holds, and our proof is complete. □

XV.9 The complex exponential

We will need to use the *exponential function* $\mathbb{C} \ni z \mapsto e^z \in \mathbb{C}$. By definition,

$$e^{x+iy} = e^x \cdot e^{iy} = e^x(\cos y + i\sin y)\,. \tag{XV.31}$$

The following well-known properties of this function (whose proofs can be found in any elementary texbook of complex function theory) will be used:

(E1) $e^{z_1+z_2} = e^{z_1}\cdot e^{z_2}$ whenever $z_1, z_2 \in \mathbb{C}$,

(E2) $e^0 = 1$,

(E3) $e^z \neq 0$ whenever $z \in \mathbb{C}$,

(E4) if $z \in \mathbb{C}$, then $e^z = 1$ if and only if $z = 2\pi ik$ for some integer k,

(E5) the series $z - 1 - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} - \frac{(z-1)^4}{4} + \ldots$ is convergent for $z \in \mathbb{C}$ such that $|z-1| < 1$, and defines a continuous function

$$\{u \in \mathbb{C} : |u-1| < 1\} \ni z \mapsto \mathrm{Ln}(z) \in \mathbb{C}$$

such that $e^{\mathrm{Ln}(z)} = z$ whenever $|z-1| < 1$.

Since $e^{\mathrm{Ln}(z)} = z$ whenever $|z-1| < 1$, the function Ln is called "the standard branch of the natural logarithm."

If follows from the above properties that

(E6) if $z_0 \in \mathbb{C}$ and $z_0 \neq 0$, then there exists continuous function
$$\{u \in \mathbb{C} : |u - z_0| < |z_0|\} \ni z \mapsto \Phi(z) \in \mathbb{C}$$
such that $e^{\Phi(z)} = z$ whenever $|z - z_0| < |z_0|$.

(E7) if $[a,b] \ni t \mapsto \gamma(t) \in \mathbb{C}$ is an arc in the complex plane (i.e., a continuous map with values in $\mathbb{C}$) such that $\gamma(t) \neq 0$ for all $t \in [a,b]$, then there exists a continuous function $[a,b] \ni t \mapsto f(t) \in \mathbb{C}$ such that $e^{f(t)} = \gamma(t)$ for all $t \in [a,b]$.

Indeed, to prove (E6), we write $z_0 = x_0 + iy_0$, $r_0 = \sqrt{x_0^2 + y_0^2}$, $x_0 = r_0 \cdot \cos\theta_0$, $y_0 = r_0 \cdot \sin\theta_0$, and observe that $r_0 > 0$, so r_0 has a natural logarithm λ_0, and then $e^{\lambda_0} = r_0$. Therefore, if we write $s_0 = \lambda_0 + i\theta_0$, (XV.31) tells us that $e^{s_0} = z_0$. Then, if $|z - z_0| < |z_0|$, we have $\left|\frac{z}{z_0} - 1\right| < 1$, so $\mathrm{Ln}\left(\frac{z}{z_0}\right)$ is well defined. We then define $\Phi(z) = s_0 + \mathrm{Ln}\left(\frac{z}{z_0}\right)$, so that
$$e^{\Phi(z)} = e^{s_0} \cdot e^{\mathrm{Ln}\left(\frac{z}{z_0}\right)} = z_0 \cdot \frac{z}{z_0} = z.$$

To prove (E7), we let $\varepsilon = \min\{\,|\gamma(t)| : t \in [a,b]\,\}$, so that $\varepsilon > 0$, and use the fact that the map γ is uniformly continuous to obtain a positive δ such that $|\gamma(t_1) - \gamma(t_2)| < \varepsilon$ whenever $t_1, t_2 \in [a,b]$ and $|t_1 - t_2| < \delta$. We then find $t_0, t_1, \ldots, t_N$ such that $a = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N = b$ and $t_j - t_{j-1} < \delta$ if $j = 1, 2, \ldots, N$. We write $z_j = \gamma(t_j)$ and $D_j = \{z \in \mathbb{C} : |z - z_j| < |z_j|\}$ for $j = 1, \ldots, N$, and use (E6) to pick, for each $j \in \{1, \ldots, N\}$, a continuous function $\Phi_j : D_j \mapsto \mathbb{C}$ such that $e^{\Phi_j(z)} = z$ whenever $z \in D_j$. We then modify the Φ_j recursively, and define new functions $\Psi_j : D_j \mapsto \mathbb{C}$, as follows: we take $\Psi_1 = \Phi_1$; if $1 < j \leq N$, and $\Psi_1, \ldots, \Psi_{j-1}$ have been defined and are such that $e^{\Psi_\ell(z)} = z$ whenever $z \in D_\ell$, $0 \leq \ell < j$, then we observe that $z_j \in D_{j-1}$, because $|z_j - z_{j-1}| = |\gamma(t_j) - \gamma(t_{j-1})| < \varepsilon \leq |z_{j-1}|$, since $t_j - t_{j-1} < \delta$; therefore $e^{\Psi_{j-1}(z_j)} = z_j = e^{\Phi_j(z_j)}$, from which it follows that $\Phi_j(z_j) = \Psi_{j-1}(z_j) + 2\pi i k_j$ for some integer k_j; we then define $\Psi_j(z) = \Phi_j(z_j) - 2\pi i k_j$ for $z \in D_j$, so that $e^{\Psi_j(z)} = z$ whenever $z \in D_j$. Using the Ψ_j, we define $f(t)$, for $t \in [a,b]$, by letting $f(t) = \Psi_j(\gamma(t))$ whenever $t_{j-1} \leq t \leq t_j$, $j \in \{1, \ldots, N\}$. Then f is well defined. (Indeed, the only values of t for which the condition $t_{j-1} \leq t \leq t_j$

does not determine j uniquely are $t_1,\ldots,t_{N-1}$. For each of these t's, if $t=t_j$, $1\le j<N$, then $t\in[t_{j-1},t_j]$ and $t\in[t_j,t_{j+1}]$, so the definition would yield both $f(t)=\Psi_j(\gamma(t))=\Psi_j(z_j)$ and $f(t)=\Psi_{j+1}(\gamma(t))=\Psi_{j+1}(z_j)$, but

$$\Psi_{j+1}(z_j)=\Phi_{j+1}(z_j)-2\pi ik_{j+1}=\Big(\Psi_j(z_j)+2\pi ik_j\Big)-2\pi ik_{j+1}=\Psi_j(z_j)\,,$$

so the two values of $f(t)$ coincide.)

XV.10 Loops

We think of $\mathbb{R}^2$ as the complex plane $\mathbb{C}$. We use $\mathbb{S}^1$ to denote the unit circle $\{(x,y)\in\mathbb{R}^2:x^2+y^2=1\}=\{z\in\mathbb{C}:|z|=1\}$.

For each positive real number L, we let κ_L be the map

$$\mathbb{R}\ni t\mapsto\kappa_L(t)\stackrel{\text{def}}{=}e^{\frac{2\pi i}{L}t}\in\mathbb{S}^1\,.$$

A *loop* (or *closed curve*) in $\mathbb{R}^2$ is a continuous map $\xi:\mathbb{S}^1\mapsto\mathbb{R}^2$. Then $C^0(\mathbb{S}^1,\mathbb{R}^2)$— or $C^0(\mathbb{S}^1,\mathbb{C})$— is the set of all loops in $\mathbb{R}^2$. The *carrier*, or *locus*, of the loop ξ is the set $\{\xi(s):s\in\mathbb{C}\}$. Clearly, $|\xi|$ is compact.

If $L>1$, then $C^0(\mathbb{S}^1,\mathbb{R}^2)$ can be identified in an obvious way with $C^0_{per,L}(\mathbb{R},\mathbb{R}^2)$, the set of all continuous maps $\zeta:\mathbb{R}\mapsto\mathbb{R}^2$ that are L-periodic, i.e., such that $\zeta(t+L)=\zeta(t)$ for all $t\in\mathbb{R}$. Naturally, the identification is made by assigning to each $\xi\in C^0(\mathbb{S}^1,\mathbb{R}^2)$ the map $\zeta=\xi\circ\kappa_L\in C^0_{per,L}(\mathbb{R},\mathbb{R}^2)$.

A loop $\xi\in C^0(\mathbb{S}^1,\mathbb{R}^2)$ is said to be *simple* if it is a one-to-one map, i.e., if $\xi(z_1)\ne\xi(z_2)$ whenever $z_1,z_2\in\mathbb{S}^1$ and $z_1\ne z_2$.

A loop $\xi\in C^0(\mathbb{S}^1,\mathbb{R}^2)$ is said to be *rectifiable* if it is absolutely continuous, in the sense that for some positive L the function $\zeta=\xi\circ\kappa_L$ is locally absolutely continuous (in which case it follows easily that $\xi\circ\kappa_L$ is locally absolutely continuous for every positive L).

XV.11 Reparametrization of loops

A *reparametrization* of a loop $\xi:\mathbb{S}^1\mapsto\mathbb{C}$ is a loop of the form $\xi\circ\psi$, where $\psi:\mathbb{S}^1\mapsto\mathbb{S}^1$ is an orientation-preserving homeomorphism. ("Orientation-preserving" means, for example, that if $\bar t$ is such that $\psi(e^{2\pi i\bar t})=1$, then the imaginary part of $\psi(e^{2\pi i(\bar t+\varepsilon)})$ is positive for all sufficiently small positive ε.)

A rectifiable loop $\xi \in C^0(\mathbb{S}^1,\mathbb{C})$ is said to be *parametrized by constant times arc-length* (abbr. PCAL) if there exists a positive constant C such that $\|\dot{\zeta}(t)\| = C$ for all $t \in \mathbb{R}$, where $\zeta(t) \stackrel{\text{def}}{=} \xi(e^{2\pi it})$.

To any rectifiable loop $\xi \in C^0(\mathbb{S}^1,\mathbb{C})$ we can associate another loop $\gamma_\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ obtained from ξ by *reparametrization by constant times arc-length*. Precisely, γ_ξ is defined as follows. Let $\zeta(t) = \xi(e^{2\pi it})$, for $t \in \mathbb{R}$. Write $\zeta(t) = x(t) + iy(t)$ for $t \in \mathbb{R}$, and define a function $s_\xi : \mathbb{R} \mapsto \mathbb{R}$ (the "arc-length along ζ") by letting

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

Then s_ξ is a monotonically nondecreasing continuous function from $\mathbb{R}$ to $\mathbb{R}$. If $L = s_\xi(1)$ (so that L is the length of ξ) then s_ξ maps $[0,1]$ onto $[0,L]$. Furthermore, if $0 \le t_1 < t_2 \le 1$ and $s_\xi(t_1) = s_\xi(t_2)$, then

$$\int_{t_1}^{t_2} \sqrt{\dot{x}(u)^2 + \dot{y}(u)^2}\, du = s_\xi(t_2) - s_\xi(t_1) = 0\,,$$

so $\dot{x}(u) \equiv \dot{y}(u) \equiv 0$ for almost all $t_1 \le u \le t_2$, and then $\xi(t_2) = \xi(t_1)$. Therefore there exists a unique map $\theta_\xi : [0,L] \mapsto \mathbb{C}$ such that $\theta_\xi(s) = \zeta(t)$ whenever $s = s_\xi(t)$, $t \in [0,1]$. In particular, this map satisfies the identities $\theta_\xi(L) = \zeta(1) = \zeta(0) = \theta_\xi(0)$, since $L = s_\xi(1)$. So we can define a continuous map $\gamma_\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ by letting $\gamma_\xi(e^{2\pi it}) = \theta_\xi(Lt)$ for $0 \le t \le 1$.

If $0 \le s_1 < s_2 \le L$, then we can pick $t_1, t_2 \in [0,1]$ such that $s_1 = s_\xi(t_1)$ and $s_2 = s_\xi(t_2)$. Then $t_1 < t_2$, and

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

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

Formally, the identity $\theta_\xi \circ s_\xi = \zeta$ implies that

$$\frac{d}{dt}(\zeta(t)) = \frac{d}{ds}(\theta_\xi(s))\Big|_{s=s_\xi(t)} \cdot \frac{d}{dt}(s_\xi(t))\,,$$

and then

$$\|\dot{\zeta}(t)\| = \left\| \frac{d}{ds}(\theta_{\xi}(s))\Big|_{s=s_{\xi}(t)} \right\| \cdot \|\dot{\zeta}(t)\|,$$

so that

$$\left\| \frac{d}{ds}(\theta_{\xi}(s)) \right\| = 1 \text{ for almost all } s \in [0,L]. \qquad \text{(XV.32)}$$

The above argument is of course nonrigorous. For a rigorous proof of (XV.32), see Proposition XV.9 below.

It then follows that

$$\left\| \frac{d}{dt}(\gamma_{\xi}(e^{2\pi it}) \right\| = L \text{ for almost all } t \in [0,1]. \qquad \text{(XV.33)}$$

Therefore

Proposition XV.12. *If $\xi \in C^0(\mathbb{S}^1,\mathbb{C})$ is a rectifiable loop of length L, and the loop γ_ξ is defined as above, then γ_ξ is parametrized by constant times arc-length (the constant being L), and $\xi = \gamma \circ \psi$, where ψ is the map*

$$\mathbb{S}^1 \ni e^{2\pi it} \mapsto \psi(e^{2\pi it}) \stackrel{\text{def}}{=} e^{2\pi i\lambda s_{\xi}(t)} \in \mathbb{S}^1,$$

and $\lambda = L^{-1}$.

XV.12 Connected components

Vaca.

XV.13 The winding number

If $\xi : \mathbb{S}^1 \mapsto \mathbb{C}$ is a loop, and $p \in \mathbb{C}\backslash|\xi|$, then the *winding number* of ξ about p in the integer $\nu(\xi,p)$ defined as follows: pick any positive L, and use (E7) to choose a continuous function $f : \mathbb{R} \mapsto \mathbb{C}$ such that $e^{f(t)} = (\xi \circ \kappa_L)(t) - p$ for all $t \in \mathbb{R}$. Then (E4) implies that the number $\frac{f(t+L)-f(t)}{2\pi i}$ necessarily has to be an integer, because

$$e^{f(t+L)} = (\xi \circ \kappa_L)(t+L) - p = (\xi \circ \kappa_L)(t) - p = e^{f(t)}.$$

It is clear that this integer depends continuously on t. Furthermore, if $f_1, f_2 : \mathbb{R} \mapsto \mathbb{C}$ are any two continuous maps such that $e^{f_1(t)} = (\xi \circ \kappa_L)(t) - p$ and $e^{f_2(t)} = (\xi \circ \kappa_L)(t) - p$ for all $t \in \mathbb{R}$, then f_1 and f_2 must differ by a constant, because $e^{f_1(t)-f_2(t)} = 1$, so that $f_1(t) - f_2(t) = 2\pi ik(t)$, where $k(t)$ is an integer that depends continuously on t. So $\frac{f(t+L)-f(t)}{2\pi i}$ is in fact independent of t and f, and it is easy to see that it is also independent of L. We define $\nu(\xi,p)\stackrel{\text{def}}{=}\frac{f(t+L)-f(t)}{2\pi i}$ for any $t\in\mathbb{R}$, where L and f are as above.

It is clear that reparametrization preserves carriers and winding numbers. Precisely,

Proposition XV.13. *If η is a reparametrization of a loop $\xi\in C^0(\mathbb{S}^1,\mathbb{C})$, then $|\eta|=|\xi|$ and $\nu(\eta,p)=\nu(\xi,p)$ for all $p\in\mathbb{C}\backslash|\xi|$.*

PROOF. The fact that $|\eta|=|\xi|$ is trivial. Let $f:\mathbb{R}\mapsto\mathbb{C}$ be continuous and such that $e^{f(t)}=\eta(e^{2\pi it})$ for all t. Let $\psi:\mathbb{S}^1\mapsto\mathbb{S}^1$ be an orientation-preserving homeomorphism such that $\eta=\xi\circ\psi$. Then there exist $a\in\mathbb{R}$ and a homeomorphism $\tilde{\psi}:[0,1]\mapsto[a,a+1]$ such that

$$e^{2\pi i\tilde{\psi}(t)}=\psi(e^{2\pi it}) \tag{XV.34}$$

for all $t\in[0,1]$, $\tilde{\psi}(0)=a$ and $\tilde{\psi}(1)=a+1$. Extend $\tilde{\psi}$ to the whole real line by defining $\tilde{\psi}(k+t)=k+\psi(t)$ for $t\in[0,1]$ and $k\in\mathbb{Z}$. Let $g=f\circ\tilde{\psi}$. Then (XV.34) holds for all $t\in\mathbb{R}$. Moreover,

$$e^{g(t)}=e^{f(\tilde{\psi}(t))}=\eta(e^{2\pi i\tilde{\psi}(t)})=\eta\Big(\psi(e^{2\pi it})\Big)=(\eta\circ\psi)(e^{2\pi it})=\xi(e^{2\pi it}).$$

Therefore $\nu(\xi,p)=\frac{g(1)-g(0)}{2\pi i}=\frac{f(\tilde{\psi}(1))-f(\tilde{\psi}(0))}{2\pi i}=\frac{f(a+1)-f(a)}{2\pi i}=\nu(\eta,p)$. $\square$

If a loop ξ is rectifiable, then we can express the winding number as an integral.

Proposition XV.14. *Let ξ be a rectifiable loop in $\mathbb{C}$. Then*

$$\nu(\xi,p)=\frac{1}{2\pi i}\int_0^L\frac{\dot{\zeta}(s)}{\zeta(s)-p}\,ds \text{ if } p\in\mathbb{C}\backslash|\xi|\,,\ L>0\,,\text{ and } \zeta=\xi\circ\kappa_L. \tag{XV.35}$$

PROOF. We can choose f by letting

$$f(t)=f_0+\int_0^t\frac{\dot{\zeta}(s)}{\zeta(s)-p}\,ds\,, \tag{XV.36}$$

where $f_0\in\mathbb{C}$ is such that $e^{f_0}=\zeta(0)-p$. Indeed, if we define $g:\mathbb{R}\mapsto\mathbb{C}$ by letting $g(t)=(\zeta(t)-p).e^{-f(t)}$, we have

$$\begin{aligned}
\dot{g}(t) &= -(\zeta(t)-p).\dot{f}(t).e^{-f(t)}+\dot{\zeta}(t).e^{-f(t)}\\
&= -(\zeta(t)-p).\frac{\dot{\zeta}(t)}{\gamma(t)-p}.e^{-f(t)}+\dot{\zeta}(t).e^{-f(t)}\\
&= -\dot{\zeta}(t).e^{-f(t)}+\dot{\zeta}(t).e^{-f(t)}\\
&= 0.
\end{aligned}$$

Therefore $g(t) \equiv$ constant. Since $g(0) = 1$, we have $g(t) \equiv 1$ for all t, so $e^{f(t)} \equiv \gamma(t) - p$. Therefore $\nu(\xi, p) = \frac{1}{2\pi i}(f(L) - f(0))$, and then (XV.36) implies that (XV.35) holds. □

Proposition XV.15. *The winding number $\nu(\xi, p)$ depends continuously on the point p and the curve ξ as long as p and ξ vary in such a way that $p \in \mathbb{C}\backslash|\xi|$. More precisely, if $\{p_j\}_{j=1}^{\infty}$ is a sequence of points of $\mathbb{C}$ that converges to $p \in \mathbb{C}$, $\{\xi_j\}_{j=1}^{\infty}$ is a sequence of loops in $\mathbb{C}$ that converges uniformly to a loop ξ, and $p \notin \mathbb{C}\backslash|\xi|$, then there exists a positive integer j_* such that $p_j \notin \mathbb{C}\backslash|\xi_j|$ and $\nu(\xi_j, p_j) = \nu(\xi, p)$ whenever $j \geq j_*$*

PROOF. Let f be a continuous map from $\mathbb{R}$ to $\mathbb{C}$ having with the property that $e^{f(t)} = \xi(e^{2\pi it}) - p \stackrel{\text{def}}{=} h(t)$ for all $t \in \mathbb{R}$. Let $h_j(t) = \xi_j(e^{2\pi it}) - p_j$. Then $h_j \to h$ as $j \to \infty$, uniformly on compact sets, so $h_j \to h$ uniformly, since the h_j are 1-periodic. Since $h(t) \neq 0$ for all t, the functions $g_j \stackrel{\text{def}}{=} \frac{h_j}{h}$ converge uniformly to the constant 1 as $j \to \infty$. Let D be the open disc $\{z \in \mathbb{C} : \|z - 1\| < 1\}$. Let $\text{Ln} : D \mapsto \mathbb{C}$ be the function defined in (E5). Fix $j_* \in \mathbb{N}$ such that $g_j(t) \in D$ whenever $t \in \mathbb{R}$ and $j \geq j_*$. Then, if $j \geq j_*$, $h_j(t)$ must be $\neq 0$ for every t, so $p_j \notin |\xi_j|$. Define

$$f_j(t) = f(t) + \text{Ln}\Big(g_j(t)\Big) \text{ for } t \in \mathbb{R},\ j \geq j_*.$$

Then

$$e^{f_j(t)} = e^{f(t)}.e^{\text{Ln}(g_j(t))} = h(t).g_j(t) = h_j(t) = \xi_j(e^{2\pi it}) - p_j\,.$$

Therefore

$$\begin{aligned} \nu(\xi_j, p_j) &= \frac{1}{2\pi i}(f_j(1) - f_j(0)) \\ &= \frac{1}{2\pi i}\Big(f(1) - f(0) + \text{Ln}(g_j(1)) - \text{Ln}(g_j(0))\Big) \\ &= \frac{1}{2\pi i}\Big(f(1) - f(0)\Big) \\ &= \nu(\xi, p) \end{aligned}$$

whenever $j \geq j_*$, since $\text{Ln}(g_j(1)) = \text{Ln}(g_j(0))$, because the function g_j is 1-periodic.)
□

XV.14 The no-retraction theorem

Theorem XV.16. *Let $R > 0$, and write*

$$D(R) = \{z \in \mathbb{C} : |z| \leq 1\}, \qquad S(R) = \{z \in \mathbb{C} : |z| = 1\}.$$

Then there does not exist a continuous map $\rho : D(R) \mapsto S(R)$ such that $\rho(z) = z$ whenever $z \in S(R)$.

PROOF. Suppose ρ exists. For $0 \le r \le 1$, define a loop $\xi_r : \mathbb{S}^1 \mapsto \mathbb{C}$ by letting $\xi_r(e^{2\pi it}) = \rho(re^{2\pi it})$. Then each ξ_r satisfies $|\xi_r(e^{2\pi it})| = 1$ for all t, so $0 \notin |\xi_r|$. Let $w(r) = \nu(\xi_r, 0)$. Therefore Proposition XV.15 implies that the integer-valued function $[0,1] \ni r \mapsto w(r) \in \mathbb{R}$ is continuous. Hence $w(r) \equiv$ constant, so $w(0) = w(1)$. Furthermore, ξ_0 is a constant map, from which it follows easily that $\nu(\xi_0, 0) = 0$. On the other hand, since $\xi_1(e^{2\pi it}) = \rho(e^{2\pi it}) = e^{2\pi it}$, the function $f : \mathbb{R} \mapsto \mathbb{C}$ given by $f(t) = 2\pi it$ satisfies $e^{f(t)} = \xi_1(e^{2\pi it})$, so $\nu(\xi_1, 0) = \frac{f(1)-f(0)}{2\pi i} = 1$. So $w(1) \neq w(0)$, and we have reached a contradiction. □

XV.15 The Jordan curve theorem

If K is a compact subset of $\mathbb{C}$, and $p \in \mathbb{C} \backslash K$, the *Borsuk map* $\beta_{p,K} : K \mapsto \mathbb{S}^1$ is defined by

$$\beta_{p,K}(z) = \frac{z-p}{|z-p|} \qquad \text{for } z \in K\,.$$

Proposition XV.17. *Let K be a compact subset of $\mathbb{C}$, let $p \in \mathbb{C} \backslash K$, and let U be the connected component of $\mathbb{C} \backslash K$ that contains p. Assume that U is bounded. Then the Borsuk map $\beta_{p,K}$ cannot be extended to a continuous map $\beta : K \cup U \mapsto \mathbb{S}^1$.*

PROOF. Assume β exists. Let R be positive and so large that $K \cup U$ is entirely contained in the disc $D(p,R) = \{z \in \mathbb{C} : |z-p| \le R\}$. Define a map $\tilde{\beta} : D(p,R) \mapsto \partial D(p,R)$ by letting

$$\tilde{\beta}(z) = \begin{cases} p + R \cdot \beta(z) & \text{if } \; z \in K \cup U\,, \\ p + R \cdot \frac{z-p}{|z-p|} & \text{if } \; z \in D(p,R) \backslash U\,. \end{cases}$$

Then $\tilde{\beta}$ is well defined, because $(D(p,R) \backslash U) \cap (K \cup U) \subseteq K$ and $\beta(z) = \frac{z-p}{|z-p|}$ if $z \in K \cup U$. Furthermore, $\tilde{\beta}$ is continuous, because both sets $D(p,R) \backslash U$ and $K \cup U$ are closed. So $\tilde{\beta}$ is a continuous map from $D(p,R)$ to $\partial D(p,R)$. Moreover, it is clear that $\tilde{\beta}(z) = z$ if $z \in \partial D(p,R)$. So $\tilde{\beta}$ is a continuous retraction from $D(p,R)$ to $\partial D(p,R)$. Since such a retraction does not exist, we have reached a contradiction.
□

Proposition XV.18. *Let K be a compact subset of $\mathbb{C}$, and let Q be a subset of $\mathbb{C} \backslash K$ that contains at least one point of each bounded connected component of $\mathbb{C} \backslash K$. Then every continuous map $f : K \mapsto \mathbb{S}^1$ can be extended to a continuous map $F : \mathbb{C} \backslash Q \mapsto \mathbb{S}^1$.*

PROOF. For each positive number r such that $r < \sqrt{2}$, we define

$$S_r = \{s \in \mathbb{S}^1 : |s-1| \leq r\},$$

and let $\rho(r) = \frac{r}{2}\sqrt{4-r^2}$. Then the set S_r can be identified with the real interval $I_r \stackrel{\text{def}}{=} [-\rho(r), \rho(r)]$ by means of the map

$$I_r \ni u \mapsto \mu_r(u) \stackrel{\text{def}}{=} \sqrt{1-u^2} + ui \in \mathbb{S}^1. \tag{XV.37}$$

Now fix α such that $0 < \alpha < \frac{\sqrt{2}}{3}$. Choose a positive number δ such that $|f(z) - f(z')| < \alpha$ whenever $z, z' \in K$ and $|z - z'| \leq 2\sqrt{2}\delta$. Let G^δ be the δ-grid of $\mathbb{C}$, i.e., the set of all rectangles

$$R^\delta_{n,m} = \{x + iy : \delta n \leq x \leq \delta(n+1), \delta m \leq x \leq \delta(m+1)\},$$

for all $n, m \in \mathbb{Z}$. For each $R \in G^\delta$ we let $V(R)$, $E(R)$ be, respectively, the set of vertices of R and the set of edges of R.

Let G^δ_K be the set of those rectangles $R \in G^\delta$ such that R intersects K. Define

$$|G^\delta_K| \stackrel{\text{def}}{=} \bigcup\{R : R \in G^\delta\}, \qquad \tilde{K} \stackrel{\text{def}}{=} K \cup |G^\delta_K|.$$

We will first extend f to a continuous map $\tilde{f} : \tilde{K} \mapsto \mathbb{S}^1$. For this purpose, we pick a point $z_R \in K \cap R$ for each $R \in G^\delta_K$. Let

$$V_K = \bigcup\{V(R) : R \in G^\delta_K\}.$$

Let $K_1 = K \cup V_K$. We then extend f to a continuous map $f_1 : K_1 \mapsto \mathbb{S}^1$ by defining $f_1(z)$ as follows, if $z \in K_1$:

1. If $z \in K$, we define $f_1(z) = f(z)$.
2. If $z \in V_K \backslash K$, we take $f_1(z)$ to be $f(z_R)$, where R is such that $z \in R$ and $R \in G^\delta_K$. (For a given z, the rectangle R exists, but need not be unique, since z could belong to two, three or four rectangles in G^δ_K. If R is not unique, we choose one for each z, in an arbitrary fashion.)

With this definition, f_1 satisfies the inequality

$$|f_1(z) - f_1(z_R)| < \alpha \quad \text{whenever} \quad z \in R \cap K_1,\ R \in G^\delta_K. \tag{XV.38}$$

Indeed, if $z \in R \cap K$, then $|z - z_R| \leq \sqrt{\delta}$, so $|f_1(z) - f_1(z_R)| < \alpha$. If $z \in V_K$ but $z \notin K$, then $f_1(z) = f(z_{R'})$ for some rectangle $R' \in G^\delta_K$ such that $R' \cap R \neq \emptyset$. Then $|z_{R'} - z_R| \leq 2\sqrt{2}\delta$, so $|f_1(z_{R'}) - f_1(z_R)| < \alpha$, and then $|f_1(z) - f_1(z_R)| < \alpha$.

Next, we let

$$E_K = \bigcup\{E(R) : R \in G_K^\delta\}\,, \qquad |E_K| = \bigcup\{e : e \in E_K\}\,.$$

Let $K_2 = K \cup |E_K| = K_1 \cup |E_K|$. We extend f_1 to a continuous map $f_2 : K_2 \mapsto \mathbb{S}^1$ by defining f_2 as follows:

1. We let $f_2(z) = f_1(z)$ if $z \in K_1$.

2. For each edge $e \in E_K$, we pick a point $z_e \in e \cap K_1$. (This is possible because $e \cap K_1 \neq \emptyset$, since the vertices of e belong to K_1.) Then the inequality $|f_1(z) - f_1(z_e)| < 2\alpha$ holds for all $z \in e \cap K_1$, because there exists $R \in G_K^\delta$ such that $e \subseteq R$, and then $|f_1(z) - f_1(z_R)| < \alpha$ for all $z \in e \cap K_1$. Then

$$e \cap K_1 \ni z \mapsto \frac{f_1(z)}{f_1(z_e)} \in \mathbb{S}^1$$

is a continuous map from $e \cap K_1$ to the set $S_{2\alpha}$ (thanks to the identity $\left|\frac{f_1(z)}{f_1(z_e)} - 1\right| = |f_1(z) - f_1(z_e)|$, which is true because $|f_1(z_e)| = 1$). Since $2\alpha < \sqrt{2}$, we know that $S_{2\alpha}$ can be identified with the real interval $I_{2\alpha}$ by means of the map $\mu_{2\alpha}$ defined in (XV.37). Then the Tietze extension theorem implies that there exists a map $\varphi_e : e \mapsto S_{2\alpha}$ such that $\varphi_e(z) = \frac{f_1(z)}{f_1(z_e)}$ if $z \in e \cap K_1$. We can then define f_2 on e by letting $f_2(z) = \varphi_e(z) \cdot f_1(z_e)$ for $z \in e$.

We now claim that f_2 satisfies the inequality

$$|f_2(z) - f_2(z_R)| < 3\alpha \quad \text{whenever} \quad z \in R \cap K_2\,,\ R \in G_K^\delta\,. \tag{XV.39}$$

Indeed, if $z \in R \cap K_1$, then $f_2(z) = f_1(z)$ and $f_2(z_R) = f_1(z_R)$, so (XV.38) implies that $|f_2(z) - f_2(z_R)| < \alpha$. If $z \in R$ but $z \notin K_1$, then z belongs to and edge $e \in E(R)$. Then (XV.38) implies that $|f_2(z_e) - f_2(z_R)| = |f_1(z_e) - f_1(z_R)| < \alpha$. On the other hand,

$$|f_2(z) - f_2(z_e)| = |f_2(z_e)| \cdot \left|\frac{f_2(z)}{f_2(z_e)} - 1\right| = |\varphi_e(z) - 1| < 2\alpha\,.$$

Therefore $|f_2(z) - f_2(z_R)| < 3\alpha$.

We now define the extension $\tilde{f} : \tilde{K} \mapsto \mathbb{S}^1$. For each rectangle $R \in G_K^\delta$, the map f_2 is defined on the compact set $K_2 \cap R$, which contains the boundary of R. Furthermore, (XV.39) implies that

$$K_2 \cap R \ni z \mapsto \frac{f_2(z)}{f_2(z_R)} \in \mathbb{S}^1$$

is a continuous map to the set $S_{3\alpha}$. Since $3\alpha < \sqrt{2}$, the set $S_{3\alpha}$ can be identified with the real interval $I_{3\alpha}$ by means of the map $\mu_{3\alpha}$ of (XV.37). Then the Tietze extension theorem implies that there exists a map $\psi_R : R^\delta \mapsto S_{3\alpha}$ such that $\psi_R(z) = \frac{f_2(z)}{f_2(z_R)}$ if $z \in K_2 \cap R$. We then define $\tilde{f}$ on R by letting $\tilde{f}(z) = \psi_R(z) \cdot f_2(z_R)$ for $z \in R$.

Let W be the set of all vertices of G^δ that do not belong to $\tilde{K}$. We let $\tilde{K}_1 = \tilde{K} \cup W$, and extend $\tilde{f}$ to a map $\tilde{f}_1 : \tilde{K}_1 \mapsto \mathbb{S}^1$ by choosing $\tilde{f}_1(z) \in \mathbb{S}^1$ in an arbitrary fashion if $z \in W$. We then let H be the set of all edges of G^δ that are not contained in $\tilde{K}$, and define $|H| = \bigcup\{h : h \in H\}$. We define $\tilde{K}_2 = \tilde{K}_1 \cup |H|$, and extend $\tilde{f}_1$ to a continuous map $\tilde{f}_2 : \tilde{K}_2 \mapsto \mathbb{S}^1$ as follows: for each $h \in H$, let v, w be the vertices of h, pick real numbers a, b such that $e^{2\pi i a} = v$ and $e^{2\pi i b} = w$, and define $\tilde{f}_2(tv + (1-t)w) = e^{2\pi i(ta+(1-t)b)}$ if $0 \le t \le 1$.

Next, we let $\mathscr{R} = G^\delta \backslash G^\delta_K$, so $\mathscr{R}$ is the set of all rectangles $R \in G^\delta$ that do not intersect K. For each $R \in \mathscr{R}$, use c_R to denote the center of R. The map $\tilde{f}_2$ is defined on the boundary ∂R. Pick, for each R, a continuous map $p_R : R\backslash\{c_R\} \mapsto \partial R$ such that $p_R(z) = z$ whenever $z \in \partial R$. We then let $\tilde{K}_3 = \tilde{K}_2 \cup \bigcup\{R\backslash\{c_R\} : R \in \mathscr{R}\}$ and extend $\tilde{f}_2$ to $\tilde{K}_3$ by letting

$$\tilde{f}_3(z) = \begin{cases} \tilde{f}_2(z) & \text{if} \quad z \in \tilde{K}_2\,, \\ \tilde{f}_2(p_R(z)) & \text{if} \quad z \in R\backslash\{c_R\} \text{ and } R \in \mathscr{R}\,. \end{cases}$$

We have now extended f to a continuous map $\tilde{f}_3 : \tilde{K}_3 \mapsto \mathbb{S}^1$, where $\tilde{K}_3$ is the whole complex plane $\mathbb{C}$, minus the discrete set $C = \{c_R : R \in \mathscr{R}\}$. Then $C \subseteq \mathbb{C}\backslash K$. If U is any bounded connected component of $\mathbb{C}\backslash K$, then $C \cap U$ is finite. Pick a point $q_U \in Q \cap U$, and define $C^0_U = (C \cap U) \cup \{q_U\}$.

The map $\tilde{f}_3$ is then continuous on $U\backslash C^0_U$. We show how to replace the restriction $\tilde{f}_{3,U}$ of $\tilde{f}_3$ to $U\backslash C^0_U$ by a map $\tilde{f}_{4,U}$ into $\mathbb{S}^1$ which is continuous on $U\backslash\{q_U\}$ and coincides with $\tilde{f}_{3,U}$ outside a compact subset of U that contains C^0_U. If $C^0_U = \{q_U\}$, then of course we can take $\tilde{f}_{4,U} = \tilde{f}_{3,U}$. Otherwise, we pick a point P of $C^0_U\backslash\{q_U\}$, and construct a polygonal curve $\xi : [a,b] \mapsto U$ such that $\xi(a) = P$, $\xi(b) = q_U$, the map ξ is one-to-one, and $\xi(t) \notin C^0_U$ whenever $a < t < b$. Let $(P_0, \ldots, P_k)$ be the ordered sequence of vertices of ξ, so that $P_0 = P$, $P_k = q_U$, and there exist $t_0, \ldots, t_k$ with the property that $a = t_0 < t_1 < \ldots < t_{k-1} < t_k = b$, $\xi(t_j) = P_j$ for $j = 0, \ldots, k$, and ξ is linear on each interval $[t_{j-1}, t_j]$.

Let $\sigma_j = \{\xi(t) : t_{j-1} \le t \le t_j\}$, so σ_j is the segment joining P_{j-1} to P_j. Then σ_0 does not contain any points of C^0_U other than P and q_U. Therefore the map $\tilde{f}_{3,U}$ is continuous on a neighborhood of σ_0, except possibly at the endpoints of σ_0. Let τ_0 be the compact neighborhood $\{z : \text{dist}(z, \sigma_0) \le \theta\}$ of σ_0, where θ is positive and small enough, so that $\tau_0 \subseteq U$ and τ_0 does not contain any points of C^0_U other than P and q_U. Then $\tilde{f}_{3,U}$ is continuous on τ_0, except possibly at P_0 and P_1. The set $\tau_0 \backslash \{P_1\}$ obviously admits a continuous retraction $\kappa_0 : \tau_0 \backslash \{P_1\} \mapsto \partial \tau_0$. Replace $\tilde{f}_{3,U}$ by $\tilde{f}_{3,U} \circ \kappa_0$ on $\tau_0 \backslash \{P_1\}$. Then the modified map $\tilde{f}^0_{3,U}$ is continuous wherever $\tilde{f}_{3,U}$ was, except possibly at P_1, and is also continuous at P_0. So the discontinuity at P_0 has been eliminated, and a possible discontinuity at P_1 has been introduced. Repeating this procedure, we replace the discontinuity at P_1 by a possible discontinuity at P_2, and continue in this way until, of all the points of ξ, only the terminal point q_U remains as a possible discontinuity of the new map. This means that the set C^0_U of points of possible discontinuity of $\tilde{f}_{3,U}$ has been reduced by eliminating one point other than q_U. We then repeat this procedure, until all points of discontinuity other than q_U have been eliminated.

By carrying out this elimination of discontinuities separately on each bounded connected component of $\mathbb{C}\backslash K$, we obtain a continuous extension $\hat{f} : \mathbb{C}\backslash(Q \cup D) \mapsto \mathbb{S}^1$ of f to the whole complex plane, minus $Q \cup D$, where D is a discrete subset which is entirely contained in the unbounded component Ω of $\mathbb{C}\backslash K$. The set D can then be completely eliminated, by first picking a positive L such that $K \subseteq \{z : |z| \leq L\}$, and then eliminating all the discontinuities in $\{z : |z| \leq L\}$ as was done for the bounded components. We then let κ be the retraction $z \mapsto L\frac{z}{|z|}$ from $\{z : |z| \geq L\}$ to $\{z : |z| = L\}$, and replace $\hat{f}$ by $\hat{f} \circ \kappa$ on $\{z : |z| \geq L\}$. The resulting map F now has no discontinuities whatsoever on Ω, so F is the desired extension. □

If K is a subset of $\mathbb{C}$, and f_0, f_1 are continuous maps from K to $\mathbb{S}^1$, a *homotopy* from f_0 to f_1 is a continuous map $h : K \times [0,1] \mapsto \mathbb{S}^1$ such that $h(z,0) = f_0(z)$ and $h(z,1) = f_1(z)$ whenever $z \in K$. We say that f_0 and f_1 are *homotopic* if there exists a homotopy from f_0 to f_1.

Proposition XV.19. *Let K be a compact subset of $\mathbb{C}$, and let C be a closed subset of $\mathbb{C}$ such that $K \subseteq C$. Let f_0, f_1 be continuous maps from K to $\mathbb{S}^1$, such that f_0 and f_1 are homotopic, and f_0 admits a continuous extension to a map $F_0 : C \mapsto \mathbb{S}^1$. Then f_1 admits a continuous extension to a map $F_1 : C \mapsto \mathbb{S}^1$.*

PROOF. Let $h : K \times [0,1] \mapsto \mathbb{S}^1$ be a continuous map that satisfies the conditions $h(z,0) = f_0(z)$ and $h(z,1) = f_1(z)$ whenever $z \in K$. The map h is uniformly continuous on $K \times [0,1] \mapsto \mathbb{S}^1$, so there exists a positive δ such that $|h(z,t) - h(z,s)| < 1$ whenever $z \in K$, $t,s \in [0,1]$, and $|t-s| \leq \delta$. Let N be a positive integer such that $N\delta > 1$. For $k = 0, \dots, N$, let g_k be the map

$$K \ni z \mapsto g_k(z) \stackrel{\text{def}}{=} h\Big(z, \frac{k}{N}\Big) \in \mathbb{S}^1.$$

Then $g_0 = f_0$ and $g_N = f_1$. We prove by induction on k that g_k has a continuous extension to C. The fact that g_0 has a continuous extension follows because $g_0 = f_0$.

Assuming that g_k has a continuous extension G_k, and $k < N$, use the inequality $|g_k(z) - g_{k+1}(z)| < 1$, which is valid for all $z \in K$, to conclude that $|\theta_k(z) - 1| < 1$ for all $z \in K$, where $\theta_k(z) \stackrel{\text{def}}{=} \frac{g_{k+1}(z)}{g_k(z)}$.

So θ_k is a continuous map from K to the set $S_1 = \{z \in \mathbb{S}^1 : |z - 1| \leq 1\}$. This set can be identified with the real interval $I_1 = [-\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2}]$, as explained in the proof of Proposition XV.18. Therefore the Tietze extension theorem yields an extension of θ_k to a continuous map $\Theta_k : C \mapsto \mathbb{S}^1$. Define $G_{k+1}(z) = \Theta_k(z) \cdot G_k(z)$ for $z \in C$. Then $G_{k+1}(z)$ is a continuous map from C to $\mathbb{S}^1$, and $G_{k+1}(z) = g_{k+1}(z)$ whenever $z \in K$.

In particular, g_N has a continuous extension. Since $g_N = f_1$, our proof is complete. □

Proposition XV.20. *Let K be a compact subset of $\mathbb{C}$. Then $\mathbb{C} \backslash K$ is connected if and only if every continuous map $f : K \mapsto \mathbb{S}^1$ is homotopic to a constant.*

PROOF. If $\mathbb{C} \backslash K$ is connected, and $f : K \mapsto \mathbb{S}^1$ is continuous, then Proposition XV.18 implies that f has a continuous extension $F : \mathbb{C} \mapsto \mathbb{S}^1$. If we define $h(z,t) = F((1-t)z)$ for $z \in K$ and $t \in [0,1]$, then h is clearly a homotopy from f to the constant map $K \ni z \mapsto F(0) \in \mathbb{S}^1$.

