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DO NOT QUOTE WITHOUT PERMISSION.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 15, 2001.

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 §13.8, 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 §4.9,
- ***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 Leibinz 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

1 Introduction **1**

Scholium I. The brachistochrone monument 9

Scholium II. Johann Bernoulli, his family and his time 10

Scholium III. Tschirnhaus 13

Scholium IV. The Marquis de L'Hôpital 13

Scholium V. A. J. Tasman's exploration of the South Pacific . . . 14

2 A preview of our main themes **19**

2.1 The setting: calculus of variations vs. optimal control, and optimal control vs. control 19

2.2 Optimal control was born in the years 1696-97, or perhaps even earlier 19

2.3 The path from the brachistochrone problem to the maximum principle 20

2.4 The clarification of the role of the momentum 21

2.5 The significance of the maximum principle 21

2.6 Missed opportunities 22

2.7 The superior power of the control-theoretic formulation of Hamilton's equations 23

2.8 Hamilton's value function: the eclipse and re-emergence of a great idea 25

2.9 The issue of mathematical rigor 26

2.10 The geometric connection, and how Johann Bernoulli missed an opportunity to discover non-Euclidean geometry 27

2.11 Differential geometry and control 28

2.12 Some open problems 28

2.13 The brachistochrone 28
2.14 A speculative conclusion 30

3 Calculus of variations, optimal control, and control 31
3.1 Some notational conventions 31
3.2 Calculus of variations and optimal control problems 33
3.3 Configuration spaces 33
3.4 Spaces of curves 38
3.5 The conventional wisdom 39
3.6 An unconventional alternative 41
3.7 Standard calculus of variations problems 41
3.8 Isoperimetric problems 44
3.9 Optimal control problems 45
3.10 Isoperimetric calculus of variations problems are standard optimal control problems 50
3.11 Optimal control is much more general and richer than the calculus of variations 52
3.12 To a controlled dynamics there is associated an interesting Lie-algebra structure 53
3.13 Control is much more general than optimal control 54
3.14 Design questions 54
3.15 Feedback 55
3.16 The brachistochrone problem truly belongs to optimal control 61
Scholium VI. Solution of the stabilization problem of §3.15 62

4 From ancient times to the maximum principle 65
4.1 The era of special problems solved by ad hoc methods 66
4.1.1 Finding the shortest path between two points. . 66
4.1.2 The classical isoperimetric problem. 67
4.1.3 Light reflection. 68
4.1.4 Light refraction. 70
4.1.5 The catenary and the brachistochrone. 71
4.2 Euler 72
4.3 Lagrange 73
4.4 Legendre 74

4.5 Hamilton misses an opportunity 75
4.6 The "control Hamiltonian" versus the classical Hamiltonian . 77
4.7 J. L. Synge on Hamilton's principal function and the 100-year eclipse of a great idea 81
4.8 Weierstrass comes close, but fails to discover Hamiltonian maximization 84
4.9 Carathéodory gets even closer, but misses too 91
4.10 A PUZZLE: What has disappeared? 92
4.11 Three toy problems 94
4.12 Two more toy problems, and the "abnormal" multiplier . . . 99
4.13 A stronger conjecture and a third toy problem 101
Scholium VII. Snell, Descartes, Fermat, Leibniz, Huygens on refraction 105
Scholium VIII. Two excerpts from G. Adair's English translation of Georges Perec's "La Disparition" 108

5 The maximum principle of optimal control theory 111
5.1 Example I: the straight-line segment 114
5.2 Example II: the isoperimetric problem 114
5.2.1 Application of the maximum principle 119
5.2.2 Finding the solutions of (IP.V.a) 120
5.2.3 Existence of area-maximizers 124
5.2.4 The final result 127
5.2.5 How to make the solution of the isoperimetric problem rigorous 127
5.3 Example III: the catenary 133