To prove the converse, suppose every continuous map $f : K \mapsto \mathbb{S}^1$ is homotopic to a constant. Since a constant map obviously admits a continuous extension to $\mathbb{C}$, Proposition XV.19 implies that every continuous map $f : K \mapsto \mathbb{S}^1$ admits a continuous extension to a map $F : \mathbb{C} \mapsto \mathbb{S}^1$. In particular, if U is a bounded connected component of $\mathbb{C} \backslash K$, and $p \in U$, then the Borsuk map $\beta_{p,K}$ would admit a continuous extension. But Proposition XV.17 says that $\beta_{p,K}$ does not have a continuous extension. Therefore $\mathbb{C} \backslash K$ has no bounded connected components. Since $\mathbb{C} \backslash K$ has only one unbounded connected component, we have established that $\mathbb{C} \backslash K$ is connected, as desired. □

Proposition XV.21. *Let K be a compact subset of $\mathbb{C}$. Assume that K is homeomorphic to the interval $[0,1]$. Then $\mathbb{C} \backslash K$ is connected.*

PROOF. It is clear that every continuous map from $[0,1]$ to $\mathbb{S}^1$ is homotopic to a constant. The conclusion then follows from Proposition XV.20. □

Proposition XV.22. *Let K be a compact subset of $\mathbb{C}$. Assume that K is homeomorphic to $\mathbb{S}^1$. Then*

1. *$\mathbb{C} \backslash K$ is not connected;*

2. *if U is any connected component of $\mathbb{C}\backslash K$, then the closure of U is the set $U \cup K$.*

PROOF. The identity map $z \mapsto z$ from $\mathbb{S}^1$ to $\mathbb{S}^1$ is not homotopic to a constant. (For example, if it was homotopic to a constant then Proposition XV.19 would yield a continuous extension to the disc $D(0,1) = \{z \in \mathbb{C} : |z| \leq 1\}$. This extension would be a continuous retraction from the disc to its boundary, contradicting the no-retraction theorem.)

Therefore not every continuous map from $\mathbb{S}^1$ to $\mathbb{S}^1$ is homotopic to a constant. So not every continuous map from K to $\mathbb{S}^1$ is homotopic to a constant. Therefore Proposition XV.20 implies that $\mathbb{C}\backslash K$ is not connected.

Now let U be a connected component of $\mathbb{C}\backslash K$, and let $\bar{U}$ denote its closure. Then $U \subseteq \bar{U} \subseteq U \cup K$. To prove our second conclusion, we have to pick $z \in K$ and show that $z \in \bar{U}$. Let B be an open disc with center z, and let B' be a compact ball with center z such that $B' \subseteq B$. Let J be a nonempty subset of K which is contained in B and is such that $K' = K\backslash J$ is homeomorphic to $[0,1]$. Let V be another connected component of $\mathbb{C}\backslash K$. Pick points $p \in U$, $q \in V$. Then p and q belong to $\mathbb{C}\backslash K'$, which is connected by Proposition XV.21. So there exists a continuous path $\xi : [0,1] \mapsto \mathbb{C}\backslash K'$ such that $\xi(0) = p$ and $\xi(1) = q$. Then ξ cannot be contained in $\mathbb{C}\backslash K$, because it joins points in two different components of $\mathbb{C}\backslash K$. Therefore $\xi(t) \in J$ for some t, and then $\xi(t) \in B'$ for some t. Let $\bar{t}$ be the smallest $t \in [0,1]$ such that $\xi(t) \in B'$. If $\bar{t} = 0$, then $p \in B'$, so $B' \cap U \neq \emptyset$ and then $B \cap U \neq \emptyset$. If $\bar{t} > 0$, then the set $\Xi = \{\xi(t) : 0 \leq t < \bar{t}\}$ is connected and entirely contained in U (because Ξ does not intersect K). Since $\xi(\bar{t}) \in B'$, $\xi(t)$ must belong to B for $t < \bar{t}$, t sufficiently close to $\bar{t}$. Therefore $B \cap U \neq \emptyset$.

We have shown that $B \cap U \neq \emptyset$ if B is an arbitrary open disc with center z. Therefore $z \in \bar{U}$, and our proof is complete. $\square$

Proposition XV.23. *Let K be a compact subset of $\mathbb{C}$, and let $f : K \mapsto \mathbb{C}$ be a continuous map. Then f is homotopic to a constant map if and only if there exists a continuous function $\varphi : K \mapsto \mathbb{R}$ such that $f(z) = e^{2\pi i\varphi(z)}$ for all $z \in K$.*

PROOF. If φ exists, then we can define $h(z,t) = e^{2\pi i(1-t)\varphi(z)}$ for $z \in K$ and $t \in [0,1]$, and it is clear that h is a homotopy from f to a constant map.

Conversely, suppose $q \in \mathbb{S}^1$ and $h : K \times [0,1] \mapsto \mathbb{S}^1$ be a continuous map that satisfies the conditions $h(z,0) = f(z)$ and $h(z,1) = q$ whenever $z \in K$. Fix α such that $0 < \alpha < \frac{1}{2}$. Then the map $[-\alpha,\alpha] \ni ue^{2\pi iu}$ is a bijection onto the compact subset $S = \{e^{2\pi iu} : u \in [-\alpha,\alpha]\}$ of $\mathbb{S}^1$. Fix β such that $0 < \beta$ and the set $V = \{z \in \mathbb{S}^1 : |z-1| \leq \beta\}$ is contained in S.

The map h is uniformly continuous on $K \times [0,1] \mapsto \mathbb{S}^1$, so there exists a positive δ such that $|h(z,t) - h(z,s)| < \beta$ whenever $z \in K$, $t, s \in [0,1]$, and $|t - s| \le \delta$. Let N be a positive integer such that $N\delta > 1$. For $j = 1, \ldots, N$, let g_j be the map

$$K \ni z \mapsto g_j(z) \stackrel{\text{def}}{=} \frac{h\left(z, \frac{j-1}{N}\right)}{h\left(z, \frac{j}{N}\right)} \in \mathbb{S}^1.$$

Let $G(z) \equiv q$. Then $f = g_1 \cdot g_2 \cdot \ldots \cdot g_N \cdot G$. Clearly, each g_k takes values in S, so there exists a continuous $\psi_j : K \mapsto \mathbb{R}$ such that $g_j(z) = e^{2\pi i \psi_j(z)}$ for all $z \in K$. Let $c \in \mathbb{R}$ be such that $q = e^{2\pi i c}$. Define $\varphi = \psi_1 + \cdots + \psi_N + c$. Then $f(z) = e^{2\pi i \varphi(z)}$ for all $z \in K$. $\square$

Proposition XV.24. *Let K be a compact subset of $\mathbb{C}$. Assume that K is the union of two compact sets K_1, K_2, such that $K_1 \cap K_2$ is connected. Let p_1, p_2 be points of $\mathbb{C} \backslash K$ that belong to the same connected component of $\mathbb{C} \backslash K_1$ and to the same connected component of $\mathbb{C} \backslash K_2$. Then p_1, p_2 belong to the same connected component of $\mathbb{C} \backslash K$.*

PROOF. Since p_1 and p_2 belong to the same connected component of $\mathbb{C} \backslash K_1$, the Borsuk maps $\beta_{p_1,K_1}, \beta_{p_2,K_1} : K_1 \mapsto \mathbb{S}^1$ are homotopic. Therefore, if we let

$$f_1(z) \stackrel{\text{def}}{=} \frac{\beta_{p_2,K_1}(z)}{\beta_{p_1,K_1}(z)} \qquad \text{for} \quad z \in K_1,$$

it follows that the map f_1 is homotopic to a constant. By Proposition XV.23, this implies that there exists a continuous function $\varphi_1 : K_1 \mapsto \mathbb{R}$ such that $f_1(z) = e^{2\pi i \varphi_1(z)}$ whenever $z \in K_1$. A similar argument shows that there exists a continuous function $\varphi_2 : K_2 \mapsto \mathbb{R}$ such that $f_2(z) = e^{2\pi i \varphi_2(z)}$ whenever $z \in K_2$, where

$$f_2(z) \stackrel{\text{def}}{=} \frac{\beta_{p_2,K_2}(z)}{\beta_{p_1,K_2}(z)} \qquad \text{for} \quad z \in K_2.$$

The function $\psi = \varphi_2 - \varphi_1 : K_1 \cap K_2 \mapsto \mathbb{R}$ is continuous and integer-valued. Since $K_1 \cap K_2$ is connected, ψ is in fact constant. After subtracting this constant from φ_2, we may assume that $\varphi_2 = \varphi_1$ on $K_1 \cap K_2$. We can the define $\varphi : K \mapsto \mathbb{R}$ by letting $\varphi(z) = \varphi_j(z)$ whenever $z \in K_j$, $j = 1, 2$. Then, if we let

$$f(z) \stackrel{\text{def}}{=} \frac{\beta_{p_2,K}(z)}{\beta_{p_1,K}(z)} \qquad \text{for} \quad z \in K,$$

we have $f(z) = e^{2\pi i\varphi(z)}$. Therefore f is homotopic to a constant, and then the maps $\beta_{p_1,K}$ and $\beta_{p_2,K}$ are homotopic.

Clearly, $\beta_{p_2,K}$ has a continuous extension to a map $\mu : \mathbb{C}\backslash\{p_2\}$. Since $\beta_{p_1,K}$ is homotopic to $\beta_{p_2,K}$, Proposition XV.19 tells us that $\beta_{p_1,K}$ also has a continuous extension to a map $\nu : \mathbb{C}\backslash\{p_2\}$.

If the connected components U_1, U_2 of p_1, p_2 in $\mathbb{C}\backslash K$ were different, p_2 would not belong to U_1, and then ν could be restricted to $K \cup U_1$, yielding a continuous extension of $\beta_{p_1,K}$ to $K \cup U_1$. On the other hand, Proposition XV.17 tells us that such an extension does not exist. Therefore $U_1 = U_2$, completing our proof. □

Theorem XV.25. *(The Jordan curve theorem.) Let K be a compact subset of $\mathbb{C}$. Assume that K is homeomorphic to $\mathbb{S}^1$. Then*

1. *$\mathbb{C}\backslash K$ has exactly two connected components;*
2. *if U is any connected component of $\mathbb{C}\backslash K$, then* $\mathrm{Closure}(U) = U \cup K$.

PROOF. The second conclusion was already proved in Proposition XV.22. Furthermore, we also know from Proposition XV.22 that $\mathbb{C}\backslash K$ has at least two connected components. Therefore, it suffices to show that $\mathbb{C}\backslash K$ cannot have two different bounded connected components.

Let $\xi : \mathbb{S}^1 \mapsto K$ be a homeomorphism, and let $\gamma(t) = \xi(e^{2\pi it})$ for $0 \le t \le 1$, so γ is a continuous map from $[0,1]$ onto K such that $\gamma(0) = \gamma(1)$ and γ is one-to-one on $[0,1[$.

Suppose U_1, U_2 are different bounded connected components of $\mathbb{C}\backslash K$. Let U_0 be the unbounded component of $\mathbb{C}\backslash K$. Fix points p_0, p_1, p_2 in U_0, U_1, U_2, respectively. Fix eight times $t_1, t_2, \ldots, t_8$ having the property that $0 < t_1 < t_2 < t_3 < t_4 < t_5 < t_6 < t_7 < t_8 < 1$. Write $J_\ell = [t_{2\ell-1}, t_{2\ell}]$ and $I_\ell =]t_{2\ell-1}, t_{2\ell}[$ for $\ell = 1,2,3,4$. For each $\ell \in \{1,2,3,4\}$, define a compact set K_ℓ by letting $K_\ell = K\backslash\gamma(I_\ell)$. Then each K_ℓ is homeomorphic to a compact interval, so the open sets $\Omega_\ell = \mathbb{C}\backslash K_\ell$ are connected by Proposition XV.21.

Since p_0 and p_1 belong to Ω_ℓ for each ℓ, there exist continuous one-to-one maps $\eta_\ell : [0,1] \mapsto \Omega_\ell$ such that $\eta_\ell(0) = p_0$ and $\eta_\ell(1) = p_1$. (We are only interested in the curves η_1 and η_3.) Since p_0 and p_1 belong to different connected components of $\mathbb{C}\backslash K$, the curves η_ℓ must intersect K. Since η_ℓ does not intersect K_ℓ, it must intersect $\gamma(J_\ell)$. If we let s_ℓ be the smallest $s \in [0,1]$ such that $\eta_\ell(s) \in \gamma(J_\ell)$, then $\eta_\ell(s_\ell) \in \gamma(J_\ell)$, and $\eta_\ell(s) \in U_0$ for $0 \le s < s_\ell$. Define $\zeta_\ell : [0,1] \mapsto \mathbb{C}$ by restricting η_ℓ to $[0,s_\ell]$ and reparametrizing time so that the interval $[0,s_\ell]$ becomes $[0,1]$. Then ζ_ℓ is a continuous one-to-one map from $[0,1]$ to $\mathbb{C}$, such that $\zeta_\ell(0) = p_0$, $\zeta_\ell(1) \in \gamma(J_\ell)$, and $\zeta_\ell(t) \in U_0$ for $0 \le t < 1$.

The sets $\zeta_3([0,1])$ and $\gamma(J_1)$ are compact and do not intersect, because $\zeta_3([0,1]) \cap K = \{\zeta_3(1)\}$, and $\zeta_3(1) \notin \gamma(J_1)$. Therefore $\zeta_1(t) \notin \zeta_3([0,1])$ if t is close enough to 1. Let $\bar{t}_1$ be the largest $t \in [0,1]$ such that $\zeta_1(t) \in \zeta_3([0,1])$. Let $\bar{t}_3$ be the unique $t \in [0,1]$ such that $\zeta_3(t) = \zeta_1(\bar{t}_1)$. If $\bar{t}_1 > 0$, replace p_0 by $\zeta_1(\bar{t}_1)$, and then replace ζ_1 and ζ_3 by their restrictions to $[t_1,1]$ and $[t_3,1]$, reparametrized so that the intervals $[t_1,1]$, $[t_3,1]$ become $[0,1]$.

We then have two continuous maps $\zeta_1, \zeta_3 : [0,1] \mapsto \mathbb{C}$ such that

1. ζ_1 and ζ_3 are one-to-one,
2. $\zeta_1(0) = \zeta_3(0) = p_0$,
3. $\zeta_1(t) \in U_0$ and $\zeta_3(t) \in U_0$ whenever $0 \leq t < 1$,
4. $\zeta_1(1) \in \gamma(J_1)$ and $\zeta_3(1) \in \gamma(J_3)$,
5. $\zeta_1(t) \neq \zeta_3(s)$ whenever $0 \leq t \leq 1$ $0 \leq s \leq 1$, and $t+s > 0$.

Let τ_1, τ_3 be the unique points in J_1, J_3 such that $\gamma(\tau_1) = \zeta_1(1)$ and $\gamma(\tau_3) = \zeta_3(1)$. We construct two simple closed curves Λ_1, Λ_2, as follows. The curve Λ_1 is the concatenation of ζ_1, followed by the restriction of γ to the interval $[\tau_1, \tau_3]$, followed by ζ_3 run backwards. The curve Λ_3 is defined similarly, except for the fact that, to go from $\gamma(\tau_1)$ to $\gamma(\tau_3)$ we move along K in the other direction (that is, we follow the map $s \mapsto \xi(e^{2\pi i(\tau_1 - s)})$ from $s = 0$ until $s = 1 + \tau_1 - \tau_3$, at which point $\xi(e^{2\pi i(\tau_1 - s)})$ becomes $\gamma(\tau_3)$).

We now show that p_1 and p_2 can be connected by a continuous path ω_1 in $\mathbb{C} \backslash |\Lambda_1|$, where Λ_1 is, of course, the carrier of Λ_1. To see this, it suffices to construct continuous paths $\theta_j : [0,1] \mapsto \mathbb{C}$, for $j = 1,2$, such that $\theta_1(0) = p_1$, $\theta_2(0) = p_2$, $\theta_1(t) \in U_1$ for $0 \leq t < 1$, $\theta_2(t) \in U_2$ for $0 \leq t < 1$, $\theta_1(1) \in \gamma(J_4)$, and $\theta_2(1) \in \gamma(J_4)$. We then define ω_1 by concatenating θ_1, a portion of γ going from $\theta_1(1)$ to $\theta_2(1)$ along $\gamma(J_4)$, and θ_2 run backwards. The resulting path ω_1 is entirely contained in $U_1 \cup U_2 \cup \gamma(J_4)$, so ω_1 does not intersect $|\Lambda_1|$.

A similar argument shows that there exists a continuous path ω_3 connecting p_1 and p_2 in $\mathbb{C} \backslash |\Lambda_3|$. Since $|\Lambda_1| \cap |\Lambda_3| = |\zeta_1| \cup |\zeta_2|$, which is homeomorphic to a compact interval, Proposition XV.24 tells us that there exists a continuous path ω connecting p_1 and p_2 in $\mathbb{C} \backslash (|\Lambda_1| \cup |\Lambda_3|)$. Since $K \subseteq |\Lambda_1| \cup |\Lambda_3|)$, ω connects p_1 and p_2 in $\mathbb{C} \backslash K$, contradicting the assumption that U_1 and U_2 are different connected components of $\mathbb{C} \backslash K$. $\square$

XV.16 The winding numbers of a simple loop

The *Jordan curve theorem* says that if ξ is a simple loop then $\mathbb{R}^2\backslash|\xi|$ has exactly two connected components. In other words, $\mathbb{R}^2\backslash|\xi|$ is the union of two disjoint connected open subsets Ω_1, Ω_2 of $\mathbb{R}^2$. If $R \in \mathbb{R}$ is such that $R > 0$ and $\|\xi(s)\| \leq R$ for all $s \in \mathbb{S}^1$, then the set $S = \{q \in \mathbb{R}^2 : \|q\| > R\}$ is open, connected, and disjoint from $|\xi|$. Therefore S is contained in one of the components Ω_1, Ω_2. The other component must then be contained in $\{q \in \mathbb{R}^2 : \|q\| \leq R\}$. So one of the components is unbounded and the other one is bounded. We use $\mathrm{Ins}(\xi)$, $\mathrm{Out}(\xi)$ (the "inside" and the "outside" of ξ) to denote, respectively, the bounded and the unbounded connected component of $\mathbb{R}^2\backslash|\xi|$. We refer to $\mathrm{Ins}(\xi)$ as *the region enclosed by the simple loop* ξ. We use $A(\xi)$ to denote the area of $\mathrm{Ins}(\xi)$.

Theorem XV.26. *Let $\xi : \mathbb{S}^1 \mapsto \mathbb{R}^2 = \mathbb{C}$ be a simple loop. Then the funcion $\mathbb{C}\backslash|\xi| \ni p \mapsto \nu(\xi,p) \in \mathbb{Z}$ is constant on* $\mathrm{Out}(\xi)$ *and on* $\mathrm{Ins}(\xi)$*, and has the value* 0 *on* $\mathrm{Out}(\xi)$ *and the value* 1 *or* -1 *on* $\mathrm{Ins}(\xi)$.

PROOF. Proposition XV.15 tells us that the winding number $\nu(\xi,p)$ depends continuously on p as long as p varies on $\mathbb{C}\backslash|\xi|$. Since this number is always an integer, the function $\mathbb{C}\backslash|\xi| \ni p \mapsto \nu(\xi,p) \in \mathbb{Z}$ must be constant on $\mathrm{Out}(\xi)$ and on $\mathrm{Ins}(\xi)$.

Fix a positive $R \in \mathbb{R}$ such that $|\xi(e^{2\pi it})| \leq R$ for all $t \in R$, and a $p_* \in \mathbb{C}$ such that $|p| > R$. Then $\left|\frac{\xi(e^{2\pi it})}{p_*}\right| < 1$ for every $t \in \mathbb{R}$, so we can define a continuous function $\sigma : \mathbb{R} \mapsto \mathbb{C}$ by letting

$$\sigma(t) = \mathrm{Ln}\Big(1 - \frac{\xi(e^{2\pi it})}{p_*}\Big) \text{ for } t \in \mathbb{R}.$$

Let $f_0 \in \mathbb{C}$ be such that $e^{f_0} = -p_*$. Define $f(t) = f_0 + \sigma(t)$. Then f is continuous and

$$e^{f(t)} = e^{\bar{f}}.e^{\sigma(t)} = (-p_*).\Big(1 - \frac{\xi(e^{2\pi it})}{p_*}\Big) = \xi(e^{2\pi it}) - p_*.$$

Therefore $\nu(\xi,p_*) = \frac{1}{2\pi i}(f(1) - f(0)) = 0$, since f is clearly 1-periodic. So $\nu(\xi,p_*) = 0$, and then $\nu(\xi,p) = 0$ for all $p \in \mathrm{Out}(\xi)$.

Now choose a point $\hat{p} \in \mathrm{Ins}(\xi)$. Fix a point $\hat{q}$ of $|\xi|$ closest to $\hat{p}$. Let $\hat{r} = |\hat{p} - \hat{q}|$. Then the open disc $\hat{D} = \{z : |z - \hat{p}| < \hat{r}\}$ does not intersect $|\xi|$ and is entirely contained in $\mathrm{Ins}(\xi)$. After a translation and a rotation of the coordinate axes of the plane, we may assume that $\hat{p} = 0$ and $\hat{q} = \hat{r}$. Fix a $\hat{t} \in \mathbb{R}$ such that, $\xi(e^{2\pi i\hat{t}}) = \hat{r}$. Fix a number $\hat{\varepsilon}$ such that $0 < \hat{\varepsilon} < \frac{1}{2}$, and two continuous functions $\rho : [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}] \mapsto]0,\infty[$, $\varphi : [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}] \mapsto \mathbb{R}$, such that $\xi(e^{2\pi it}) = \rho(t)e^{2\pi i\varphi(t)}$

whenever $t \in [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}]$, and $\varphi(\hat{t}) = 0$. Then $\rho(t) \geq \hat{r}$ for all $\in [\hat{t} - \hat{\varepsilon}, \hat{t} + \hat{\varepsilon}]$, and $\rho(\hat{t}) = \hat{r}$.

Then, if $0 < \varepsilon \leq \hat{\varepsilon}$, it cannot be the case that $\varphi(t)$ vanishes for all $t \in [\hat{t} - \varepsilon, \hat{t} + \varepsilon]$. (Indeed, suppose that $\varphi(t) = 0$ whenever $t \in [\hat{t} - \varepsilon, \hat{t} + \varepsilon]$. Then $\xi(e^{2\pi i t}) = \rho(t) \in \mathbb{R}$ for $t \in [\hat{t} - \varepsilon, \hat{t} + \varepsilon]$. Therefore the function $[\hat{t} - \varepsilon, \hat{t} + \varepsilon] \ni t \mapsto \xi(e^{2\pi i t})$ would be real-valued and would have a local minimum at $\hat{t}$, contradicting the fact that ξ is one-to-one on some neighborhood of $\hat{t}$.) It follows that there exists a sequence $\{\tau_j\}_{j=1}^{\infty}$ in $\mathbb{R}$ such that $\tau_j \to \hat{t}$ as $j \to \infty$ and $\varphi(\tau_j) \neq 0$ for all j. If we let L_j be the closed half-line $\{re^{2\pi i \varphi(\tau_j)} : r \geq 0\}$, then $L_j \cap |\xi|$ is a compact set. Furthermore, $L_j \cap |\xi|$ is nonempty, because $\xi(\tau_j) = \rho(\tau_j) e^{2\pi i \varphi(\tau_j)} \in L_j \cap |\xi|$. So $L_j \cap |\xi|$ has a point q_j closest to 0. Write $q_j = \xi(e^{2\pi i t_j})$, $t_j \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}]$. Also, let $q_j = r_j e^{2\pi i \varphi(t_j)}$, so $\hat{r} \leq r_j \leq \rho(\tau_j)$.

We claim that $\lim_{j\to\infty} t_j = \hat{t}$. Indeed, if this was not so there would exist a subsequence $\{t_{j(\ell)}\}_{\ell=1}^{\infty}$ of $\{t_j\}_{j=1}^{\infty}$ that converges to a $\tilde{t} \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}] \backslash \{\hat{t}\}$. Let $\tilde{q} = \xi(e^{2\pi i \tilde{t}})$. Then $\tilde{q} = \lim_{\ell\to\infty} q_{j(\ell)}$. Since $e^{2\pi i \varphi(\tau_j)} \to 1$ as $j \to \infty$, and $e^{2\pi i \varphi(t_j)} = e^{2\pi i \varphi(\tau_j)}$ (because $q_j \in L_j$), we conclude that $e^{2\pi i \varphi(t_j)} \to 1$ as $j \to \infty$, so $\tilde{q}$ is a positive real number $\tilde{r}$, and $\tilde{r} = \lim_{\ell\to\infty} r_{j(\ell)}$. Since $\hat{r} \leq r_{j(\ell)} \leq \rho(\tau_{j(\ell)})$, we have $\tilde{r} = \hat{r}$, and then $\tilde{q} = \hat{q}$. Since $|\tilde{t} - \bar{t}| \leq \frac{1}{2}$, the fact that $\xi(e^{2\pi i \tilde{t}}) = \tilde{q} = \hat{q} = \xi(e^{2\pi i \bar{t}})$ implies $\tilde{t} = \bar{t}$, which is a contradiction, proving our claim.

By passing to a subsequence we may assume that $|t_j - \hat{t}| < \hat{\varepsilon}$ for all j, and either $t_j < \hat{t}$ for all j or $t_j > \hat{t}$ for all j. Then $q_j = \rho(t_j) e^{2\pi i \varphi(t_j)}$—so $r_j = \rho(t_j)$ and $\varphi(t_j) = \varphi(\tau_j)$—and $re^{2\pi i \varphi(t_j)} \notin |\xi|$ whenever $0 \leq r < r_j$. By passing to a subsequence, we assume that either $\varphi(t_j) < 0$ for all j or $\varphi(t_j) > 0$ for all j.

Under our conditions, one of the following possibilities occurs:

I. $t_j < \hat{t}$ and $\varphi(t_j) < 0$ for all j,

II. $t_j < \hat{t}$ and $\varphi(t_j) > 0$ for all j,

III. $t_j > \hat{t}$ and $\varphi(t_j) < 0$ for all j,

IV. $t_j > \hat{t}$ and $\varphi(t_j) > 0$ for all j.

We will discuss Case I, and prove that $\nu(\xi, 0) = 1$. (The other three cases are similar. In Cases II and III one can prove that $\nu(\xi, 0) = -1$, and in Case IV that $\nu(\xi, 0) = 1$.)

Write $\varphi_j = -\varphi(t_j)$, so $\varphi_j > 0$. Fix a positive real number δ such that $\delta < \hat{r} = \mathrm{dist}(0, |\xi|)$.

Let $\eta_j : [0,3] \mapsto \mathbb{C}$ be the curve given by

$$\eta_j(u) = \begin{cases} \Big((1-u)r_j + u\delta\Big)e^{-2\pi i\varphi_j} & \text{if} \quad 0 \le u \le 1\,, \\ \delta e^{-2\pi i((2-u)\varphi_j+1-u)} & \text{if} \quad 1 \le u \le 2\,, \\ (u-2)\hat{r} + (3-u)\delta & \text{if} \quad 2 \le u \le 3\,. \end{cases}$$

Then, as u goes from 0 to 3, $\eta_j(u)$ starts at the point $\xi(e^{2\pi i t_j}) = r_j e^{-2\pi i\varphi_j}$, moves along the straight-line segment M_j^1 from $r_j e^{-2\pi i\varphi_j}$ to $\delta e^{-2\pi i\varphi_j}$, then follows an arc M_j^2 of length $\delta.(2\pi - \varphi(t_j))$ of the circle $\{z : |z| = \delta\}$, clockwiset from $\delta e^{-2\pi i\varphi_j}$ to δ, and finally moves along the straight-line segment M_3 from δ to $\hat{r}$, ending at $\hat{r} = \xi(e^{2\pi i\hat{t}})$. It is clear that η_j is a one-to-one map. Furthermore, $\eta_j(u) \notin |\xi|$ whenever $0 < u < 3$.

Define $\hat{\zeta}_j : [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}] \mapsto \mathbb{C}$ by letting

$$\hat{\zeta}_j(t) = \begin{cases} \xi(e^{2\pi i t}) & \text{if} \quad t \in [\hat{t} - \frac{1}{2}, \hat{t} + \frac{1}{2}] \backslash [t_j, \hat{t}]\,, \\ \eta_j\Big(\frac{3(t-t_j)}{\hat{t}-t_j}\Big) & \text{if} \quad t \in [t_j, \hat{t}]\,. \end{cases}$$

Then $\hat{\zeta}_j$ is continuous, and $\hat{\zeta}_j(\hat{t} + \frac{1}{2}) = \hat{\zeta}_j(\hat{t} + \frac{1}{2})$, so we can extend $\hat{\zeta}_j$ to a 1-periodic map ζ_j from $\mathbb{R}$ to $\mathbb{C}$. Write $\zeta_j(t) = \omega_j(e^{2\pi i t})$ for $t \in \mathbb{R}$. Then ω_j is a simple loop. Furthermore, $|\omega_j| \subseteq \{z : |z| \le R\}$, so $p_* \in \mathrm{Out}(\omega_j)$, and then $\nu(\omega_j, p_*) = 0$.

Let $I_j = \{\xi(e^{2\pi i t}) : t_j < t < \hat{t}\}$. Let $K_j = |\xi| \backslash I_j$. Let $U_j = \mathbb{C} \backslash K_j$. Then K_j is a compact subset of $\mathbb{C}$ which is homeomorphic to the interval $[0,1]$. Therefore K_j does not disconnect U_j. Since p_* and 0 obviously belong to U_j, there exists a continuous path $\gamma_j : [0,1] \mapsto U_j$ such that $\gamma_j(0) = p_*$ and $\hat{\gamma}_j(1) = 0$. This path cannot be contained in $\mathbb{C} \backslash |\xi|$, because p_* and 0 belong to different connected components of $\mathbb{C} \backslash |\xi|$. Therefore the set $G_j = \{s \in [0,1] : \hat{\gamma}_j(s) \in |\xi|\}$ is nonempty. Since G_j is obviously compact, it has a smallest member s_j. Then $\gamma_j(s_j) \in |\xi|$ and $\gamma_j(s) \notin |\xi|$ whenever $0 \le s < s_j$. Since γ_j is contained in U_j, $\gamma_j(s_j)$ cannot belong to K_j, so $\gamma_j(s_j) \in I_j$. Therefore $\gamma_j(s_j) = \xi(e^{2\pi i v_j})$ for some $v_j \in I_j$. Pick an $\ell(j)$ such that $t_j < t_{\ell(j)} < \hat{t}$. Define a curve $\mu_j : [0,3] \mapsto \mathbb{C}$ by letting

$$\mu_j(u) = \begin{cases} \gamma_j(s_j u) & \text{if} \quad 0 \le u \in \le 1\,, \\ \xi(e^{2\pi i((2-u)v_j + (u-1)t_{\ell(j)})}) & \text{if} \quad 1 \le u \le 2\,, \\ (3-u)r_j e^{2\pi i\varphi(t_{\ell(j)})} & \text{if} \quad 2 \le u \le 3\,. \end{cases}$$

Then μ_j is continuous, $\mu_j(0) = p_*$, and $\mu_j(3) = 0$. Moreover, μ_j never goes through $|\omega_j|$. Therefore $\nu(\omega_j, 0) = \nu(\omega_j, p_*) = 0$.

We now relate $\nu(\omega_j,0)$ to $\nu(\xi,0)$. Let $f:\mathbb{R}\mapsto\mathbb{C}$ be a continuous map such that $e^{f(t)}=\xi(e^{2\pi it})$ for all $t\in\mathbb{R}$. Define $g_j:[0,3]\mapsto\mathbb{C}$ by

$$g_j(u)=\begin{cases}\ln\Big((1-u)r_j+u\delta\Big)-2\pi i\varphi_j & \text{if} \quad 0\le u\le 1\,,\\ \ln\delta-2\pi i((2-u)\varphi_j+1-u) & \text{if} \quad 1\le u\le 2\,,\\ \ln\Big((u-2)\hat{r}+(3-u)\delta\Big)+2\pi i & \text{if} \quad 2\le u\le 3\,.\end{cases}$$

Then g_j is continuous and $e^{g_j(u)}=\eta_j(u)$ for $0\le u\le 3$. Define

$$h_j(t)=\begin{cases} f(t) & \text{if} \quad \hat{t}-\frac{1}{2}\le t\le t_j,\,,\\ g_j\Big(\frac{3(t-t_j)}{\hat{t}-t_j}\Big)+f(t_j)-g_j(0) & \text{if} \quad t_j\le t\le\hat{t}\,,\\ f(t)+f(t_j)-g_j(0)+g_j(3)-f(\hat{t}) & \text{if} \quad \hat{t}\le t\le\hat{t}+\frac{1}{2}\,.\end{cases}$$

Then h_j is continuous and $e^{h_j(t)}=\zeta_j(t)$ whenever $t\in[\hat{t}-\frac{1}{2},\hat{t}+\frac{1}{2}]$. Therefore $h_j(\hat{t}+\frac{1}{2})-h_j(\hat{t}-\frac{1}{2})=2\pi i\,\nu(\omega_j,0)=0$. Then

$$f\Big(\hat{t}+\tfrac{1}{2}\Big)+f(t_j)-g_j(0)+g_j(3)-f(\hat{t})=h_j\Big(\hat{t}+\tfrac{1}{2}\Big)=h_j\Big(\hat{t}-\tfrac{1}{2}\Big)=f\Big(\hat{t}-\tfrac{1}{2}\Big)\,,$$

so that

$$f\Big(\hat{t}+\tfrac{1}{2}\Big)-f\Big(\hat{t}-\tfrac{1}{2}\Big)=f(t_j)-g_j(0)+g_j(3)-f(\hat{t})\,.$$

It follows from the definition of g_j that

$$g_j(0)=\ln r_j-2\pi i\varphi_j \qquad \text{and} \qquad g_j(3)=\ln\hat{r}+2\pi i\,.$$

Since $\lim_{j\to\infty}r_j=\hat{r}$ (because $\hat{r}\le r_j\le\rho(\tau_j)$, $\tau_j\to\hat{t}$, and $\rho(\hat{t})=\hat{r}$), and $\lim_{j\to\infty}\varphi_j=\lim_{j\to\infty}(-\varphi(\tau_j))=-\varphi(\hat{t})=0$, we can conclude that $\lim_{j\to\infty}(g_j(3)-g_j(0))=2\pi i$. Therefore

$$f\Big(\hat{t}+\tfrac{1}{2}\Big)-f\Big(\hat{t}-\tfrac{1}{2}\Big)=\lim_{j\to\infty}\Big(f(t_j)-g_j(0)+g_j(3)-f(\hat{t})\Big)=2\pi i\,,$$

since $\lim_{j\to\infty}t_j=\hat{t}=2\pi i$. Hence

$$\nu(\xi,0)=\frac{1}{2\pi i}\Big(f\Big(\hat{t}+\frac{1}{2}\Big)-f\Big(\hat{t}-\frac{1}{2}\Big)\Big)=1\,.$$

Therefore $\nu(\xi,z)=1$ for every $z\in\mathrm{Ins}(\xi)$, and our proof is complete. ☐

XV.17 Green's formula

If $\xi\in C^0(\mathbb{S}^1,\mathbb{C})$ is a loop in $\mathbb{C}$), the *winding number function* of ξ is the function

$$\mathbb{C}\backslash|\xi|\ni z\mapsto\nu_\xi(z)\stackrel{\text{def}}{=}\nu(\xi,z)\in\mathbb{Z}\,.$$

We want to prove that ν_ξ is a well defined member of $L^1(\mathbb{R}^2,\mathbb{R})$ if ξ is rectifiable. For this purpose, we have to know first of all that the set $|\xi|$ has measure zero, so the fact that ν_ξ is not well defined on $|\xi$ does not matter, and ν_ξ is well defined as a measurable function from $\mathbb{R}^2$ to $\mathbb{R}$ up to equality almost everywhere.

Proposition XV.27. *Let $\xi \in C^0(\mathbb{S}^1,\mathbb{C})$ be a rectifiable loop. Then the 2-dimensional Lebesgue measure of $|\xi|$ is equal to zero.*

PROOF. In view of Proposition vaca, we may assume that ξ is Lipschitz. Let C be a Lipschitz constant for the map $\mathbb{R} \ni t \mapsto \eta(t) = \xi(e^{2\pi it})$. For each $n,k \in \mathbb{N}$ such that $1 \le k \le n$, let $D_{k,n}$ be the disc with center $\eta\left(\frac{2k-1}{2n}\right)$ and radius $\frac{C}{2n}$. Then $|\xi| \subseteq \bigcup_{k=1}^n D_{k,n}$, so

$$\operatorname{meas}(|\xi|) \subseteq \sum_{k=1}^{n} \operatorname{meas}(D_{k,n}) = n\pi\frac{C^2}{4n^2} = \frac{\pi C^2}{4n}\,.$$

Since n is arbitrary, $\operatorname{meas}(|\xi|) = 0$, as stated. □

Proposition XV.28. *Let $C > 0$, and let $\xi \in C^0(\mathbb{S}^1,\mathbb{C})$ be a Lipschitz loop with Lipschitz constant C. Then*

1. *The winding number function ν_ξ is Lebesgue integrable.*
2. *There exists a sequence $\{\xi_j\}_{j=1}^\infty$ of polygonal loops in $\mathbb{C}$ with Lipschitz constant C such that, as $j \to \infty$,*
   - *2.a. $\xi_j \to \xi$ uniformly,*
   - *2.b. the functions ν_{ξ_j} converge to ν_ξ in $L^1(\mathbb{R}^2,\mathbb{R})$.*

PROOF. Let $\eta(t) = \xi(e^{2\pi it})$, so $\eta : \mathbb{R} \mapsto \mathbb{C}$ is a 1-periodic function. For $j = 1,2,\ldots$, we define $\eta_j : \mathbb{R} \mapsto \mathbb{C}$ by letting

$$\eta_j(t) = (k-2^jt)\xi\left(\frac{k-1}{2^j}\right) + (2^jt+1-k)\xi\left(\frac{k}{2^j}\right) \quad \text{if} \quad k \in \mathbb{Z}\,,\ \frac{k-1}{2^j} \le t \le \frac{k}{2^j}\,.$$

Then each η_j is 1-periodic, and linear on each interval $I_{j,k} \stackrel{\text{def}}{=} \left[\frac{k-1}{2^j}, \frac{k}{2^j}\right]$. If we define $\xi_j : \mathbb{S}^1 \mapsto \mathbb{C}$ by letting $\xi_j(e^{2\pi it}) = \eta_j(t)$, then ξ_j is a polygonal loop with 2^j sides which is Lipschitz with constant C. Furthermore, the ξ_j converge uniformly to ξ as $j \to \infty$.

For $k=0,1,\ldots,2^j$, let $A_k^j = \xi\left(\frac{k}{2^j}\right)$, so $A_j^0,\ldots,A_j^{2^j-1}$ are the vertices of ξ_j, $A_j^{2^j} = A_j^0$, and the segments $\sigma_j^k = \overline{A_j^{k-1}A_j^k}$, for $k=1,\ldots,2^j$, are the sides of ξ_j. Then ξ_{j+1} is obtained from ξ_j by replacing each side σ_j^k by the concatenation of the two segments σ_{j+1}^{2k-1} and σ_{j+1}^{2k}. That means that the polygonal ξ_{j+1} is obtained from ξ_j by adjoining to it the 2^j triangles $T_j^k = A_j^{k-1}A_{j+1}^{2k-1}A_j^k$, for $k=1,\ldots,2^j$. Therefore

$$\nu_{\xi_{j+1}} = \nu_{\xi_j} + \sum_{k=1}^{2^j} \nu_{T_j^k}\,.$$

For given j,k, The function $\nu_{T_j^k}$ has the value 1 or -1 inside T_j^k, and 0 outside. Therefore, $\nu_{T_j^k}$ is integrable, and its L^1 norm is equal to the area inside T_j^k, which is bounded by $C^2 2^{-2j-2}$. It follows that

$$\left\|\nu_{\xi_{j+1}} - \nu_{\xi_j}\right\|_{L^1} \le C^2 2^{-j-2}\,.$$

This shows that the sequence $\{\nu_{\xi_j}\}_{j=1}^\infty$ converges in $L^1(\mathbb{R}^2,\mathbb{R})$ to a limit $\hat{\nu}$.

On the other hand, if $z \notin |\xi|$, then $z \notin |\xi_j|$ if j is large enough, and $\lim_{j\to\infty} \nu_{\xi_j}(z) = \nu_\xi(z)$. In view of Proposition XV.27, $\nu_{\xi_j}(z) \to \nu_\xi(z)$ for almost all $z \in \mathbb{C}$. Therefore $\nu_\xi = \hat{\nu}$, and our conclusion is proved. □

Remark XV.29. The results of Propositions XV.27 and XV.28 are also true, with almost identical proof, if the curve ξ is Hölder continuous with exponent α, as long as $\alpha > \frac{1}{2}$. Indeed, to prove Proposition XV.27 in this case it suffices to let $D_{k,n}$ be the disc with center $\eta\left(\frac{2k-1}{2n}\right)$ and radius $\left(\frac{C}{2n}\right)^\alpha$. To prove Proposition XV.28, we use the same ξ_j s, and observe that the area inside T_j^k is bounded by $C^2 2^{-2\alpha j-2}$. Then

$$\left\|\nu_{\xi_{j+1}} - \nu_{\xi_j}\right\|_{L^1} \le C^2 2^{-(2\alpha-1)j-2}\,,$$

from which it still follows that $\{\nu_{\xi_j}\}_{j=1}^\infty$ converges in $L^1(\mathbb{R}^2,\mathbb{R})$. □

Theorem XV.30. *(Green's formula) Let $\xi \in C^0(\mathbb{S}^1,\mathbb{C})$ be a rectifiable loop. Then*

$$A(\xi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \nu_\xi(x+iy)\,dx\,dy\,. \qquad \text{(XV.40)}$$

PROOF. Both sides of (XV.40) are invariant under time reparametrization, so we may replace ξ by its arc-length reparametrization, and assume that ξ is a Lipschitz loop with Lipschitz constant 1.

By Proposition XV.28, there exists a sequence $\{\xi_j\}_{j=1}^{\infty}$ of polygonal loops that are Lipschitz with constant 1, converge uniformly to ξ, and are such that the functions v_{ξ_j} converge to v_ξ in L^1. Then Proposition vaca implies that $A(\xi_j) \to A(\xi)$. Therefore the validity of (XV.40) for all rectifiable loops ξ follows if (XV.40) is true for all polygonal loops.

To prove (XV.40) for polygonal loops, it suffices to observe that each polygonal loop can be obtained from a triangle by successive adjunctions of triangles, and both sides of (XV.40) are additive with respect to adjunction. Hence it suffices to prove (XV.40) for triangles. Finally, the proof of (XV.40) for triangles is an elementary direct calculation. □

Corollary XV.31. *Let $\xi \in C^0(\mathbb{S}^1, \mathbb{C})$ be a simple rectifiable loop. Let $\omega = 1$ if ξ is oriented cuonterclockwise, and $\omega = -1$ if ξ is oriented clockwise. Then the area of the region* $\mathrm{Ins}(\xi)$ *is equal to* $\omega A(\xi)$.

Chapter XVI

The value function

XVI.1 Introduction

The purpose of this note is to present some new results, together with a number of particularly simple and user-friendly versions of results obtained in recent years by the author and M. Malisoff, on the uniqueness of solutions of the Hamilton-Jacobi-Bellman equation (HJBE) for deterministic finite-dimensional optimal control problems under non-standard hypotheses. Our approach is completely control-theoretic and totally self-contained, using the systematic construction of special trajectories of various kinds, and not involving any PDE methods. We will *not* assume that the Lagrangian is positive, or that the dynamics is Lipschitz-continuous.

We will consider autonomous *Lagrangian optimization problems* involving a *state variable* x which takes values in an open subset Ω of $\mathbb{R}^n$, a *control variable* u taking values in a *control space* U, and a *target set* $\mathscr{T}$, which is a closed subset of the closure of Ω disjoint from Ω itself. The dynamics is given by an ordinary differential equation

$$\dot{x} = f(x,u)\,, \tag{XVI.1}$$

the cost functional to be minimized is

$$J = \int_{\tau_-(\xi)}^{\tau_+(\xi)} L(\xi(t),\eta(t))\,dt\,, \tag{XVI.2}$$

(where $\tau_-(\xi)$, $\tau_+(\xi)$ are, respectively, the initial and terminal times of the trajectory ξ), and the minimization is supposed to be, for each initial state $x \in \Omega$, over the set $\mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}$ of all pairs $\Xi = (\xi, \xi_0)$ such that

(i) Ξ consists of a trajectory ξ of (XVI.1) (i.e., a locally absolutely continuous function ξ that satisfies $\dot{\xi}(t) = f(\xi(t),\eta(t))$ for almost all t) corresponding to some U-valued control η, and a "running cost" function ξ_0 corresponding to ξ and η (i.e., a locally absolutely continuous function ξ_0 such that $\dot{\xi}_0(t) = L(\xi(t), \eta(t))$ for almost all t);

(ii) ξ starts at x, and "ends at the target" in a sense to be defined precisely later.

We will refer to a pair $\Xi = (\xi, \xi_0)$ for which (i) above holds as an "augmented trajectory" of our system, because it really is a trajectory of the "augmented control system"

$$\dot{x} = f(x,u), \qquad \dot{x}_0 = L(x,u) \tag{XVI.3}$$

obtained from (XVI.1) by "adding the cost as an extra variable" in a well known way. We will write $J(\Xi)$, rather than just J, for the left-hand side of (XVI.2), because it is easy to see that the natural arguments for our cost functional J are really augmented trajectories, since (i) one cannot just regard J as a functional of ξ only, because the integral of (XVI.2) involves the control η as well as ξ, but on the other hand (ii) the cost is completely determined once we know ξ and a running cost ξ_0, because in that case $L(\xi(t), \eta(t)) = \dot{\xi}_0(t)$.

The infimum $V(x)$ of the costs $J(\Xi)$ of all augmented trajectories $\Xi \in \mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}$ is the *value* of our problem at x. (If the set $\mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}$ is itself empty, then of course $V(x) = +\infty$.) The function $V : \Omega \mapsto \mathbb{R} \cup \{-\infty, +\infty\}$ is the *value function* of our problem.

Our goal is to prove uniqueness theorems, showing that a viscosity solution of the HJBE that satisfies an appropriate boundary condition is necessarily the value function. "Uniqueness" is to be understood as "uniqueness within a class defined by some additional properties," such as the class of all functions that are continuous and bounded below.

We will work with a class of systems which is sufficiently general to capture some interesting phenomena not commonly addressed in the literature, and at the same time restricted enough to make it possible to prove strong theorems. In particular, we will asume that the sets $F_{\hat{\Sigma}}(x,U) = \{(f(x,u), L(x,u)) : u \in U\}$ are closed and convex, but will not require them to be compact, and will instead impose a "local coerciveness" condition, according to which, locally, an inequality of the form $\|f(x,u)\|^r \leq L(x,u) + C$, with $C > 0$ and $r > 1$, holds uniformly with respect to u. We will also require $f(x,u)$ and $L(x,u)$ to be continuous with respect to x, with the continuity being uniform with respect to u locally in x.

On the other hand, we will most definitely *not* require that the dynamics $f(x,u)$ be Lipschitz-continuous with respect to x, since one of the main purposes of this work is to clarify the exact role of the Lipschitz-continuity assumptions often made in the viscosity literature. The answer we will propose is as follows:

(a) Without any Lipschitz-continuity hypotheses, one can prove, for continuous viscosity solutions V of the HJBE, an *existence theorem for trajectories*, asserting that, starting at every point of x of Ω, there is a maximally defined "augmented trajectory of steepest descent," that is, a maximally defined pair $\Xi = (\xi, \xi_0)$ defined on a interval I such that $0 = \min I$, and having the property that the inequality

$$V(\xi(t)) + \xi_0(t) \geq V(\xi(s)) + \xi_0(s) \tag{XVI.4}$$

holds whenever $s,t \in I$ and $s \leq t$. (This is Theorem XVI.19 below.)

(b) As a trivial corollary of the existence of steepest descent trajectories (applied to $-V$ and $-L$), we get the existence of "DPI trajectories" (where "DPI" is an acronym for "Dynamic Programming Inequality"), i.e., augmented trajectories $\Xi = (\xi, \xi_0)$ along which the inequality

$$V(\xi(t)) + \xi_0(t) \leq V(\xi(s)) + \xi_0(s) \tag{XVI.5}$$

(that is, the exact opposite of (XVI.4)) is satisfied.

(c) The existence result of (b) says that for every "sufficiently nice" (e.g., piecewise constant) control η *there exists a trajectory of* η *with the given initial condition along which the DPI holds.* This is almost, but not quite, what is needed to prove that V is bounded above by the value function.

(d) The gap between the existence result for DPI trajectories and what would actually be needed to prove that V is bounded above by the value function is that to achieve the latter goal one needs the DPI to hold for *all* trajectories, and it is not enough to have just one DPI trajectory for every intial condition and every control.

(e) The gap described in (d) clearly does not exist when there is uniqueness of trajectories for every given control and initial condition.

(f) In particular, the gap does not exist when, for every admissible control η, the corresponding time-varying vector field $(x,t) \mapsto f(x, \eta(t))$ satisfies a Lipschitz-Carathéodory condition that guarantees uniqueness of trajectories.

(g) Naturally, the Lipschitz-Carathéodory condition can be replaced by weaker conditions that guarantee uniqueness, such as the requirement that a bound

$$\langle f(x,\eta(t)) - f(y,\eta(t)), x-y\rangle \leq k(t)\|x-y\|^2\,, \qquad \text{(XVI.6)}$$

with k integrable, hold locally.

(h) Even more generally, the only property that really matters is that, if we pick a sequence $\{\eta_j\}_{j=1}$ of piecewise constant controls such that our augmented trajectory Ξ can be approximated by augmented trajectories Ξ_j corresponding to the η_j—with, say, the same initial condition—then the Ξ_j converge to Ξ uniformly *no matter how the* Ξ_j *are chosen*. We call such trajectories "uniquely limiting," and use this concept in the statement of our main theorem.

The following important issues will not be discussed here:

(1) Whether the value function itself satisfies the conditions of our main theorem, i.e., whether it is a continuous viscosity solution of the HJBE and whether it is bounded below.

(2) What happens when the sets $F(x,U)$ are not closed and convex. (This would require considering relaxed controls, and using trajectories for which the steepest descent property holds approximately rather than exactly. It turns out to be possible to extend our results under fairly general conditions, provided our system has appropriate local controllability properties.)

Remark XVI.1. *The approach followed here owes a great deal to the book [125] by A.I. Subbotin. We point out, however, that Subbotin considers viscosity solutions of PDEs of the form* $F(x,u(x),Du(x)) = 0$*, where the Hamiltonian* $F(x,u,p)$ *is required to be globally Lipschitz with respect to the momentum variable* p *(cf. Equation (2.2) in page 9 of [125]). A somewhat weaker hypothesis is also considered later, in which the Lipschitz requirement is replaced by the condition that for any* $\Lambda > 0$ *there exists a positive constant* $\mu(\Lambda)$ *such that the estimate*

$$|F(x,z,s) - F(x,z,p)| \leq \mu(\Lambda)(1+\|s-p\|)$$

holds for all $s \in \mathbb{R}^n$ *such that* $\|s\| \leq \Lambda$ *and all* $p \in \mathbb{R}^n$ *(cf. page 37 of [125]). In particular, even with the weakened requirements, these hypotheses are not sufficient to cover, for example, coercive problems of the kind discussed here, such as linear quadratic optimal control. (For example, for the optimal control problem of minimizing the integral* $\frac{1}{2}\int(x^2+u^2)$*, with a scalar state* x *and a scalar control* u*,* *and dynamics $\dot{x} = u$, the function F is given by $F(x,z,p) = \frac{1}{2}(p^2 - x^2)$. Therefore $F(x,z,s) - F(x,z,p) = \frac{1}{2}(s^2 - p^2) = \frac{1}{2}(s+p)(s-p)$, and for the desired estimate to be satisfied the sum $s+p$ would have to be bounded by a constant $\mu(\Lambda)$ for all $s \in \mathbb{R}^n$ such that $\|s\| \leq \Lambda$ and all $p \in \mathbb{R}^n$, and such a bound obviously does not hold.* □

XVI.2 The main theorem

XVI.2.1 Systems and augmented systems.

If n is a positive integer, an *n-dimensional control system* is a triple $\Sigma = (\Omega, U, f)$ such that Ω (the *state space* of Σ) is an open subset of $\mathbb{R}^n$, U (the *control space* of Σ) is a nonempty set, and f (the *dynamics* of Σ) is a map $\Omega \times U \ni (x,u) \mapsto f(x,u) \in \mathbb{R}^n$.

An *n-dimensional augmented control system* is a 4-tuple $\hat{\Sigma} = (\Omega, U, f, L)$ such that $\Sigma = (\Omega, U, f, L)$ is an n-dimensional control system and L (the *Lagrangian* of $\hat{\Sigma}$) is a map $\Omega \times U \ni (x,u) \mapsto L(x,u) \in \mathbb{R}$. (In that case, the state space, control space, and dynamics of Σ are also called the state space, control space, and dynamics of $\hat{\Sigma}$.)

An augmented control system $\hat{\Sigma} = (\Omega, U, f, L)$ is *continuous* if the maps $\Omega \ni x \mapsto f(x,u) \in \mathbb{R}^n$ and $\Omega \ni x \mapsto L(x,u) \in \mathbb{R}$ are continuous for each fixed u. We call $\hat{\Sigma}$ *uniformly continuous* on a subset S of Ω if there exists a function $\omega :]0,\infty[\mapsto [0,\infty]$ such that $\lim_{s\downarrow 0} \omega(s) = 0$, having the property that

$$\|f(x,u) - f(y,u)\| + |L(x,u) - L(y,u)| \leq \omega(\|x-y\|)$$

whenever $x, y \in S$ and $u \in U$. We call $\hat{\Sigma}$ *locally uniformly continuous* if it is uniformly continuous on every compact subset of Ω, and *globally uniformly continuous* if it is uniformly continuous on Ω. We say that $\hat{\Sigma}$ is *Lipschitz continuous* if the maps $\Omega \ni x \mapsto f(x,u) \in \mathbb{R}^n$ and $\Omega \ni x \mapsto L(x,u) \in \mathbb{R}$ are Lipschitz continuous for each fixed u. We call $\hat{\Sigma}$ *uniformly Lipschitz continuous* on a subset S of Ω if there exists a positive constant C such that

$$\|f(x,u) - f(y,u)\| + |L(x,u) - L(y,u)| \leq C\|x-y\|$$

whenever $x, y \in S$ and $u \in U$. We call $\hat{\Sigma}$ *locally uniformly Lipschitz continuous* if it is uniformly Lipschitz continuous on every compact subset of Ω, and *globally uniformly Lipschitz continuous* if it is uniformly Lipschitz continuous on Ω.

Remark XVI.2. *Naturally, the concepts of continuity, uniform continuity, Lipschitz continuity, and uniform Lipschitz continuity, also make sense for a control* *system $\Sigma = (\Omega, U, f)$, by taking the same definitions given above and omitting the parts that refer to L. In order to avoid having to make a similar remark for other concepts to be introduced in the future, we adopt the convention that any concept X that we define for an augmented control system $\hat{\Sigma} = (\Omega, U, f, L)$ is automatically understood to apply to a control system $\Sigma = (\Omega, U, f)$, in the sense that "X of Σ" means "X of the augmented system $\hat{\Sigma} = (\Omega, U, f, 0)$."* $\square$

The augmented control system $\hat{\Sigma} = (\Omega, U, f, L)$ is *coercive* on a subset S of Ω if there exist real constants r, A, C, such that $A > 0$, $C > 0$, $r > 1$, and

$$\|f(x,u)\|^r \leq AL(x,u) + C \text{ for all } x \in S\,,\ u \in U\,. \tag{XVI.7}$$

We call $\hat{\Sigma}$ *locally coercive* if it is coercive on every compact subset of Ω, and *globally coercive* if it is coercive on Ω.

Remark XVI.3. *If $\hat{\Sigma}$ is coercive on a set S, then it is always possible to choose C, r, such that $C > 0$, $r > 1$, and (XVI.7) holds with $A = 1$. Indeed, let A, C, r be such that $A > 0$, $C > 0$, $r > 1$, and (XVI.7) holds. Pick ρ such that $1 < \rho < r$, and let $s = \frac{\rho}{r-\rho}$. Let $x \in \Omega$, $u \in U$. Then, if $\|f(x,u)\|^{r-\rho} > A$, we have*

$$\begin{aligned} \|f(x,u)\|^\rho &= \frac{\|f(x,u)\|^r}{\|f(x,u)\|^{r-\rho}} \\ &\leq A^{-1}(AL(x,u)+C) \\ &= L(x,u) + \frac{C}{A} \\ &\leq L(x,u) + \frac{C}{A} + A^s\,, \end{aligned}$$

while on the other hand, if $\|f(x,u)\|^{r-\rho} \leq A$, we find that

$$\begin{aligned} \|f(x,u)\|^\rho &= (\|f(x,u)\|^{r-\rho})^s \\ &\leq A^s \\ &\leq A^{-1}\|f(x,u)\|^r + A^s \\ &\leq A^{-1}(AL(x,u)+C) + A^s \\ &= L(x,u) + \frac{C}{A} + A^s. \end{aligned}$$

Hence (XVI.7) holds if r, C, A are replaced by ρ, $\frac{C}{A} + A^s$, and 1. $\square$

For an augmented control system $\hat{\Sigma} = (\Omega, U, f, L)$, we define a map $F_{\hat{\Sigma}} : \Omega \times U \mapsto \mathbb{R}^{n+1}$ (called the *augmented dynamics* of $\hat{\Sigma}$) by letting $F_{\hat{\Sigma}}(x,u) = (f(x,u), L(x,u))$

for $x \in \Omega$, $u \in U$. We say that $\hat{\Sigma}$ *satisfies the convexity and upper semicontinuity condition* if, for each $x \in \Omega$,

$$F_{\hat{\Sigma}}(x,U) = \bigcap_{\varepsilon>0} \overline{\mathrm{co}}\Big(\bigcup\{F_{\hat{\Sigma}}(x',U) : \|x'-x\| \le \varepsilon\}\Big)\,, \tag{XVI.8}$$

where "$\overline{\mathrm{co}}$" stands for "closed convex hull," and $F_{\hat{\Sigma}}(x,U) \stackrel{\mathrm{def}}{=} \{(f(x,u),L(x,u)) : u \in U\}$.

Remark XVI.4. *If $\hat{\Sigma} = (\Omega,U,f,L)$ is locally uniformly continuous, then $\hat{\Sigma}$ satisfies the convexity and upper semicontinuity condition if and only if the set $F_{\hat{\Sigma}}(x,U)$ is closed and convex for every $x \in \Omega$. Indeed, the "only if" assertion is trivial, since the right-hand side of (XVI.8) is obviously closed and convex. To prove the "if" part, we fix $x \in \Omega$ and assume that $F_{\hat{\Sigma}}(x,U)$ is closed and convex. We choose a δ such that $\delta > 0$ and $B_\delta(x) \stackrel{\mathrm{def}}{=} \{x' \in \mathbb{R}^n : \|x'-x\| \le \delta\} \subseteq \Omega$ and a function $\omega :]0,\infty[\mapsto [0,\infty]$ such that $\lim_{s\downarrow 0}\omega(s) = 0$ and $\|F_{\hat{\Sigma}}(y,u) - F_{\hat{\Sigma}}(z,u)\| \le \omega(\|y-z\|)$ whenever $u \in U$ and $y,z \in B_\delta(x)$. We then pick*

$$v \in \bigcap_{\varepsilon>0} \overline{\mathrm{co}}\Big(\bigcup\{F_{\hat{\Sigma}}(x',U) : \|x'-x\| \le \varepsilon\}\Big)$$

and prove that $v \in F_{\hat{\Sigma}}(x,U)$. We let $\varepsilon_k = 2^{-k}$, and use the fact that $v \in \overline{\mathrm{co}}\Big(\bigcup\{F_{\hat{\Sigma}}({}',U) : \|x'-x\| \le \varepsilon_k\}\Big)$ to find, for each sufficiently large k, a member v_k of the convex hull of $\bigcup\{F_{\hat{\Sigma}}(x',U) : \|x'-x\| \le \varepsilon_k\}$ such that $\|v-v_k\| < \varepsilon_k$. We then write $v_k = \sum_{j=0}^{n+1}\alpha_{k,j}F_{\hat{\Sigma}}(x_{k,j},u_{k,j})$, where the $x_{k,j}$ belong to Ω and satisfy $\|x_{k,j}-x\| \le \varepsilon_k$, the $u_{k,j}$ belong to U, and the $\alpha_{k,j}$ are nonnegative and satisfy $\sum_{j=0}^{n+1}\alpha_{k,j} = 1$. Then $\|F_{\hat{\Sigma}}(x,u_{k,j}) - F_{\hat{\Sigma}}(x,u_{k,j})\| \le \omega(\varepsilon_k)$. Therefore, if we let $w_k = \sum_{j=0}^{n+1}\alpha_{k,j}F_{\hat{\Sigma}}(x,u_{k,j})$, we have $\|w_k - v_k\| \le \omega(\varepsilon_k)$. Hence $w_k \to v$ as $k \to \infty$. Since $F_{\hat{\Sigma}}(x,U)$ is convex, the w_k belong to $F_{\hat{\Sigma}}(x,U)$. Since $F_{\hat{\Sigma}}(x,U)$ is closed, $v \in F_{\hat{\Sigma}}(x,U)$, and the proof is complete. □

A *target* for an augmented control system $\hat{\Sigma} = (\Omega,U,f,L)$ is a closed subset $\mathscr{T}$ of $\mathbb{R}^n$ such that $\mathscr{T} \subseteq \mathrm{Closure}\,\Omega$ and $\mathscr{T} \cap \Omega = \emptyset$.

XVI.2.2 Trajectories

A *trajectory* of $\hat{\Sigma} = (\Omega,U,f,L)$ is a locally absolutely continuous curve

$$I \ni t \mapsto \xi(t) \in \Omega\,, \tag{XVI.9}$$

defined on a nonempty subinterval I of $\mathbb{R}$, having the property that $\dot{\xi}(t) \in f(\xi(t),U)$ for almost every $t \in I$. An *augmented trajectory* of $\hat{\Sigma}$ is a locally absolutely continuous curve

$$I \ni t \mapsto \Xi(t) = (\xi(t), \xi_0(t)) \in \Omega \times \mathbb{R}, \tag{XVI.10}$$

defined on a subinterval I of $\mathbb{R}$, having the property that $\dot{\Xi}(t) \in F_{\hat{\Sigma}}(\xi(t),U)$ for almost every $t \in I$.

The *initial time*, or *starting time* of a trajectory ξ (resp. an augmented trajectory $\Xi = (\xi, \xi_0)$) with domain I is the time $\tau_-(\xi) \stackrel{\text{def}}{=} \min I$ (resp. $\tau_-(\Xi) \stackrel{\text{def}}{=} \min I$)), if the minimum exists, i.e., if I is bounded below and its infimum belongs to I. If the initial time of ξ (resp. Ξ) exists, then (i) the point $x_-(\xi) \stackrel{\text{def}}{=} \xi(\tau_-(\xi))$ (resp. $x_-(\Xi) \stackrel{\text{def}}{=} \xi(\tau_-(\Xi))$) is the *starting point*, or *initial point*, of ξ (resp. Ξ), and (ii) the ordered pair $\partial_-(\xi) \stackrel{\text{def}}{=} (\tau_-(\xi), x_-(\xi))$ (resp. $\partial_-(\Xi) \stackrel{\text{def}}{=} (\tau_-(\Xi), x_-(\Xi))$) is the *initial condition* of ξ (resp. Ξ). If $\partial_-(\xi) = (t,x)$ (resp. $\partial_-(\Xi) = (t,x)$), we say that ξ (resp. Ξ) *starts at x at time t*.

If $\mathscr{T}$ is a target for $\hat{\Sigma} = (\Omega, U, f, L)$, then a trajectory ξ or augmented trajectory $\Xi = (\xi, \xi_0)$ with domain I *ends at* $\mathscr{T}$ if the limit

$$\xi(\uparrow) \stackrel{\text{def}}{=} \lim_{t \uparrow \sup I} \xi(t)$$

exists and belongs to $\mathscr{T}$.

For each $x \in \Omega$, we let $\mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}$ be the set of all augmented trajectories $\Xi = (\xi, \xi_0)$ of $\hat{\Sigma}$ such that

(i) $\partial_-(\Xi) = (0,x)$,

(ii) $\xi_0(0) = 0$,

(iii) Ξ ends at the target,

(iv) the limit $\xi_0(\uparrow) \stackrel{\text{def}}{=} \lim_{t \uparrow \sup \operatorname{domain} \Xi} \xi_0(t)$ exists.

If $\Xi = (\xi, \xi_0) \in \mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}$ then the *cost* of Ξ is the number

$$J(\Xi) \stackrel{\text{def}}{=} \xi_0(\uparrow)\,. \tag{XVI.11}$$

The *value function* of the optimal control problem defined by $\hat{\Sigma}$ and the target $\mathscr{T}$ is the function $\mathscr{V}^{\hat{\Sigma}}_{\mathscr{T}} : \Omega \cup \mathscr{T} \mapsto \mathbb{R} \cup \{-\infty, +\infty\}$ given by

$$\mathscr{V}^{\hat{\Sigma}}_{\mathscr{T}}(x) = \begin{cases} \inf\{J(\Xi) : \Xi \in \mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}\} & \text{if} \quad x \in \Omega \\ 0 & \text{if} \quad x \in \mathscr{T}\,. \end{cases}$$

If $V : \Omega \mapsto \mathbb{R}$ is a function, then an *augmented trajectory of* $\hat{\Sigma}$ *of steepest descent with respect to* V is an augmented trajectory $\Xi = (\xi, \xi_0)$ of $\hat{\Sigma}$ such that

$$\begin{aligned} \xi_0(s) + V(\xi(s)) &\geq \xi_0(t) + V(\xi(t)) \\ &\text{whenever } s,t \in \text{domain}\,\Xi \text{ and } s \leq t\,. \end{aligned} \tag{XVI.12}$$

We use $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V}$ to denote the set of all augmented trajectories of $\hat{\Sigma}$ of steepest descent with respect to V, and $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ to denote the set of all $\Xi \in \mathscr{S}\mathscr{D}_{\hat{\Sigma},V}$ such that $\partial_-(\Xi) = (0,x)$.

If $x \in \Omega$, a *maximal augmented trajectory of* $\hat{\Sigma}$ *from* x *of steepest descent with respect to* V is a $\Xi \in \mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ that cannot be extended to a $\tilde{\Xi} \in \mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ defined on an interval which is strictly larger that the domain of Ξ. We use $\mathscr{M}\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ to denote the set of all maximal augmented trajectories of $\hat{\Sigma}$ from x of steepest descent with respect to V.

Remark XVI.5. *We have set up our definitions in such a way that the notion of "maximal augmented trajectory of steepest descent" always means "maximal within the class* $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ *of augmented trajectories* Ξ *of steepest descent with initial condition* $\partial_-(\Xi) = (0,x)$*, for a fixed* $x \in \Omega$*." (In other words, there is no such thing as a "maximal augmented trajectory;" there are only "maximal augmented trajectories* from a given point x*.") Since all the members of* $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ *start at time* 0 *at the point* x*, the only way that a* $\Xi \in \mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$*, defined on an interval* I *(that necessarily starts at time* 0*), could fail to belong to* $\mathscr{M}\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$*, would be for* Ξ *to be "extendable to the right," that is, such that there exists an extension* $\tilde{\Xi} \in \mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ *which is defined on an interval* $\tilde{I}$ *that also starts at* 0 *and is strictly larger than* I*. Naturally, it may also happen that* Ξ *can be extended on the left, to an augmented trajectory* $\hat{\Xi}$ *of steepest descent defined on an interval* $\hat{I} =]\text{-}\varepsilon, 0] \cup I$ *for some positive* ε*. But such an extension will of course no longer start at time* 0*, so its existence does not affect the possibility that* Ξ *might belong to* $\mathscr{M}\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$.
□

Definition XVI.6. . *If* Ω *is an open subset of* $\mathbb{R}^n$*, and* $\xi : I \mapsto \Omega$ *is a curve, we say that* ξ *is* right-unbounded *if*

(i) the interval I *is open on the right (that is, if* $\tau = \sup I$*, then either (a)* $\tau = +\infty$ *or (b)* τ *is finite and does not belong to* I*),*

and

(ii) if τ is finite, then for every compact subset K of Ω there exists a τ_K such that $0 \le \tau_K < \tau$ and $\xi(t) \notin K$ whenever $\tau_K < t < \tau$.

(Equivalently, condition (ii) asserts that $\lim_{t\uparrow\tau} \xi(t) = \infty_\Omega$*, where ∞_Ω is the point at infinity of the one-point compactification of Ω.)* $\square$

The following observation is completely trivial given our definitions, but we state it explicitly as a separate result for future reference. We emphasize that this trivial result is valid *under no technical hypotheses whatsoever on $\hat{\Sigma}$ or V*. The reader is warned that the result is *not* a true "existence theorem" for trajectories of steepest descent, even though at first sight it may appear to be, because the member Ξ of $\mathscr{MSD}_{\hat{\Sigma},V,x}$ whose existence it asserts could very well turn out to be the trivial trajectory Ξ_x^{triv}, where Ξ_x^{triv} is the map $\Xi : \{0\} \mapsto \Omega \times \mathbb{R}$ such that $\Xi(0) = (x,0)$. The true "existence theorem," yielding the existence of a *nontrivial* maximal trajectory of steepest descent and, in fact, asserting the stronger conclusion that *every* maximal trajectory of steepest descent is right-unbounded in the sense of Definition XVI.6. This will be proved later (cf. Theorem (XVI.19) and, naturally, will depend on our technical hypotheses on $\hat{\Sigma}$ and V.

Proposition XVI.7. *If $\hat{\Sigma} = (\Omega, U, f, L)$ is an augmented control system, $V : \Omega \mapsto \mathbb{R}$ is a function, and $x \in \Omega$, then the set $\mathscr{MSD}_{\hat{\Sigma},V,x}$ of maximal augmented trajectories of $\hat{\Sigma}$ from x of steepest descent with respect to V is nonempty.*

Proof. Fix x. Let $\mathscr{Z}$ be the set of all pairs (I,Ξ) such that I is a subinterval of $[0,\infty[$, $0 \in I$, and $\Xi = (\xi,\xi_0) : I \mapsto \Omega \times \mathbb{R}$ is an augmented trajectory of $\hat{\Sigma}$ which is of steepest descent with respect to V and such that $\xi(0) = x$. We partially order $\mathscr{Z}$ by stipulating that, if $(I_i,\Xi_i) \in \mathscr{Z}$ for $i = 1,2$, then $(I_1,\Xi_1) \preceq (I_2,\Xi_2)$ iff $I_1 \subseteq I_2$ and Ξ_1 is the restriction of Ξ_2 to I_1.

It is clear that $\mathscr{Z} \neq \emptyset$, because the pair $(\{0\},\Xi_x^{triv})$—where Ξ_x^{triv} is the map defined above, in the paragraph preceding the statement of our proposition—belongs to $\mathscr{Z}$. If Z is a totally ordered subset of $\mathscr{Z}$, we show that Z has an upper bound (I_*,Ξ_*) in $\mathscr{Z}$. This conclusion is trivial if $Z = \emptyset$, for in that case we can take $(I_*,\Xi_*) = (\{0\},\Xi_x^{triv})$. Assume that $Z \neq \emptyset$. Let I_* be the union of the intervals I for all the members (I,Ξ) of Z. Then I_* is obviously a subinterval of $[0,\infty[$, and $0 \in I$. If $t \in I_*$, we define $\Xi_*(t) = \Xi(t)$, where (I,Ξ) is any member of Z such that $t \in I$. Write $\Xi_* = (\xi_*,\xi_{0,*})$. Then Ξ_* is obviously well defined, and is an augmented trajectory of $\hat{\Sigma}$ such that $\xi_*(0) = x$. If $t \in I_*$, then we can pick $(I,\Xi) \in Z$ such that $t \in I$, and then $\Xi_*(s) = \Xi(s)$ for all $s \in I$, and in particular for all $s \in [0,t]$, since $[0,t] \subseteq I$. Furthemore, if we write $\Xi = (\xi,\xi_0)$, then the fact that $\Xi \in \mathscr{SD}_{\hat{\Sigma},V}$ implies that $V(x) \ge \xi_0(t) + V(\xi(t)) = \xi_{0,*}(t) + V(\xi_*(t))$. Since t is an arbitrary member of I_*, we have shown that $\Xi_* \in \mathscr{S}\mathscr{D}_{\hat{\Sigma},V}$. Therefore $(I_*, \xi_*) \in \mathscr{Z}$. Furthermore, it is clear that (I_*, Ξ_*) is an upper bound for Z. So we have shown that every totally ordered subset of $\mathscr{Z}$ has an upper bound, and that $\mathscr{Z} \neq \emptyset$. Therefore Zorn's Lemma implies that $\mathscr{Z}$ has a maximal element (I_*, Ξ_*). Clearly, such a maximal element is a member of $\mathscr{M}\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$, and our proof is complete. □

An *augmented arc* is an augmented trajectory whose domain is a compact interval. If $\Xi = (\xi, \xi_0)$ is an augmented arc with domain $[a,b]$, then an *improvement* of Ξ is an augmented arc $\Xi' = (\xi', \xi_0')$, with domain $[a', b']$, such that $\xi'(a') = \xi(a)$, $\xi'(b') = \xi(b)$, and $\xi_0'(b') - \xi_0'(a') \leq \xi_0(b) - \xi_0(a)$.

If $\hat{\Sigma} = (\Omega, U, f, L)$ is an augmented control system, then an augmented arc $\Xi = (\xi, \xi_0)$ of $\hat{\Sigma}$ with domain $[a,b]$ is *uniquely limiting* if there exists a sequence $\{\eta_j\}_{j=1}^{\infty}$ of piecewise constant functions $\eta_j : [a,b] \mapsto U$ such that

(*) *if* $\{\Xi_j\}_{j=1}^{\infty}$ *is an arbitrary sequence of maximally defined augmented trajectories of* $\hat{\Sigma}$ *such that* $a \in \mathrm{domain}(\Xi_j)$ *and* $\Xi_j(a) = \Xi(a)$ *for every* j*, then* $[a,b] \subseteq \mathrm{domain}(\Xi_j)$ *if* j *is large enough, and* $\Xi_j \to \Xi$ *uniformly on* $[a,b]$ *as* $j \to \infty$.

Example XVI.8. *Suppose* $\Xi = (\xi, \xi_0)$ *is an augmented arc of* $\hat{\Sigma}$ *with domain* $[a,b]$ *such that*

(#) there exist a positive number δ, a function $\eta : [a,b] \mapsto U$, and a function $\varphi : [a,b] \mapsto [0, \infty]$, such that

(i) $\dot{\Xi}(t) = \Big(f(x, \eta(t)), L(x, \eta(t))\Big)$ for almost every $t \in [a,b]$,

(ii) the map $t \mapsto \Big(f(x, \eta(t)), L(x, \eta(t))\Big)$, on the compact set
$$I_x \stackrel{\mathrm{def}}{=} \{t : a \leq t \leq b \wedge \|x - \xi(t)\| \leq \delta\},$$
is measurable for each $x \in \Omega$,

(iii) the map $x \mapsto \Big(f(x, \eta(t)), L(x, \eta(t))\Big)$, on the compact set
$$I^t \stackrel{\mathrm{def}}{=} \{x \in \Omega : \|x - \xi(t)\| \leq \delta\},$$
is measurable for each $t \in [a,b]$,

(iv) φ is integrable,

(v) the inequality
$$\Big\langle f(x, \eta(t)) - f(x', \eta(t)), x - x' \Big\rangle \leq \varphi(t) \|x - x'\|^2$$

holds whenever $t \in [a,b]$, $\|x - \xi(t)\| \le \delta$, and $\|x' - \xi(t)\| \le \delta$,

(vi) the inequality $|L(x,\eta(t))| \le \varphi(t)$ holds whenever $t \in [a,b]$ and $\|x - \xi(t)\| \le \delta$.

Then Ξ is uniquely limiting. The proof is essentially as follows. By dividing $[a,b]$ into small intervals, we can assume that there is a fixed compact ball B such that $B \subseteq \Omega$, ξ is entirely contained in the interior of B, and the bound of (v) *holds whenever $t \in [a,b]$ and $x, x' \in B$. We then write $F^u(x) = (f(x,u,L(x,u))$ for each $x \in B$, $u \in U$, and observe that the set $\mathscr{F} = \{F^{\eta(t)} : t \in [a,b]\}$ is a subset of the separable Banach space $C^0(B,\mathbb{R}^{n+1})$. Then $[a,b] \ni t \mapsto F^{\eta(t)}$ is an L^1 $C^0(B,\mathbb{R}^{n+1})$-valued map. Therefore one can approximate this map in L^1 by piecewise constant $\mathscr{F}$-valued maps. In other words, one can find a sequence $\{\eta_j\}_{j=1}^{\infty}$ of piecewise constant $\mathscr{G}$-valued functions (where $\mathscr{G} = \{\eta(t) : t \in [a,b]\}$), and integrable functions $k_j : [a,b] \mapsto [0,+\infty]$, such that*

$$\|f(x,\eta(t)) - f(x,\eta_j(t))\| + |L(x,\eta(t)) - L(x,\eta_j(t))| \le k_j(t)$$

whenever $a \le t \le b$, $x \in B$, and $j \in \mathbb{N}$, and $\lim_{j\to\infty} \int_a^b k_j(t)\,dt = 0$. Then, if $a \le c \le d \le b$, and $\zeta : [c,d] \mapsto B$, $\theta : [c,d] \mapsto B$, are trajectories of η, η_j, respectively, Gronwall's inequality yields the bound

$$\|\zeta(t) - \theta(t)\| \le e^{\int_a^b \varphi(s)ds}\Big(\|k_j\|_{L^1} + \|\zeta(c) - \theta(c)\|\Big)$$

if $t \in [c,d]$. If we apply this with $c = a$ and $\zeta = \xi$, letting θ be any trajectory ξ_j of η_j starting at $\xi(a)$ at time a, and defined on some subinterval $[a,d]$ of $[a,b]$, we see that, as long as j is large enough, the maximum of the $\|\xi(t) - \xi_j(t)\|$ is bounded by a small constant. This guarantees that ξ_j actually exists on the whole interval $[a,b]$, and then the Gronwall bound implies that $\xi_j \to \xi$ uniformly as $j \to \infty$. Then the integrals $\int_a^t L(\xi_j(s),\eta_j(s))\,ds$ differ from $\int_a^t L(\xi_j(s),\eta(s))\,ds$ by less than $\|k_j\|_{L^1}$, in view of the bound

$$|L(x,\eta(t)) - L(x,\eta_j(t))| \le k_j(t)\,,$$

and $\int_a^t L(\xi_j(s),\eta(s))\,ds \to \int_a^t L(\xi(s),\eta(s))\,ds$ as $j \to \infty$, because $L(\xi_j(s),\eta(s)) \to L(\xi(s),\eta(s))$ for each s, and $|L(\xi_j(s),\eta(s))| \le \varphi(s)$. $\square$

An augmented trajectory $\Xi = (\xi, \xi_0)$ with domain I is *locally uniquely limiting* if for every compact subinterval I' of I the restriction of Ξ to I' is uniquely limiting.

An augmented trajectory $\Xi = (\xi, \xi_0)$ with domain I is *almost uniquely limiting* if there exists a finite subset B of I such that the restriction of Ξ to every subinterval of $I \backslash B$ is locally uniquely limiting.

If Ω is an open subset of $\mathbb{R}^n$, we say that a function $V : \Omega \mapsto \mathbb{R}$ *satisfies the inequality*

$$\sup\{-\langle \nabla V(x), f(x,u)\rangle - L(x,u) : u \in U\} \geq 0 \tag{XVI.13}$$

on Ω *in the viscosity sense* if

(V_+) *whenever $x \in \Omega$ and $p \in \mathbb{R}^n$ is a subdifferential of V at x, it follows that*

$$\sup\{-\langle p, f(x,u)\rangle - L(x,u) : u \in U\} \geq 0\,.$$

(We recall that, if Ω is open in $\mathbb{R}^n$, then a *subdifferential* of a function $V : \Omega \mapsto \mathbb{R}$ at a point $\bar{x} \in \Omega$ is a vector $p \in \mathbb{R}^n$ such that

$$\liminf_{x \to \bar{x}} \frac{V(x) - V(\bar{x}) - p \cdot (x - \bar{x})}{\|x - \bar{x}\|} \geq 0\,.)$$

Similarly, we say that V *satisfies the inequality*

$$\sup\{-\langle \nabla V(x), f(x,u)\rangle - L(x,u) : u \in U\} \leq 0 \tag{XVI.14}$$

on Ω *in the viscosity sense* if

(V_-) *whenever $x \in \Omega$ and $p \in \mathbb{R}^n$ is a superdifferential of V at x, it follows that*

$$\sup\{-\langle p, f(x,u)\rangle - L(x,u) : u \in U\} \leq 0\,.$$

(A *superdifferential* of V at x_* is a vector p such that $-p$ is a subdifferential of $-V$ at x_*.)

We say that V *satisfies the equation*

$$\sup\{-\langle \nabla V(x), f(x,u)\rangle - L(x,u) : u \in U\} = 0 \tag{XVI.15}$$

on Ω *in the viscosity sense* if it satisfies (XVI.13) and (XVI.14) in the viscosity sense.

Remark XVI.9. *The definition of "viscosity solution" given here is known to be equivalent to the more common one involving test functions, cf. [6].* □

Our main result is the following theorem:

Theorem XVI.10. *Let $\hat{\Sigma} = (\Omega, U, f, L)$ be an augmented control system, let $\mathscr{T}$ be a target for $\hat{\Sigma}$, and let $V : \Omega \cup \mathscr{T} \mapsto \mathbb{R}$ be a function. Assume that*

(1) $\hat{\Sigma}$ is locally uniformly continuous, locally coercive, and such that $F_{\hat{\Sigma}}(x,U)$ is closed and convex for every $x \in \Omega$.

(2) V is continuous.

(3) V satisfies (XVI.15) on Ω in the viscosity sense.

(4) V vanishes on $\mathscr{T}$.

(5) Every augmented arc has an almost locally uniquely limiting improvement.

(6) Whenever $x \in \Omega$, $\Xi = (\xi, \xi_0) \in \mathscr{MSD}_{\hat{\Sigma},V,x}$, and ξ is right-unbounded, it follows that $\Xi \in \mathscr{A}^{\hat{\Sigma}}_{x,\mathscr{T}}$.

Then $V \equiv \mathscr{V}^{\hat{\Sigma}}_{\mathscr{T}}$.