6 The control Hamiltonian and the Hamiltonian formulation of relativistic mechanics 141

7 The value function and the comeback of Hamilton's great idea 147
7.1 The true meaning of the momentum variable 149

8 From principle to theorem 153
8.1 Principles and versions 153

8.2 Euler . . . 155
8.3 Lagrange . . . 155
8.4 The modern justification of Lagrange's derivation; the Lavrentiev phenomenon . . . 158
8.5 Legendre . . . 161
8.6 The needle variations approach . . . 163
8.7 The classical version of the maximum principle . . . 163

9 Recent versions of the maximum principle: the nonsmooth connection 167
9.1 The nonsmooth analysis approach . . . 167
9.2 High-order necessary conditions . . . 168
9.3 The "combined systems" problem . . . 168
9.4 A unified nonsmooth approach . . . 169

10 The geometric connection 173
10.1 Birds, bees, and the birth of geometry and optimal control . . 176
10.2 Line segments . . . 179
10.3 The parallel postulate . . . 180
10.4 Circles and the isoperimetric problem . . . 181
10.5 Shortest paths in geometrical optics . . . 182
10.6 Johann Bernoulli misses an opportunity to solve the problem of Euclid's fifth postulate . . . 184
10.7 The missing link: curve minimization as a rich source of new geometries . . . 187
10.8 Covariance and invariance, from Lagrange to Riemann 188
10.9 Invariants and Klein's Erlangen program . . . 191

11 Beyond the maximum principle, geometrically 193
11.1 Covariance and invariance in control: Lie brackets 194
11.2 Control, Lie brackets, zoology and medieval bestiaries 197
11.3 Using Lie brackets to understand optimal trajectories 198
11.4 Implementing Lagrange's covariance idea in optimal control . 200
11.5 Why the language of Lie brackets is natural for optimal control 201

12 Some open problems 203

13 The brachistochrone, 300 years ago and today 207

13.1 Johann Bernoulli's challenge 207

Scholium IX. The cycloid's double role as the brachistochrone and the tautochrone 209

13.2 Why was the brachistochrone so important in 1696-97? 210

13.3 Formulation of the brachistochrone problem 211

13.4 Johann Bernoulli's solution 213

13.5 Remarks on Johann Bernoulli's argument 216

13.6 The calculus of variations approach 217

13.7 The brachistochrone as an optimal control problem 218

13.8 Optimality proofs and the existence problem 220

13.9 Hestenes' "variation of the brachistochrone problem" 222

13.10 The reflected brachistochrone problem 224

13.11 A modest claim 225

13.12 Hamilton's principal function for the brachistochrone problem, and the differential-geometric connection 226

13.13 What is Hamilton's principal function for the brachistochrone? 227

13.14 Lie brackets, integrals of motion, and the differential-geometric perspective 228

13.15 The diamond Lie algebra 229

13.16 Energy conservation 230

13.17 Naïve application of the maximum principle 231

13.18 Application of the maximum principle on manifolds 231

13.19 The complete synthesis and Hamilton's principal function for the 4-dimensional brachistochrone problem 233

13.20 Deriving Snell' law from the maximum principle 234

14 Conclusion 237

Appendix: Some formal definitions 239

A.1. Controlled dynamical systems 239

A.2. Cost functionals, optimal control problems, and the value 240

A.3. Generalized solutions 240

A.4. The value function (i.e., Hamilton's principal function) . 242

A.5. Systems on manifolds, control bundles, partially defined vector fields 244

14.1 Example II: the isoperimetric problem 269
14.1.1 Problem statement 269
14.1.2 Problem reformulations 270
14.1.3 Finding the solutions of (IP") 273
14.2 Rigorous solution of the isoperimetric problem 277
14.2.1 Existence of area-maximizers 277
14.2.2 The final result 279
14.2.3 Polygonal loops 279
14.2.4 Rectifiable curves 280
14.2.5 Absolutely continuous curves 280
14.2.6 Parametrization by arc length 282
14.2.7 Changing variables 286
14.2.8 The complex exponential 290
14.2.9 Loops 291
14.2.10 Reparametrization of loops 292
14.2.11 Connected components 294
14.2.12 The winding number 294
14.2.13 The no-retraction theorem 296
14.2.14 The Jordan curve theorem 297
14.2.15 The winding numbers of a simple loop 307
14.3 Green's formula 311

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.)