Remark XVI.11. *Condition (6) was essentially introduced by M. Malisoff, cf. especially [97].* □

XVI.3 Examples

Example XVI.12. (Linear-quadratic optimal control.) *Consider the standard linear-quadratic optimal control problem, in which x, u take values in $\mathbb{R}^n$, $\mathbb{R}^m$, respectively, the dynamical law is*

$$\dot{x} = Ax + Bu\,, \tag{XVI.16}$$

the Lagrangian is given by

$$L(x,u) = x^\dagger Rx + u^\dagger Su\,,$$

the square matrices R, S are strictly positive definite, and the pair (A,B) is stabilizable. We take the target set $\mathscr{T}$ to consist of the origin of $\mathbb{R}^n$. (In order to satisfy the condition that $F_{\hat{\Sigma}}(x,U)$ is convex for every $x \in \Omega$, we add a new scalar nonnegative control variable v, in such a way that the dynamical law (XVI.16) remains unchanged but the Lagrangian L is replaced by $\tilde{L}$, where $\tilde{L}(x,u,v) \stackrel{\text{def}}{=} L(x,u) + v$.) The crucial technical issue here is the fact that the Lagrangian is not bounded away from zero. The hypotheses of our main theorem (including the coerciveness, *which follows from the positive definiteness of S) are easily verified as long as V is bounded below. The only nontrivial point is the verification of condition (6). To prove that this holds, let* $\Xi : [0,\tau[\mapsto \mathbb{R}^n \times \mathbb{R}$ *be a right-unbounded maximal trajectory of steepest descent with respect to V that does not end at the target, and write* $\Xi = (\xi, \xi_0)$ *in the usual way. Then* τ *has to be infinite, because if* τ *was finite then the boundedness of the cost (arising from the fact that* $\Xi = (\xi, \xi_0)$ *is of steepest descent and V is bounded below) would trivially imply an* L^2 *bound on the control, from which it would follow that* Ξ *can be extended to the closed interval* $[0,\tau]$*, and then the assumption that* Ξ *does not end at the target would enable us to use Proposition XVI.7 and Theorem XVI.19 (applied with* $\Omega = \mathbb{R}^n \backslash \{0\}$*) to extend* Ξ *even further, contradicting maximality. So* τ *is infinite. On the other hand, the fact that V is bounded below and* Ξ *is of steepest descent implies that the integral*

$$\int_0^\infty \Big(\xi(t)^\dagger R \xi(t) + \eta(t)^\dagger S \eta(t)\Big) dt$$

is finite, if η *is an open-loop control that generates* Ξ*. But then* ξ *and* η *are square-integrable, so the condition that* $\dot{\xi} = A\xi + B\eta$ *implies that* ξ *is square-integrable and has a square-integrable derivative, and then Barbalat's lemma implies that* ξ *ends at the target, as desired.* □

Example XVI.13. (Fuller's problem, cf., e.g., [**?**].) *This is the optimal control problem for the dynamical law*

$$\begin{aligned} \dot{x} &= y\,, \\ \dot{y} &= u\,, \end{aligned}$$

with control constraint $-1 \le u \le 1$*. The target set* $\mathscr{T}$ *consists of the origin of* $\mathbb{R}^2$*. The Lagragian is* $L(x,y,u) = x^2$*. The crucial technical issue here is the fact that the Lagrangian is not bounded away from zero, and in fact has a whole line of zeros. The hypotheses of our main theorem are easily verified as long as V is bounded below. The only nontrivial point is the verification of condition (6). To prove that this holds, let* $\Xi : [0,\tau[\mapsto \mathbb{R}^3$ *be a right-unbounded maximal trajectory of steepest descent with respect to V that does not end at the target, and write* $\Xi = (\xi, \xi_0)$ *in the usual way. Then* τ *has to be infinite, because if* τ *was finite then the boundedness of the control would trivially imply that* Ξ *can be extended to the closed interval* $[0,\tau]$*, and then the assumption that* Ξ *does not end at the target enables us to use Proposition XVI.7 and Theorem XVI.19 (applied with* $\Omega = \mathbb{R}^2 \backslash \{(0,0)\}$*) to extend* Ξ *even further, contradicting maximality. So* τ *is infinite. On the other hand, the fact that V is bounded below and* Ξ *is of steepest descent implies that the integral* $\int_0^\infty x(t)^2 dt$ *is finite, if we write* $\xi(t) = (x(t), y(t))$*. But then*

$x(\cdot)$ is a square-integrable function on $[0,\infty[$ whose second derivative is bounded. By a straightforward generalization of Barbalat's lemma, this implies that both $x(\cdot)$ and $y(\cdot)$ go to zero, i.e., that ξ ends at the target, as desired. □

Example XVI.14. (The reflected brachistochrone problem.) *This is the minimum time problem for the dynamical law*

$$\begin{aligned} \dot{x} &= u\sqrt{|y|}\,, \\ \dot{y} &= v\sqrt{|y|}\,, \end{aligned}$$

with control constraint $u^2+v^2 \leq 1$. The target set $\mathscr{T}$ consists of a single point $B \in \mathbb{R}^2$. The crucial technical issue here is the fact that the dynamical law is not Lipschitz-continuous with respect to the state. The hypotheses of our main theorem are easily verified. The only nontrivial point is the verification of condition (5). To prove that this holds, we pick an arbitrary integral arc $\xi : [a,b] \mapsto \mathbb{R}^2$, and observe that either (i) $\xi(t)$ never belongs to the x axis $X = \{(x,y) \in \mathbb{R}^2 : y = 0\}$ or (ii) there exist $t_-, t_+ \in [a,b]$ such that $t_- \leq t_+$, $\xi(t) \notin X$ whenever $a \leq t < t_-$ or $t_+ > t \leq b$, $\xi(t_-) \in X$, and $\xi(t_+) \in X$. If (i) holds, then ξ satisfies the conditions of Example XVI.8, so ξ is uniquely limiting. If (ii) holds and $t_+ = t_-$, then the restriction of ξ to each of the intervals $[a,t_-[+,,]t_+,b]$, is locally uniquely limiting, so ξ is almost locally uniquely limiting. Finally, if (ii) holds and $t_+ < t_-$, then the restriction $\tilde{\xi}$ of ξ to the interval $[t_-,t_+]$ is such that the set $S = \{t \in [a,b] : \xi(t) \notin X\}$ is the union of a finite or countable infinite collection $\mathscr{I}$ of pairwise disjoint relatively open subintervals of $[a,b]$. If $I \in \mathscr{I}$, then the restriction $\tilde{\xi}_I$ of $\tilde{\xi}$ to I is entirely contained in the open upper half-plane or in the open lower half-plane. By reflecting $\tilde{\xi}_I$ with respect to X, if necessary, we may assume that $\tilde{\xi}_I$ is entirely contained in the open upper half-plane for every $I \in \mathscr{I}$. Then $\tilde{\xi}$ is a trajectory of our system entirely contained in the closed upper half-plane. It is well known that the problem in the closed upper half-plane H_+ is the famous "brachistochrone problem," whose time-optimal trajectories $\zeta : [\alpha,\beta] \mapsto H_+$ are cycloids such that $\zeta(t)$ is an interior point of H_+ whenever $\alpha < t < \beta$. It follows that we can always replace $\tilde{\xi}$ by a cycloid ζ, thereby obtaining an almost locally uniquely limiting improvement of ξ. □

Example XVI.15. (An example with a continuous non-Lipschitz dynamics where uniqueness fails.) *Let $\varphi : [0,1] \mapsto \mathbb{R}$ be a nonnegative continuous function such that (a) the set $\{x \in [0,1] : \varphi(x) = 0\}$ is exactly the Cantor set, and (b) $\int_0^1 \frac{dx}{\varphi(x)} < \infty$. (For example, we may take φ to be given by $\varphi(x) = \mathrm{dist}(x,C)^\rho$, where C is the Cantor set and ρ is a positive number such that $\rho < 1 - \log_3 2$. An explicit calculation shows that $\int_0^1 \frac{dx}{\varphi(x)} = (1-\rho)^{-1} 2^\rho \sum_{j=1}^{\infty} \theta^j$, where $\theta = \frac{2}{3} \times 3^\rho$. Our choice of ρ guarantees that $\theta < 1$, so the integral is finite.) Extend φ to a function defined on*

$\mathbb{R}$ by making it periodic of period 1. Then consider the optimal control problem on $\mathbb{R}$ whose dynamics is $\dot{x} = u\varphi(x)$, $|u| \le 1$, and where the goal is to reach the origin in minimum time. It is easy to see that the optimal trajectory from each point x exists and is obtained by "moving towards the target as fast as possible." Precisely, this means the we use the control $u = -1$ as long as we are to the right of the origin, and we use $u = 1$ if we are to the left. This, however, does not suffice to specify the optimal trajectories, because of the lack of uniqueness of solutions. The complete specification of the optimal trajectories is as follows. Suppose $\bar{x} < 0$. Define a function $\tau : [\bar{x}, 0] \mapsto \mathbb{R}$ by letting $\tau(x) = \int_{\bar{x}}^{x} \frac{dy}{\varphi(y)}$. Then τ is aboslutely continuous, strictly increasing, and such that $\tau(\bar{x}) = 0$. Therefore τ maps the interval $[\bar{x}, 0]$ homeomorphically onto the interval $[0, \tau(0)]$. Let ξ be the inverse function, so ξ maps $[0, \tau(0)]$ homeomorphically onto $[\bar{x}, 0]$. Then ξ is absolutely continuous, and $\dot{\xi}(t) = \varphi(\xi(t))$ for almost all $t \in [0, \tau(0)]$. So ξ is a trajectory of our system which goes from $\bar{x}$ to 0 in time $\tau(0)$, and it is easy to see that ξ is the optimal trajectory from $\bar{x}$ to 0. It follows that optimal time to go from $\bar{x}$ to 0 is $\tau(0)$, that is, $\int_{\bar{x}}^{0} \frac{dy}{\varphi(y)}$. A similar contruction applies when $\bar{x} > 0$. Then the value function $\bar{V}$ for our problem is given by

$$\bar{V}(x) = \int_{\min(x,0)}^{\max(x,0)} \frac{dy}{\varphi(y)}.$$

The HJBE for our problem is

$$|V'(x)|\varphi(x) - 1 = 0. \tag{XVI.17}$$

The function $\bar{V}$ is a solution of this equation on $\mathbb{R}\backslash\{0\}$ in the viscosity sense. (This follows from the fact that, for problems such as this one, the value function is automatically a viscosity solution of the HJBE. In addition, one can also verify this directly. Let $\mathscr{O}$ be the set of points where $\varphi(x) > 0$. Then on $\mathscr{O}$ the function $\bar{V}$ is smooth, and its derivative is $\frac{1}{\varphi}$ when $x < 0$, and $-\frac{1}{\varphi}$ when $x > 0$, so (XVI.17) holds. At points x where $\varphi(x) = 0$, the viscosity solution requirements say that $-1 \ge 0$ whenever p is a subdifferential of $\bar{V}$ at x, and $-1 \le 0$ whenever p is a superdifferential of $\bar{V}$ at x. The second condition is trivially true. To verify the first condition, we need to show that it is satisfied vacuously, i.e., that there are no subdifferentials of $\bar{V}$ at x. But this easy. Suppose, say, that $x < 0$. The difference quotient $\frac{1}{h}(V(x+h) - V(x))$ is equal, if $h > 0$, to $-\frac{1}{h}\int_{x}^{x+h} \frac{dy}{\varphi(y)}$, which is bounded above by $\zeta(h) = -\frac{1}{\max\{\varphi(y): y \in [x, x+h]\}}$. Since $\varphi(x) = 0$, $\zeta(h)$ goes to $-\infty$ as $h \to 0$. This shows that the right derivative of $\bar{V}$ at x is equal to ∞, from which it follows easily that there exist no subdifferentials of $\bar{V}$ at x. A similar argument shows that if $x > 0$ the left derivative of $\bar{V}$ at x is equal to $+\infty$, from which it follows once again that there are no subdifferentials of $\bar{V}$ at x.

We now show that there exist nonnegative continuous functions $\hat{V}:\mathbb{R}\mapsto\mathbb{R}$ *other than* $\bar{V}$ *that satisfy the HJBE on* $\mathbb{R}\backslash\{0\}$ *and are such that* $\hat{V}(0)=0$*. To see this, we let* W *be a continuous monotonically nondecreasing real-valued function on* $[0,\infty[$ *such that (a)* $W(0)=0$*, (b)* W *is constant on each connected component of the set* $\{x:x>0\wedge\varphi(x)>0\}$*, and (c)* $W(x)<W(y)$ *whenever* $0\leq x<y$ *and the interval* $[x,y]$ *contains a zero of* φ*. (Such a function is easily constructed using the well known Cantor function.) We then extend* W *to all of* $\mathbb{R}$ *by defining* $W(x)=W(-x)$ *when* $x<0$*. Using* W*, we define* $\hat{V}=\bar{V}+W$*. Then* $\hat{V}$ *is continuous,* $\hat{V}(0)=0$*, and* $\hat{V}(x)>\bar{V}(x)$ *whenever* $x\neq 0$*. Let us show that* $\hat{V}$ *is also a solution of the HJBE for our problem on* $\mathbb{R}\backslash\{0\}$*. Near points* x *such that* $\varphi(x)>0$*, the functions* $\bar{V}$ *and* $\hat{V}$ *differ by a constant, so the fact that* $\bar{V}$ *satisfies the HJBE implies that the same is true for* $\hat{V}$*. If* $x\neq 0$ *but* $\varphi(x)=0$*, the viscosity solution requirements say that* $-1\geq 0$ *whenever* p *is a subdifferential of* $\hat{V}$ *at* x*, and* $-1\leq 0$ *whenever* p *is a superdifferential of* $\hat{V}$ *at* x*, and the second one of these conditions is trivially true. As for the first condition, if* $x<0$ *then we have already shown that the right derivative of* $\bar{V}$ *at* x *is equal to* ∞*, and this clearly implies that the right derivative of* $\hat{V}$ *at* x *is equal to* ∞ *as well, since* $\hat{V}=\bar{V}+W$ *and* W *is monotonically nonincreasing near* x*. Hence there exist no subdifferentials of* $\hat{V}$ *at* x*. A similar argument applies if* $x>0$*, and we conclude that first one of the viscosity requirements is satisfied vacuously.*

It follows that for our example the value function is not the unique continuous nonnegative function that vanishes at the target and satisfies the HJBE. In the example, the reason for the failure of uniqueness is easy to understand, and clearly related to the non-uniqueness of trajectories. Notice that the spurious value function $\hat{V}$ *is bounded below by the true value function, so what goes wrong is the other inequality, which is related to the dynamic programming inequality (DPI). And, indeed, the DPI fails, and this makes it impossible to draw the conclusion that* $\hat{V}\leq\bar{V}$*. Furthermore, the failure of the DPI happens exactly as described in our general analysis: given any control* $u(\cdot)$ *and any initial condition* x_0*, it is easy to construct a maximal trajectory* ξ *for* $u(\cdot)$ *starting at* x_0 *along which the DPI for* $\hat{V}$ *holds. (It suffices to follow the only possible trajectory for* $u(\cdot)$ *as long as* $\varphi\neq 0$*, and stopping at* $\bar{x}$ *and staying there for ever as soon as we reach the first point* $\bar{x}$ *where* φ *vanishes.) This* ξ *is not, however, the only trajectory for* $u(\cdot)$ *starting at* x_0*. And the fact that the DPI holds along* ξ *does not imply that that the DPI holds for every trajectory for* $u(\cdot)$ *that starts at* x_0*. (Indeed, if for example* $x_0<0$ *and* $u(t)\equiv 1$*, then in addition to the* ξ *given by our construction we could also consider* ξ_{opt}*, the optimal trajectory described earlier. The DPI for* $\hat{V}$ *clearly fails along* ξ_{opt}*, because if it was true it would imply that* $\hat{V}(x_0)\leq\bar{V}(x_0)$*, whereas we know that* $\hat{V}(x_0)>\bar{V}(x_0)$*.)* $\square$

XVI.4 The main technical lemma

XVI.4.1 Infinitesimal steepest descent.

Let $\hat{\Sigma} = (\Omega, U, f, L)$ be an augmented control system. For every $x \in \Omega$ and every positive number δ such that $\mathrm{dist}(x, \mathbb{R}^n \backslash \Omega) > \delta$, we let $\Phi_{\delta,\hat{\Sigma}}(x)$ be the closed convex hull of all the vectors $F_{\hat{\Sigma}}(x', u)$, for all pairs (x', u) such that $x' \in \Omega$, $\|x' - x\| \leq \delta$, and $u \in U$. Then $\Phi_{\delta,\hat{\Sigma}}(x)$ is a closed convex subset of $\mathbb{R}^{n+1}$. Clearly, $\Phi_{\delta,\hat{\Sigma}}(x) \subseteq \Phi_{\delta',\hat{\Sigma}}(x)$ whenever $0 < \delta \leq \delta'$.

Let $V : \Omega \mapsto \bar{\mathbb{R}}$ be a real-valued function, and let $x_* \in \Omega$. We say that V *satisfies the infinitesimal steepest descent condition for* $\hat{\Sigma}$ *at* x_* if

(ISD) *there exist sequences* $\{x_j\}_{j=1}^{\infty}$, $\{v_j\}_{j=1}^{\infty}$, $\{\lambda_j\}_{j=1}^{\infty}$, $\{h_j\}_{j=1}^{\infty}$, $\{\gamma_j\}_{j=1}^{\infty}$ *in* Ω, $\mathbb{R}^n$, $\mathbb{R}$, $\mathbb{R}$, *and* $\mathbb{R}$, *respectively, such that*

(1) $h_j > 0$, $\gamma_j > 0$, $(v_j, \lambda_j) \in F_{\hat{\Sigma}}(x_*, U)$, *and* $\|x_j - x_* - h_j v_j\| \leq h_j \gamma_j$ *for all* j,

(2) $h_j \downarrow 0$, $x_j \to x_*$, *and* $\gamma_j \downarrow 0$ *as* $j \to \infty$,

and

(3) $V(x_j) \leq V(x_*) - h_j \lambda_j + h_j \gamma_j$ for all j.

We say that V *satisfies the weak infinitesimal steepest descent condition for* $\hat{\Sigma}$ *at* x_* if

(WISD) *there exist sequences* $\{x_j\}_{j=1}^{\infty}$, $\{v_j\}_{j=1}^{\infty}$, $\{\lambda_j\}_{j=1}^{\infty}$, $\{h_j\}_{j=1}^{\infty}$, $\{\delta_j\}_{j=1}^{\infty}$, $\{\gamma_j\}_{j=1}^{\infty}$ *in* Ω, $\mathbb{R}^n$, $\mathbb{R}$, $\mathbb{R}$, $\mathbb{R}$, *and* $\mathbb{R}$, *respectively, such that*

(1) $h_j > 0$, $\delta_j > 0$, $\gamma_j > 0$, $(v_j, \lambda_j) \in \Phi_{\delta_j,\hat{\Sigma}}(x_*)$, $\|x_j - x_*\| \leq \delta_j$, *and* $\|x_j - x_* - h_j v_j\| \leq h_j \gamma_j$ *for all* j,

(2) $h_j \downarrow 0$, $\delta_j \downarrow 0$, *and* $\gamma_j \downarrow 0$ *as* $j \to \infty$,

and

(3) $V(x_j) \leq V(x_*) - h_j \lambda_j + h_j \gamma_j$ for all j.

We say that V *satisfies (XVI.13) on* Ω *in the ISD sense* if V satisfies the infinitesimal steepest descent condition for $\hat{\Sigma}$ at x for every $x \in \Omega$. We say that V *satisfies (XVI.13) on* Ω *in the WISD sense* if V satisfies the weak infinitesimal steepest descent condition for $\hat{\Sigma}$ at x for every $x \in \Omega$.

Theorem XVI.16. *Let* $\hat{\Sigma} = (\Omega, U, f, L)$ *be an n-dimensional augmented control system and let* $V : \Omega \mapsto \mathbb{R}$ *be a continuous function. Then*

(1) Condition (WISD) holds at every point $x_* \in \Omega$ *where (ISD) holds. In particular, if* V *satisfies (XVI.13) on* Ω *in the ISD sense then* V *satisfies (XVI.13) on* Ω *in the WISD sense.*

(2) If $\hat{\Sigma}$ *is locally uniformly continuous and such that* $F_{\hat{\Sigma}}(x, U)$ *is convex for every* $x \in \Omega$*, then (ISD) holds at every point* x_* *where (WISD) holds, and in particular if* V *satisfies (XVI.13) on* Ω *in the WISD sense then* V *satisfies (XVI.13) on* Ω *in the ISD sense.*

(3) If $\hat{\Sigma}$ *is locally coercive, and and such that* $F_{\hat{\Sigma}}(x, U)$ *is closed and convex for every* $x \in \Omega$*, then*

(3.i) if V *satisfies (XVI.13) on* Ω *in the ISD sense then* V *satisfies (XVI.13) on* Ω *in the viscosity sense;*

(3.ii) if V *satisfies (XVI.13) on* Ω *in the viscosity sense it follows that if* V *satisfies (XVI.13) on* Ω *in the WISD sense.*

In particular, if $\hat{\Sigma}$ *is locally uniformly continuous, locally coercive, and such that* $F_{\hat{\Sigma}}(x, U)$ *is closed and convex for every* $x \in \Omega$*, then the three concepts of solution of (XVI.13) on* Ω *(viscosity, ISD, and WISD) are equivalent.*

Proof. We first prove (1). We assume that V x_* is a point of Ω where (ISD) holds, and prove that (WISD) is true as well. Let $\{x_j\}_{j=1}^{\infty}$, $\{v_j\}_{j=1}^{\infty}$, $\{\lambda_j\}_{j=1}^{\infty}$, $\{h_j\}_{j=1}^{\infty}$, $\{\gamma_j\}_{j=1}^{\infty}$ be sequences that satisfy the properties of (ISD). Define $\delta_j = \|x_j - x_*\|$. Then all the conclusions of (WISD) are true, so (WISD) holds, and (1) is proved.

Next, we prove (2). We assume that $\hat{\Sigma}$ is locally uniformly continuous and (WISD) holds at a point $x_* \in \Omega$, and prove that (ISD) holds at x_* as well. To do this, we pick sequences $\{x_j\}_{j=1}^{\infty}$, $\{v_j\}_{j=1}^{\infty}$, $\{\lambda_j\}_{j=1}^{\infty}$, $\{h_j\}_{j=1}^{\infty}$, $\{\delta_j\}_{j=1}^{\infty}$, $\{\gamma_j\}_{j=1}^{\infty}$ with the properties specified in (WISD). We pick δ such that

$$B_\delta(x_*) = \{x \in \mathbb{R}^n : \|x - x_*\| \leq \delta\} \subseteq \Omega,$$

and a function $\omega :]0, +\infty[\mapsto [0, +\infty]$ such that $\lim_{s \downarrow 0} \omega(s) = 0$ and

$$\|f(x,u)-f(x',u)\|+|L(x,u)-L(x',u)|\leq \omega(\|x-x'\|)$$

whenever $x,x'\in B_\delta(x_*)$. We then pass to a subsequence, if necessary, and assume that $\delta_j\leq\delta$ for all j. We use the fact that $(v_j,\lambda_j)\in\Phi_{\delta_j,\hat{\Sigma}}(x_*)$ to find a (w_j,ℓ_j) of the form $\sum_{k=0}^{n+1}\alpha_{j,k}F_{\hat{\Sigma}}(x_{j,k},u_{j,k})$ such that $\alpha_{j,k}\geq 0$, $\sum_{k=0}^{n+1}\alpha_{j,k}=1$, $\|x_{j,k}-x_*\|\leq\delta_j$, and $\|w_j-v_j\|+|\ell_j-\lambda_j|\leq 2^{-j}$. We then define

$$\begin{aligned}\tilde{v}_j &= \sum_{k=0}^{n+1}\alpha_{j,k}f(x_*,u_{j,k})\,,\\ \tilde{\lambda}_j &= \sum_{k=0}^{n+1}\alpha_{j,k}L(x_*,u_{j,k})\,,\end{aligned}$$

and conclude that $\|\tilde{v}_j-w_j\|\leq\omega(\delta_j)$ and $\|\tilde{\lambda}_j-\ell_j\|\leq\omega(\delta_j)$. We then let

$$\tilde{\gamma}_j=\gamma_j+2^{-j}+\omega(\delta_j)\,,$$

so $\tilde{\gamma}_j\to 0$ as $j\to\infty$. It is then clear that $\|x_j=x_*-h_j\tilde{v}_j\|\leq h_j\tilde{\gamma}_j$ and $V(x_j)\leq V(x_*)-h\tilde{\lambda}_j+h_j\tilde{\gamma}_j$ for all j. Finally, we have $(\tilde{v}_j,\tilde{\lambda}_j)=\sum_{k=0}^{n+1}\alpha_{j,k}F_{\hat{\Sigma}}(x_*,u_{j,k})$, so $(\tilde{v}_j,\tilde{\lambda}_j)\in F_{\hat{\Sigma}}(x_*,U)$, since $F_{\hat{\Sigma}}(x_*,U)$ is convex. This completes the proof that (WISD) implies (ISD).

We now turn to the proof of (3), for which purpose we assume that $\hat{\Sigma}$ is locally coercive and such that $F_{\hat{\Sigma}}(x,U)$ is closed and convex for every $x\in\Omega$. To prove (3.i), we assume in addition that V satisfies (XVI.13) on Ω in the ISD sense, and show that V satisfies (XVI.13) on Ω in the viscosity sense. To prove this, we pick $x_*\in\Omega$ and a subdifferential p of V at x_*, and show that

$$\sup\{-p\cdot f(x_*,u)-L(x_*,u):u\in U\}\geq 0\,. \qquad \text{(XVI.18)}$$

Using (ISD), we pick sequences $\{x_j\}_{j=1}^\infty$, $\{v_j\}_{j=1}^\infty$, $\{\lambda_j\}_{j=1}^\infty$, $\{h_j\}_{j=1}^\infty$, $\{\gamma_j\}_{j=1}^\infty$ in Ω, $\mathbb{R}^n$, $\mathbb{R}$, $\mathbb{R}$, and $\mathbb{R}$, respectively, such that $h_j>0$, $\gamma_j>0$, $(v_j,\lambda_j)\in F_{\hat{\Sigma}}(x_*,U)$, $\|x_j-x_*-h_jv_j\|\leq h_j\gamma_j$, and $V(x_j)\leq V(x_*)-h_j\lambda_j+h_j\gamma_j$ for all j, and $h_j\downarrow 0$, $x_j\to x_*$, and $\gamma_j\downarrow 0$ as $j\to\infty$. Let $J=\{j:x_j=x_*\}$. Then for $j\in J$ the inequality $\|x_j-x_*-h_jv_j\|\leq h_j\gamma_j$ implies $\|v_j\|\leq\gamma_j$, so $\lim_{j\to\infty,j\in J}v_j=0$. On the other hand, the inequality $V(x_j)\leq V(x_*)-h_j\lambda_j+h_j\gamma_j$ implies $h_j\lambda_j\leq h_j\gamma_j$, i.e., $\lambda_j\leq\gamma_j$. Since the sequence $\{\lambda_j\}_{j=1}^\infty$ is bounded below (for example, because the local coercivity implies a bound $\lambda_j\geq\|v_j\|^r-C$), we may assume, after replacing J by a smaller infinite set, if necessary, that $\lambda=\lim_{j\to\infty,j\in J}\lambda_j$ exists. Since $\lambda_j\leq\gamma_j$, λ must be ≤ 0. Furthermore, the vector $(0,\lambda)$ is a limit of vectors $(v_j,\lambda_j)\in F_{\hat{\Sigma}}(x_*)$, so $(0,\lambda)\in F_{\hat{\Sigma}}(x_*,U)$, since $F_{\hat{\Sigma}}(x_*,U)$ is closed. Hence there exists $\bar{u}\in U$ such that

$f(x_*,\bar{u}) = 0$ and $L(x_*,\bar{u}) \leq 0$. But then $-p \cdot f(x_*,\bar{u}) - L(x_*,\bar{u}) \geq 0$, so (XVI.18) holds.

We now consider the case when the set J is finite. In this case, after passing to a subsequence, if necessary, we may assume that J is empty, i.e., that $x_j \neq x_*$ for all j. Since p is a subdifferential of V at x_*, we have

$$\lim \inf_{x \to x_*, x \neq x_*} \frac{V(x) - V(x_*) - p \cdot (x - x_*)}{\|x - x_*\|} \geq 0 . \tag{XVI.19}$$

Since the x_j converge to x_* and are different from x_*, (XVI.19) implies

$$\liminf_{j \to \infty} \frac{V(x_j) - V(x_*) - p \cdot (x_j - x_*)}{\|x_j - x_*\|} \geq 0 . \tag{XVI.20}$$

Since $V(x_j) \leq V(x_*) - h_j\lambda_j + h_j\gamma_j$, (XVI.20) implies

$$\liminf_{j \to \infty} \frac{-h_j\lambda_j + h_j\gamma_j - p \cdot (x_j - x_*)}{\|x_j - x_*\|} \geq 0 . \tag{XVI.21}$$

Now, $x_j - x_* = x_j - x_* - h_jv_j + h_jv_j = h_j(w_j + v_j)$, where $w_j = h_j^{-1}(x_j - x_* - h_jv_j)$. Hence

$$\liminf_{j \to \infty} \frac{-\lambda_j + \gamma_j - p \cdot (w_j + v_j)}{\|w_j + v_j\|} \geq 0 . \tag{XVI.22}$$

Hence, given a positive ε there exists a $j(\varepsilon)$ such that

$$-\lambda_j + \gamma_j - p \cdot (w_j + v_j) \geq -\varepsilon\|w_j + v_j\| \text{ whenever } j \geq j(\varepsilon) .$$

Therefore

$$-\lambda_j - p \cdot v_j \geq -\varepsilon\|v_j\| - \varepsilon\|w_j\| + p \cdot w_j - \gamma_j \text{ if } j \geq j(\varepsilon) .$$

Since $w_j \to 0$, there is—for each ε—a $j'(\varepsilon)$ such that

$$-\lambda_j - p \cdot v_j \geq -\varepsilon\|v_j\| - \varepsilon \text{ whenever } j \geq j'(\varepsilon) . \tag{XVI.23}$$

The coercivity bound yields $\|v_j\|^r \leq \lambda_j + C$, so $-\lambda_j \leq -\|v_j\|^r + C$. Hence

$$-\|v_j\|^r + C - p \cdot v_j \geq -\varepsilon\|v_j\| - \varepsilon \text{ whenever } j \geq j'(\varepsilon) ,$$

so

$$-\|v_j\|^r + \varepsilon\|v_j\| + \|p\|.\|v_j\| \geq -\varepsilon - C \text{ if } j \geq j'(\varepsilon) . \tag{XVI.24}$$

Now, if the sequence $\{\|v_j\|\}_{j=1}^{\infty}$ was unbounded, then we could pick an infinite subset J of $\mathbb{N}$ such that $\|v_j\| \to +\infty$ as $j \to \infty$ via values in J. But then, taking for example $\varepsilon = 1$, we would contradict (XVI.24), because the number $-\|v_j\|^r + \|v_j\| + \|p\|.\|v_j\|$ is equal to $-\|v_j\|^r\Big(1-(1+\|p\|)\|v_j\|^{1-r}\Big)$, which goes to $-\infty$ as $j \to \infty$ via values in J. Therefore the sequence $\{\|v_j\|\}_{j=1}^{\infty}$ is bounded. Pick a constant K such that $\|v_j\| \le K$ for all j. Then, for each ε, if u_ε is such that $v_{j'(\varepsilon)} = f(x_*, u_\varepsilon)$ and $\lambda_{j'(\varepsilon)} = L(x_*, u_\varepsilon)$, (XVI.23) implies

$$-p \cdot f(x_*, u_\varepsilon) - L(x_*, u_\varepsilon) \ge -\varepsilon(K+1)\,.$$

Hence

$$\sup\{-p \cdot f(x_*, u) - L(x_*, u) : u \in U\} \ge -\varepsilon(K+1)\,.$$

Since ε is arbitrary, we see that (XVI.18) holds. This concludes the proof of (3.i).

We now proceed to proving (3.ii). We assume that V satisfies (XVI.13) on Ω in the viscosity sense, pick an $x_* \in \Omega$, and prove that condition (WISD) holds. We do this by assuming that the sequences whose existence is asserted by (WISD) do not exist and deriving a contradiction. We will assume, as we clearly may without loss of generality, that $x^* = 0$ and $V(x^*) = 0$, i.e., $V(0) = 0$. In particular, this implies of course that $0 \in \Omega$.

Since (WISD) is not satisfied, there must exist a $\bar{\gamma}$ such that $0 < \bar{\gamma}$,

$$\{x \in \mathbb{R}^n : \|x\| \le \bar{\gamma}\} \subseteq \Omega\,, \tag{XVI.25}$$

and

$$V(h(v+v')) + h\lambda > h\bar{\gamma} \tag{XVI.26}$$

whenever $0 < h \le \bar{\gamma}$, $(v,\lambda) \in \Phi_{\bar{\gamma},\hat{\Sigma}}(0)$, $h\|v+v'\| \le \bar{\gamma}$, and $\|v'\| \le \bar{\gamma}$. (Indeed, if $\bar{\gamma}$ did not exist, then for each sufficiently large natural number j we could define $\gamma_j = 2^{-j}$, and find h_j, v_j, λ_j, v'_j, x_j, such that $0 < h_j \le \gamma_j$, $(v_j, \lambda_j) \in \Phi_{\gamma_j,\hat{\Sigma}}(0)$, $h_j\|v_j + v'_j\| \le \gamma_j$, $\|v'_j\| \le \gamma_j$, and $V(h_j(v_j + v'_j)) + h_j\lambda_j \le \gamma_j h_j$. But then, if we take $x_j = h_j(v_j + v'_j)$, $\delta_j = \gamma_j$, the sequence $\{(x_j, v_j, \lambda_j, h_j, \delta_j, \gamma_j)\}_{j=1}^{\infty}$ satisfies conditions (1), (2) and (3) of (WISD), contradicting the fact that V does not satisfy (WISD).)

By making $\bar{\gamma}$ smaller, if necessary, we can assume that there exist real numbers C, r, such that $C > 0$, $r > 1$ and $\|f(x,u)\|^r \le L(x,u) + C$ whenever $\|x\| \le \bar{\gamma}$ and $u \in U$. It then follows, if we let

$$\psi(v,\lambda) = \|v\|^r - \lambda - C \text{ for } (v,\lambda) \in \mathbb{R}^n \times \mathbb{R},$$

that

$$\psi(w) \leq 0 \text{ whenever } w = F(x,u)\,, \|x\| \leq \bar{\gamma}, u \in U\,. \tag{XVI.27}$$

Since ψ is convex, the inequality $\psi(w) \leq 0$ holds whenever $w \in \Phi_{\bar{\gamma},\hat{\Sigma}}(0)$, and then

$$\|v\|^r \leq \lambda + C \quad \text{whenever } (v,\lambda) \in \Phi_{\bar{\gamma},\hat{\Sigma}}(0)\,. \tag{XVI.28}$$

If $(v,\lambda) \in \Phi_{\bar{\gamma},\hat{\Sigma}}(0)$, $v' \in \mathbb{R}^n$, and $\|v'\| \leq \bar{\gamma}$, then the inequality $(\alpha+\beta)^r \leq 2^r(\alpha^r + \beta^r)$, valid for nonnegative α, β, implies

$$\begin{aligned} \|v+v'\|^r &\leq (\|v\| + \|v'\|)^r \\ &\leq 2^r(\|v\|^r + \|v'\|^r) \\ &\leq 2^r(\lambda + C + \bar{\gamma}^r)\,. \end{aligned}$$

Pick $\tilde{r}$ such that $1 < \tilde{r} < r$. Let $A = 2^{\frac{r}{r-\tilde{r}}}$ Then

$$\begin{aligned} \|v+v'\|^{\tilde{r}} &= \|v+v'\|^{\tilde{r}-r} \cdot \|v+v'\|^r \\ &\leq A^{\tilde{r}-r}\|v+v'\|^r \\ &\leq A^{\tilde{r}-r}2^r(\lambda + C + \bar{\gamma}^r) \\ &= 2^{-r}2^r(\lambda + C + \bar{\gamma}^r) \\ &= \lambda + C + \bar{\gamma}^r \end{aligned}$$

if $\|v+v'\| \geq A$. On the other hand, if $\|v+v'\| \leq A$, then $\|v+v'\|^{\tilde{r}} \leq A^{\tilde{r}} \leq \|v\|^r + A^{\tilde{r}} \leq \lambda + C + A^{\tilde{r}}$. Therefore, if we let $\tilde{C} = C + \max(A^{\tilde{r}}, \bar{\gamma}^r)$, and then relabel $\tilde{r}$, $\tilde{C}$ as our new r and C, we have shown that

$$\begin{aligned} &\Big((v,\lambda) \in \Phi_{\bar{\gamma},\hat{\Sigma}}(0) \wedge v' \in \mathbb{R}^n \wedge \|v'\| \leq \bar{\gamma}\Big) \\ &\qquad \Rightarrow \|v+v'\|^r \leq \lambda + C\,. \end{aligned} \tag{XVI.29}$$

For $0 < \delta \leq \bar{\gamma}$, we define

$$\begin{aligned} &\Xi_\delta(0) \stackrel{\text{def}}{=} \Big\{(v+v', \lambda+\lambda') : \\ &(v,\lambda) \in \Phi_{\delta,\hat{\Sigma}}(0), v' \in \mathbb{R}^n\,, \|v'\| \leq \bar{\gamma}, \lambda' \in \mathbb{R}, \lambda' \geq 0\Big\}\,. \end{aligned}$$

Then

(a) *If* $0 < \delta \leq \bar{\gamma}$, *then* $\Xi_\delta(0)$ *is a closed, convex, nonempty subset of* $\mathbb{R}^{n+1}$

(The fact that $\Xi_\delta(0)$ is nonempty follows because $F_{\hat{\Sigma}}(0,U) \subseteq \Phi_{\delta,\hat{\Sigma}}(0) \subseteq \Xi_\delta(0)$, and $F_{\hat{\Sigma}}(0,U) \neq \emptyset$ because $U \neq \emptyset$. The fact that $\Xi_\delta(0)$ is closed follows because, if a sequence $\{(v_j+v'_j, \lambda_j+\lambda'_j)\}_{j=1}^\infty$ with the property that $(v_j,\lambda_j) \in \Phi_{\delta,\hat{\Sigma}}(0)$, $v'_j \in \mathbb{R}^n$, $\|v'_j\| \leq \bar{\gamma} \in \mathbb{R}^n$, $\lambda'_j \geq 0$ converges to a limit $(\hat{v},\hat{\lambda})$, then the sequence $\{v'_j\}_{j=1}^\infty$ is bounded, so we may assume after passing to a subsequence that $\lim_{j\to\infty} v'_j = v'$ exists, and then of course $\|v'\| \leq \bar{\gamma} \in \mathbb{R}^n$, and $\lim_{j\to\infty} v_j = v$ exists as well, since $v_j + v'_j \to \hat{v}$, and then $\hat{v} = v + v'$. Furthermore, the bound $\|v_j\|^r \leq \lambda_j + C$ implies that $\lambda_j \geq \|v_j\|^r - C \geq -C$, so

$$\lambda'_j = (\lambda_j + \lambda'_j) - \lambda_j \leq (\lambda_j + \lambda'_j) + C\,.$$

Hence the sequence $\{\lambda'_j\}_{j=1}^\infty$ is bounded above, because $\{\lambda_j + \lambda'_j\}_{j=1}^\infty$ is convergent. Since $\lambda'_j \geq 0$, the sequence $\{\lambda'_j\}_{j=1}^\infty$ is bounded, so we may assume it is convergent to a limit λ', after passing to a subsequence. Clearly, then, $\lambda' \geq 0$, and the limit $\lim_{j\to\infty}\lambda_j = \lambda$ exists as well, and satisfies $\hat{\lambda} = \lambda + \lambda'$. Since $(v,\lambda) = \lim_{j\to\infty}(v_j,\lambda_j)$, $(v_j,\lambda_j) \in \Phi_{\delta,\hat{\Sigma}}(0)$, and $\Phi_{\delta,\hat{\Sigma}}(0)$ is closed, we see that $(v,\lambda) \in \Phi_{\delta,\hat{\Sigma}}(0)$. Since $\|v'\| \leq \bar{\gamma} \in \mathbb{R}^n$, and $\lambda' \geq 0$, we see that $(\hat{v},\hat{\lambda}) \in \Xi_\delta(0)$. The convexity of $\Xi_\delta(0)$ is trivial.)

Let $\Lambda = \{0\} \times [\bar{\gamma}, +\infty[$, so $\Lambda \subseteq \mathbb{R}^{n+1}$. We let $\Psi_\delta(0)$ be the convex hull of $\Lambda \cup \Xi_\delta(0)$. We show that

(b) *If* $0 < \delta \leq \bar{\gamma}$ *then* $\Psi_\delta(0)$ *is a nonempty closed convex subset of* $\mathbb{R}^{n+1}$.

(c) *There exist real numbers* r, C *such that* $r > 1$ *and*

$$\|v\|^r \leq \lambda + C \quad \text{whenever } (v,\lambda) \in \Psi_{\bar{\gamma}}(0)\,. \tag{XVI.30}$$

(d) *The inequality*

$$V(hv) + h\lambda \geq h\bar{\gamma} \tag{XVI.31}$$

holds whenever $0 < h \leq \bar{\gamma}$, $(v,\lambda) \in \Psi_{\bar{\gamma}}(0)$, *and* $h\|v\| \leq \bar{\gamma}$.

(e) $(0,\ell) \notin \Psi_{\bar{\gamma}}(0)$ *whenever* $\ell < \bar{\gamma}$

We will prove the above assertions in order, except for the statement that $\Psi_\delta(0)$ is closed if $0 < \delta \leq \bar{\gamma}$, which will be proved last.

The fact that $\Psi_\delta(0)$ is convex is trivial, and the fact that $\Psi_\delta(0)$ is nonempty follows from (a), because $\Xi_\delta(0) \subseteq \Psi_\delta(0)$.

To prove (c), we choose r, C such that $r > 1$, $C > 0$, and (XVI.29) holds, and observe that (XVI.29) trivially implies that the inequality $\|v\|^r \leq \lambda + C$ is true whenever $(v,\lambda) \in \Xi_{\bar\gamma}(0)$. Since $C > 0$, the inequality is also true whenever $(v,\lambda) \in \Lambda$. Hence $\|v\|^r \leq \lambda + C$ whenever $(v,\lambda) \in \Xi_{\bar\gamma}(0) \cup \Lambda$, from which it follows that $\|v\|^r \leq \lambda + C$ whenever $(v,\lambda) \in \Psi_\delta(0)$, since the function $(v,\lambda) \mapsto \|v\|^r - \lambda - C$ is convex.

To prove (d), we observe that (XVI.26) trivially implies that $V(hv) + h\lambda > h\bar\gamma$ whenever $(v,\lambda) \in \Xi_{\bar\gamma}(0)$, $0 < h \leq \bar h$, and $h\|v\| \leq \bar\gamma$. (Indeed, if $(v,\lambda) \in \Xi_{\bar\gamma}(0)$, $0 < h \leq \bar h$, and $h\|v\| \leq \bar\gamma$, then $(v,\lambda) = (\bar v,\bar\lambda) + (v',\lambda')$, with $(\bar v,\bar\lambda) \in \Phi_{\bar\gamma,\hat\Sigma}(0)$, $v' \in \mathbb{R}^n$, $\|v'\| \leq \bar\gamma$, $\lambda' \in \mathbb{R}$, and $\lambda' \geq 0$. Then we can apply (XVI.26), with $\bar v$, $\bar\lambda$ in the roles of v, λ, and conclude that $V(h(\bar v + v')) + h\bar\lambda > h\bar\gamma$, since $h\|\bar v + v'\| \leq \bar\gamma$. Therefore $V(hv) + h\bar\lambda > h\bar\gamma$, and then *a fortiori* $V(hv) + h\lambda > h\bar\gamma$, since $\lambda = \bar\lambda + \lambda'$ and $\lambda' \geq 0$.) Assume that $(v,\lambda) \in \Psi_{\bar\gamma}(0)$. Then we can write $(v,\lambda) = \alpha(v',\lambda') + (1-\alpha)(0,\ell)$, with $(v',\lambda') \in \Xi_{\bar\gamma}(0)$, $\ell \geq \bar\gamma$, and $0 \leq \alpha \leq 1$. Let h be such that $0 < h \leq \bar\gamma$ and $h\|v\| \leq \bar\gamma$. If $\alpha = 0$, then $hv = 0$ and $\lambda = \ell \geq \bar\gamma$, so $V(hv) + h\lambda = V(0) + h\ell = h\ell \geq h\bar\gamma$. If $\alpha > 0$, define $\tilde h = \alpha h$. Then $0 < \tilde h \leq \bar\gamma$ and $\tilde h\|v'\| = \alpha h\|v'\| = h\|v\| \leq \bar\gamma$, since $v = \alpha v'$. Therefore $V(\tilde h v') + \tilde h\lambda' \geq \tilde h\bar\gamma$, since $(v',\lambda') \in \Psi_{\bar\gamma}(0)$. On the other hand, $(1-\alpha)h\ell \geq (1-\alpha)h\bar\gamma$, since $\ell \geq \bar\gamma$. Therefore

$$\begin{aligned}
V(hv) + h\lambda &= V(h\alpha v') + h\alpha\lambda' + h(1-\alpha)\ell \\
&= V(\tilde h v') + \tilde h\lambda' + h(1-\alpha)\ell \\
&\geq \tilde h\bar\gamma + h(1-\alpha)\bar\gamma \\
&= h\alpha\bar\gamma + h(1-\alpha)\bar\gamma \\
&= h\bar\gamma,
\end{aligned}$$

completing the proof of (d).

Statement (e) now follows easily: if $(0,\ell) \in \Psi_{\bar\gamma}(0)$, then we can apply (d) taking $h = \bar\gamma$, $v = 0$, $\lambda = \ell$, and conclude that $\bar\gamma\ell \geq \bar\gamma^2$, so that $\ell \geq \bar\gamma$.

We now prove that $\Psi_\delta(0)$ is closed if $0 < \delta \leq \bar\gamma$. Let $\{w_j\}_{j=1}^\infty$ be a sequence of points of $\Psi_\delta(0)$ that converges to a limit $w \in \mathbb{R}^{n+1}$. We will show that $w \in \Psi_\delta(0)$. Let $w_j = \alpha_j(v_j,\lambda_j) + (1-\alpha_j)(0,\ell_j)$, where $(v_j,\lambda_j) \in \Xi_\delta(0)$, $\ell_j \geq \bar\gamma$, and $0 \leq \alpha_j \leq 1$. Let $w = (v,\lambda)$. By passing to a subsequence, if necessary, we may assume that the α_j converge to a limit $\hat\alpha$. If the sequence $\{(v_j,\lambda_j)\}$ is bounded, then we may pass to a subsequence and assume that $(\hat v,\hat\lambda) = \lim_{j\to\infty}(v_j,\lambda_j)$ exists. Then $(\hat v,\hat\lambda) \in \Xi_\delta(0)$. Furthermore, the limit $\lim_{j\to\infty}\alpha_j(v_j,\lambda_j)$ exists, so $\mu = \lim_{j\to\infty}(1-\alpha_j)\ell_j$ exists as well, and $\mu \geq 0$. Clearly, $w = \hat\alpha(\hat v,\hat\lambda) + (0,\mu)$. If $\hat\alpha = 1$, then $w = (\hat v,\hat\lambda) + (0,\mu)$, so $w \in \Xi_\delta(0)$—and *a fortiori* $w \in \Psi_\delta(0)$—because $(\hat v,\hat\lambda) \in \Xi_\delta(0)$ and $\mu \geq 0$. If $\hat\alpha < 1$, then $\ell = \lim_{j\to\infty}\ell_j$ exists and satisfies $\ell = \frac{\mu}{1-\hat\alpha}$ and $\ell \geq \bar\gamma$. Then $w = \hat\alpha(\hat v,\hat\lambda) + (1-\hat\alpha)(0,\ell)$, and $(\hat v,\hat\lambda)$, $(0,\ell)$, belong to $\Xi_\delta(0)$ and Λ, respectively, so $w \in \Psi_\delta(0)$. Now suppose that the sequence $\{(v_j,\lambda_j)\}$ is unbounded. Then

(XVI.30) implies that $\{\lambda_j\}$ is unbounded. Since the λ_j are bounded below, we may assume, after passing to a subsequence, that $\lambda_j \to +\infty$ as $j \to \infty$. On the other hand, $\alpha_j v_j \to v$ and $\alpha_j\lambda_j + (1-\alpha_j)\ell_j \to \lambda$. Since both sequences $\{\alpha_j\lambda_j\}$, $\{(1-\alpha_j)\ell_j\}$ are bounded below, we may pass to a subsequence and assume that the limits $\mu = \lim_{j\to\infty}(1-\alpha_j)\ell_j$ and $\nu = \lim_{j\to\infty}\alpha_j\lambda_j$ exist. But then

$$\begin{aligned}\alpha_j\|v_j\| \le \alpha_j(\lambda_j + C)^{1/r} \quad &= \quad \alpha_j\lambda_j^{1/r}\Big(1+\frac{C}{\lambda_j}\Big)^{1/r}\\ &= \quad \alpha_j\lambda_j\lambda_j^{1/r-1}\Big(1+\frac{C}{\lambda_j}\Big)^{1/r}\\ &\underset{j\to\infty}{\longrightarrow} \quad 0\,,\end{aligned}$$

since $\lambda_j \underset{j\to\infty}{\longrightarrow} +\infty$, $\alpha_j\lambda_j \underset{j\to\infty}{\longrightarrow} \nu$, and $r>1$. Then $v = \lim_{j\to\infty}\alpha_j v_j = 0$. This implies, in particular, that $\bar\gamma\alpha_j\|v_j\| \le \bar\gamma$ if j is large enough. So we can apply (d) with $h = \bar\gamma\alpha_j$ and (v_j,λ_j) in the role of (v,λ), and conclude that $V(\bar\gamma\alpha_j v_j) + \bar\gamma\alpha_j\lambda_j \ge \bar\gamma\alpha_j\bar\gamma$. On the other hand, $(1-\alpha_j)\ell_j \ge (1-\alpha_j)\bar\gamma$, because $\ell_j \ge \bar\gamma$. Therefore

$$V(\bar\gamma\alpha_j v_j) + \bar\gamma\alpha_j\lambda_j + \bar\gamma(1-\alpha_j)\ell_j \ge \bar\gamma\alpha_j\bar\gamma + \bar\gamma(1-\alpha_j)\bar\gamma = \bar\gamma^2$$

for large enough j. If we let $j\to\infty$, and use the facts that $\alpha_j v_j \underset{j\to\infty}{\longrightarrow} 0$, V is continuous, and $V(0)=0$, we find that $\bar\gamma\lambda = \lim_{j\to\infty}(\bar\gamma\alpha_j\lambda_j + \bar\gamma(1-\alpha_j)\ell_j) \ge \bar\gamma^2$, so $\lambda \ge \bar\gamma$. Therefore $w = (0,\lambda)$ and $\lambda \ge \bar\gamma$, so $w \in \Lambda$ and then $w \in \Psi_\delta(0)$.

We have now completed the proofs of (b), (c), (d) and (e). Let w^* be the member of $\Psi_{\bar\gamma}(0)$ such that

$$\|w^*\| \le \|w\| \qquad \text{whenever} \qquad w \in \Psi_{\bar\gamma}(0)\,.$$

(The existence and uniqueness of w^* follows from the fact that $\Psi_{\bar\gamma}(0)$ is closed and convex.) Since $(0,0) \notin \Psi_{\bar\gamma}(0)$, it follows that $w^* \ne (0,0)$. Furthermore, the inequality

$$\langle w^*, w\rangle \ge \|w^*\|^2 \tag{XVI.32}$$

holds whenever $w \in \Psi_{\bar\gamma}(0)$, because if $w \in \Psi_{\bar\gamma}(0)$ then $\|w^* + t(w-w^*)\|^2 \ge \|w^*\|^2$ whenever $0 \le t \le 1$, since $w^* + t(w-w^*) \in \Psi_{\bar\gamma}(0)$ for such t, and then

$$\|w^*\|^2 + t^2\|w-w^*\|^2 + 2t\langle w^*, w-w^*\rangle \ge \|w^*\|^2\,,$$

so $t^2\|w-w^*\|^2 + 2t\langle w^*, w-w^*\rangle \ge 0$, which implies $t\|w-w^*\|^2 + 2\langle w^*, w-w^*\rangle \ge 0$ if $0 < t \le 1$, and then $\langle w^*, w-w^*\rangle \ge 0$ (since we can let $t \downarrow 0$), so (XVI.32) holds.

We now let Q be the set of all vectors $q \in \mathbb{R}^{n+1}$ such that $q = hw$ for some h, w such that $h \in [0, \infty[$ and $w \in \Psi_{\bar{\gamma}}(0)$. We will show that

(f) *Q is a closed convex cone such that $Q \backslash \{(0,0)\} \neq \emptyset$.*

(g) $\langle w^*, q \rangle \geq 0$ *for all* $q \in Q$ *and* $\langle w^*, q \rangle > 0$ *for all* $q \in Q \backslash \{(0,0)\}$.

(h) *there exist real constants κ_-, κ_+, such that $0 < \kappa_- \leq \kappa_+$ and $\kappa_- \|q\| \leq \langle w^*, q \rangle \leq \kappa_+ \|q\|$ whenever $q \in Q$.*

Indeed, Q is obviously a convex cone. The fact that $Q \backslash \{(0,0)\} \neq \emptyset$ follows because $\Psi_{\bar{\gamma}}(0) \subseteq Q$, $\Psi_{\bar{\gamma}}(0) \neq \emptyset$, and $(0,0) \notin \Psi_{\bar{\gamma}}(0)$. To show that Q is closed, we pick a sequence $\{q_j\}_{j=1}^{\infty}$of points of Q that converges to a limit $q \in \mathbb{R}^{n+1}$, and show that $q \in Q$. Write $q_j = h_j w_j$, $h_j \geq 0$, $w_j \in \Psi_{\bar{\gamma}}(0)$. If $q = (0,0)$ then $q \in Q$, so we may assume that $q \neq (0,0)$ and that $q_j \neq (0,0)$ for all j. Then $h_j \neq 0$ as well. If the sequence $\{w_j\}_{j=1}^{\infty}$ is bounded, then we may pass to a subsequence and assume that the w_j converge to a limit w, which must belong to $\Psi_{\bar{\gamma}}(0)$ because $\Psi_{\bar{\gamma}}(0)$ is closed. In particular, w and the w_j are $\neq (0,0)$. But then $h_j = \frac{\|q_j\|}{\|w_j\|} \underset{j \to \infty}{\longrightarrow} \frac{\|q\|}{\|w\|} \stackrel{\text{def}}{=} h$. Therefore $q = hw$, so $q \in Q$. Now suppose that the sequence $\{w_j\}_{j=1}^{\infty}$ is unbounded. Write $w_j = (v_j, \lambda_j)$, and use (c) to conclude that $\|v_j\|^r \leq \lambda_j + C$ for all j. Then the sequence $\{\lambda_j\}_{j=1}^{\infty}$ is unbounded, and we may assume, after passing to a subsequence, that $\lambda_j \underset{j \to \infty}{\longrightarrow} +\infty$. Since $q = \lim_{j \to \infty}(h_j v_j, h_j \lambda_j)$, the sequence $\{h_j \lambda_j\}_{j=1}^{\infty}$ converges to a finite limit μ, so $h_j \underset{j \to \infty}{\longrightarrow} 0$. Then $h_j \|v_j\| \leq h_j (\lambda_j + C)^{1/r} = h_j \lambda_j \lambda_j^{\frac{1}{r}-1} \left(1 + \frac{C}{\lambda_j}\right)^{1/r} \underset{j \to \infty}{\longrightarrow} 0$. So $h_j v_j \underset{j \to \infty}{\longrightarrow} 0$. Therefore $q = (0, \mu)$, so $q = h(0, \bar{\gamma})$, where $h = \frac{\mu}{\bar{\gamma}}$. Since $\mu \geq 0$ and $(0, \bar{\gamma}) \in \Psi_{\bar{\gamma}}(0)$, it is now clear that $q \in Q$. This completes the proof of (f).

The fact that $\langle w^*, q \rangle \geq 0$ for all $q \in Q$ follows trivially from the definition of Q, because if $q \in Q$ then $q = hw$ for some $w \in \Psi_{\bar{\gamma}}(0)$ and some nonnegative h, so $\langle w^*, q \rangle = h \langle w^*, w \rangle \geq h \|w^*\|^2 \geq 0$. Furthermore, if $q \neq (0,0)$ then $h \neq 0$, so $h > 0$, and then $\langle w^*, q \rangle \geq h \|w^*\|^2 > 0$, since $w^* \neq 0$. This proves (g).

Let $K = \{q \in Q : \|q\| = 1\}$. Then K is compact, so the continuous function $K \ni q \mapsto \langle w^*, q \rangle \in \mathbb{R}$ attains a minimum value κ_- and a maximum value κ_+ on K. Clearly, $\kappa_- > 0$, because $\langle w^*, q \rangle > 0$ for all $q \in K$. Furthermore, $\kappa_- \|q\| \leq \langle w^*, q \rangle \leq \kappa_+ \|q\|$ for all $q \in Q$, because the inequalities hold when $\|q\| = 1$ and involve functions of q that are positively homogeneous of degree 1. This proves (h).

Next, we define a function $\sigma : Q \mapsto \mathbb{R}$ by letting $\sigma(q)$ be, if $q \in Q$, the largest $h \in \mathbb{R}$ such that $q = hw$ for some $w \in \Psi_{\bar{\gamma}}(0)$. (The existence of such a largest h is trivial if $q=0$, for in that case the only possible value of h is 0, since $(0,0) \notin \Psi_{\bar{\gamma}}(0)$. If $q \in Q$ and $q \neq 0$, let $H = \{h \in \mathbb{R} : h > 0, h^{-1}q \in \Psi_{\bar{\gamma}}(0)\}$. Then H must be bounded, for otherwise $(0,0)$ would be a limit of points of $\Psi_{\bar{\gamma}}(0)$, and then $(0,0)$ would have to belong to $\Psi_{\bar{\gamma}}(0)$. If $h = \sup H$, then the fact that $\Psi_{\bar{\gamma}}(0)$ is closed implies that $h \in H$, so $\sigma(q)$ exists and is equal to h.) We prove the following properties of σ.

(i) *σ is strictly positive on $Q \backslash \{(0,0)\}$.*

(j) *σ is positively homogeneous of degree* 1 *(that is, $\sigma(rq) = r\sigma(q)$ whenever $q \in Q$ and $r \geq 0$).*

(k) *There exists a constant $\kappa \in \mathbb{R}$ such that $\sigma(q) \leq \kappa \|q\|$ whenever $q \in Q$.*

Statements (i) and (j) are immediate consequences of the definition of σ. To prove (k), we assume it is not true, and find a sequence $\{q_m\}_{m=1}^{\infty}$ of points of Q such that $\sigma(q_m) > m\|q_m\|$ for all m. We then write $q_m = \sigma(q_m)w_m$, with $w_m \in \Psi_{\bar{\gamma}}(0)$, and use (j) to conclude that $\sigma(w_m) = 1$ and $\sigma(w_m) > m\|w_m\|$ for all m. Then $\|w_m\| < \frac{1}{m}$, so $w_m \to (0,0)$ as $m \to \infty$. Since $w_m \in \Psi_{\bar{\gamma}}(0)$, and $\Psi_{\bar{\gamma}}(0)$ is closed, we conclude that $(0,0) \in \Psi_{\bar{\gamma}}(0)$, contradicting (e). This completes the proof of (k).

Next, we define

$$\begin{aligned} \gamma^* &\stackrel{\text{def}}{=} \frac{\bar{\gamma}}{\max(1,\kappa)} \\ Q^* &\stackrel{\text{def}}{=} \left\{ q \in Q : \|q\| \leq \gamma^* \right\}. \end{aligned}$$

Then Q^* is a convex, compact subset of $\mathbb{R}^{n+1}$ such that $(0,0) \in Q^*$ but Q^* contains at least one point other than $(0,0)$ (because of (f)). In addition, if $q = (x, x_0) \in Q^*$, then $\|x\| \leq \gamma^* \leq \bar{\gamma}$, so (XVI.25) tells us that $x \in \Omega$, and then $q \in \Omega \times \mathbb{R}$. Hence $Q^* \subseteq \Omega \times \mathbb{R}$.

We then define a function $W : \Omega \times \mathbb{R} \mapsto \mathbb{R}$ by letting $W(x,x_0) = V(x) + x_0$ for $(x,x_0) \in \Omega \times \mathbb{R}$, and observe that $W(q)$ is defined whenever $q \in Q^*$, because $Q^* \subseteq \Omega \times \mathbb{R}$.

We then claim that

$$W(q) \geq \bar{\gamma}\sigma(q) \quad \text{whenever } q \in Q^*. \tag{XVI.33}$$

To prove (XVI.33), we first observe that the inequality is clearly true if $q = 0$. Let us pick $q \in Q^* \backslash \{0\}$ and write $q = hw$, where $h = \sigma(q)$ and $w = (v, \lambda) \in \Psi_{\bar{\gamma}}(0)$.

Then $0 < h$, since $q \neq 0$, and $h \leq \bar{\gamma}$, because $\sigma(q) \leq \kappa\|q\| \leq \kappa\gamma^* \leq \bar{\gamma}$. Furthermore, $\|q\| \leq \bar{\gamma}$ (because $\gamma^* \leq \bar{\gamma}$), and $q = (hv, h\lambda)$, so $h\|v\| \leq \bar{\gamma}$. It then follows from (XVI.26) that $W(q) = V(hv) + h\lambda \geq h\bar{\gamma} = \bar{\gamma}\sigma(q)$, completing the proof of (XVI.33).

It follows from (XVI.33) and (i) that

$$W(q) > 0 \quad \text{whenever } q \in Q^* \backslash \{(0,0)\}\,. \tag{XVI.34}$$

Now fix a number ρ such that $0 < \rho \leq \kappa_- \gamma^*$, and define

$$Q^\# = \{q \in Q^* : \langle w^*, q\rangle \geq \rho\}\,. \tag{XVI.35}$$

If q is any member of Q such that $\|q\| = \gamma^*$, then $q \in Q^*$, and in addition $\langle w^*, q\rangle \geq \kappa_- \|q\| = \kappa_- \gamma^* \geq \rho$, so $q \in Q^\#$. Hence $Q^\#$ is nonempty, and it is clear that $Q^\#$ is compact and convex. If $q \in Q^\#$, then $q \in Q^* \backslash \{(0,0)\}$, so $W(q) > 0$. It follows that $\bar{\mu} > 0$, if we let $\bar{\mu} = \min\{W(q) : q \in Q^\#\}$. Since $W(0,0) = 0$, we may pick a μ such that $0 < \mu < \bar{\mu}$, and use the Clarke-Ledyaev mean value theorem (cf. [45, 46, 47]) to conclude that if β is any positive number, and we use $\mathscr{N}_\beta$ to denote the β-neighborhood of the set $\mathscr{N}_0 \stackrel{\text{def}}{=} \{hq : q \in Q^\#, 0 \leq h \leq 1\}$, then there exists a subdifferential (π, π_0) of W at some point $q^\#$ belonging to $\mathscr{N}_\beta$ such that

$$\langle (\pi, \pi_0), q\rangle > \mu \quad \text{for all} \quad q \in Q^\#\,. \tag{XVI.36}$$

Write $q^\# = (x^\#, x_0^\#)$. Then, if we write $q = (x, x_0)$ for q near $q^\#$, we have

$$\liminf_{q \to q^\#} \frac{W(q) - W(q^\#) - \pi \cdot (x - x^\#) - \pi_0(x_0 - x_0^\#)}{\|x - x^\#\| + |x_0 - x_0^\#|} \geq 0.$$

Taking $q = (x, x_0^\#)$, this implies

$$\liminf_{x \to x^\#} \frac{V(x) - V(x^\#) - \pi \cdot (x - x^\#)}{\|x - x^\#\|} \geq 0\,,$$

so π is a subdifferential of V at $x^\#$.

Taking $q = (x^\#, x_0)$, we get

$$\liminf_{x_0 \to x_0^\#} \frac{x_0 - x_0^\# - \pi_0(x_0 - x_0^\#)}{|x_0 - x_0^\#|} \geq 0\,,$$

so $\pi_0 = 1$.

Now, if $u \in U$, and we let $w = (f(x^\#,u), L(x^\#,u))$, then $w \in \Phi_{\bar{\gamma},\hat{\Sigma}}(0)$, so $w \in Q$. Let $q = \frac{\gamma^* w}{\|w\|}$. Then $\|q\| = \gamma^*$, so $q \in Q^*$. Furthermore, $\langle w^*, q\rangle \geq \kappa_- \|q\| = \kappa_- \gamma^* \geq \rho$, so $q \in Q^\#$. Therefore

$$\langle (\pi, \pi_0), q\rangle > \mu\,, \qquad \text{(XVI.37)}$$

that is,

$$\frac{\gamma^*}{\|w\|}\Big(\langle \pi, f(x^\#,u)\rangle + L(x^\#,u)\Big) > \mu\,. \qquad \text{(XVI.38)}$$

Therefore

$$\langle \pi, f(x^\#,u)\rangle + L(x^\#,u) > \frac{\|w\|\mu}{\gamma^*} \geq \frac{\|w^*\|\mu}{\gamma^*}\,, \qquad \text{(XVI.39)}$$

so

$$-\langle \pi, f(x^\#,u)\rangle - L(x^\#,u) < -\frac{\|w^*\|\mu}{\gamma^*}\,. \qquad \text{(XVI.40)}$$

Since (XVI.40) is true for every $u \in U$, we can conclude that

$$\sup\Big\{-\langle \pi, f(x^\#,u)\rangle - L(x^\#,u) : u \in U\Big\} \leq \frac{\|w^*\|\mu}{\gamma^*} < 0\,. \qquad \text{(XVI.41)}$$

But π is a subdifferential of V at $x^\#$, and then (XVI.41) contradicts the fact that V is a solution of (XVI.13) on Ω in the viscosity sense. This contradiction establishes (3.ii) and completes our proof. □

XVI.5 The compactness theorem.

If $\hat{\Sigma} = (\Omega, U, f, L)$ is an augmented control system, and ε is a positive number, an *ε-approximate augmented trajectory of* $\hat{\Sigma}$ is a locally absolutely continuous curve $I \ni t \mapsto \Xi(t) = (\xi(t), \xi_0(t)) \in \mathbb{R}^{n+1}$ having the property that there exists a measurable function $I \ni t \mapsto v(t) \in \mathbb{R}^n$ such that

(1) $\|v(t)\| \leq \varepsilon$ for almost all $t \in I$,

(2) $\dot{\Xi}(t) - (v(t), 0) \in F_{\hat{\Sigma}}(\xi(t), U)$ for almost all $t \in I$.

Remark XVI.17. *Roughly speaking, an ε-approximate augmented trajectory of $\hat{\Sigma}$ is an augmented trajectory of the "ε-extended system" $\hat{\Sigma}^\varepsilon = (\Omega, U^\varepsilon, f^\varepsilon, L^\varepsilon)$ whose control space U^ε is the Cartesian product $U \times \{v \in \mathbb{R}^n : \|v\| \leq \varepsilon\}$, and whose dynamics f^ε and Lagrangian L^ε are given by*

$$f^\varepsilon(x,u,v) = f(x,u) + v\,, \quad L^\varepsilon(x,u,v) = L(x,u)\,.$$

More precisely, a curve $\Xi = (\xi, \xi_0) : I \mapsto \mathbb{R}^{n+1}$ is an ε-approximate augmented trajectory of $\hat{\Sigma}$ if and only if Ξ is locally absolutely continuous and there exist functions $I \ni t \mapsto \eta(t) \in U$, $I \ni t \mapsto v(t) \in \{v \in \mathbb{R}^n : \|v\| \leq \varepsilon\}$ such that v is measurable and $\dot{\Xi}(t) \in F_{\hat{\Sigma}^\varepsilon}(\xi(t), U^\varepsilon)$ for almost every $t \in I$. The definition of an augmented trajectory of $\hat{\Sigma}^\varepsilon$ is exactly the same, except that in that case the requirement that v be measurable is omitted. □

Theorem XVI.18. *Let $\hat{\Sigma} = (\Omega, U, f, L)$ be an n-dimensional locally uniformly continuous, locally coercive augmented control system such that $F_{\hat{\Sigma}}(x,U)$ is closed and convex for every $x \in \Omega$. Let K be a compact subset of Ω, let T be a positive time, and let $k \in \mathbb{R}$. Let $\{\varepsilon_j\}_{j=1}^\infty$ be a sequence of positive numbers such that $\varepsilon_j \to 0$ as $j \to \infty$, and let $\Xi^j = (\xi^j, \xi_0^j)$ be, for each j, an ε_j-approximate augmented trajectory of $\hat{\Sigma}$, defined on the interval $[0,T]$, such that $\xi^j(t) \in K$ for all $t \in [0,T]$, $\xi_0^j(0) = 0$, and $\xi_0^j(T) \leq k$. Then there exist a subsequence $\{\Xi^{j(k)}\}_{k=1}^\infty$ of the sequence $\{\Xi^j\}_{j=1}^\infty$ and an augmented trajectory $\Xi^\infty = (\xi^\infty, \xi_0^\infty)$ of $\hat{\Sigma} = (\Omega, U, f, L)$ such that*

(i) $\xi_0^\infty(0) = 0$,

(ii) the sequence $\{\xi^{j(k)}\}_{k=1}^\infty$ converges uniformly to ξ^∞, and

(iii) $\liminf_{k\to\infty} \inf\left\{\Delta^{j(k),\infty}(t,s) : 0 \leq s \leq t \leq T\right\} \geq 0$, *where, if $j, j' \in \mathbb{N} \cup \{+\infty\}$, we define*

$$\Delta^{j,j'}(t,s) \stackrel{\text{def}}{=} (\xi_0^j(t) - \xi_0^j(s)) - (\xi_0^{j'}(t) - \xi_0^{j'}(s))\,. \tag{XVI.42}$$

Proof. Pick constants r, C, such that $r > 1$, $C > 0$, and $\|f(x,u)\|^r \leq L(x,u) + C$ whenever $x \in K$, $u \in U$, and a function $\omega :]0, +\infty[\mapsto [0, +\infty]$ such that $\lim_{s\downarrow 0} \omega(s)$ and $\|F_{\hat{\Sigma}}(x,u) - F_{\hat{\Sigma}}(x',u)\| \leq \omega(s)$ whenever $x, x' \in K$ and $\|x - x'\| \leq s$. Choose, for each j, a measurable function $v^j : [0,T] \mapsto \mathbb{R}^n$ and a function $\eta^j : [0,T] \mapsto U$ such that $\|v(t)\| \leq \varepsilon_j$, $\dot{\xi}^j(t) = f(\xi^j(t), \eta^j(t)) + v(t)$, and $\dot{\xi}_0^j(t) = L(\xi^j(s), \eta^j(s))$ for almost all $t \in [0,T]$.

Let $C' = 2^r(C+\varepsilon^r)$. Then for each j the coerciveness condition implies the inequality

$$\|\dot{\xi}^j(t)\|^r \le 2^r L(\xi^j(t),\eta^j(t)) + C' \text{ for almost all } t\,,$$

because

$$\begin{aligned}\|\dot{\xi}^j(t)\|^r &= \|f(\xi^j(t),\eta^j(t)) + v(t)\|^r \\ &\le \big(\|f(\xi^j(t),\eta^j(t))\| + \|v(t)\|\big)^r \\ &\le 2^r\|f(\xi^j(t),\eta^j(t))\|^r + 2^r\|v(t)\|^r \\ &\le 2^r(L(\xi^j(t),\eta^j(t)) + C) + 2^r\varepsilon^r \\ &= 2^r L(\xi^j(t),\eta^j(t)) + C'\,.\end{aligned}$$

from which it follows that $\int_0^T \|\dot{\xi}^j(t)\|^r dt \le 2^r k + C'T$. Then the sequence $\{\dot{\xi}^j(t)\}_{j=1}^\infty$ is uniformly bounded in L^r, so we may assume, after passing to a subsequence if necessary, that the weak L^r-limit $\zeta = \text{w-lim}_{j\to\infty}\dot{\xi}^j$ exists. Then the ξ^j converge uniformly as $j\to\infty$ to a limit ξ^∞ such that $\xi^\infty(t) - \xi^\infty(s) = \int_s^t \zeta(r)\,dr$ for all $s,t \in [0,T]$. After passing to a subsequence once more, if necessary, we assume that

$$\|\xi^j(t) - \xi^\infty(t)\| \le 2^{-j} \text{ for all } j\in\mathbb{N},\ t\in[0,T]\,. \tag{XVI.43}$$

Let $\theta^j(t) = \dot{\xi}_0^j(t) + C$. Then the θ^j are nonnegative, because $\dot{\xi}_0^j(t) = L(\xi^j(t),\eta^j(t)) \ge -C$. Furthermore, the sequence $\{\theta^j\}_{j=1}^\infty$ is bounded in $L^1([0,T],\mathbb{R})$, because

$$\begin{aligned}\|\theta^j\|_{L^1} &= \int_0^T \theta^j(t)\,dt \\ &= \int_0^T (\dot{\xi}_0^j(t) + C)\,dt \\ &\le k + CT\,.\end{aligned}$$

The space $L^1([0,T],\mathbb{R})$ can be embedded in the usual way in $C^0([0,T],\mathbb{R})^\dagger$ (the dual of $C^0([0,T],\mathbb{R})$), which is the space of finite Borel measures on $[0,T]$, by means of the map $\psi \mapsto \mu_\psi$ that assigns to each function $\psi \in L^1([0,T],\mathbb{R})$ the Borel measure μ_ψ such that

$$\mu_\psi(\varphi) = \int_0^T \psi(t)\varphi(t)dt \text{ for every } \varphi \in C^0([0,T],\mathbb{R})\,.$$

Then we may assume, after passing to a subsequence for a third time, if necessary, that the weak* limit μ^∞ of the measures $\mu^j \stackrel{\text{def}}{=} \mu_{\theta^j}$ defined by the θ^j exists as $j\to\infty$.

The measure μ^∞ then has a decomposition

$$\mu^\infty = \mu^{\infty,ac} + \mu^{\infty,at} + \mu^{\infty,sing}$$

into the sum of an absolutely continuous part, an atomic part, and a singular part. Since the measure μ^∞ is positive, because it is a limit of positive measures, the three components $\mu^{\infty,ac}$, $\mu^{\infty,at}$, $\mu^{\infty,sing}$ are positive as well. Let θ^∞ be the Radon-Nikodym derivative of μ^∞, so θ^∞ is an integrable function on $[0,T]$ such that

$$\mu^{\infty,ac}(\varphi) = \int_0^T \theta^\infty(t)\varphi(t)\,dt \text{ for } \varphi \in C^0([0,T],\mathbb{R}) .$$

Define

$$\xi_0^\infty(t) = -Ct + \int_0^t \theta^\infty(s)ds \qquad \text{(XVI.44)}$$

and then set $\Xi^\infty = (\xi^\infty, \xi_0^\infty)$.

We will show that Ξ^∞ is an augmented trajectory of the system $\hat{\Sigma}$. To see this, we observe first of all that by construction the functions $\xi^\infty : [0,T] \mapsto \mathbb{R}^n$ and $\xi_0^\infty : [0,T] \mapsto \mathbb{R}$ are absolutely continuous, and their derivatives at t are equal to $\zeta(t)$ and $\theta^\infty(t)$, respectively, for all t in a subset G of $[0,T]$ such that $[0,T]\backslash G$ has measure zero. Let A be the set of atoms of $\mu^{\infty,at}$, and let B be a subset of $[0,T]$ of Lebesgue measure zero such that $\mu^{\infty,sing}([0,T]\backslash B) = 0$. Then the set $G' = G\backslash(A \cup B \cup \{T\})$ has measure T and $\mu^{\infty,at}(G') = \mu^{\infty,sing}(G') = 0$. Let G'' be the set of points of density of G' that are Lebesgue points of ζ and θ^∞, so G'' has measure T as well. (Recall that a *Lebesgue point* of a scalar- or vector-valued integrable function σ defined on an interval $[a,b]$ is a point $t \in]a,b[$ such that $\lim_{h\downarrow 0} \frac{1}{h}\int_{t-h}^{t+h} \|\sigma(s) - \sigma(t)\|\,ds = 0$.) Let $t \in G''$, and fix an h such that $0 < h < T - t$. Let $E_{t,h} = [t,t+h] \cap G'$, so the Lebesgue measure $|E_{t,h}|$ of $E_{t,h}$ satisfies $\lim_{h\downarrow 0} h^{-1}|E_{t,h}| = 1$. Using the facts that the Borel measure $\mu^{\infty,at} + \mu^{\infty,sing}$ is regular and $(\mu^{\infty,at} + \mu^{\infty,sing})(E_{t,h}) = 0$, we can find a relatively open subset $\tilde{U}_{t,h}$ of $[0,T]$ such that $E_{t,h} \subseteq \tilde{U}_{t,h}$ and $(\mu^{\infty,at} + \mu^{\infty,sing})(\tilde{U}_{t,h}) \leq h^2$. We then let $U_{t,h} = \tilde{U}_{t,h} \cap]t,t+h[$, so $U_{t,h}$ is an open subset of $\mathbb{R}$, $U_{t,h} \subseteq]t,t+h[$, and $(\mu^{\infty,at} + \mu^{\infty,sing})(U_{t,h}) \leq h^2$. Using the regularity of Lebesgue measure we can find a compact subset $K_{t,h}$ of $E_{t,h}\backslash\{t,t+h\}$ such that $|K_{t,h}| \geq |E_{t,h}| - h^2$. Then of course $\lim_{h\downarrow 0} h^{-1}|K_{t,h}| = 1$, and $K_{t,h} \subseteq U_{t,h}$ for each h. Let $\tilde{\varphi}_{t,h}(s) = \text{dist}(s, \mathbb{R}\backslash U_{t,h})$, so $\tilde{\varphi}_{t,h} : \mathbb{R} \mapsto \mathbb{R}$ is continuous, $\tilde{\varphi}_{t,h}(s) = 0$ whenever $s \notin U_{t,h}$, and $\tilde{\varphi}_{t,h}(s) > 0$ whenever $s \in K_{t,h}$. If we let

$$\begin{aligned}
\beta_{t,h} &= \min\{\tilde{\varphi}_{t,h}(s) : s \in K_{t,h}\} , \\
\hat{\varphi}_{t,h}(s) &= \min(\tilde{\varphi}_{t,h}(s), \beta_{t,h}) , \\
\varphi_{t,h}(s) &= \beta_{t,h}^{-1}\hat{\varphi}_{t,h}(s) ,
\end{aligned}$$

then $\varphi_{t,h}$ is a continuous real-valued function on $\mathbb{R}$ such that $0 \le \varphi_{t,h}(s) \le 1$ for all s, $\varphi_{t,h}(s) = 1$ for all $s \in K_{t,h}$, and $\varphi_{t,h}(s) = 0$ for all $s \in \mathbb{R}\backslash U_{t,h}$. In particular, $\varphi_{t,h}(s) = 0$ whenever $s \notin]t, t+h[$.

Let $a_{t,h} = \int_t^{t+h} \varphi_{t,h}(s)\,ds = \int_{-\infty}^{+\infty} \varphi_{t,h}(s)\,ds$. Then

$$|E_{t,h}| - h^2 \le a_{t,h} \le h\,,$$

from which it follows that

$$\lim_{h\downarrow 0} \frac{a_{t,h}}{h} = 1\,.$$

Let $\psi_{t,h} = a_{t,h}^{-1}\varphi_{t,h}$. Then

$$\int_t^{t+h} \psi_{t,h}(s)\,ds = 1\,.$$

If $h > 0$, write

$$\delta_j(h) = 2^{-j} + h^{1/\rho}(2^r|k| + C'T)^{1/r}\,.$$

Then, if $s \in [t, t+h]$, we have

$$\begin{aligned}
&\|\xi^j(s) - \xi^\infty(t)\| \\
\le\ & \|\xi^j(t) - \xi^\infty(t)\| + \|\xi^j(s) - \xi^j(t)\| \\
\le\ & 2^{-j} + \int_t^s \|\dot{\xi}^j(\tau)\|\,dt \\
\le\ & 2^{-j} + (s-t)^{1/\rho}\Big(\int_t^s \|\dot{\xi}^j(\tau)\|^r\,dt\Big)^{1/r} \\
\le\ & 2^{-j} + h^{1/\rho}\Big(\int_0^T \|\dot{\xi}^j(\tau)\|^r\,dt\Big)^{1/r} \\
\le\ & 2^{-j} + h^{1/\rho}\Big(\int_0^T (2^r L(\xi^j(\tau), \eta^j(\tau)) + C')\,d\tau\Big)^{1/r} \\
\le\ & 2^{-j} + h^{1/\rho}(2^r|k| + C'T)^{1/r} \\
=\ & \delta_j(h)\,.
\end{aligned}$$

For almost all $s \in [t, t+h]$, the derivative $\dot{\Xi}^j(s)$ exists and is equal to $F_{\hat{\Sigma}}(\xi^j(s), \eta^j(s)) + (v^j(s), 0)$. Hence

$$\|\dot{\Xi}^j(s) - F_{\hat{\Sigma}}(\xi^\infty(t), \eta^j(s))\| \le \omega(\delta_j(h)) + \varepsilon_j\,,$$

from which it follows that

$$\mathrm{dist}(\dot{\Xi}^j(s), F_{\hat{\Sigma}}(\xi^\infty(t), U) \le \omega(\delta_j(h)) + \varepsilon_j\,.$$

Therefore the average

$$A^j_{t,h} = \int_t^{t+h} \psi_{t,h}(s)\dot{\Xi}^j(s)\,ds = \int_0^T \psi_{t,h}(s)\dot{\Xi}^j(s)\,ds$$

also satisfies

$$\mathrm{dist}(A^j_{t,h}, F_{\hat{\Sigma}}(\xi^\infty(t),U) \leq \omega(\delta_j(h)) + \varepsilon_j\,, \tag{XVI.45}$$

because $F_{\hat{\Sigma}}(\xi^\infty(t),U)$ is closed and convex. As $j \to \infty$, the vector functions $\dot{\xi}^j$ converge weakly in L^r to ζ, so

$$\begin{aligned} \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^j(s)\,ds \quad &\to \quad \int_t^{t+h} \psi_{t,h}(s)\zeta(s)\,ds \\ &= \quad \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^\infty(s)\,ds \end{aligned}$$

as $j \to \infty$, $j \in J$. The integral $\int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^\infty(s)\,ds$ satisfies

$$\begin{aligned} \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^\infty(s)\,ds \quad &= \quad \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^\infty(t)\,ds + \mathscr{E}_{t,h} \\ &= \quad \dot{\xi}^\infty(t) + \mathscr{E}_{t,h} \end{aligned}$$

where

$$\mathscr{E}_{t,h} = \int_t^{t+h} \psi_{t,h}(s)(\dot{\xi}^\infty(s) - \dot{\xi}^\infty(t))\,ds\,,$$

and we have used the fact that $\int_t^{t+h} \psi_{t,h}(s)\,ds = 1$.

The error term $\mathscr{E}_{t,h}$ satisfies

$$\begin{aligned} |\mathscr{E}_{t,h}| \quad &\leq \quad \max\big\{\psi_{t,h}(s) : s \in [0,T]\big\} \\ &\quad \times \int_t^{t+h} \|\dot{\xi}^\infty(s) - \dot{\xi}^\infty(t)\|\,ds \\ &= \quad h a^{-1}_{t,h}\hat{\mathscr{E}}_{t,h} \\ &= \quad \alpha_{t,h}\hat{\mathscr{E}}_{t,h}\,, \end{aligned}$$

where $\alpha_{t,h} = h a^{-1}_{t,h}$, so $\alpha_{t,h} \to 1$ as $h \downarrow 0$, and

$$\hat{\mathscr{E}}_{t,h} = \frac{1}{h}\int_t^{t+h} \|\dot{\xi}^\infty(s) - \dot{\xi}^\infty(t)\|\,ds\,.$$

Then

$$\begin{aligned}\lim \sup_{j\to\infty, j\in J} & \left\| \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^j(s)\,ds - \dot{\xi}^\infty(t)\right\| \\ \leq \quad & \alpha_{t,h}\hat{\mathscr{E}}_{t,h}\,. \end{aligned} \tag{XVI.46}$$

To analyze the behavior of the integrals

$$I^j_{t,h} \stackrel{\text{def}}{=} \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}^j_0(s)\,ds\,,$$

we write $\dot{\xi}^j_0(s) = \theta^j(s) - C$, so

$$\begin{aligned} I^j_{t,h} &= -Ch + \int_t^{t+h} \psi_{t,h}(s)\theta^j(s)\,ds \\ &= -Ch + \int_0^T \psi_{t,h}(s)\theta^j(s)\,ds \\ &= -Ch + \int_{[0,T]} \psi_{t,h}d\mu^j\,. \end{aligned}$$

Hence

$$I^j_{t,h} \to -Ch + \int_{[0,T]} \psi_{t,h}(s)\,d\mu^\infty(s) \text{ as } j\to\infty\,. \tag{XVI.47}$$

Write

$$\begin{aligned}\int_{[0,T]} \psi_{t,h}(s)\,d\mu^\infty(s) \quad &= \quad \int_0^T \psi_{t,h}(s)\theta^\infty(s)\,ds \\ & \qquad + \int_{[0,T]} \psi_{t,h}(s)\,d\hat{\mu}^\infty(s)\,,\end{aligned}$$

where $\hat{\mu}^\infty = \mu^{\infty,at} + \mu^{\infty,sing}$. Since $\dot{\xi}^\infty_0(s) = \theta^j(s) - C$, (XVI.47) implies

$$\begin{aligned} I^j_{t,h} \quad \to \quad & \int_{[0,T]} \psi_{t,h}(s)\dot{\xi}^\infty_0(s)\,ds \\ & + \int_{[0,T]} \psi_{t,h}(s)\,d\hat{\mu}^\infty(s) \text{ as } j\to\infty\,. \end{aligned} \tag{XVI.48}$$

The integral $\int_{[0,T]} \psi_{t,h}(s)\,d\hat{\mu}^\infty(s)$ that occcurs in (XVI.48) is a nonnegative number, and is bounded above by $\max\{\psi_{t,h}(s) : s \in [0,T]\}$ times $\hat{\mu}^\infty(U_{t,h})$, since $\psi_{t,h}$ vanishes outside $U_{t,h}$. Therefore

$$0 \leq \int_{[0,T]} \psi_{t,h}(s)\,d\hat{\mu}^\infty(s) \leq h^2 a^{-1}_{t,h} = h\alpha_{t,h}\,, \tag{XVI.49}$$

On the other hand, the integral $\int_{[0,T]} \psi_{t,h}(s)\dot{\xi}_0^\infty(s)\,ds$ satisfies

$$\begin{aligned}\int_{[0,T]} \psi_{t,h}(s)\dot{\xi}_0^\infty(s)\,ds &= \int_t^{t+h} \psi_{t,h}(s)\dot{\xi}_0^\infty(s)\,ds\\ &= \dot{\xi}_0^\infty(t))\,ds + E_{t,h}\\ &= \dot{\xi}_0^\infty(t) + E_{t,h}\,,\end{aligned} \tag{XVI.50}$$

where

$$E_{t,h} = \int_t^{t+h} \psi_{t,h}(s)(\dot{\xi}_0^\infty(s) - \dot{\xi}_0^\infty(t))\,ds\,,$$

and we have used the fact that $\int_t^{t+h} \psi_{t,h}(s)\,ds = 1$.

The error term $E_{t,h}$ satisfies

$$\begin{aligned}|E_{t,h}| &\leq \max\left\{\psi_{t,h}(s) : s \in [0,T]\right\}\\ &\qquad \times \int_t^{t+h} |\dot{\xi}_0^\infty(s) - \dot{\xi}_0^\infty(t)|\,ds\\ &= ha_{t,h}^{-1}\hat{E}_{t,h}\\ &= \alpha_{t,h}\hat{E}_{t,h}\,,\end{aligned}$$

where

$$\hat{E}_{t,h} = \frac{1}{h}\int_t^{t+h} |\dot{\xi}_0^\infty(s) - \dot{\xi}_0^\infty(t)|\,ds\,.$$

It follows from (XVI.48), (XVI.49), (XVI.50), and the bound $|E_{t,h}| \leq \alpha_{t,h}\hat{E}_{t,h}$, that

$$\lim \sup_{j\to\infty, j\in J} \left| I_{t,h}^j - \dot{\xi}_0^\infty(t) \right| \leq \alpha_{t,h}(h + \hat{E}_{t,h})\,. \tag{XVI.51}$$

If we now combine (XVI.46) and (XVI.51), we find that

$$\limsup_{j\to\infty, j\in J} \left\| A_{t,h}^j - \dot{\Xi}^\infty(t) \right\| \leq \alpha_{t,h}(h + \hat{E}_{t,h} + \mathscr{E}_{t,h})\,. \tag{XVI.52}$$

Then (XVI.45) implies that

$$\begin{aligned}&\mathrm{dist}(\dot{\Xi}^\infty(t), F_{\hat{\Sigma}}(\xi^\infty(t)))\\ &\qquad \leq \alpha_{t,h}(h + \hat{E}_{t,h} + \mathscr{E}_{t,h}) + \limsup_{j\to\infty}(\omega(\delta_j(h)) + \varepsilon_j)\,.\end{aligned}$$

Hence, given any j_*, we have

$$\begin{aligned}&\mathrm{dist}(\dot{\Xi}^\infty(t),F_{\hat{\Sigma}}(\xi^\infty(t)))\\&\quad\leq \alpha_{t,h}(h+\hat{E}_{t,h}+\mathscr{E}_{t,h})+\sup\{\omega(\delta_j(h))+\varepsilon_j : j\geq j_*\}\,.\end{aligned}$$

Given any positive number β, we can find a positive γ such that $\omega(s)<\beta$ whenever $0<s\leq\gamma$, and then find j_*, h_* such that $\delta_j(h)\leq\gamma$ whenever $j\geq j_*$ and $0<h\leq h_*$, and $\varepsilon_j<\beta$ whenever $j\geq j_*$. Then we can pick h such that $0<h\leq h_*$, and $\alpha_{t,h}(h+\hat{E}_{t,h}+\mathscr{E}_{t,h})<\beta$. Then

$$\mathrm{dist}(\dot{\Xi}^\infty(t),F_{\hat{\Sigma}}(\xi^\infty(t)))<3\beta\,.$$

Since β was arbitrary, we conclude that $\mathrm{dist}(\dot{\Xi}^\infty(t),F_{\hat{\Sigma}}(\xi^\infty(t)))=0$, so $\dot{\Xi}^\infty(t)\in F_{\hat{\Sigma}}(\xi^\infty(t))$, because $F_{\hat{\Sigma}}(\xi^\infty(t))$ is closed. Since this is true for almost all $t\in[0,T]$, and Ξ^∞ is absolutely continuous, we have shown that Ξ^∞ is an augmented trajectory of $\hat{\Sigma}$.

By construction, the ξ^j converge uniformly to ξ^∞. Also, it is clear from (XVI.44) that $\xi_0^\infty(0)=0$. To conclude our proof, we have to show that

$$\liminf{}_{j\to\infty}\inf\left\{\Delta^{j,\infty}(t,s):0\leq s\leq t\leq T\right\}\geq 0\,, \qquad \text{(XVI.53)}$$

where $\Delta^{j,\infty}(t,s)$ is the quantity defined in (XVI.42). Suppose that $\liminf_{j\to\infty}\inf\left\{\Delta^{j,\infty}(t,s):0\leq s\leq t\leq T\right\}<0$. Pick a number β such that $\beta>0$ and

$$\liminf_{j\to\infty}\inf\left\{\Delta^{j,\infty}(t,s):0\leq s\leq t\leq T\right\}\leq -3\beta\,.$$

Then there exists a subsequence $\{\Xi^{j(k)}\}_{k=1}^\infty$ of $\{\Xi^j\}_{j=1}^\infty$ such that $\inf\left\{\Delta^{j,\infty}(t,s):0\leq s\leq t\leq T\right\}\leq -2\beta$ for all k. We can then choose, for each k, members s_k and t_k of $[0,T]$ such that $s_k\leq t_k$ and

$$\left(\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k)\right)-\left(\xi_0^\infty(t_k)-\xi_0^\infty(s_k)\right)\leq -\beta\,. \qquad \text{(XVI.54)}$$

By passing to a subsequence, if necessary, we may assume that the s_k and the t_k converge to limits s, t. Clearly, then, $s\leq t$.