COMMENT NO. 3. Something has to be done to change the macros for table of contents, so that there will be more space, for example, after 13.10 or 13.14 above.

Chapter 1

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

[1] Cf. Chapter 13 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. §13.1 in Page 208 below.

[3] We will speculate on these reasons in §13.2 (in Page 210) below.

[4] Though not unanimously, cf. *infra*, Chapter 3, esp. Remark 3.5.2.

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

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

So it is not surprising that the story of the brachistochrone is told in many books, and that the importance of this curve is stressed in most histories of mathematics. The brachistochrone has received honors not commonly bestowed on mathematical curves. For example, it is depicted in a stained-glass window of the Academy building (the main venue of the university) in Groningen (cf. page 13). 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 9).

COMMENT NO. 4.

PUT PICTURE OF THE BRACHISTOCHRONE MONUMENT IN GRONINGEN SOMEWHERE AROUND HERE

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. §13.8 in page 220 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 [121]!

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 14). 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

[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

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 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 13, 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

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

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.

- 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. 5. 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]

[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. §14.1 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

COMMENT NO. 6. 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.

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 ([121]) 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,

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.

- 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] 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 13 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. 7. 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

[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.

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

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. 8. 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.

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, [6], 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." ([6], p. 136.)

FIGURE 1. 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. §5.3, in page 133). 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." ([6], 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. 9. 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. [6], 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. [60]). 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. 10. 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" ([49], 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. §10.6 in Page 184).

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,

[13]The discovery took place in late November 1642. The first European to see the island was 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

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 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.

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 [68]), 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.

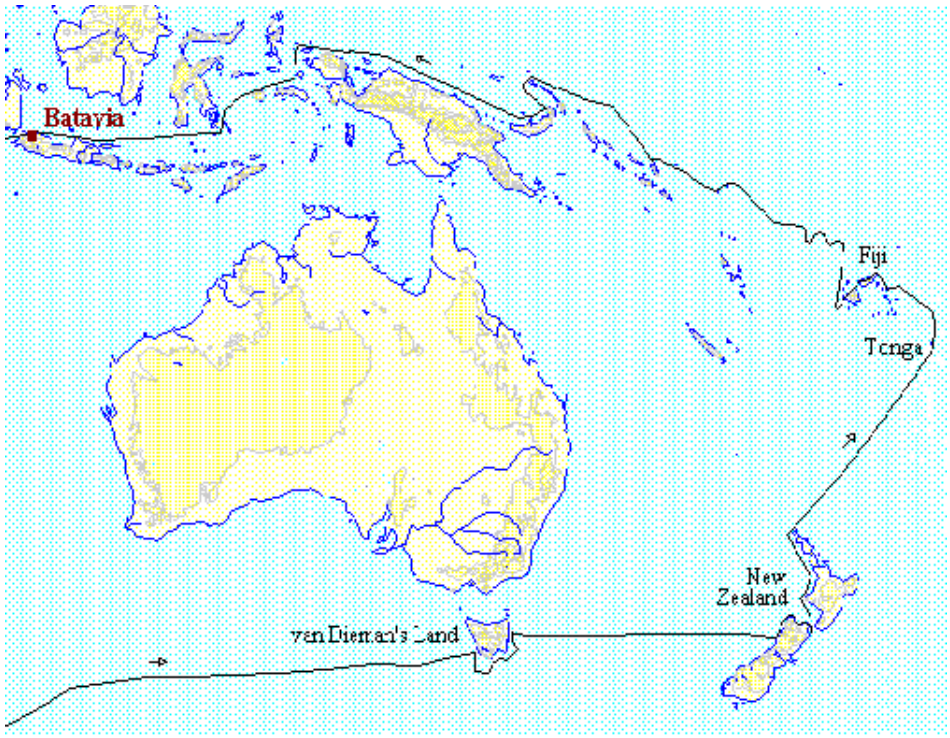


FIGURE 2. Tasman's 1642 trip

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 10, 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]. ♢

[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* ([94]), 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.*

Chapter 2

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.