If $s=t$, then

$$\begin{aligned}\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k) &= \int_{s_k}^{t_k}\dot{\xi}_0^{j(k)}(v)\,dv\\ &= \int_{s_k}^{t_k}\left(\theta^{j(k)}(v)-C\right)dv\\ &\geq -C(t_k-s_k)\,,\end{aligned}$$

so the fact that $\lim_{k\to\infty}(\xi_0^\infty(t_k) - \xi_0^\infty(s_k)) = 0$ implies the inequalities

$$\begin{aligned}
&\liminf_{k\to\infty}\Big(\big(\xi_0^{j(k)}(t_k) - \xi_0^{j(k)}(s_k)\big) - \big(\xi_0^\infty(t_k) - \xi_0^\infty(s_k)\big)\Big) \\
&= \liminf_{k\to\infty}\big(\xi_0^{j(k)}(t_k) - \xi_0^{j(k)}(s_k)\big) - \lim_{k\to\infty}\big(\xi_0^\infty(t_k) - \xi_0^\infty(s_k)\big) \\
&= \liminf_{k\to\infty}\big(\xi_0^{j(k)}(t_k) - \xi_0^{j(k)}(s_k)\big) \\
&\geq \liminf_{k\to\infty}(-C(t_k - s_k)) \\
&= \lim_{k\to\infty}(-C(t_k - s_k)) \\
&= 0,
\end{aligned}$$

which clearly contradict (XVI.54).

Now assume that $s < t$. Fix a positive number γ such that $2\gamma < t - s$, and let $\Phi_{s,t,\gamma}$ be the set of all continuous nonnegative functions $\varphi : \mathbb{R} \mapsto \mathbb{R}$ that vanish outside the interval $[s+\gamma, t-\gamma]$ and are such that $\varphi(v) \leq 1$ for all $v \in \mathbb{R}$. Then

$$
\begin{aligned}
&\liminf_{k\to\infty}\Big(\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k)\Big)\\
&= \liminf_{k\to\infty}\int_{s_k}^{t_k}\dot{\xi}_0^{j(k)}(v)\,dv\\
&= \liminf_{k\to\infty}\int_{s_k}^{t_k}(\theta^{j(k)}(v)-C)\,dv\\
&= \liminf_{k\to\infty}\Big(-C(t_k-s_k)+\int_{s_k}^{t_k}\theta^{j(k)}(v)\,dv\Big)\\
&= -C(t-s)+\liminf_{k\to\infty}\int_{s_k}^{t_k}\theta^{j(k)}(v)\,dv\\
&\geq -C(t-s)+\liminf_{k\to\infty}\int_{s_k}^{t_k}\varphi(v)\theta^{j(k)}(v)\,dv\\
&= -C(t-s)+\liminf_{k\to\infty}\int_{[0,T]}\varphi(v)d\mu^{j(k)}(v)\\
&= -C(t-s)+\lim_{k\to\infty}\int_{[0,T]}\varphi(v)d\mu^{j(k)}(v)\\
&= -C(t-s)+\int_{[0,T]}\varphi(v)d\mu^{\infty}(v)\\
&\geq -C(t-s)+\int_{[0,T]}\varphi(v)d\mu^{\infty,ac}(v)\\
&= -C(t-s)+\int_{s}^{t}\varphi(v)\theta^{\infty}(v)\,dv\\
&= \int_{s}^{t}\varphi(v)(\theta^{\infty}(v)-C)\,dv\\
&= \int_{s}^{t}\varphi(v)\dot{\xi}_0^{\infty}(v)\,dv\,,
\end{aligned}
$$

where, for the first inequality, we have used the fact that $[s+\gamma,t-\gamma]\subseteq[s_k,t_k]$ when k is large enough. Therefore

$$\liminf_{k\to\infty}\Big(\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k)\Big)\geq\int_{s}^{t}\varphi(v)\dot{\xi}_0^{\infty}(v)\,dv$$

for every $\varphi\in\Phi_{s,t,\gamma}$. Hence

$$\begin{aligned}
&\liminf_{k\to\infty}\Big(\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k)\Big)\\
&\qquad\geq\quad \sup\Big\{\int_s^t \varphi(v)\dot{\xi}_0^\infty(v)\,dv : \varphi\in\Phi_{s,t,\gamma}\Big\}\\
&\qquad=\quad \int_{s+\gamma}^{t-\gamma}\dot{\xi}_0^\infty(v)\,dv\\
&\qquad=\quad \xi_0^\infty(t-\gamma)-\xi_0^\infty(s+\gamma)\,.
\end{aligned}$$

Since γ is arbitrary, we can let $\gamma\downarrow 0$, and conclude that

$$\begin{aligned}
&\liminf_{k\to\infty}\Big(\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k)\Big)\\
&\qquad\geq\quad \xi_0^\infty(t)-\xi_0^\infty(s)\\
&\qquad=\quad \lim_{k\to\infty}\Big(\xi_0^\infty(t_k)-\xi_0^\infty(s_k)\Big)\,,
\end{aligned}$$

so

$$\liminf_{k\to\infty}\Big(\big(\xi_0^{j(k)}(t_k)-\xi_0^{j(k)}(s_k)\big)-\big(\xi_0^\infty(t_k)-\xi_0^\infty(s_k)\big)\Big)\geq 0\,,$$

contradicting (XVI.54). This completes the proof of (XVI.53). ☐

XVI.6 Trajectories of steepest descent.

We recall from §XVI.2 that $\mathscr{M}\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ denotes the set of all maximal augmented trajectories of $\hat{\Sigma}$ from x of steepest descent with respect to V and that, as explained in Proposition XVI.7 and the remarks preceding its statement, $\mathscr{M}\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ is always nonempty for trivial reasons, because the trivial trajectory Ξ_x^{triv} always belongs to $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$, and once we know that $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}\neq\emptyset$ it follows immediately from Zorn's Lemma that $\mathscr{S}\mathscr{D}_{\hat{\Sigma},V,x}$ must have a maximal element.

The truly nontrivial and useful result is the statement that maximal steepest descent trajectories not only exist but are "large," in the sense that they are "right-unbounded." Precisely, if $\xi : I\mapsto\Omega$ is a curve, we say that ξ is *right-unbounded* if (i) the interval I is open on the right (that is, if $\tau=\sup I$, then either (a) $\tau=+\infty$ or (b) τ is finite and does not belong to I), and (ii) if τ is finite, then for every compact subset K of Ω there exists a τ_K such that $0\leq\tau_K<\tau$ and $\xi(t)\notin K$ whenever $\tau_K<t<\tau$. (Equivalently, condition (ii) asserts that $\lim_{t\uparrow\tau}\xi(t)=\infty_\Omega$, where ∞_Ω is the point at infinity of the one-point compactification of Ω.)

Theorem XVI.19. *Let $\hat{\Sigma} = (\Omega, U, f, L)$ be a locally coercive, locally uniformly continuous augmented control system such that $F_{\hat{\Sigma}}(x,U)$ is closed and convex for every $x \in \Omega$. Let $V : \Omega \mapsto \mathbb{R}$ be a continuous function that satisfies (XVI.13) on Ω in the viscosity sense. Let $x_* \in \Omega$, and let $\Xi = (\xi, \xi_0)$ be a maximal augmented trajectory of $\hat{\Sigma}$ from x_* of steepest descent with respect to V, defined on an interval I. Then ξ is right-unbounded.*

Proof. We assume that the conclusion is not true and derive a contradiction. Pick a $\Xi = (\xi, \xi_0) \in \mathscr{MSD}_{\hat{\Sigma},V,x_*}$ that violates the conclusion. This means, to begin with, that Ξ is defined on a bounded interval I, and, in addition, this interval is either of the form $[0,\tau]$, with $0 \le \tau < +\infty$ (the "right-closed case"), or of the form $[0,\tau[$, with $0 < \tau < +\infty$ (the "right-open case"). Furthermore, in the right-open case there exist a compact subset K of Ω and a sequence $\{t_j\}_{j=1}^{\infty}$ such that $t_j \in I$ and $\xi(t_j) \in K$ for all j, and $\lim_{j\to\infty} t_j = \tau$. In order to treat the right-open and right-closed cases together, we also choose a compact subset K of Ω and a sequence $\{t_j\}_{j=1}^{\infty}$ in I in the right-closed case, subject to the only requirements that $\xi(t) \in K$ for all $t \in [0,\tau]$ and $t_j \to \tau$ as $j \to \infty$. (For example, we could just choose $K = \{\xi(t) : 0 \le t \le \tau\}$, $t_j = \tau$ for all j.) Then in both cases the t_j belong to I and converge to τ, and the $\xi(t_j)$ belong to K. In addition, we pick a function $\eta : I \mapsto U$ such that $\dot{\Xi}(t) = F(\xi(t), \eta(t))$ for almost all $t \in I$.

Fix a positive number δ such that the compact set

$$K_\delta = \{x \in \mathbb{R}^n : \operatorname{dist}(x,K)\| \le \delta\}$$

is contained in Ω. Then use the fact that $\hat{\Sigma}$ is locally coercive to choose r, C such that $r > 1$, $C > 0$, and $\|f(x,u)\|^r \le L(x,u) + \tilde{C}$ for all $x \in K_\delta$, $u \in U$. Let

$$\bar{V} = \max\{|V(x)| : x \in K_\delta\}.$$

If $\rho = \frac{r}{r-1}$, so that $\frac{1}{r} + \frac{1}{\rho} = 1$, and we define

$$\bar{\delta}(\sigma) \stackrel{\text{def}}{=} 2\Big(\sigma^{1/\rho}(2|\bar{V}| + (C+1)\sigma)^{1/r} + \sigma\Big),$$

then $\bar{\delta}(\sigma)$ goes to zero as $\sigma \downarrow 0$. Therefore we can pick σ such that $\bar{\delta}(\sigma) < \delta$.

Now suppose that we are in the right-open case. Given any j, the point $x_j = \xi(t_j)$ belongs to K. Let

$$S_j = \Big\{t \in I : t \ge t_j \wedge (\forall s)\big((t_j \le s \le t) \Rightarrow \xi(s) \in K_\delta\big)\Big\}.$$

Then S_j is a subinterval of I, whose left endpoint is t_j.

If $t, t' \in S_j$, and $t' \leq t$, then the steepest descent property of Ξ implies that

$$\xi_0(t) + V(\xi(t)) \leq \xi_0(t') + V(\xi(t')),$$

so

$$\xi_0(t) - \xi_0(t') \leq V(\xi(t)) - V(\xi(t')) \leq 2|\bar{V}|,$$

since both $\xi(t)$ and $\xi(t')$ belong to K_δ. Furthermore, the inequality

$$\|f(\xi(s), \eta(s))\|^r \leq L(\xi(s), \eta(s)) + C$$

is true for all $s \in S_j$, since $\xi(s) \in K_\delta$ for all such s. Therefore

$$\begin{aligned}
&\|\xi(t) - \xi(t')\| \\
&= \left\| \int_{t'}^{t} \dot{\xi}(s)\, ds \right\| \\
&\leq \int_{t'}^{t} \|\dot{\xi}(s)\|\, ds \\
&= \int_{t'}^{t} \|f(\xi(s), \eta(s))\|\, ds \\
&\leq (t - t')^{1/\rho} \Big(\int_{t'}^{t} \|f(\xi(s), \eta(s))\|^r\, ds \Big)^{1/r} \\
&\leq (t - t')^{1/\rho} \Big(\int_{t'}^{t} (L(\xi(s), \eta(s)) + C)\, ds \Big)^{1/r} \\
&= (t - t')^{1/\rho} (C(t - t') + \xi_0(t) - \xi_0(t'))^{1/r} \\
&= (t - t')^{1/\rho} (C(t - t') + 2|\bar{V}|))^{1/r} \\
&\leq \frac{1}{2} \bar{\delta}(t - t').
\end{aligned}$$

Now pick j so large that $\tau - t_j < \sigma$, and apply the above inequality with $t' = t_j$. Then

$$\|\xi(t) - \xi(t_j)\| \leq \frac{1}{2} \bar{\delta}(\sigma) \leq \frac{1}{2} \delta \;\text{ for all }\; t \in S_j. \tag{XVI.55}$$

Hence, if we let $\bar{t} = \sup S_j$, it is impossible that $\bar{t} < \tau$, because if $\bar{t} < \tau$ then $\bar{t} \in I$, and (XVI.55) implies that $\|\xi(\bar{t}) - \xi(t_j)\| \leq \frac{1}{2}\delta$, from which it follows, by continuity, that there exists a positive α such that $\|\xi(t) - \xi(t_j)\| < \delta$ for $\bar{t} \leq t \leq \bar{t} + \alpha$; but then $\xi(t) \in K_\delta$ if $\bar{t} \leq t \leq \bar{t} + \alpha$, since $\xi(t) \in K$; therefore $\bar{t} + \alpha \in S_j$, contradicting the definition of $\bar{t}$.

Therefore $\bar{t} = \tau$, and this implies that $S_j = [t_j, \tau[$, so $\xi(t) \in K_\delta$ for all $t \in [t_j, \tau[$. Then the bound $L(\xi(s), \eta(s)) + C \geq \|f(\xi(s), \eta(s))\|^r$ holds for almost all $s \in [t_j, \tau[$, showing in particular that the function

$$[t_j, \tau[\, \ni s \mapsto \psi(s) \stackrel{\text{def}}{=} L(\xi(s), \eta(s)) + C$$

which is measurable because $L(\xi(s), \eta(s)) = \dot{\xi}_0(s)$, is nonnegative. Hence, to prove that ψ is Lebesgue-integrable on $[t_j, \tau[$, it suffices to show that the integrals $\int_{t_j}^t (L(\xi(s), \eta(s)) + C)\, ds$, for $t_j \leq t < \tau$, are bounded above by a fixed constant. But, if $t_j \leq t < \tau$, then

$$\begin{aligned} \int_{t_j}^t (L(\xi(s), \eta(s)) + C)\, ds &= \xi_0(t) - \xi_0(t_j) + C(t - t_j) \\ &\leq 2|V| + C\sigma \,. \end{aligned}$$

Therefore ψ is Lebesgue-integrable on $[t_j, \tau[$, and then $\dot{\xi}_0$ is also Lebesgue-integrable on $[t_j, \tau[$, because $\dot{\xi}_0 = \psi - C$. Since $\dot{\xi}_0$ is Lebesgue-integrable on $[0, t_j]$, we conclude that $\dot{\xi}_0$ is Lebesgue-integrable on $[0, \tau[$. Hence the limit $x_0^\# = \lim_{t \uparrow \tau} \xi_0(t)$ exists, and the extended function $\xi_0^\# : [0, \tau] \mapsto \mathbb{R}$ defined by $\xi_0^\#(t) = \xi_0(t)$ if $0 \leq t < \tau$, $\xi_0^\#(\tau) = x_0^\#$, is absolutely continuous.

In addition, the bound

$$\|\dot{\xi}(s)\|^r = \|f(\xi(s), \eta(s)\|^r \leq L(\xi(s), \eta(s)) + C = \psi(s)\,,$$

valid on $[t_j, \tau[$, shows that $\dot{\xi}$ in L^r on $[s_j, \tau[$, so *a fortiori* $\dot{\xi}$ is Lebesgue-integrable on $[s_j, \tau[$, and then the limit $x^\# = \lim_{t \uparrow \tau} \xi(t)$ exists and belongs to K_δ (because $x^\# = \lim_{\ell \to \infty} \xi(t_\ell)$ and the $\xi(t_\ell)$ belong to K_δ), and the extended function $\xi^\# : [0, \tau] \mapsto \Omega$ defined by $\xi^\#(t) = \xi(t)$ if $0 \leq t < \tau$, $\xi^\#(\tau) = x_0^\#$, is absolutely continuous.

Hence we have constructed an absolutely continuous extension $\Xi^\# = (\xi^\#, \xi_0^\#)$ of Ξ to the closed interval $[0, \tau]$. Clearly, $\Xi^\#$ is also an augmented trajectory of $\hat{\Sigma}$ starting at x_* at time 0, and the fact that Ξ is of steepest descent with respect to V and V is continuous implies that $\Xi^\#$ is of steepest descent with respect to V as well. Therefore Ξ is not maximal, because we have constructed an extension to a strictly larger interval. It follows that the "right-open case" cannot arise at all.

We now analyze the right-closed case, and show that it cannot arise either. We do this by constructing an extension $\Xi^\#$ of Ξ to a $\Xi^\# \in \mathscr{MSD}_{\hat{\Sigma}, V, x_*}$ defined on the interval $[0, \tau + \sigma]$. This will, of course, contradict the assumed maximality of Ξ, and conclude our proof.

To construct $\Xi^{\#}$, we construct ε-approximate augmented trajectories $\Xi^{\varepsilon}:[\tau,\tau+\sigma]\mapsto\Omega\times\mathbb{R}$ of $\hat{\Sigma}$ that are "ε-approximately of steepest descent," and such that $\Xi^{\varepsilon}(\tau)=\Xi(\tau)$. We then pass to the limit as $\varepsilon\downarrow 0$, using the compactness theorem (XVI.18), and get an exact augmented trajectory $\tilde{\Xi}:[\tau,\tau+\sigma]\mapsto\Omega\times\mathbb{R}$ of $\hat{\Sigma}$ such that $\tilde{\Xi}(\tau)=\Xi(\tau)$, which is exactly of steepest descent.

Write $x^{\#}=\xi(\tau)$, $x_0^{\#}=\xi_0(\tau)$. Fix an ε such that $0<\varepsilon<1$. Let $\mathscr{Z}_\varepsilon$ be the set of all triples (I',Ξ',S') such that

(1) I' is a subinterval of $[\tau,\tau+\sigma]$ such that $\tau\in I'$,

(2) $\Xi'=(\xi',\xi_0'):I'\mapsto K_\delta\times\mathbb{R}$ is an ε-approximate augmented trajectory of $\hat{\Sigma}$,

(3) $\Xi'(\tau)=(x^{\#},x_0^{\#})$,

(4) S' is a strongly ε-dense subset of I' such that $\xi_0'(s)+V(\xi'(s))+\varepsilon(t-s)\geq\xi_0'(t)+V(\xi'(t))$ for all $s,t\in S'$ such that $s\leq t$.

(We say that a subset S of an interval J is *strongly ε-dense* if for every $t\in J$ there exist $s,s'\in S$ such that $s\leq t\leq s'$ and $s'-s\leq\varepsilon$.)

We partially order $\mathscr{Z}_\varepsilon$ by stipulating that, if $(I_i,\Xi_i,S_i)\in\mathscr{Z}_\varepsilon$ for $i=1,2$, then $(I_1,\Xi_1,S_1)\preceq(I_2,\Xi_2,S_2)$ iff $I_1\subseteq I_2$, Ξ_1 is the restriction of Ξ_2 to I_1, and $S_1=S_2\cap I_1$.

It is clear that $\mathscr{Z}_\varepsilon\neq\emptyset$, because the triple $(\{\tau\},\Xi_{x^{\#},\tau},\{\tau\})$—where $\Xi_{x^{\#},\tau}$ is the map $\Xi:\{\tau\}\mapsto\Omega\times\mathbb{R}$ such that $\Xi(\tau)=(x^{\#},\xi_0(\tau))$—belongs to $\mathscr{Z}_\varepsilon$. If Z is a totally ordered subset of $\mathscr{Z}_\varepsilon$, we show that Z has an upper bound (I'_*,Ξ'_*,S'_*) in $\mathscr{Z}_\varepsilon$. This conclusion is trivial if $Z=\emptyset$, for in that case we can take $(I'_*,\Xi'_*,S'_*)=(\{\tau\},\Xi_{x^{\#},\tau},\{\tau\})$. Assume that $Z\neq\emptyset$. Let I'_* be the union of the intervals I' for all the members (I',Ξ',S') of Z. Then I'_* is clearly a subinterval of $[\tau,\tau+\sigma]$, and $\tau\in I'$. If $t\in I'_*$, we define $\Xi'_*(t)=\Xi'(t)$, where (I',Ξ',S') is any member of Z such that $t\in I$. Write $\Xi'_*=(\xi'_*,\xi'_{0,*})$. Then Ξ'_* is obviously well defined, and is an ε-approximate augmented trajectory of $\hat{\Sigma}$ such that $\xi'_*(\tau)=x^{\#}$ and $\xi'_{0,*}(\tau)=x_0^{\#}$. We then let

$$S'_*=\bigcup_{(I',\Xi',S')\in Z}S'. \tag{XVI.56}$$

We want to prove that the triple (I'_*,Ξ'_*,S'_*) is an upper bound for Z in $\mathscr{Z}_\varepsilon$. If we show that

$$(I'_*,\Xi'_*,S'_*)\in\mathscr{Z}_\varepsilon\,, \tag{XVI.57}$$

then the fact that (I'_*,Ξ'_*,S'_*) is an upper bound for Z is immediate, so all we really need is to prove (XVI.57).

It is evident that the triple (I'_*, Ξ'_*, S'_*) satisfies the first three of the four conditions in the definition of $\mathscr{Z}_\varepsilon$. Let us show that it satisfies the fourth one as well. If $t \in I'_*$, we can find $(I'', \Xi'', S'') \in Z$ such that $t \in I''$. Since S'' is a strongly ε-dense subset of I'', there exist $s_1, s_2 \in S''$ such that $s_1 \le t \le s_2$ and $s_2 - s_1 \le \varepsilon$. Then s_1 and s_2 belong to S'_*, and this establishes that S'_* is a strongly ε-dense subset of I'_*. Now, if s_1, s_2 are members of S'_* such that $s_1 \le s_2$, then we can find (using the fact that Z is totally ordered) a member (I', Ξ', S') of Z such that s_1 and s_2 belong to I'. It then follows easily that s_1 and s_2 must belong to S'. Then, if we write $\Xi' = (\xi', \xi'_0)$, the fact that $s_1 \in S'$, $s_2 \in S'$, $s_1 \le s_2$, and $(I', \Xi', S') \in \mathscr{Z}_\varepsilon$, imply that

$$\xi'_0(s_1) + V(\xi'(s_1)) + \varepsilon(s_2 - s_1) \ge \xi'_0(s_2) + V(\xi'(s_2)) \,.$$

Hence

$$\xi'_{0,*}(s_1) + V(\xi'_*(s_1)) + \varepsilon(s_2 - s_1) \ge \xi'_{0,*}(s_2) + V(\xi'_*(s_2)) \,.$$

Since this is true for any two members s_1, s_2 of S'_* such that $s_1 \le s_2$, we conclude that the fourth condition holds as well, and the proof of (XVI.57) is complete.

We have shown that every totally ordered subset of $\mathscr{Z}_\varepsilon$ has an upper bound in $\mathscr{Z}_\varepsilon$. Therefore Zorn's Lemma implies that $\mathscr{Z}_\varepsilon$ has a maximal element (I', Ξ', S'). We claim that $I' = [\tau, \tau + \sigma]$. Suppose this was not true. Then either

(A) $I' = [\tau, \zeta[$ for some ζ such that $\tau < \zeta \le \tau + \sigma$,

or

(B) $I' = [\tau, \zeta]$ for some ζ such that $\tau \le \zeta < \tau + \sigma$.

We shall exclude both possibilities.

Write $\Xi' = (\xi', \xi'_0)$. Let $\mathscr{B}_\varepsilon = \{v \in \mathbb{R}^n : \|v\| \le \varepsilon\}$. Let $I' \ni t \mapsto (\eta'(t), v(t)) \in U \times \mathscr{B}_\varepsilon$ be a function such that the function $v(\cdot)$ is measurable, and the equalities $\dot{\xi}'(t) = f(\xi'(t), \eta'(t)) + v(t)$ and $\dot{\xi}'_0(t) = L(\xi'(t), \eta'(t))$ hold for almost all $t \in I'$.

Let $t \in I'$. Then there must exist a $\bar{t} \in S'$ such that $\bar{t} \ge t$. Also, τ must belong to S', because τ is the leftmost point of I'. Then

$$\xi'_0(\tau) + V(\xi'(\tau)) + \varepsilon(t - \tau) \ge \xi'_0(\bar{t}) + V(\xi'(\bar{t})) \,,$$

so (since $\varepsilon < 1$)

$$\begin{aligned} \xi'_0(\bar{t}) - \xi'_0(\tau) &\le \varepsilon(t - \tau) + V(\xi'(\tau)) - V(\xi'(\bar{t})) \\ &\le 2|\bar{V}| + \sigma \,. \end{aligned}$$

We then have (using the fact that $\|v(s)\| < 1$, which is true because $\varepsilon < 1$),

$$
\begin{aligned}
&\int_\tau^t \|\dot{\xi}'(s)\| ds \\
&= \int_\tau^t \|f(\xi'(s), \eta'(s)) + v(s)\| ds \\
&\le \int_\tau^{\bar{t}} \|f(\xi'(s), \eta'(s)) + v(s)\| ds \\
&\le \int_\tau^{\bar{t}} \|f(\xi'(s), \eta'(s))\| ds + \int_\tau^{\bar{t}} \|v(s))\| ds \\
&\le (\bar{t} - \tau)^{1/\rho} \Big(\int_\tau^{\bar{t}} \|f(\xi'(s), \eta'(s))\|^r ds \Big)^{1/r} + \bar{t} - \tau \\
&\le \sigma^{1/\rho} \Big(\int_\tau^{\bar{t}} \big(L(\xi'(s), \eta'(s)) + C \big) ds \Big)^{1/r} + \sigma \\
&\le \sigma^{1/\rho} \Big(\xi_0'(\bar{t}) - \xi_0'(\tau) + C(\bar{t} - \tau) \Big)^{1/r} + \sigma \\
&\le \sigma^{1/\rho} \big(2|\bar{V}| + \sigma + C\sigma \big)^{1/r} + \sigma \\
&= \frac{1}{2} \bar{\delta}(\sigma) \\
&\le \frac{1}{2} \delta \, .
\end{aligned}
$$

Also, if we let $\theta(s) = \dot{\xi}_0'(s) + C = L(\xi'(s), \eta'(s)) + C$, then θ is nonnegative and

$$
\begin{aligned}
\int_\tau^t \theta(s) ds &\le \int_\tau^{\bar{t}} \theta(s) ds \\
&= \xi_0'(\bar{t}) - \xi_0'(\tau) + C(\bar{t} - \tau) \\
&\le 2|\bar{V}| + (C+1)\sigma \, .
\end{aligned}
$$

Since t is an arbitrary member of I', the above inequalities imply that the functions $\dot{\xi}'$ and θ are integrable on I'. Since I' is bounded, $\dot{\xi}'$ is integrable as well. This implies, in particular, that the limits $\lim_{t \uparrow \tau+\zeta} \xi'(t)$ and $\lim_{t \uparrow \tau+\zeta} \xi_0'(t)$ exist. Hence, if (A) holds, we can extend Ξ' to the closed interval $\tilde{I}' \stackrel{\text{def}}{=} [\tau, \tau + \zeta]$, and the result is a curve $\tilde{\Xi}'$ in $\mathbb{R}^{n+1}$ which is obviously an ε-approximate augmented trajectory of $\hat{\Sigma}$. If we then define $\tilde{S}' = S' \cup \{\tau + \zeta\}$, then $\tilde{S}'$ is a strongly ε-dense subset of $\tilde{I}'$. Now, if s_1, s_2 are members of $\tilde{S}'$ such that $s_1 \le s_2$ we have to prove that $\tilde{\xi}_0'(s_1) + V(\tilde{\xi}'(s_1)) + \varepsilon(s_2 - s_1) \ge \tilde{\xi}_0'(s_2) + V(\tilde{\xi}'(s_2))$. This is clearly true if

$s_2 < \tau + \zeta$ or if $s_1 = s_2$. So the only remaining case for us to consider is when $s_1 < s_2 = \tau + \zeta$. But in that case we can take a sequence $\{v_\ell\}_{\ell=1}^\infty$ of points of I' such that $v_\ell \uparrow \tau + \zeta$ as $\ell \to \infty$, and $v_\ell > s_1$ for all ℓ. Then we can pick, for each ℓ, a $w_\ell \in S'$ such that $v_\ell \le w_\ell$. On the other hand, $s_1 \in S'$, since $s_1 < \tau + \zeta$. Therefore $\tilde{\xi}_0'(s_1) + V(\tilde{\xi}'(s_1)) + \varepsilon(w_\ell - s_1) \ge \tilde{\xi}_0'(w_\ell) + V(\tilde{\xi}'(w_\ell))$ for all ℓ. If we let $\ell \to \infty$ and use the continuity of $\tilde{\xi}_0'$ and V, we find that

$$\tilde{\xi}_0'(s_1) + V(\tilde{\xi}'(s_1)) + \varepsilon(s_2 - s_1) \ge \tilde{\xi}_0'(s_2) + V(\tilde{\xi}'(s_2)),$$

as desired. This completes the proof that the extension $(\tilde{I}', \tilde{\Xi}', \tilde{S}')$ of (I', Ξ', S') is also in $\mathscr{X}_\varepsilon$, a fact that of course contradicts the maximality of (I', Ξ', S') if (A) holds. We have thus derived a contradiction from the assumption that (A) is true. Hence (A) is excluded.

We are thus left with Case (B), that is, the possibility that $I' = [\tau, \tau + \zeta]$ and $\zeta < \sigma$. We now proceed to exclude this case. The integral calculation done above shows that

$$\int_\tau^{\tau+\zeta} \|\dot{\xi}'(t)\| dt \le \frac{\delta}{2},$$

so

$$\|\xi'(\tau + \zeta) - \xi'(\tau)\| \le \frac{\delta}{2}. \tag{XVI.58}$$

Since $\xi'(\tau) \in K$, (XVI.58) implies that, if $q = \xi'(\tau + \zeta)$, then $\text{dist}(q, K) \le \frac{\delta}{2}$. In particular, q is an interior point of K_δ. Using Theorem XVI.16, we construct sequences $\{x_j\}_{j=1}^\infty$, $\{v_j\}_{j=1}^\infty$, $\{\lambda_j\}_{j=1}^\infty$, $\{h_j\}_{j=1}^\infty$, $\{\gamma_j\}_{j=1}^\infty$ in Ω, $\mathbb{R}^n$, $\mathbb{R}$, $\mathbb{R}$, and $\mathbb{R}$, respectively, such that

(1) the inequalities

$$\begin{aligned} h_j &> 0, \\ \gamma_j &> 0, \\ \|x_j - q - h_j v_j\| &\le h_j \gamma_j, \\ V(x_j) &\le V(q) - h_j \lambda_j + h_j \gamma_j, \end{aligned}$$

hold for all j,

(2) $(v_j, \lambda_j) \in F_{\hat{\Sigma}}(q, U)$ for all j,

(3) $h_j \downarrow 0$, $\gamma_j \downarrow 0$, and $x_j \to q$ as $j \to \infty$.

We let $(v_j,\lambda_j) = F_{\hat{\Sigma}}(q,u_j)$, $u_j \in U$, and write $\tilde{v}_j = h_j^{-1}(x_j - q)$, so that

$$\|\tilde{v}_j - v_j\| \le \gamma_j\,. \tag{XVI.59}$$

Choose j so large that the following conditions are fulfilled:

$$h_j < \varepsilon\,, \tag{XVI.60}$$

$$\gamma_j < \frac{\varepsilon}{2}\,, \tag{XVI.61}$$

$$\|x_j - q\| < \beta\,, \tag{XVI.62}$$

where β is a positive number such that $\beta < \frac{\delta}{2}$ and $\omega(\beta) < \frac{\varepsilon}{2}$.

Then

$$h_j\|\tilde{v}_j\| \le \|x_j - q\| \le \beta \le \frac{\delta}{2}\,.$$

Define a new trajectory $\xi^{\#} : I^{\#} \mapsto \mathbb{R}^n$, where

$$I^{\#} \stackrel{\text{def}}{=} [\tau, \tau+\zeta+h_j]\,,$$

by letting

$$\xi^{\#}(s) = \begin{cases} \xi'(s) & \text{if} \quad \tau \le s \le \tau+\zeta\,, \\ q+(s-\tau-\zeta)\tilde{v}_j & \text{if} \quad \tau+\zeta \le s \le \tau+\zeta+h_j\,, \end{cases}$$

and augment it by defining

$$\xi_0^{\#}(s) = \begin{cases} \xi_0'(s) & \text{if} \quad \tau \le s \le \tau+\zeta\,, \\ q_0+\int_{\tau+\zeta}^{s} L(\xi^{\#}(v),u_j)\,dv & \text{if} \quad \tau+\zeta \le s \le \tau+\zeta+h_j, \end{cases}$$

where $q_0 = \xi_0'(\tau+\zeta)$. Let $\Xi^{\#} = (\xi^{\#}, \xi_0^{\#})$. Then $\Xi^{\#}$ is clearly absolutely continuous, and its restriction to $[\tau, \tau+\zeta]$ is an ε-approximate augmented trajectory of $\hat{\Sigma}$. If s belongs to $[\tau+\zeta, \tau+\zeta+h_j]$, then

$$\|\xi^{\#}(s) - q\| \le h_j\|\tilde{v}_j\| \le \frac{\delta}{2}\,.$$

Therefore $\xi^{\#}(s) \in K_\delta$. Then

$$\begin{aligned} \dot{\Xi}^{\#}(s) &= (v_j, L(\xi^{\#}(s),u_j)) \\ &= (f(q,u_j), L(\xi^{\#}(s),u_j)) \\ &= F_{\hat{\Sigma}}(\xi^{\#}(s),u_j) + (w_j(s),0)\,, \end{aligned}$$

where

$$w_j(s) = f(q,u_j) - f(\xi^\#(s),u_j)\,.$$

Since

$$\|\xi^\#(s) - q\| \le h_j\|v_j\| \le \beta\,,$$

we have

$$\|w_j(s)\| \le \omega(\beta) \le \varepsilon\,.$$

Therefore $\Xi^\#$ is an ε-approximate augmented trajectory of $\hat{\Sigma}$.

Clearly, $\xi^\#(\tau+\zeta+h_j) = x_j$. Therefore,

$$\begin{aligned}
&V(\xi^\#(\tau+\zeta+h_j))\\
&\quad= V(x_j)\\
&\quad\le V(q) - h_j\lambda_j + h_j\gamma_j\\
&\quad= V(\xi^\#(\tau+\zeta)) - h_jL(q,u_j) + h_j\gamma_j\\
&\quad= V(\xi^\#(\tau+\zeta)) - \int_{\tau+\zeta}^{\tau+\zeta+h_j} L(\xi^\#(v),u_j)\,dv\\
&\qquad\quad - \int_{\tau+\zeta}^{\tau+\zeta+h_j}\Big(L(q,u_j) - L(\xi^\#(v),u_j)\Big)\,dv + h_j\gamma_j\\
&\quad= V(\xi^\#(\tau+\zeta)) + \xi_0^\#(\tau+\zeta) - \xi_0^\#(\tau+\zeta+h_j) + E\,,
\end{aligned}$$

where

$$E = -\int_{\tau+\zeta}^{\tau+\zeta+h_j}\Big(L(q,u_j) - L(\xi^\#(v),u_j)\Big)\,dv + h_j\gamma_j\,.$$

Then

$$E \le h_j\omega(\beta) + h_j\gamma_j \le \varepsilon h_j\,.$$

Therefore

$$\begin{aligned}
&V(\xi^\#(\tau+\zeta+h_j)) \le\\
&\quad V(\xi^\#(\tau+\zeta)) + \xi_0^\#(\tau+\zeta) - \xi_0^\#(\tau+\zeta+h_j) + \varepsilon h_jE\,,
\end{aligned}$$

that is,

$$\begin{aligned} &V(\xi^{\#}(\tau+\zeta)) + \xi_0^{\#}(\tau+\zeta) + \varepsilon h_j \\ &\geq \quad V(\xi^{\#}(\tau+\zeta+h_j)) + \xi_0^{\#}(\tau+\zeta+h_j)\,. \end{aligned} \tag{XVI.63}$$

This last inequality implies that, if we define

$$S^{\#} = S' \cup \{\tau+\zeta+h_j\}\,,$$

then we can easily show that $S^{\#}$ is a strongly ε-dense subset of $I^{\#}$ such that

$$\begin{aligned} &V(\xi^{\#}(s_1)) + \xi_0^{\#}(s_1) + \varepsilon(s_2 - s_1) \\ &\qquad \geq \quad V(\xi^{\#}(s_2)) + \xi_0^{\#}(s_2) \end{aligned} \tag{XVI.64}$$

whenever $s_1, s_2 \in S^{\#}$ and $s_2 \geq s_1$. The strong ε-density follows because S' is a strongly ε-dense subset of I' and $h_j < \varepsilon$, since $\tau+\zeta$ necessarily belongs to S'. Inequality (XVI.64) is clearly true if $s_1 = s_2$ or both s_1 and s_2 belong to S'. To verify that it holds in the remaining case, that is, when $s_1 \in S'$ and $s_2 = \tau+\zeta+h_j$, if suffices to use once again the fact that $\tau+\zeta \in S'$, so

$$\begin{aligned} &V(\xi^{\#}(s_1)) + \xi_0^{\#}(s_1) + \varepsilon(\tau+\zeta - s_1) \\ &\qquad \geq \quad V(\xi^{\#}(\tau+\zeta) + \xi_0^{\#}(\tau+\zeta)\,. \end{aligned} \tag{XVI.65}$$

If we add (XVI.63) and (XVI.65), and cancel the sum $V(\xi^{\#}(\tau+\zeta)) + \xi_0^{\#}(\tau+\zeta)$ that appears on both sides of the result, we get

$$\begin{aligned} &V(\xi^{\#}(s_1)) + \xi_0^{\#}(s_1) + \varepsilon(\tau+\zeta+h_j - s_1) \\ &\geq \quad V(\xi^{\#}(\tau+\zeta+h_j)) + \xi_0^{\#}(\tau+\zeta+h_j)\,, \end{aligned} \tag{XVI.66}$$

that is,

$$\begin{aligned} &V(\xi^{\#}(s_1)) + \xi_0^{\#}(s_1) + \varepsilon(s_2 - s_1) \\ &\qquad \geq \quad V(\xi^{\#}(s_2)) + \xi_0^{\#}(s_2)\,. \end{aligned} \tag{XVI.67}$$

It then follows that $(I^{\#}, \Xi^{\#}, S^{\#})$ belongs to $\mathscr{X}_\varepsilon$. Since $(I', \Xi', S') \preceq (I^{\#}, \Xi^{\#}, S^{\#})$ but $(I', \Xi', S') \neq (I^{\#}, \Xi^{\#}, S^{\#})$, we have arrived at a contradiction, which this time has arisen from the assumption that (B) holds. Hence (B) is excluded as well.

It now follows that $I' = [\tau, \tau+\sigma]$. In other words, we have shown that there exists an ε-approximate augmented trajectory $\Xi_\varepsilon = (\xi_\varepsilon, \xi_{0,\varepsilon})$ of $\hat{\Sigma}$ which is defined on $[\tau, \tau+\sigma]$, and is such that there is a strongly ε-dense subset S_ε of $[\tau, \tau+\sigma]$ having the property that

$$V(\xi_\varepsilon(s_1)) + \xi_{0,\varepsilon}(s_1) + \varepsilon(s_2 - s_1) \geq V(\xi_\varepsilon(s_2)) + \xi_{0,\varepsilon}(s_2) \qquad \text{(XVI.68)}$$

whenever $s_1, s_2 \in S_\varepsilon$ and $s_1 \leq s_2$.

Clearly, the points τ, $\tau+\sigma$ must belong to S_ε. If we aplly (XVI.68) with $s_1 = \tau$, $s_2 = \tau+\sigma$, we find the bound

$$\xi_{0,\varepsilon}(\tau+\sigma) - \xi_{0,\varepsilon}(\tau) \leq 2|\bar{V}| + \sigma\,.$$

Therefore, if $\{\varepsilon_j\}_{j=1}^\infty$ is a sequence of positive numbers that converges to 0, we can apply Theorem (XVI.18) with $T = \sigma$ and $\Xi^j = (\xi^j, \xi_0^j) = (\check{\xi}_{\varepsilon_j}, \check{\xi}_{0,\varepsilon_j})$, where $\check{\xi}_{\varepsilon_j}(s) = \xi_{\varepsilon_j}(s+\tau)$ and $\check{\xi}_{0,\varepsilon_j}(s) = \xi_{0,\varepsilon_j}(s+\tau) - \xi_{0,\varepsilon_j}(\tau)$ for $s \in [0, \sigma$. Then, after passing to a subsequence, we may assume that there exists an augmented trajectory $\Xi^\infty = (\xi^\infty, \xi_0^\infty)$ of $\hat{\Sigma}$ such that

(i) $\xi_0^\infty(\tau) = 0$,

(ii) the sequence $\{\xi^j\}_{j=1}^\infty$ converges uniformly to ξ^∞, and

(iii) $\liminf_{j\to\infty} \inf\left\{\Delta^{j,\infty}(t,s) : 0 \leq s \leq t \leq T\right\} \geq 0$, where, if $j, j' \in \mathbb{N} \cup \{+\infty\}$, we define $\Delta^{j,j'}$ as in (XVI.42).

We now show that Ξ^∞ is an augmented trajectory of steepest descent of $\hat{\Sigma}$ from $x^\#$ with respect to V. For this purpose, we pick $s_1, s_2 \in [0, \sigma]$ such that $s_1 \leq s_2$, and prove that

$$V(\xi^\infty(s_1)) + \xi_0^\infty(s_1) \geq V(\xi^\infty(s_2)) + \xi_0^\infty(s_2)\,. \qquad \text{(XVI.69)}$$

For this purpose we pick, for each sufficiently large j, points s_1^j, s_2^j in S_{ε_j} such that

$$s_1^j \leq s_1 + \tau \leq s_1 + \varepsilon_j < s_2^j - \varepsilon_j \leq s_2 + \tau \leq s_2^j\,.$$

(We assume that $s_1 < s_2$, because (XVI.69) is trivially true if $s_1 = s_2$.)