2.1 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 3, the subject matters of the calculus of variations and optimal control, and explain how the latter is much 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 3 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.

2.2 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 3), 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 13—to what Leibniz called a "splendid" problem.

2.3 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 4 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 5.

2.4 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

$$\pi(t) = \pi_\xi(t) \stackrel{\text{def}}{=} \frac{\partial L}{\partial \dot{x}}(\xi(t), \dot{\xi}(t), t)\,, \tag{2.1}$$

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 (2.1).* 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 (2.1) 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" (2.1).

2.5 The significance of the maximum principle

The maximum principle was a major achievement that, as we shall explain in Chapters 4 and 5, 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 8, 9, 10, and 11.)

COMMENT NO. 11. JAN: In the "missed opportunities" section we have to say something about Dyson's famous article [47]. 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.

2.6 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.

2.7 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 4, 5 and 6 in three ways:

1. In Chapter 4 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 4, 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 4—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 §4.9, 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 6, 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 6), 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.

2.8 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 7. 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 7, 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. 12. On mathematical rigor: we have to check carefully that we really do all the things promised here in §2.9!

2.9 The issue of mathematical rigor

The progression described in Chapter 4 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 8 and 9 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. 13. 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 5, where it will be stated as a *principle* (i.e., not yet a complete theorem,

with precise hypotheses and conclusions), Chapters 8 and 9 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 [98] 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. 14. Need a box on "Principles and Versions," either here or in the rigorous chapters.

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

In Chapter 10 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. §10.2). 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 §11, which discusses the application of differential-geometric ideas to control, we include in Chapter 10 a general discussion of the concepts of covariance and invariance.

2.11 Differential geometry and control

In Chapter 11 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. 15. JOB FOR HJS: Expand and complete this section

2.12 Some open problems

In Chapter 12 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.

2.13 The brachistochrone

In Chapter 13 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 §13.12 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 [114, 115, 119, 120, 123, 124, 125]). Hestenes' problem is "an interesting variation of the brachistochrone" dealing with "an idealized rocket ship moving in a vertical plane (Hestenes [66], 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.[1] But all three problems are covered by the more recent nonsmooth result.

[1]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.

2.14 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.

Chapter 3

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.

3.1 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 (4.5) 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 $\mathbb{R}^3$.

3.2 Calculus of variations and optimal control problems

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

$$J(\xi) = \int_a^b L(\xi(t), \dot{\xi}(t), t)\, dt \tag{3.1}$$

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.

3.3 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]

$$q = (q^1, q^2, \ldots, q^n) \tag{3.2}$$

[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.

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

$$\dot{q} = (\dot{q}^1, \dot{q}^2, \ldots, \dot{q}^n) \in \mathbb{R}^n \,. \tag{3.3}$$

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 (3.2) and (3.3).

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 (3.2) is the ordinary matrix product[5]

$$p \cdot q = p_1 q^1 + p_2 q^2 + \ldots + p_n q^n = \sum_{j=1}^{n} p_j q^j \,. \tag{3.4}$$

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

[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 (3.4), but we will not go that far.

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 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 "cutoff" 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].

---

[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-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 $d(a, c) \leq d(a, b) + d(b, c)$),

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

---

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 $r \geq 0$.

[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.

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 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

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

[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.

[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 $\mathcal{F}$ 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 $E \times S$.

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 3.3.1 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 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 3.3.2 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.* ♢

3.4 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 (3.1) 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 4.11.3 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 §8.4 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 $\mathcal{C}$ 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 $\mathcal{C}$.

3.5 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

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

his group (cf. [98]), or perhaps a few years earlier, with the work of McShane and Hestenes, cf. [90, 65]. 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 3.5.1 The article by L. Markus in [48] 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 non-mathematical 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 took place in 1696-97. For example, G. A. Bliss wrote in 1925 (cf. [14]):

> 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. [107], 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."

[14]that is, Jakob

D. J. Struik, in [109], 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 3.5.2 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. [74]), 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." ♢

3.6 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.*

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.

3.7 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.***

[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 [33] on alternate realities, Casti indeed observes that "classical optimal control theory dates back at least to the well-known 'brachistochrone' problem."

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 §3.4, "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":

Given	Q, L, a, b, $\bar{q}$, $\hat{q}$, Ξ,
minimize	$I = \int_a^b L(\xi(t), \dot{\xi}(t), t)dt$,
subject to	$\xi(a) = \bar{q}$, $\xi(b) = \hat{q}$,
and	$\xi \in \Xi$.

(3.5)

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 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 (3.5) in the equivalent form

$$\boxed{\begin{array}{rl} \text{Given} & Q,\, L,\, a,\, b,\, \bar{q},\, \hat{q},\, \Xi\,, \\ \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)dt\,, \\ \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\ & \dot{\xi}(t) = \eta(t) \text{ for } a \le t \le b\,, \\ \text{and} & \xi \in \Xi\,. \end{array}} \tag{3.6}$$

The vector variable q is the *state*, or *configuration*, of the problem (3.5) or (3.6). 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 3.7.1 A minimization problem of the form (3.5)—or (3.6)—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

$$\xi(a) \in S_1 \quad \text{and} \qquad \xi(b) \in S_2\,, \tag{3.7}$$

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

$$(\xi(a), \xi(b)) \in S\,, \tag{3.8}$$

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. ♢

3.8 Isoperimetric problems

In addition to the standard problems (3.5) or (3.6), the more general class of "isoperimetric problems" was considered from the early days of the calculus of variations. These are problems of the form

$$\begin{array}{rl} \text{Given} & Q,\, L^0,\, L^1,\, \ldots\,,\, L^m,\, \ell^1,\, \ldots\,,\, \ell^m,\, a,\, b,\, \bar{q},\, \hat{q},\, \Xi\,, \\ \text{minimize} & \int_a^b L^0(\xi(t), \dot{\xi}(t), t)\, dt\,, \\ \text{subject to} & \int_a^b L^1(\xi(t), \dot{\xi}(t), t)\, dt = \ell^1\,, \\ & \qquad \vdots \\ & \int_a^b L^m(\xi(t), \dot{\xi}(t), t)\, dt = \ell^m\,, \\ & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\ \text{and} & \xi \in \Xi\,. \end{array} \tag{3.9}$$

We will show later—in §3.10— 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]

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

3.9 Optimal control problems

The distinctive feature of classical calculus of variations problems is that the minimization of (3.5) or (3.6) takes place in the space of "all curves" (but cf. §3.4), 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 $\mathcal{C}$ 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

$$\boxed{\dot{q} = f(q, u, t)} \tag{3.10}$$

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

$$\dot{\xi}(t) = f(\xi(t), \eta(t), t) \tag{3.11}$$

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

Even more precisely: it may happen that a member ξ of $\mathcal{C}$ does not uniquely determine the control η for which (3.10) holds (cf. Example 3.9.5 below). It is therefore better to take $\mathcal{C}$ as a set of *trajectory-control pairs* $(\xi(\cdot), \eta(\cdot))$.

An ordinary differential equation such as (3.10) defines a *control system*, or *controlled dynamics*, or *controlled dynamical system*[17].

As will be explained below, controlled dynamical systems of the form (3.10) are interesting objects to study even in the absence of a minimization problem. Thus ***control theory is a much richer subject than optimal***

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

control, which deals with only one of many kinds of important control problems.