Then

$$\begin{aligned} & V(\xi_{\varepsilon_j}(s_1^j)) - V(\xi_{\varepsilon_j}(s_2^j)) + \varepsilon_j(s_2^j - s_1^j) \\ & \quad \geq \quad \xi_{0,\varepsilon_j}(s_2^j) - \xi_{0,\varepsilon_j}(s_1^j)\,, \end{aligned} \tag{XVI.70}$$

that is,

$$\begin{aligned} & V(\xi^j(s_1^j-\tau)) - V(\xi^j(s_2^j-\tau)) + \varepsilon_j(s_2^j - s_1^j) \\ \geq \quad & \xi_0^j(s_2^j-\tau) - \xi_0^j(s_1^j-\tau) \\ = \quad & \xi_0^\infty(s_2^j-\tau) - \xi_0^\infty(s_1^j-\tau) + \Delta^{j,\infty}(s_2^j-\tau, s_1^j-\tau) \\ \geq \quad & \xi_0^\infty(s_2^j-\tau) - \xi_0^\infty(s_1^j-\tau) \\ & \quad + \inf\{\Delta^{j,\infty}(t,s) : 0 \leq s \leq t \leq \sigma\} \end{aligned}$$

Then

$$\begin{aligned} & V(\xi^\infty(s_1)) - V(\xi^\infty(s_2)) \\ = \quad & \lim_{j\to\infty} \Big(V(\xi^j(s_1^j-\tau)) - V(\xi^j(s_2^j-\tau)) + \varepsilon_j(s_2^j - s_1^j) \Big) \\ \geq \quad & \liminf_{j\to\infty} \Big(\xi_0^j(s_2^j-\tau) - \xi_0^j(s_1^j-\tau) \Big) \\ = \quad & \liminf_{j\to\infty} \Big(\xi_0^\infty(s_2^j-\tau) - \xi_0^\infty(s_1^j-\tau) \\ & \quad + \Delta^{j,\infty}(s_2^j-\tau, s_1^j-\tau) \Big) \\ \geq \quad & \liminf_{j\to\infty} \Big(\xi_0^\infty(s_2^j-\tau) - \xi_0^\infty(s_1^j-\tau) \\ & \quad + \inf\{\Delta^{j,\infty}(t,s) : 0 \leq s \leq t \leq \sigma\} \Big) \\ = \quad & \lim_{j\to\infty} \Big(\xi_0^\infty(s_2^j-\tau) - \xi_0^\infty(s_1^j-\tau) \Big) \\ & \quad + \liminf_{j\to\infty} \Big(\inf\{\Delta^{j,\infty}(t,s) : 0 \leq s \leq t \leq \sigma\} \Big) \\ = \quad & \xi_0^\infty(s_2) - \xi_0^\infty(s_1) \\ & \quad + \liminf_{j\to\infty} \Big(\inf\{\Delta^{j,\infty}(t,s) : 0 \leq s \leq t \leq \sigma\} \Big) \\ \geq \quad & \xi_0^\infty(s_2) - \xi_0^\infty(s_1)\,. \end{aligned}$$

We have thus proved (XVI.69), thereby establishing that Ξ^∞ is an augmented trajectory of $\hat{\Sigma}$ from $x^\#$ of steepest descent with respect to V. If we now concatenate Ξ and Ξ^∞ in the obvious way, by defining

$$\xi^\#(s) = \begin{cases} \xi(s) & \text{if} \quad 0 \le s \le \tau\,, \\ \xi^\infty(s-\tau) & \text{if} \quad \tau \le s \le \tau+\sigma\,, \end{cases}$$

$$\xi_0^\#(s) = \begin{cases} \xi_0(s) & \text{if} \quad 0 \le s \le \tau\,, \\ \xi_0^\infty(s-\tau)+\xi_0(\tau) & \text{if} \quad \tau \le s \le \tau+\sigma\,, \end{cases}$$

then $\Xi^\#$ is an augmented trajectory of $\hat{\Sigma}$ from x_* of steepest descent with respect to V, defined on $[0,\tau+\sigma]$. This contradicts the maximality of Ξ, and concludes our proof. □

XVI.7 The dynamic programming inequality.

Theorem XVI.20. *Let Ω be an open subset of $\mathbb{R}^n$, and let $f:\Omega \mapsto \mathbb{R}^n$, $L:\Omega \mapsto \mathbb{R}$ be continuous maps. Let $V:\Omega \mapsto \mathbb{R}$ be a continuous function that satisfies*

$$-\nabla V(x)\cdot f(x) - L(x) \le 0 \qquad \text{(XVI.71)}$$

on Ω in the viscosity sense. Then for every $x_* \in \Omega$ there exists a curve ξ in Ω, defined on an interval I that contains 0, such that

(1) ξ is an integral curve of f (that is, ξ is locally absolutely continuous and $\dot{\xi}(t) = f(\xi(t))$ for almost every $t \in I$, from which it follows that ξ is continuously differentiable and $\dot{\xi}(t) = f(\xi(t))$ for every $t \in I$),

(2) $\xi(0) = x_*$,

(3) $V(\xi(s)) \le V(\xi(t)) + \int_s^t L(\xi(v))\,dv$ whenever $s,t \in I$ and $s \le t$,

(4) ξ is right-unbounded.

Proof. Let U be a set consisting of a single point $\bar{u}$. Let $\tilde{\Sigma}$ be the augmented control system $(\Omega,U,\tilde{f},\tilde{L})$, where $\tilde{f}(x,\bar{u}) = f(x)$, $\tilde{L}(x,\bar{u}) = -L(x)$. Then $\tilde{\Sigma}$ satisfies all the hypotheses of Theorem XVI.19.

Let $\mathscr{V} = -V$. We claim that $\mathscr{V}$ satisfies

$$\sup\{-\nabla\mathscr{V}(x)\cdot\tilde{f}(x,u) - \tilde{L}(x,u) : u \in U\} \ge 0$$

in the viscosity sense. To prove this, we have to pick a point $x \in \Omega$ and a subdifferential p of $\mathscr{V}$ at x, and show that

$$\sup\{-p\cdot \tilde{f}(x,u) - \tilde{L}(x,u) : u \in U\} \geq 0\,,$$

i.e., that

$$-p\cdot f(x) + L(x) \geq 0\,. \tag{XVI.72}$$

But, if p is a subdifferential of $-V$ at x, and we let $\pi = -p$, then it follows that π is a superdifferential of V at x. Since V satisfies (XVI.71) in the viscosity sense, this implies that $-\pi\cdot f(x) - L(x) \leq 0$. But then $p\cdot f(x) - L(x) \leq 0$, so $-p\cdot f(x) + L(x) \geq 0$, and (XVI.72) has been proved.

We can therefore apply the trivial Proposition XVI.7 to the augmented system $\tilde{\Sigma}$ and the function $\mathscr{V}$, and conclude that there exists a maximal augmented trajectory $\Xi = (\xi, \xi_0)$ of $\tilde{\Sigma}$ from x_* of steepest descent with respect to $\mathscr{V}$, and then use the nontrivial Theorem XVI.19 to conclude that ξ is right-unbounded. The fact that ξ is a trajectory of $\tilde{\Sigma}$ means, of course, that ξ is an integral curve of f. The steepest descent condition says that

$$\mathscr{V}(\xi(s)) \geq \mathscr{V}(\xi(t)) + \int_s^t (-L(\xi(v)))\,dv$$

whenever $0 \leq s \leq t < \tau$. But this says precisely that

$$V(\xi(s)) \leq V(\xi(t)) + \int_s^t L(\xi(v))\,dv$$

whenever $0 \leq s \leq t < \tau$. Hence ξ satisfies all the desired properties, and our proof is complete. □

The following result is then a trivial corollary of Theorem XVI.20.

Theorem XVI.21. *Let $\hat{\Sigma} = (\Omega, U, f, L)$ be an n-dimensional augmented system such that the map $\Omega \ni x \mapsto (f(x,u), L(x,u)$ is continuous for each $u \in U$. Let $V : \Omega \mapsto \mathbb{R}$ be a continuous function that satisfies (XVI.14) on Ω in the viscosity sense. Then for every $x_* \in \Omega$ and every piecewise constant function $\eta : [0,\infty[\mapsto U$ there exists a curve $\xi : I \mapsto \Omega$, defined on a subinterval I of $[0,\infty[$, such that*

(1) $0 \in I$ and $\xi(0) = x_*$,

(2) ξ is a trajectory for the control η (that is, ξ is locally absolutely continuous and $\dot{\xi}(t) = f(\xi(t), \eta(t))$ for almost every $t \in I$)

(3) $V(\xi(s)) \leq V(\xi(t)) + \int_s^t L(\xi(v), \eta(v))\,dv$ whenever $s,t \in I$ and $s \leq t$,

(4) ξ is right-unbounded. □

Theorem XVI.21 has the following immediate consequence.

Theorem XVI.22. *Let $\hat{\Sigma} = (\Omega, U, f, L)$ be an n-dimensional augmented system such that the map $\Omega \ni x \mapsto (f(x,u), L(x,u)$ is continuous for each $u \in U$. Let $V : \Omega \mapsto \mathbb{R}$ be a continuous function that satisfies (XVI.14) on Ω in the viscosity sense. Let $\Xi(\xi, \xi_0) : I \mapsto \Omega \times \mathbb{R}$ be a locally uniquely limiting augmented trajectory of $\hat{\Sigma}$. Then the dynamic programming inequality*

$$V(\xi(s)) \leq V(\xi(t)) + \xi_0(t) - \xi_0(s)$$

holds for all $s, t \in I$ such that $s \leq t$. □

XVI.8 Proof of Theorem XVI.10

According to Theorem XVI.22, the dynamic programming inequality holds along every almost locally uniquely limiting augmented trajectory. The hypothesis that every augmented arc has an almost locally uniquely limiting improvement then implies that the dynamic programming inequality holds along every augmented trajectory. If we apply the inequality to an augmented arc Ξ that starts at a point x and ends at the target, and use the fact that $V = 0$ on the target, we find that $V(x) \leq J(\Xi)$. Hence V is bounded above by the value function $\mathscr{V}_{\mathscr{T}}^{\hat{\Sigma}}$.

To prove that $V \geq \mathscr{V}_{\mathscr{T}}^{\hat{\Sigma}}$, we pick $x \in \Omega$ and use Proposition XVI.7 to conclude that there exists a maximal augmented trajectory $\Xi = (\xi, \xi_0)$ of $\tilde{\Sigma}$ from x of steepest descent with respect to $\mathscr{V}$, and then use Theorem XVI.19 to conclude that ξ is right-unbounded. We then invoke our hypotheses to conclude that Ξ ends at the target. Then $V(x) \geq J(\Xi)$, so V is bounded below by the value function, and our proof is complete. □

Chapter XVII

Conclusion

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

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

Appendix: Some formal definitions.

COMMENT NO. 44. This appendix is very weak. It's got to be completely rewritten and improved.

A.1. Controlled dynamical systems

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

It is clear how to define the *restrictions* $\eta\lceil I'$, $\gamma\lceil I'$ of an open-loop control η and of a TCP $\gamma = (\xi, \eta)$ to a subinterval I' of $D(\eta)$. Also, one can define the *concatenation* $\eta_1 * \eta_2$ of two controls η_1, η_2, provided that their domains I_1, I_2 are such that $I_1 \cup I_2$ is an interval, $I_1 \cap I_2$ is a null set (i.e., $\mathrm{card}(I_1 \cap I_2) \leq 1$) and I_1 lies to the left of I_2. If $\gamma_1 = (\xi_1, \eta_1)$, $\gamma_2 = (\xi_2, \eta_2)$ are two TCP's with domains I_1, I_2 such that $\eta_1 * \eta_2$ is defined, then the concatenation $\gamma_1 * \gamma_2$ is defined—in an obvious way—if $\xi_1(t) = \xi_2(t)$ for $t \in I_1 \cap I_2$, and the curve ξ whose graph is the union of the graphs of of ξ_1 and ξ_2 is locally absolutely continuous on $I_1 \cup I_2$. (For example, if $I_i = [a_i, b_i]$ are compact intervals, then $\gamma_1 * \gamma_2$ is defined iff $b_1 = a_2$ and $\xi_1(b_1) = \xi_2(a_2)$.)

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

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

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

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

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

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

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

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

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

A.3. Generalized solutions

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

Example XVII.1. For a trivial example of this situation, consider the problem of reaching $0 \in \mathbb{R}$ in "minimum time" from an initial point $x \in \mathbb{R}$, when the dynamics is $\dot{x} = u$, $u \in U$, with $U = \mathbb{R}$. It is clear that when $x \neq 0$ the value of this problem is zero, but this value is not attained as a true minimum by any trajectory-control pair arising from an "ordinary" control, where "ordinary" means, for example, Lebesgue integrable. On the other hand, if we allow the "controls" to contain impulses, then we can take, for $x \in \mathbb{R}$, $\eta_x = -x.\delta$, where "δ" is Dirac's Delta function, and the control η_x obviously steers x to 0 in time 0.

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

Remark XVII.2. In the previous example the true solution of the minimization problem was a distribution in the sense of L. Schwartz's theory. The reader should not extrapolate from this example and imagine that for general optimal control problems the appropriate class of generalized controls is that of Schwartz distributions. It turns out that one needs a rather different class of generalized objects. This is discussed in detail in papers by A. Bressan, F. Rampazzo, and their coworkers, cf. [26, 29, 30, 103, 104, 114]. □

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

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

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

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

and

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

We emphasize that a "controlled dynamical system with cost" (CDSWC) $\mathscr{C}$ is not at all the same as an "optimal control problem." A CDSWC gives rise to ***many*** optimal control problems, such as for example the fixed-endpoint problems $P_{\mathscr{C}}(x,x')$ described before, as well as other problems obtained by making different choices of $\mathscr{B}$. (For example, one could choose $\mathscr{B}$ to be the set of all TCP's in $\mathscr{A}$ that start at some given subset A and end at some other subset B of S, obtaining the problem of "joining A to B with minimum cost by a TCP γ of $\dot{x} = f(x,u)$, $u \in U$, $\gamma \in \mathscr{A}$.") An optimal control problem has a *value*, which is an extended real number, and if the set of trajectories is $\neq \emptyset$ then the problem may have a *solution*—i.e., a minimizer—or at least a minimizing sequence or generalized solution, which *realizes* this value. A CDSWC $\mathscr{C}$ has a *value function* $V_{\mathscr{C}}$. The object that realizes the value is something called an "optimal synthesis," by which it is meant—roughly—a family $\{\gamma_{x,x'}\}_{(x,x')\in C}$ in $\mathscr{A}$ that selects, for each $(x,x') \in C$ (where $C = \{(x,x') \in S \times S : \mathscr{B}(x,x') \neq \emptyset\}$, an—ordinary or generalized—solution $\gamma_{x,x'}$ of the problem $P_{\mathscr{C}}(x,x')$. (This is only a first approximation to the concept of a synthesis. In the control literature one looks for syntheses having extra regularity properties, i.e., in selecting solutions $\gamma_{x,x'}$ that depend "nicely"—e.g. continuously, or smoothly, or piecewise smoothly—on x and x'.)

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

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

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

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

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

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

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

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

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

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

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

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

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

The above definitions can be extended in various ways. For example, one can allow S to be a subset of a manifold M other than $\mathbb{R}^n$. In that case, the condition that f be a map from $S \times U$ to $\mathbb{R}^n \cup \{\#\}$ is replaced by the requirement that f be a—possibly partially defined—map on $S \times U$ such that $f(x,u)$ is a tangent vector to M at x whenever $x \in S$, $u \in U$, and $f(x,u)$ is defined.

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

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

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

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

Chapter XVIII

Lots of figures that could be useful

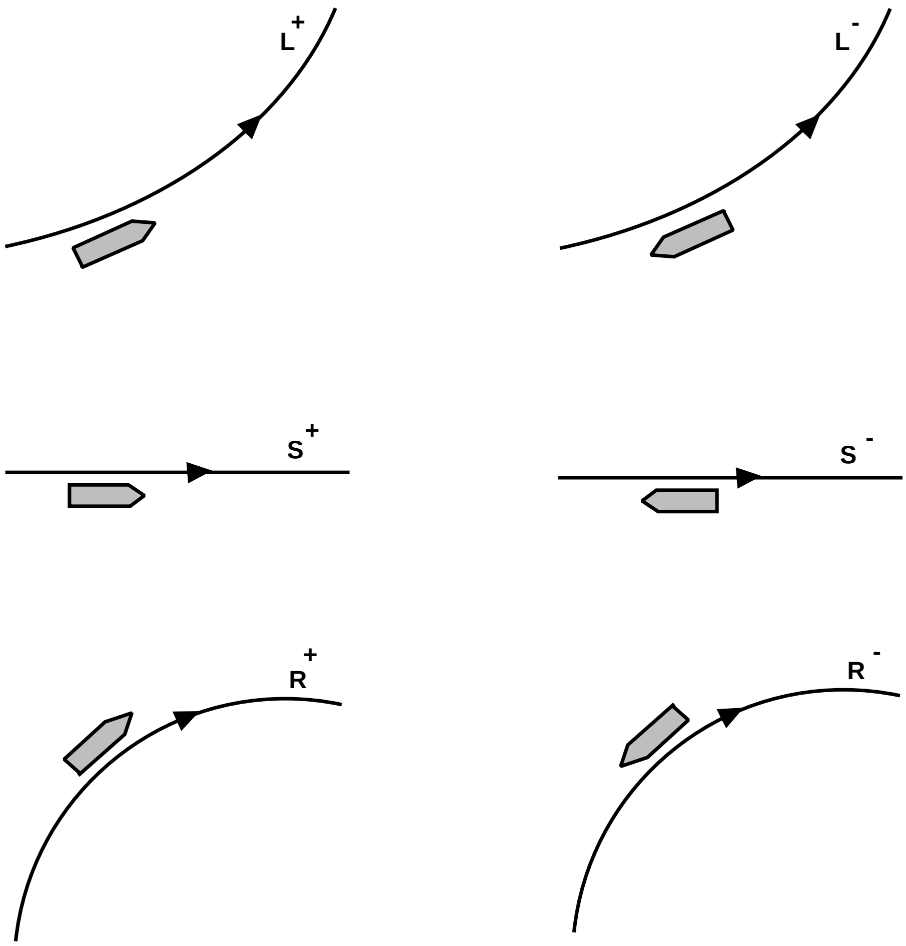


FIGURE 12. Snell's law

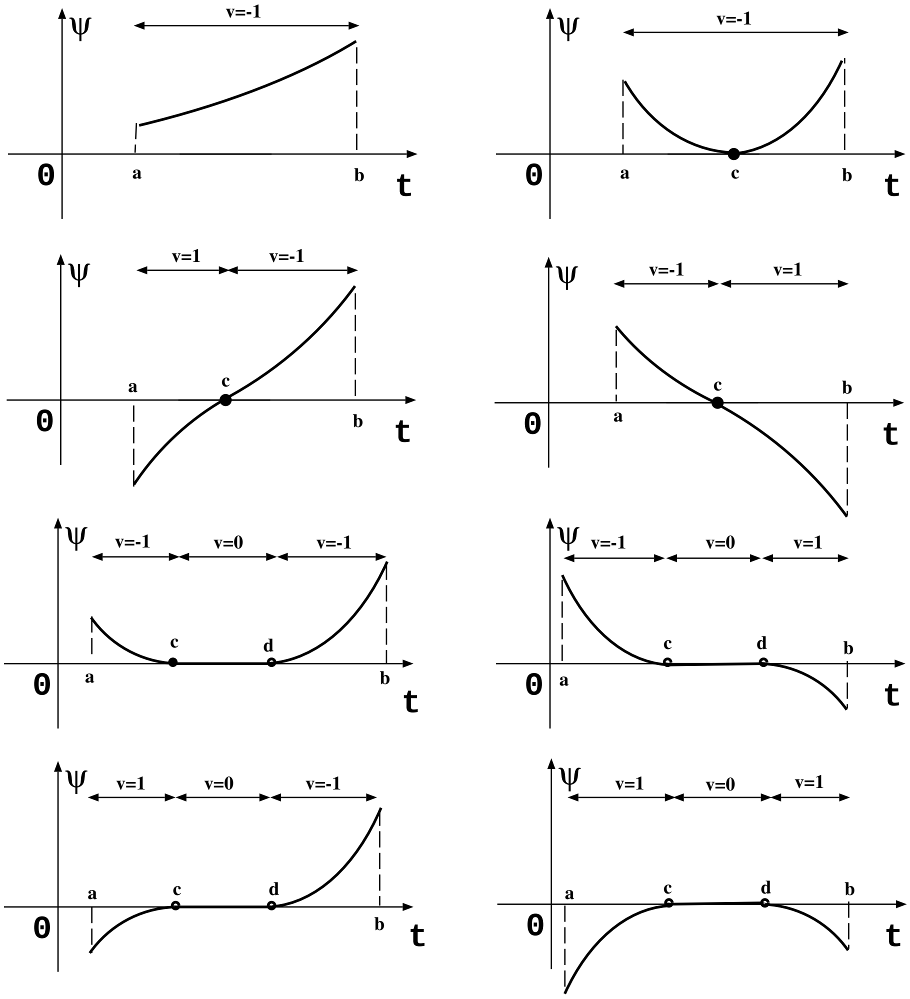


FIGURE 13. Snell's law

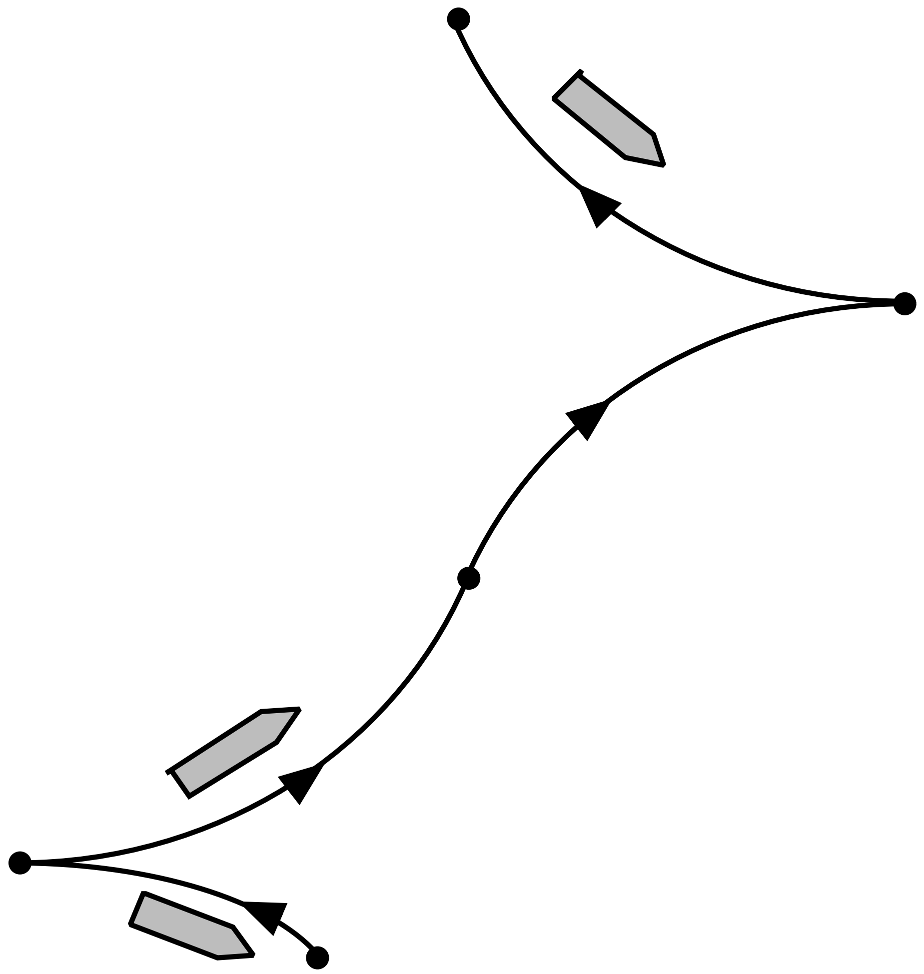


FIGURE 14. Snell's law

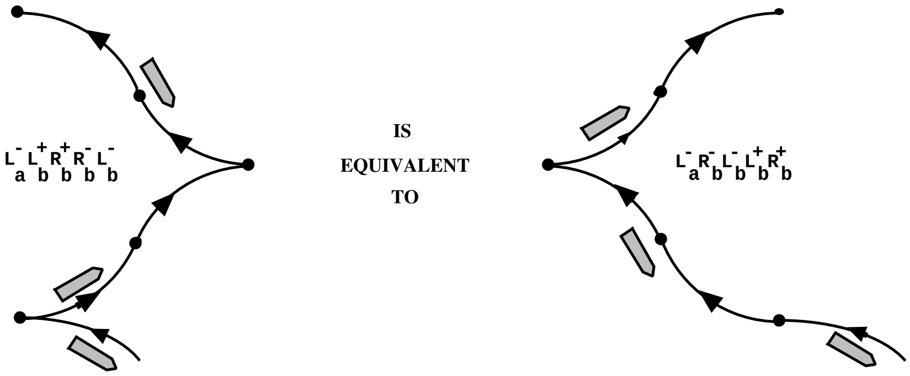


FIGURE 15. Snell's law

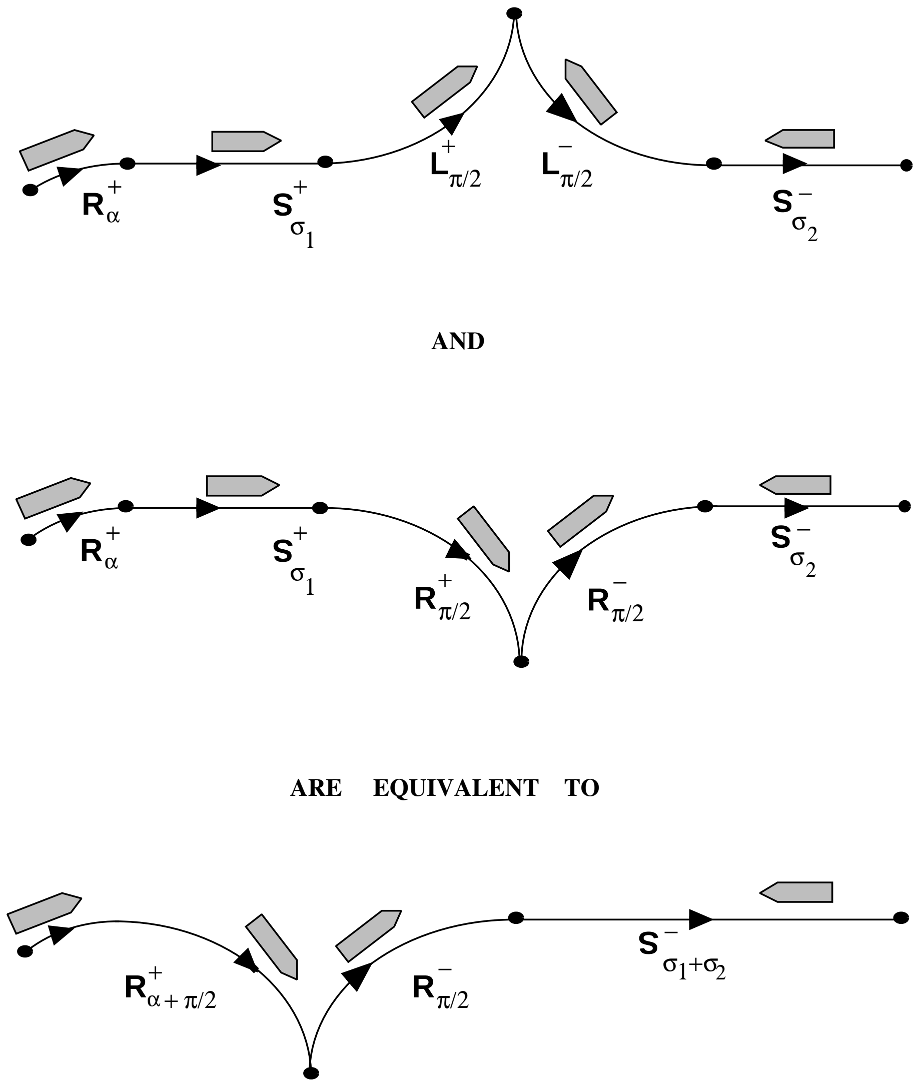


FIGURE 16. Snell's law

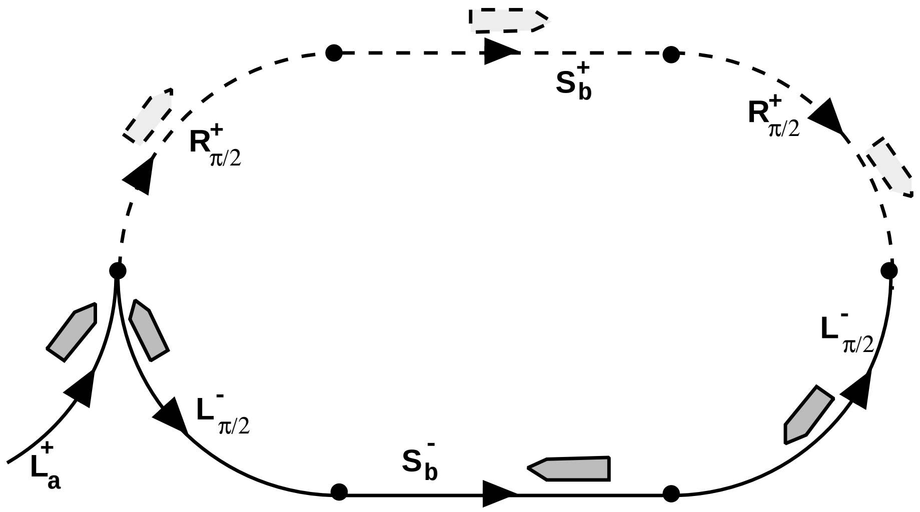


FIGURE 17. Snell's law

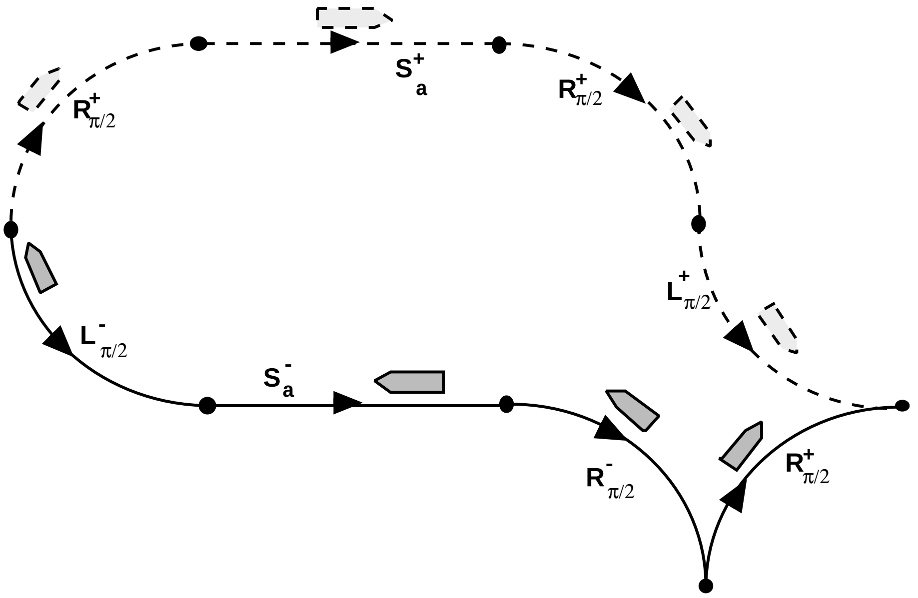


FIGURE 18. Snell's law

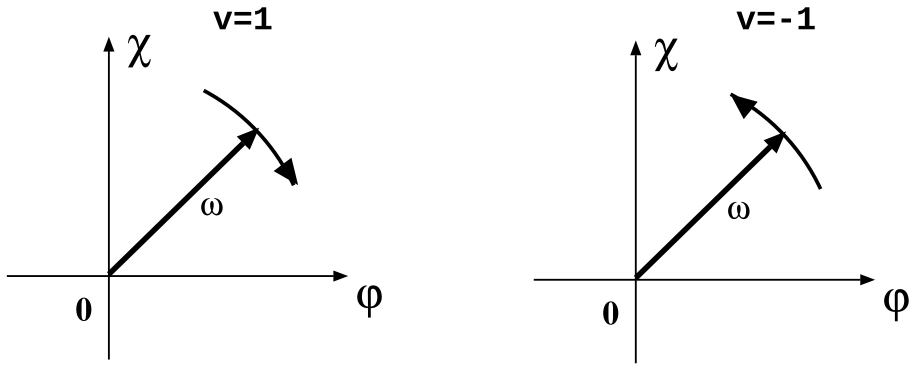


FIGURE 19. Snell's law

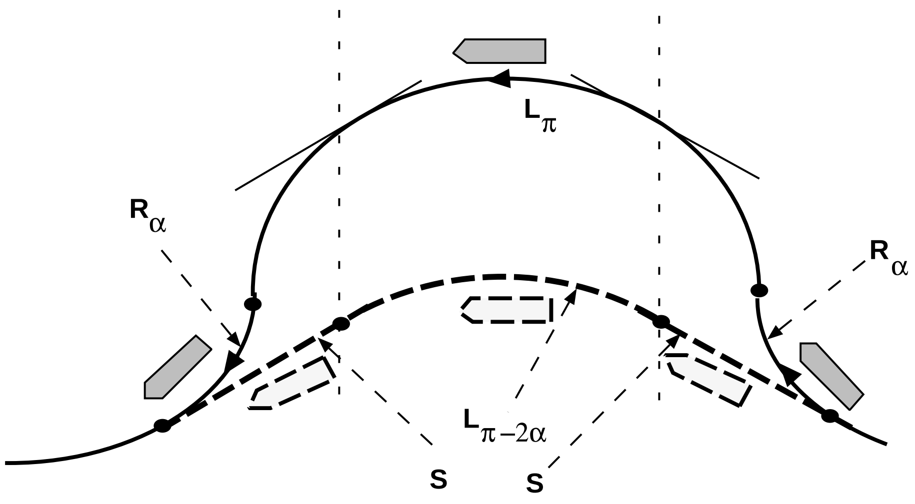
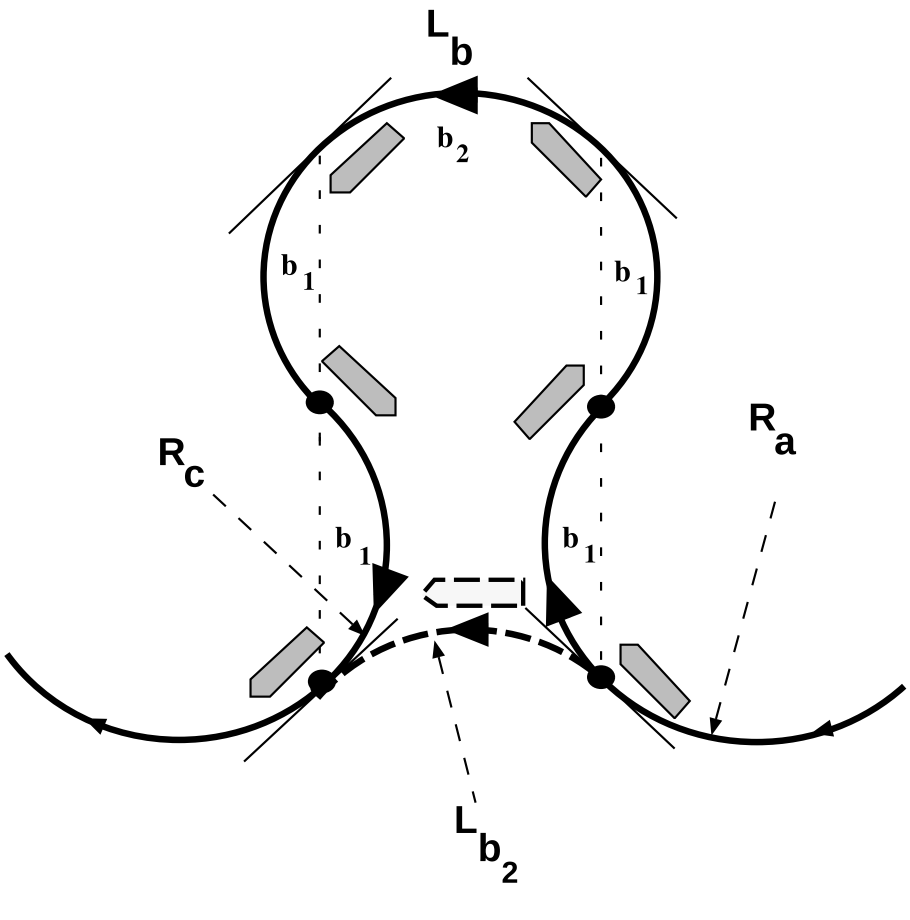


FIGURE 20. Snell's law



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

FIGURE 21. Snell's law

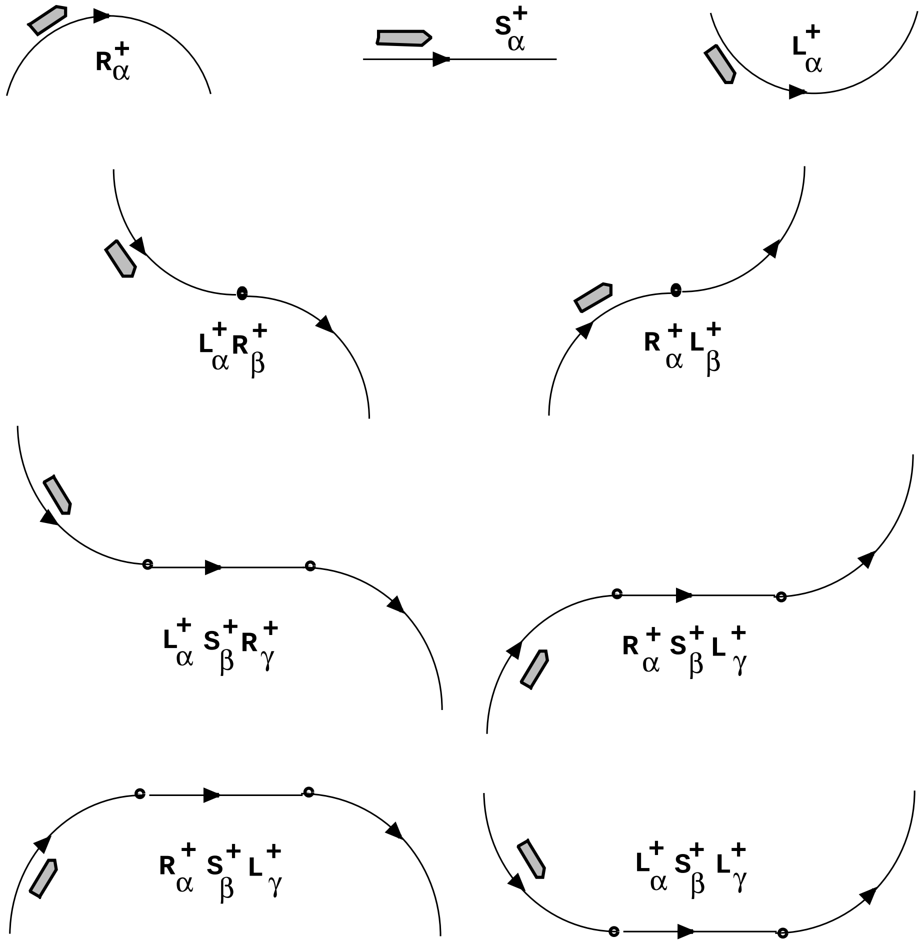


FIGURE 22. Snell's law

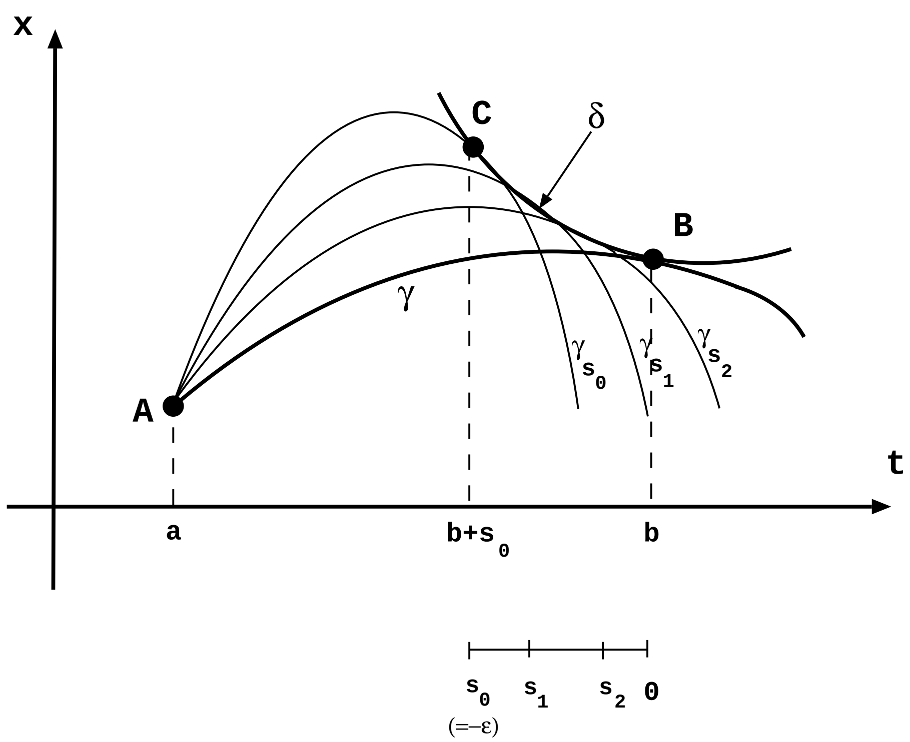


FIGURE 23. Snell's law

Bibliography

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

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

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

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

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

[6] Bardi, M., and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[7] Bardi, Martino, and Francesca Da Lio, "On the Bellman equation for some unbounded control problems." *NoDEA Nonlinear Differential Equations Appl.* **4** (1997), pp. 276-285.

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

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

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

[11] Beltrami, E., *Saggio di interpetazione della geometria non-euclidea.* Giornale di Matematiche **VI**, 1868, pp. 284-312 (in Itialian). Eng. transl. in Stillwell, [123], pp. 7–34.

[12] S. Benett, *A History of Control Engineering.* Peter Pelegrinus, 1979.

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

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

[15] 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

[16] 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.

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

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

[19] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[34] Cambini, A., and S. Querci, *Equazione differenziale del primo ordine con secondo membro discontinuo rispetto all'incognita. Rend. Ist. Mat. Univ. Trieste* **1**, 1969, pp. 89-97.

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

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

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

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

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

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

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

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

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

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

[45] Clarke, F. H., and Yu. S. Ledyaev, "Mean value inequalities." *Proc. Amer. Math. Soc.* **122** (1994), no. 4, pp. 1075–1083.

[46] Clarke, F. H., and Yu. S. Ledyaev, "Mean value inequalities in Hilbert space." *Trans. Amer. Math. Soc.* **344** (1994), no. 1, pp. 307–324.

[47] Clarke, F. H., Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory.* Graduate Texts in Mathematics **178**, Springer-Verlag, New York, 1998.

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

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

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

[51] Doyle, J. C., K. Glover, P. Khargonekar and B. A. Francis, *State space solutions to standard $\mathfrak{H}_2$ and $\mathfrak{H}_\infty$ control problems.* IEEE Transactions on Automatic Control, Vol. 34, 1989, pp. 831-847.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[91] Leibniz, G. W., *Nova methodus pro maximis et minimis, itemque tangentibus, qua nec fractas nec irrationales quantitates moratur, et singulari pro illis calculus genus.* Acta Eruditorum **3**, 1684, pp. 467-473. Engl. transl. in Struik, [124], pp. 272-280.

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

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

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

[95] Malisoff, M., "Viscosity Solutions of the Bellman Equation for Optimal Control Problems with Exit Times." Ph.D. Thesis, Rutgers University, New Brunswick, NJ, January 2000.

[96] Malisoff, M., “Viscosity solutions of the Bellman equation for exit time optimal control problems with vanishing Lagrangians.” *SIAM J. Control Optim.*, to appear.

[97] Malisoff, M., “Viscosity solutions of the Bellman equation for exit time optimal control problems with non-Lipschitz dynamics.” *ESAIM: Control, Optimisation and Calculus of Variations*, **6** (2001), pp. 415-441.

[98] Malisoff, M., “Bounded-from-below solutions of the Hamilton-Jacobi equation for optimal control problems with exit times: Vanishing Lagrangians, eikonal equations, and shape-from-shading.” submitted.

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

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

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

[102] O. Mayr, *The Origins of Feedback Control.* MIT Press, 1970.

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

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

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

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

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

[108] Perec, G., *A Void*, English translation of [107], by G. Adair, Harper-Collins, London, 1994.

[109] Pillai, H. K., and S. Shankar, *A behavioral approach to control of distributed systems. SIAM Journal on Control and Optimization* **37**, 1999, pp. 388-408.

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

[111] Polderman, J. W., and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach.* Springer-Verlag, 1998.