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

$$\boxed{\begin{array}{ll} \text{Given} & Q,\, U,\, f,\, L,\, a,\, b,\, \bar{q},\, \hat{q},\, \mathcal{C}\,, \\ \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)dt\,, \\ \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\ & (\xi, \eta) \in \mathcal{C}\,, \\ \text{and} & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \leq t \leq b\,. \end{array}} \tag{3.12}$$

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:

$$\boxed{\begin{array}{ll} \text{Given} & Q,\, U,\, f,\, L,\, S\,,\, \mathcal{C}\,, \\ \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)dt\,, \\ \text{subject to} & (\xi(a), a, \xi(b), b) \in S\,, \\ & (\xi, \eta) \in \mathcal{C}\,, \\ \text{and} & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \leq t \leq b\,. \end{array}} \tag{3.13}$$

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

Example 3.9.1 Consider the following controlled dynamics in $\mathbb{R}^3$:

$$\begin{array}{rcll} \dot{x}_1 & = & u_1 \cdot \cos\theta\,, & \\ \dot{x}_2 & = & u_1 \cdot \sin\theta\,, & |u_1| \leq 1\,,\ |u_2| \leq 1 \\ \dot{\theta} & = & u_2\ . & \end{array} \tag{3.14}$$

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 (3.14) 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."

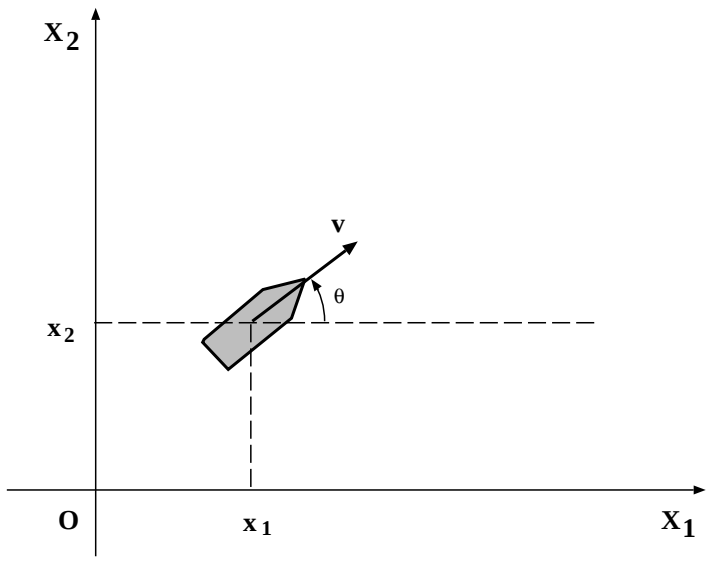


FIGURE 3. The unicycle

The dynamical behavior described by (3.14) 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.

[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 $\mathbb{R}^3$.

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, 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?

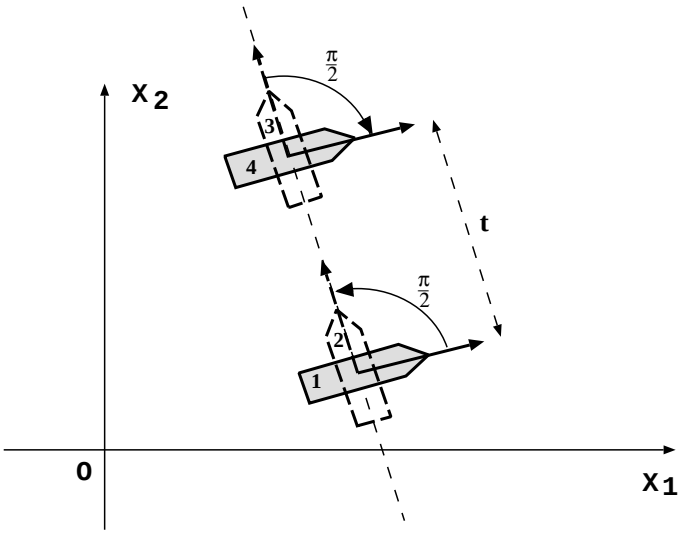


FIGURE 4. Moving sideways

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 $\mathbf{t}$, 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 $\mathbf{t}$ units of length at full speed to configuration 3 (that is, use $u = 1$, $v = 0$) during $\mathbf{t}$ 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 (3.14) 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 3.9.2 Consider the following controlled dynamics in $\mathbb{R}^3$:

$$\begin{aligned} \dot{x}_1 &= \cos\theta\,, \\ \dot{x}_2 &= \sin\theta\,, \qquad |u| \leq 1\,, \\ \dot{\theta} &= u\,. \end{aligned} \tag{3.15}$$

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 3.9.1 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 4.

Example 3.9.3 For the control systems described in Examples 3.9.1 and 3.9.2, 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 (3.13). The dynamical law f is given by the right-hand sides of the equations of (3.14) and (3.15). 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 3.9.4 The minimum time problem for the system (3.15), in which the vehicle can only move forwards, is known as ***Dubins' problem***. It was

solved by L. Dubins in his 1957 paper [45]. The problem for the system (3.14), 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 [100].

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 3.9.5 Consider a control system Σ defined by the dynamical equation

$$\dot{q} = uq\,, \tag{3.16}$$

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. $\diamondsuit$

Example 3.9.6 Consider a one-dimensional optimal control problem whose dynamical behavior is defined by the same equation (3.16) as in Example 3.9.5. Suppose we wish to minimize the integral cost $\int_0^1 \eta(t)^2\,dt$, subject to the endpoint 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. $\diamondsuit$

3.10 Isoperimetric calculus of variations problems are standard optimal control problems

Exactly as we did in §3.8 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 (3.13) and a collection of m additional integral constraints $I^1 = 0, \ldots, I^m = 0$, where

$$I^j = \int_a^b L^j(\xi(t), \eta(t), t)\, dt\,, \tag{3.17}$$

we could consider the corresponding "isoperimetric problem":

Given	Q, U, f, L^0, L^1, $\ldots$, L^m, ℓ^1, $\ldots$, ℓ^m, S, $\mathcal{C}$,
minimize	$I = \int_a^b L^0(\xi(t), \eta(t), t)\, dt$,
subject to	$(\xi(a), a, \xi(b), b) \in S$,
	$(\xi, \eta) \in \mathcal{C}$,
	$\dot{\xi}(t) = f(\xi(t), \eta(t), t)$ for $a \le t \le b$,
and	$\int_a^b L^j(\xi(t), \eta(t), t)\, dt = \ell^j$, for $j = 1, \ldots, m$.

(3.18)

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 (3.18). We can then augment the state space Q of (3.12) by adding a new m-vector $r = (r^1, \ldots, r^m)$ of state variables, subject to the dynamical laws

$$\dot{r}^j = L^j(q, u, t)\,, \quad j = 1, \ldots, m\,. \tag{3.19}$$

If we let

$$Q^* = Q \times \mathbb{R}^m\,, \qquad L^*(q, r, u, t) = L^0(q, u, t)\,, \tag{3.20}$$

$$q^* = \begin{bmatrix} q \\ r^1 \\ \vdots \\ r^m \end{bmatrix}, \quad f^*(q^*, u, t) = \begin{bmatrix} f(q, u, t) \\ L^1(q, u, t) \\ \vdots \\ L^m(q, u, t) \end{bmatrix}, \tag{3.21}$$

and

$$S^* = \left\{ \left(\begin{bmatrix} \bar{q} \\ \bar{r} \end{bmatrix}, a, \begin{bmatrix} \hat{q} \\ \hat{r} \end{bmatrix}, b \right) : (\bar{q}, a, \hat{q}, b) \in S,\ \bar{r} = 0,\ \hat{r} = \begin{bmatrix} \ell^1 \\ \vdots \\ \ell^m \end{bmatrix} \right\}, \tag{3.22}$$

then we see that (3.18) is equivalent to a standard optimal control problem (3.13), 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.