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

[113] Pucci, A., Sistemi di equazioni differenziali con secondo membro discontinuo rispetto all'incognita." *Rend. Ist. Mat. Univ. Trieste* **3**, 1971, pp. 75-80.

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

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

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

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

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

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

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

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

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

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

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

[125] Subbotin, A. I., *Generalized solutions of first-order PDEs: the dynamical optimization perspective.* Birkhäuser, Boston, 1995.

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

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

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

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

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

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

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

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

[134] 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.

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

[136] Sussmann, H. J., "From the Brachystochrone problem to the maximum principle." In *Proc. 35th I.E.E.E. Conference on Decision and Control, Kobe, Japan, Dec. 1996,* I.E.E.E. publications, 1996, pp. 1588-1594.

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

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

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

[140] 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.

[141] 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.

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

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

[144] Sussmann, H.J., "A nonsmooth hybrid maximum principle." In *Stability and Stabilization of Nonlinear Systems, Ghent, Belgium, 1999*. D. Aeyels, F. Lamnabhi-Lagarrigue & A.J. van der Schaft Editors. Lect. Notes Control and Information Sciences **246**, Springer-Verlag, 1999, pp. 325–354.

[145] 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.

[146] Sussmann, H.J., "New theories of set-valued differentials and new versions of the maximum principle of optimal control theory." In *Nonlinear Control in the Year 2000*, A. Isidori, F. Lamnabhi-Lagarrigue and W. Respondek Editors, Springer-Verlag, 2000, pp. 487-526.

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

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

[149] H.L. Trentelman, H. L., A. A. Stoorvogel, and M. L. J. Hautus, *Control Theory for Linear Systems.* Springer-Verlag, 2001.

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

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

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

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

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

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

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

[157] J.C. Willems, Paradigms and puzzles in the theory of dynamical systems, *IEEE Transactions on Automatic Control,* volume 36, pages 259-294, 1991.

[158] Willems, J. C., *On interconnections, control, and feedback.* IEEE Transactions on Automatic Control **42**, 1997, pp. 326-339.

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

XVIII.1 The optimal control and separation problems

The data. The basic data for our optimal control problem will be

(D.1) a ***control system***

$$\Sigma: \quad \dot{x} = f(x,u,t)\,,\ x \in \Omega,\ t \in I,\ u \in U\,,\ u(\cdot) \in \mathscr{U}\,, \tag{A.1}$$

(which will be also referred to, somewhat more concisely, as "the control system $\Sigma = (\Omega, U, I, f, \mathscr{U})$;" the letters x, u, t are, respectively, the "state variable," the "control variable," and the "time variable" of the system);

(D.2) a ***Lagrangian*** L.

(D.3) an ***endpoint constraint set*** $\mathscr{S}$;

(D.4) a ***reference trajectory-control pair*** $\gamma_* = (\xi_*, \eta_*)$;

The assumptions; controls, trajectories, endpoint conditions. We will assume that

(X.A1.1) *n is a positive integer, and Ω is an open subset of $\mathbb{R}^n$, known as the "state space" of the system;*

(X.A1.2) *U is a set, known as the "control set" (or "control space") of the system;*

Under these conditions, we can define the concept of a "U-control."

Definition XVIII.1. If U is a set, a ***U-control*** is a map $\eta : [a_\eta, b_\eta] \mapsto U$, which is defined on a compact subinterval domain$(\eta) = [a_\eta, b_\eta]$ of $\mathbb{R}$. □

We will also assume that

(X.A1.3) *I is a subinterval of the real line, known as the "time interval" of the system;*

(X.A1.4) *the "controlled dynamical law" f is a map*

$$\Omega \times U \times I \ni (x,u,t) \mapsto f(x,u,t) \in \mathbb{R}^n\,;$$

(X.A1.5) *the "class of admissible controls" $\mathscr{U}$ is a set of U-controls whose domains are subintervals of I;*

(X.A1.6) *$\mathscr{S}$ is a subset of* $\Omega \times \Omega$;

We can then define the concepts of "trajectory," "full trajectory," "trajectory-control pair," and "boundary value."

Remark XVIII.2. In Definitions XVIII.12 and XVIII.14 below we use the relatively more advanced mathematical ideas of "local absolute continuity" and "equality almost everywhere." Those readers who wish to see a proof of the maximum principle for classical systems (with piecewise continuous controls) that avoids these notions can refer instead to Definition X.46, where trajectories for classical systems are defined in a way which is equivalent to that of Definition XVIII.12 but does not involve the more advanced concepts. □

Definition XVIII.3. Given a control system Σ as in (A.7), a ***Σ-trajectory*** for a control $\eta : [a_\eta, b_\eta] \mapsto U$ is a locally absolutely continuous curve $\xi : I_\xi \mapsto \Omega$, defined on some nonempty subinterval I_ξ of $[a_\eta, b_\eta]$ such that

$$\dot{\xi}(t) = f(\xi(t), \eta(t), t) \quad \text{for almost all} \quad t \in I_\xi\,. \tag{A.2}$$

A ***full Σ-trajectory*** for η is a Σ-trajectory ξ whose domain I_ξ is the whole interval $[a_\eta, b_\eta]$.

We use $TCP(\Sigma;\eta)$ to denote the set of all full Σ-trajectories for η.

A ***trajectory-control pair*** for Σ is a pair $\gamma = (\xi, \eta)$ such that $\eta \in \mathscr{U}$ and ξ belongs to $TCP(\Sigma;\eta)$.

We use $TCP(\Sigma)$ to denote the set of all trajectory-control pairs of Σ. □

Definition XVIII.4. If $\xi : [a,b] \mapsto \Omega$ the ***endpoint value*** (or ***endpoint condition***, or ***boundary value***, or ***boundary condition***) ***of*** ξ is the pair

$$\partial\xi \stackrel{\text{def}}{=} (\xi(a), \xi(b))\,.$$

We use $\mathbf{EPS}(\Sigma)$ to denote the set $\{\partial\xi : (\xi, \eta) \in TCP(\Sigma)\}$ of all endpoint values of all full trajectories ξ corresponding to controls $\eta \in \mathscr{U}$. □

We will then make three additional assumptions:

(X.A1.7) $\gamma_* = (\xi_*, \eta_*) \in TCP(\Sigma)$;

(X.A1.8) $\partial\xi_* \in \mathscr{S}$ *(cf. Definition XVIII.13 below);* U-control η

(X.A1.9) *the "Lagrangian" L is a function*

$$\Omega \times U \times I \ni (x,u,t) \mapsto L(x,u,t) \in \mathbb{R}.$$

Minimizers. Suppose we are given a system Σ as in (A.7), and a Lagrangian L as in (X.A1.9).

Definition XVIII.5. An ***L-augmented Σ-trajectory*** for a U-control η is an ordered pair $\Xi = (\xi^0, \xi)$ such that ξ is a Σ-trajectory for η, $\xi^0 : I_\xi \mapsto \mathbb{R}$ is a locally absolutely continuous function, and

$$\dot{\xi}^0(t) = L(\xi(t), \eta(t), t) \qquad \text{for almost all} \qquad t \in I_\xi \,. \tag{A.3}$$

A ***full L-augmented Σ-trajectory*** for η is an augmented trajectory $\Xi = (\xi^0, \xi)$ such that $I_\xi = [a_\eta, b_\eta]$. We will use the expression $TCP(\Sigma, L; \eta)$ to denote the set of all full L-augmented Σ-trajectories for η.

An ***L-augmented trajectory-control pair for*** Σ is a pair $\Gamma = (\Xi, \eta)$ such that $\Xi \in TCP(\Sigma, L; \eta)$. We write $TCP(\Sigma, L)$ to denote the set of all trajectory-control pairs for Σ. □

Definition XVIII.6. The ***cost functional*** for the system Σ and the Lagrangian L is the function $\mathscr{J}_{\Sigma,L} : TCP(\Sigma, L) \mapsto \mathbb{R}$ that assigns to each L-augmented trajectory-control pair $\Gamma = (\Xi, \eta) = (\xi^0, \xi, \eta)$ for Σ the number $\mathscr{J}_{\Sigma,L}(\Gamma)$ given by

$$\begin{aligned} \mathscr{J}_{\Sigma,L}(\Gamma) &= \mathscr{J}_{\Sigma,L}(\Xi, \eta) \\ &\stackrel{\text{def}}{=} \mathscr{J}_{\Sigma,L}(\xi, \eta) \\ &= \xi^0(b_\eta) - \xi^0(a_\eta) \\ &= \int_{a_\eta}^{b_\eta} L(\xi(t), \eta(t), t)\, dt \,. \end{aligned}$$

A ***minimizer*** for the optimal control problem defined by Σ, L, and $\mathscr{S}$, is an L-augmented trajectory-control pair $\Gamma_* = (\xi^0_*, \xi_*, \eta_*)$ of Σ such that $\partial\xi_* \in \mathscr{S}$ and

(MIN) $\mathscr{J}_{\Sigma,L}(\xi^0_*, \xi_*, \eta_*) \le \mathscr{J}_{\Sigma,L}(\xi^0, \xi, \eta)$ *for all* $(\xi^0, \xi, \eta) \in TCP(\Sigma, L)$ *such that* $\partial\xi \in \mathscr{S}$,

i.e., Γ_* minimizes the cost functional $\mathscr{J}_{\Sigma,L}$ among all L-augmented trajectory-control pairs (ξ^0, ξ, η) of Σ such that $\partial\xi \in \mathscr{S}$. □

Remark XVIII.7. In Definition XVIII.15 we introduced the notation $\mathscr{J}_{\Sigma,L}(\xi, \eta)$ as an alternative for $\mathscr{J}_{\Sigma,L}(\Xi, \eta)$, if $\Xi = (\xi^0, \xi)$. This is justified because, if the pair (ξ, η) is fixed, then any two functions ξ^0_1, ξ^0_2, such that (ξ^0_1, ξ, η) and (ξ^0_1, ξ, η) belong to $TCP(\Sigma, L)$ must differ by a constant, so $\xi^0_1(b_\eta) - \xi^0_1 = \xi^0_2(b_\eta) - \xi^0_2(a_\eta)$, and then $\mathscr{J}_{\Sigma,L}(\xi^0_1, \xi, \eta) = \mathscr{J}_{\Sigma,L}(\xi^0_1, \xi, \eta)$.

If we think of $\mathscr{J}_{\Sigma,L}$ as a functional $(\xi,\eta) \mapsto \mathscr{J}_{\Sigma,L}(\xi,\eta)$, then the domain of $\mathscr{J}_{\Sigma,L}$ should be the set $TCP_L(\Sigma)$ of all "L-admissible" trajectory-control pairs of Σ, where a trajectory-control pair $(\xi,\eta) \in TCP(\Sigma)$ is said to be ***L-admissible*** if the function $[a_\eta, b_\eta] \ni t \mapsto L(\xi(t),\eta(t))$ is Lebesgue integrable. ☐

The optimal control problem. We will try to solve the following

OPTIMAL CONTROL PROBLEM (OCP)
Given data n, $\Sigma = (\Omega, U, I, f, \mathscr{U})$, L, $\mathscr{S}$, $\gamma_* = (\xi_*, \eta_*)$, such that conditions (X.A1.1-2-3-4-5-6-7-8-9) hold; **find**: a necessary condition for the pair $\gamma_* = (\xi_*, \eta_*)$ to be a minimizer.

The separation problem. The data for the separation problem are the same as those for the optimal control problem, except that there is no Lagrangian, and instead of looking for minimizers we look for trajectories that satisfy a separation condition.

Definition XVIII.8. The ***endpoint value set*** (or ***endpoint condition set***, or ***boundary value set***, or ***boundary condition set***) for a control system Σ is the set $\mathbf{EPS}(\Sigma)$ of all endpoint values $\partial\xi$, for all pairs $(\xi,\eta) \in TCP(\Sigma)$. ☐

Definition XVIII.9. If A, B are subsets of a metric space M (cf. §X.2.1), we say that A is ***separated*** from B at a point $m \in M$ if $A \cap B = \{m\}$ We say that A is ***locally separated*** from B at a point $m \in M$ if $A \cap B \cap \mathscr{N} = \{m\}$ for some neighborhood $\mathscr{N}$ of m in M. ☐

SEPARATION PROBLEM (SP)
Given: data n, $\Sigma = (\Omega, U, I, f, \mathscr{U})$, $\mathscr{S}$, $\gamma_* = (\xi_*, \eta_*)$, such that conditions (X.A1.1-2-3-4-5-6-7-8) hold; **find**: a necessary condition for the endpoint value set $\mathbf{EPS}(\Sigma)$ to be locally separated from $\mathscr{S}$ at $\partial\xi_*$.

Reduction of the optimal control problem to the separation problem. Having solved the SP, we show how to derive as a corollary a solution of the OCP. This will be an easy matter, for reasons that we now briefly outline. (The detailed proof, with a precise specification of all relevant technical conditions, will be given later.)

Suppose that the L-augmented pair $\Gamma_* = (\Xi_*, \eta_*) = (\xi_*^0, \xi_*, \eta_*)$ is a minimizer for the OCP with data $n, \Sigma = (\Omega, U, I, f, \mathscr{U})$, L, and $\mathscr{S}$. We let $\hat{\Sigma}$ be the "augmented control system" $(\hat{\Omega}, U, I, \hat{f}, \mathscr{U})$, where $\hat{\Omega} = \mathbb{R} \times \Omega$ and

$$\hat{f}(x^0, x, u, t) = (L(x,u,t), f(x,u,t)) \quad \text{if} \quad (x^0, x) \in \hat{\Omega},\ u \in U,\ t \in I. \tag{A.4}$$

That is, $\hat{\Sigma}$ is the system

$$\hat{\Sigma}: \qquad \dot{x}^0 = L(x,u,t),\ \dot{x} = f(x,u,t),\ (x^0, x) \in \mathbb{R} \times \Omega,\ u \in U,\ u(\cdot) \in \mathscr{U}, \tag{A.5}$$

obtained from Σ by adding the "running cost" x^0 as a new state variable.

If we write $a_* = a_{\eta_*}$, $b_* = b_{\eta_*}$, $x_* = \xi_*(a_*)$, $\tilde{x}_* = \xi_*(b_*)$, $x_* = \xi_*^0(a_*)$, $\tilde{x}_* = \xi_*^0(b_*)$, $c_* = \tilde{x}_*^0 - x_*^0 = \mathscr{J}_{\Sigma,L}(\Gamma_*)$, and define $\hat{\mathscr{S}}$ to be the set of all ordered pairs $\Big((x^0, x), (\tilde{x}^0, \tilde{x})\Big) \in \hat{\Omega} \times \hat{\Omega}$ such that

(a) $(x, \tilde{x}) \in \mathscr{S}$,

(b) $\tilde{x}^0 - x^0 \le c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2$

(c) $x^0 = x_*^0$,

then it is clear that Ξ_* is a full $\hat{\Sigma}$-trajectory for η_* and $\partial\Xi_* \in \hat{\mathscr{S}}$.

Furthermore, if $Q \in \mathbf{EPS}(\hat{\Sigma}) \cap \hat{\mathscr{S}}$, then there must exist a trajectory-control pair $\Gamma = (\Xi, \eta)$ of $\hat{\Sigma}$ such that $Q = \partial\Xi \in \hat{\mathscr{S}}$. If we write $a = a_\eta$, $b = b_\eta$, $\Xi = (\xi^0, \xi)$, $x = \xi(a)$, $\tilde{x} = \xi(b)$, $x^0 = \xi^0(a)$, $\tilde{x}^0 = \xi^0(b)$, $c = \tilde{x}^0 - x^0 = \mathscr{J}_{\Sigma,L}(\Gamma)$, then the fact that $\partial\Xi \in \hat{\mathscr{S}}$ implies that $\partial\xi \in \mathscr{S}$ and

$$c \le c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2. \tag{A.6}$$

Since Γ_* is a minimizer, and $\partial\xi \in \mathscr{S}$, we have $\mathscr{J}_{\Sigma,L}(\Gamma_*) \le \mathscr{J}_{\Sigma,L}(\Gamma)$, i.e., $c_* \le c$. Therefore $c_* \le c \le c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2$, from which it follows that $x = x_*$, $\tilde{x} = \tilde{x}_*$, and $c = c_*$. Since $x^0 = x_*^0$, we can conclude that $\tilde{x}^0 = \tilde{x}_*^0$, so that $\partial\Xi = \partial\Xi_*$. Hence $Q = \partial\Xi_*$.

We have shown that $\mathbf{EPS}(\hat{\Sigma}) \cap \hat{\mathscr{S}} = \{\partial\Xi_*\}$. Therefore $\mathbf{EPS}(\hat{\Sigma})$ and $\hat{\mathscr{S}}$ are locally separated at $\partial\Xi_*$. We can the apply the necessary condition for separation given by the solution of the SP, and obtain a necessary condition for minimization for OCP. We will carry out the details later, after we have described the precise technical hypotheses for our results. $\square$

XVIII.2 Proof of a classical maximum principle

As explained in the previous section, the optimal control problem can be reduced to the separation problem. We will therefore focus our efforts from now on the SP.

So we assume we are given data n, U, f, Σ, η_*, ξ_*, a, b, $\mathscr{S}$ as in the Separation Problem box (page 471), and begin by outlining the construction of packets of needle variations. For simplicity, we will present the outline only for the fixed time interval case, although when we do the technical details rigorously we will take care of both cases.

The main purpose of the construction of packets of needle variations is to compute, for our given reference trajectory-control pair (ξ_*, η_*), a "first-order approximation" $\mathscr{P}$ of the set $\mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ near the point $\xi_*(b)$. The set $\mathscr{P}$ should be a convex cone in $\mathbb{R}^n$, and the fact that it is a "first-order approximation" of $\mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ near $\xi_*(b)$ requires a precise definition to be given later, but means, to begin with, that each member of $\mathscr{P}$ is the tangent vector e at $\varepsilon = 0$ to a curve $\gamma : [0, \bar{\varepsilon}] \mapsto \mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ such that $\gamma(0) = \xi_*(b)$.

If we have another convex cone $\mathscr{Q}$ which is "tangent" to $\mathscr{S}$ in a similar way, then it is natural to guess that a necessary condition for separation of the sets $\mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ and $\mathscr{S}$ should be the linear separation of the cones $\mathscr{P}$ and $\mathscr{Q}$, i.e., the existence of a nontrivial linear functional $\lambda : \mathbb{R}^n \mapsto \mathbb{R}$ such that $\langle \lambda, v \rangle \geq 0$ for all $v \in \mathscr{P}$ and $\langle \lambda, v \rangle \leq 0$ for all $v \in \mathscr{Q}$.

It turns out, however, that if we use the naïve concept of "tangent cone" hinted at in the previous discussion, it is not true that "if two sets $\mathscr{R}$ and $\mathscr{S}$ are separated at a point x_*, and have tangent cones $\mathscr{P}$ and $\mathscr{Q}$ at x_*, then the cones $\mathscr{P}$ and $\mathscr{Q}$ are linearly separated."

Roughly speaking, a (fixed time interval) "variation" of the reference control η_* is a "curve" $[0, \bar{\varepsilon}] \ni \varepsilon \mapsto \eta^{\varepsilon}$ in the space $A_{[a,b]}(\Sigma)$ of all control functions $[a,b] \ni t \mapsto \eta(t) \in U$ defined on $[a,b]$ that starts at η_*. That is, a ***variation*** of η_* is a one-parameter family $V = \{(\xi^{\varepsilon}, \eta^{\varepsilon}) : \varepsilon \in [0, \bar{\varepsilon}\}$ of trajectory-control pairs of Σ such that $(\xi^0, \eta^0) = (\xi_*, \eta_*)$.

Roughly speaking, a (fixed time interval) "variation" of the reference trajectory-control pair (ξ_*, η_*) is a "curve" $[0, \bar{\varepsilon}] \ni \varepsilon \mapsto (\xi^{\varepsilon}, \eta^{\varepsilon})$ in the space $TCP_{[a,b]}(\Sigma)$ of all trajectory-control pairs of Σ defined on $[a,b]$ that starts at (ξ_*, η_*). That is, a ***variation*** of (ξ_*, η_*) is a one-parameter family $V = \{\eta^{\varepsilon} : \varepsilon \in [0, \bar{\varepsilon}\}$ of controls defined on $[a,b]$ such that $\eta^0 = \eta_*$.

Each control function $\eta : [a,b] \mapsto U$ gives rise to a ***time-varying vector field*** f_η, defined by letting $f_\eta(x,t) = f(x, \eta(t), t)$.

We assume that we are given

of the given at various times τ in the domain $[a,b]$ of ξ_*. The effects of these variations are then propagated to the terminal time bb by means of the differentials $D\Phi_{b,\tau}(\xi_*(\tau))$ of the reference flow maps $\Phi_{b,\tau}$. This produces a convex cone (the "Pontryagin cone") which in some sense is a first-order approximation to the attainable set near the point $\xi_*(b)$. The proof is then concluded by means of a topological argument, based on the Brower fixed point theorem, to pass from the separation of two sets to the separation of their approximating cones.

Each needle variation consists of a family $V = \{\xi^\varepsilon\}_{\varepsilon\in[0,\bar\varepsilon]}$ of trajectories, which are defined on $[a,b]$ if V is an "equal time variation," or on intervals $[a^\varepsilon, b^\varepsilon]$ if V is a "variable time variation," and are such that $\xi^0 = \xi_*$ and $\xi^\varepsilon(a^\varepsilon) = \xi_*(a)$ afor all ε. Under suitable technical conditions, the terminal points $\xi^\varepsilon(b^\varepsilon)$ of the arcs ξ^ε describe a continuous curve $\gamma_{_V} : [0,\bar\varepsilon] \mapsto X$ (where X is the state space of our control system), which is differentiable at $\varepsilon = 0$. The "effect" of the variation V at the terminal point is then the vector $e_{_V} = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\gamma_{_V}(\varepsilon)$. This vector is, by construction, "tangent" to the set $\mathscr{R}$

COMMENT NO. 45. THE MAP IN PAGE 16 COMES FROM THE WEB SITE

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

IS THERE A COPYRIGHT PROBLEM?

COMMENT NO. 46. THE PICTURE IN PAGE 12 COMES FROM THE WEB SITE

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

IS THERE A COPYRIGHT PROBLEM?

COMMENT NO. 47. Re: Page 41. Maybe we should put in some stuff about transversality conditions, after all!

COMMENT NO. 48. Re: Page 41. Say something about dynamic programming here.

COMMENT NO. 49. Re: Footnote 17 in page 42. Do we really do that in the appendix?

COMMENT NO. 50. Re: Remark II.9 in page 46: must put more stuff on the Dubins-Reeds-Shepp problem.

COMMENT NO. 51. Re: Page 42. Check that we do indeed show in §II.10 what we say we do.

COMMENT NO. 52. Re: Page 99 Put explanation of why (II) is equivalent to (I).

COMMENT NO. 53. Re: Page 100. Put elementary proof of Z1 and Z2

COMMENT NO. 54. The Euler footnote 8 in page 103 has to be expanded, especially adding stuff about Euler after 1727, and in particular on Euler's friendship with Daniel B., and how it affected the relation with Johann B. (I am thinking of Euler passing on to Johann the manuscript he got from Daniel on the hydrodinamics law, which was then plagiarized by Johann. **Maybe this footnote should be changed into a Scholium?**

COMMENT NO. 55. Re: Page 104. We need something biographical on Lagrange. It should include the moving story of his great love for a much younger woman when he was old. (I think her name was Renée).

COMMENT NO. 56. Re: Page 105 We need some biographical stuff on Legendre.

COMMENT NO. 57. Re: page 106. Need biographical sketch for Hamilton

COMMENT NO. 58. Re: page 106. We should clarify the following issue: was Hamilton the first person to write "Hamilton's equations"? I have heard somewhere (I wish I could remember where) that it was Cauchy. Notice that in the text I talk about the "classical Hamiltonian" but I avoid saying "the Hamiltonian that Hamilton wrote" for this reason. We need some research here.

COMMENT NO. 59. Re: Page 115 Need biographical stuff about Weierstrass

COMMENT NO. 60. Re: Page 120. Need biographical stuff about Caratheorody

XVIII.3 The optimal control and separation problems

The data. The basic data for our optimal control problem will be

(D.1) a ***control system***

$$\Sigma: \quad \dot{x} = f(x,u,t)\,,\ x \in \Omega\,,\ t \in I\,,\ u \in U\,,\ u(\cdot) \in \mathscr{U}\,, \qquad \text{(A.7)}$$

(which will be also referred to, somewhat more concisely, as "the control system $\Sigma = (\Omega, U, I, f, \mathscr{U})$;" the letters x, u, t are, respectively, the "state variable," the "control variable," and the "time variable" of the system);

(D.2) a ***Lagrangian*** L.

(D.3) an ***endpoint constraint set*** $\mathscr{S}$;

(D.4) a ***reference trajectory-control pair*** $\gamma_* = (\xi_*, \eta_*)$;

The assumptions; controls, trajectories, endpoint conditions. We will assume that

(X.A1.1) *n is a positive integer, and Ω is an open subset of $\mathbb{R}^n$, known as the "state space" of the system;*

(X.A1.2) *U is a set, known as the "control set" (or "control space") of the system;*

Under these conditions, we can define the concept of a "U-control."

Definition XVIII.10. If U is a set, a ***U-control*** is a map $\eta : [a_\eta, b_\eta] \mapsto U$, which is defined on a compact subinterval domain$(\eta) = [a_\eta, b_\eta]$ of $\mathbb{R}$. □

We will also assume that

(X.A1.3) *I is a subinterval of the real line, known as the "time interval" of the system;*

(X.A1.4) *the "controlled dynamical law" f is a map*

$$\Omega \times U \times I \ni (x,u,t) \mapsto f(x,u,t) \in \mathbb{R}^n\,;$$

(X.A1.5) *the "class of admissible controls" $\mathscr{U}$ is a set of U-controls whose domains are subintervals of I;*

(X.A1.6) *$\mathscr{S}$ is a subset of* $\Omega \times \Omega$;

We can then define the concepts of "trajectory," "full trajectory," "trajectory-control pair," and "boundary value."

Remark XVIII.11. In Definitions XVIII.12 and XVIII.14 below we use the relatively more advanced mathematical ideas of "local absolute continuity" and "equality almost everywhere." Those readers who wish to see a proof of the maximum principle for classical systems (with piecewise continuous controls) that avoids these notions can refer instead to Definition X.46, where trajectories for classical systems are defined in a way which is equivalent to that of Definition XVIII.12 but does not involve the more advanced concepts. □

Definition XVIII.12. Given a control system Σ as in (A.7), a ***Σ-trajectory*** for a control $\eta : [a_\eta, b_\eta] \mapsto U$ is a locally absolutely continuous curve $\xi : I_\xi \mapsto \Omega$, defined on some nonempty subinterval I_ξ of $[a_\eta, b_\eta]$ such that

$$\dot{\xi}(t) = f(\xi(t), \eta(t), t) \quad \text{for almost all} \quad t \in I_\xi\,. \tag{A.8}$$

A ***full Σ-trajectory*** for η is a Σ-trajectory ξ whose domain I_ξ is the whole interval $[a_\eta, b_\eta]$.

We use $TCP(\Sigma;\eta)$ to denote the set of all full Σ-trajectories for η.

A ***trajectory-control pair*** for Σ is a pair $\gamma = (\xi, \eta)$ such that $\eta \in \mathscr{U}$ and ξ belongs to $TCP(\Sigma;\eta)$.

We use $TCP(\Sigma)$ to denote the set of all trajectory-control pairs of Σ. □

Definition XVIII.13. If $\xi : [a,b] \mapsto \Omega$ the ***endpoint value*** (or ***endpoint condition***, or ***boundary value***, or ***boundary condition***) ***of*** ξ is the pair

$$\partial\xi \stackrel{\text{def}}{=} (\xi(a), \xi(b))\,.$$

We use $\mathbf{EPS}(\Sigma)$ to denote the set $\{\partial\xi : (\xi,\eta) \in TCP(\Sigma)\}$ of all endpoint values of all full trajectories ξ corresponding to controls $\eta \in \mathscr{U}$. □

We will then make three additional assumptions:

(X.A1.7) $\gamma_* = (\xi_*, \eta_*) \in TCP(\Sigma)$;

(X.A1.8) $\partial\xi_* \in \mathscr{S}$ *(cf. Definition XVIII.13 below);* U-control η

(X.A1.9) *the "Lagrangian" L is a function*

$$\Omega \times U \times I \ni (x,u,t) \mapsto L(x,u,t) \in \mathbb{R}.$$

Minimizers. Suppose we are given a system Σ as in (A.7), and a Lagrangian L as in (X.A1.9).

Definition XVIII.14. An ***L-augmented Σ-trajectory*** for a U-control η is an ordered pair $\Xi = (\xi^0, \xi)$ such that ξ is a Σ-trajectory for η, $\xi^0 : I_\xi \mapsto \mathbb{R}$ is a locally absolutely continuous function, and

$$\dot{\xi}^0(t) = L(\xi(t), \eta(t), t) \qquad \text{for almost all} \qquad t \in I_\xi\,. \tag{A.9}$$

A ***full L-augmented Σ-trajectory*** for η is an augmented trajectory $\Xi = (\xi^0, \xi)$ such that $I_\xi = [a_\eta, b_\eta]$. We will use the expression $TCP(\Sigma, L; \eta)$ to denote the set of all full L-augmented Σ-trajectories for η.

An ***L-augmented trajectory-control pair for*** Σ is a pair $\Gamma = (\Xi, \eta)$ such that $\Xi \in TCP(\Sigma, L; \eta)$. We write $TCP(\Sigma, L)$ to denote the set of all trajectory-control pairs for Σ. ☐

Definition XVIII.15. The ***cost functional*** for the system Σ and the Lagrangian L is the function $\mathscr{I}_{\Sigma,L} : TCP(\Sigma, L) \mapsto \mathbb{R}$ that assigns to each L-augmented trajectory-control pair $\Gamma = (\Xi, \eta) = (\xi^0, \xi, \eta)$ for Σ the number $\mathscr{I}_{\Sigma,L}(\Gamma)$ given by

$$\begin{aligned}
\mathscr{I}_{\Sigma,L}(\Gamma) &= \mathscr{I}_{\Sigma,L}(\Xi, \eta) \\
&\stackrel{\text{def}}{=} \mathscr{I}_{\Sigma,L}(\xi, \eta) \\
&= \xi^0(b_\eta) - \xi^0(a_\eta) \\
&= \int_{a_\eta}^{b_\eta} L(\xi(t), \eta(t), t)\, dt\,.
\end{aligned}$$

A ***minimizer*** for the optimal control problem defined by Σ, L, and $\mathscr{S}$, is an L-augmented trajectory-control pair $\Gamma_* = (\xi^0_*, \xi_*, \eta_*)$ of Σ such that $\partial\xi_* \in \mathscr{S}$ and

(MIN) $\mathscr{I}_{\Sigma,L}(\xi^0_*, \xi_*, \eta_*) \le \mathscr{I}_{\Sigma,L}(\xi^0, \xi, \eta)$ *for all* $(\xi^0, \xi, \eta) \in TCP(\Sigma, L)$ *such that* $\partial\xi \in \mathscr{S}$,

i.e., Γ_* minimizes the cost functional $\mathscr{I}_{\Sigma,L}$ among all L-augmented trajectory-control pairs (ξ^0, ξ, η) of Σ such that $\partial\xi \in \mathscr{S}$. ☐

Remark XVIII.16. In Definition XVIII.15 we introduced the notation $\mathscr{I}_{\Sigma,L}(\xi, \eta)$ as an alternative for $\mathscr{I}_{\Sigma,L}(\Xi, \eta)$, if $\Xi = (\xi^0, \xi)$. This is justified because, if the pair (ξ, η) is fixed, then any two functions ξ^0_1, ξ^0_2, such that (ξ^0_1, ξ, η) and (ξ^0_1, ξ, η) belong to $TCP(\Sigma, L)$ must differ by a constant, so $\xi^0_1(b_\eta) - \xi^0_1 = \xi^0_2(b_\eta) - \xi^0_2(a_\eta)$, and then $\mathscr{I}_{\Sigma,L}(\xi^0_1, \xi, \eta) = \mathscr{I}_{\Sigma,L}(\xi^0_1, \xi, \eta)$.

If we think of $\mathscr{J}_{\Sigma,L}$ as a functional $(\xi,\eta) \mapsto \mathscr{J}_{\Sigma,L}(\xi,\eta)$, then the domain of $\mathscr{J}_{\Sigma,L}$ should be the set $TCP_L(\Sigma)$ of all "L-admissible" trajectory-control pairs of Σ, where a trajectory-control pair $(\xi,\eta) \in TCP(\Sigma)$ is said to be ***L-admissible*** if the function $[a_\eta, b_\eta] \ni t \mapsto L(\xi(t),\eta(t))$ is Lebesgue integrable. $\square$

The optimal control problem. We will try to solve the following

OPTIMAL CONTROL PROBLEM (OCP)
Given data n, $\Sigma = (\Omega, U, I, f, \mathscr{U})$, L, $\mathscr{S}$, $\gamma_* = (\xi_*, \eta_*)$, such that conditions (X.A1.1-2-3-4-5-6-7-8-9) hold; **find**: a necessary condition for the pair $\gamma_* = (\xi_*, \eta_*)$ to be a minimizer.

The separation problem. The data for the separation problem are the same as those for the optimal control problem, except that there is no Lagrangian, and instead of looking for minimizers we look for trajectories that satisfy a separation condition.

Definition XVIII.17. The ***endpoint value set*** (or ***endpoint condition set***, or ***boundary value set***, or ***boundary condition set***) for a control system Σ is the set $\mathbf{EPS}(\Sigma)$ of all endpoint values $\partial\xi$, for all pairs $(\xi,\eta) \in TCP(\Sigma)$. $\square$

Definition XVIII.18. If A, B are subsets of a metric space M (cf. §X.2.1), we say that A is ***separated*** from B at a point $m \in M$ if $A \cap B = \{m\}$ We say that A is ***locally separated*** from B at a point $m \in M$ if $A \cap B \cap \mathscr{N} = \{m\}$ for some neighborhood $\mathscr{N}$ of m in M. $\square$

SEPARATION PROBLEM (SP)
Given: data n, $\Sigma = (\Omega, U, I, f, \mathscr{U})$, $\mathscr{S}$, $\gamma_* = (\xi_*, \eta_*)$, such that conditions (X.A1.1-2-3-4-5-6-7-8) hold; **find**: a necessary condition for the endpoint value set $\mathbf{EPS}(\Sigma)$ to be locally separated from $\mathscr{S}$ at $\partial\xi_*$.

Reduction of the optimal control problem to the separation problem. Having solved the SP, we show how to derive as a corollary a solution of the OCP. This will be an easy matter, for reasons that we now briefly outline. (The detailed proof, with a precise specification of all relevant technical conditions, will be given later.)

Suppose that the L-augmented pair $\Gamma_* = (\Xi_*, \eta_*) = (\xi_*^0, \xi_*, \eta_*)$ is a minimizer for the OCP with data $n, \Sigma = (\Omega, U, I, f, \mathscr{U}), L$, and $\mathscr{S}$. We let $\hat{\Sigma}$ be the "augmented control system" $(\hat{\Omega}, U, I, \hat{f}, \mathscr{U})$, where $\hat{\Omega} = \mathbb{R} \times \Omega$ and

$$\hat{f}(x^0, x, u, t) = (L(x, u, t), f(x, u, t)) \quad \text{if} \quad (x^0, x) \in \hat{\Omega},\ u \in U,\ t \in I. \tag{A.10}$$

That is, $\hat{\Sigma}$ is the system

$$\hat{\Sigma}: \qquad \dot{x}^0 = L(x, u, t),\ \dot{x} = f(x, u, t),\ (x^0, x) \in \mathbb{R} \times \Omega,\ u \in U,\ u(\cdot) \in \mathscr{U}, \tag{A.11}$$

obtained from Σ by adding the "running cost" x^0 as a new state variable.

If we write $a_* = a_{\eta_*}$, $b_* = b_{\eta_*}$, $x_* = \xi_*(a_*)$, $\tilde{x}_* = \xi_*(b_*)$, $x_* = \xi_*^0(a_*)$, $\tilde{x}_* = \xi_*^0(b_*)$, $c_* = \tilde{x}_*^0 - x_*^0 = \mathscr{J}_{\Sigma,L}(\Gamma_*)$, and define $\hat{\mathscr{S}}$ to be the set of all ordered pairs $\Big((x^0, x), (\tilde{x}^0, \tilde{x})\Big) \in \hat{\Omega} \times \hat{\Omega}$ such that

(a) $(x, \tilde{x}) \in \mathscr{S}$,

(b) $\tilde{x}^0 - x^0 \leq c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2$

(c) $x^0 = x_*^0$,

then it is clear that Ξ_* is a full $\hat{\Sigma}$-trajectory for η_* and $\partial \Xi_* \in \hat{\mathscr{S}}$.

Furthermore, if $Q \in \mathbf{EPS}(\hat{\Sigma}) \cap \hat{\mathscr{S}}$, then there must exist a trajectory-control pair $\Gamma = (\Xi, \eta)$ of $\hat{\Sigma}$ such that $Q = \partial \Xi \in \hat{\mathscr{S}}$. If we write $a = a_\eta$, $b = b_\eta$, $\Xi = (\xi^0, \xi)$, $x = \xi(a)$, $\tilde{x} = \xi(b)$, $x^0 = \xi^0(a)$, $\tilde{x}^0 = \xi^0(b)$, $c = \tilde{x}^0 - x^0 = \mathscr{J}_{\Sigma,L}(\Gamma)$, then the fact that $\partial \Xi \in \hat{\mathscr{S}}$ implies that $\partial \xi \in \mathscr{S}$ and

$$c \leq c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2. \tag{A.12}$$

Since Γ_* is a minimizer, and $\partial \xi \in \mathscr{S}$, we have $\mathscr{J}_{\Sigma,L}(\Gamma_*) \leq \mathscr{J}_{\Sigma,L}(\Gamma)$, i.e., $c_* \leq c$. Therefore $c_* \leq c \leq c_* - \|x - x_*\|^2 - \|\tilde{x} - \tilde{x}_*\|^2$, from which it follows that $x = x_*$, $\tilde{x} = \tilde{x}_*$, and $c = c_*$. Since $x^0 = x_*^0$, we can conclude that $\tilde{x}^0 = \tilde{x}_*^0$, so that $\partial \Xi = \partial \Xi_*$. Hence $Q = \partial \Xi_*$.

We have shown that $\mathbf{EPS}(\hat{\Sigma}) \cap \hat{\mathscr{S}} = \{\partial \Xi_*\}$. Therefore $\mathbf{EPS}(\hat{\Sigma})$ and $\hat{\mathscr{S}}$ are locally separated at $\partial \Xi_*$. We can the apply the necessary condition for separation given by the solution of the SP, and obtain a necessary condition for minimization for OCP. We will carry out the details later, after we have described the precise technical hypotheses for our results. □

So we assume we are given data n, U, f, Σ, η_*, ξ_*, a, b, $\mathscr{S}$ as in the Separation Problem box (page 471), and begin by outlining the construction of packets of needle variations. For simplicity, we will present the outline only for the fixed time interval case, although when we do the technical details rigorously we will take care of both cases.

The main purpose of the construction of packets of needle variations is to compute, for our given reference trajectory-control pair (ξ_*, η_*), a "first-order approximation" $\mathscr{P}$ of the set $\mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ near the point $\xi_*(b)$. The set $\mathscr{P}$ should be a convex cone in $\mathbb{R}^n$, and the fact that it is a "first-order approximation" of $\mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ near $\xi_*(b)$ requires a precise definition to be given later, but means, to begin with, that each member of $\mathscr{P}$ is the tangent vector e at $\varepsilon = 0$ to a curve $\gamma : [0, \bar{\varepsilon}] \mapsto \mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ such that $\gamma(0) = \xi_*(b)$.

If we have another convex cone $\mathscr{Q}$ which is "tangent" to $\mathscr{S}$ in a similar way, then it is natural to guess that a necessary condition for separation of the sets $\mathscr{R}^{\Sigma}_{[a,b]}(\xi_*(a))$ and $\mathscr{S}$ should be the linear separation of the cones $\mathscr{P}$ and $\mathscr{Q}$, i.e., the existence of a nontrivial linear functional $\lambda : \mathbb{R}^n \mapsto \mathbb{R}$ such that $\langle \lambda, v \rangle \geq 0$ for all $v \in \mathscr{P}$ and $\langle \lambda, v \rangle \leq 0$ for all $v \in \mathscr{Q}$.

It turns out, however, that if we use the naïve concept of "tangent cone" hinted at in the previous discussion, it is not true that "if two sets $\mathscr{R}$ and $\mathscr{S}$ are separated at a point x_*, and have tangent cones $\mathscr{P}$ and $\mathscr{Q}$ at x_*, then the cones $\mathscr{P}$ and $\mathscr{Q}$ are linearly separated."

Roughly speaking, a (fixed time interval) "variation" of the reference control η_* is a "curve" $[0, \bar{\varepsilon}] \ni \varepsilon \mapsto \eta^{\varepsilon}$ in the space $A_{[a,b]}(\Sigma)$ of all control functions $[a,b] \ni t \mapsto \eta(t) \in U$ defined on $[a,b]$ that starts at η_*. That is, a ***variation*** of η_* is a one-parameter family $V = \{(\xi^{\varepsilon}, \eta^{\varepsilon}) : \varepsilon \in [0, \bar{\varepsilon}\}$ of trajectory-control pairs of Σ such that $(\xi^0, \eta^0) = (\xi_*, \eta_*)$.

Roughly speaking, a (fixed time interval) "variation" of the reference trajectory-control pair (ξ_*, η_*) is a "curve" $[0, \bar{\varepsilon}] \ni \varepsilon \mapsto (\xi^{\varepsilon}, \eta^{\varepsilon})$ in the space $TCP_{[a,b]}(\Sigma)$ of all trajectory-control pairs of Σ defined on $[a,b]$ that starts at (ξ_*, η_*). That is, a ***variation*** of (ξ_*, η_*) is a one-parameter family $V = \{\eta^{\varepsilon} : \varepsilon \in [0, \bar{\varepsilon}\}$ of controls defined on $[a,b]$ such that $\eta^0 = \eta_*$.

Each control function $\eta : [a,b] \mapsto U$ gives rise to a ***time-varying vector field*** f_{η}, defined by letting $f_{\eta}(x,t) = f(x, \eta(t), t)$.

We assume that we are given

of the given at various times τ in the domain $[a,b]$ of ξ_*. The effects of these variations are then propagated to the terminal time bb by means of the differentials $D\Phi_{b,\tau}(\xi_*(\tau))$ of the reference flow maps $\Phi_{b,\tau}$. This produces a convex cone (the "Pontryagin cone") which in some sense is a first-order approximation to the attainable set near the point $\xi_*(b)$. The proof is then concluded by means of a topological argument, based on the Brower fixed point theorem, to pass from the separation of two sets to the separation of their approximating cones.

Each needle variation consists of a family $V = \{\xi^\varepsilon\}_{\varepsilon\in[0,\bar{\varepsilon}]}$ of trajectories, which are defined on $[a,b]$ if V is an "equal time variation," or on intervals $[a^\varepsilon, b^\varepsilon]$ if V is a "variable time variation," and are such that $\xi^0 = \xi_*$ and $\xi^\varepsilon(a^\varepsilon) = \xi_*(a)$ afor all ε. Under suitable technical conditions, the terminal points $\xi^\varepsilon(b^\varepsilon)$ of the arcs ξ^ε describe a continuous curve $\gamma_{_V} : [0,\bar{\varepsilon}] \mapsto X$ (where X is the state space of our control system), which is differentiable at $\varepsilon = 0$. The "effect" of the variation V at the terminal point is then the vector $e_{_V} = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\gamma_{_V}(\varepsilon)$. This vector is, by construction, "tangent" to the set $\mathscr{R}$