3.11 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 (3.13) 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.

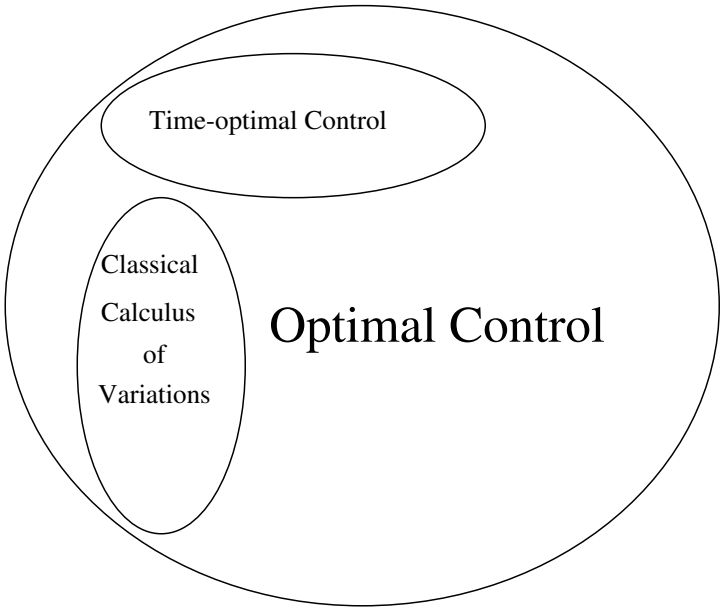


FIGURE 5.

3.12 To a controlled dynamics there is associated an interesting Lie-algebra structure

More recently, it has become apparent that, as R. Hermann (cf. [63]) and others (cf. Brockett [28], Hermes [64], Krener [76], Lobry [86], Jurdjevic-Sussmann [111, 70]) had argued since the 1960s and early 70s, a controlled dynamics such as (3.10) 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.*

3.13 Control is much more general than optimal control

In modern terms, one can think of a controlled dynamical system (3.10) 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 (3.10) 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 (3.10) 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 (3.10) 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***.

3.14 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.

3.15 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 (3.10). 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 3.15.1 Consider the dynamical behavior described by the equation

$$\dot{q} = u\,, \qquad q \in \mathbb{R}^n\,. \tag{3.23}$$

Here the vector u is arbitrary but subject to a bound $||u|| \leq c$, 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

$$\eta(t) = -c\frac{\bar{q}}{||\bar{q}||}\,. \tag{3.24}$$

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

$$u = \omega(q) = -c\frac{q}{||q||}\,. \tag{3.25}$$

There is an enormous difference between the solutions given by (3.24) and (3.25). 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 (3.25) 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 3.15.2 For a mathematically richer example, consider the equation

$$\ddot{x} + ax = u\,, \tag{3.26}$$

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 $x(0) = \bar{q}^1$, $\bar{x}(0) = \bar{q}^2$. 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 3.15.3 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 (3.26) 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

[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.

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

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

that is,

$$\left\{ \begin{aligned} \dot{x} &= w \\ \dot{w} &= h(x, y) - ax\,, \\ \dot{y} &= g(x, y)\,, \end{aligned} \right. \tag{3.27}$$

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 (3.27) becomes a system of first-order differential equations. The precise technical requirement imposed on the system (3.27) is that 0 be a globally asymptotically stable equilibrium.)

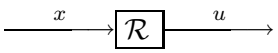
One should think of y as the state variable of a "controller" or "regulator" system that we, the controllers, are running, for example by solving the equation $\dot{y}(t) = g(x(t), y(t))$ on line with a digital computer, or by feeding x as a physical input —e.g., a voltage—into some physical device such 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 (3.26) 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

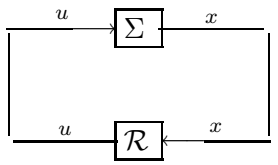
(3.28)

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



, (3.29)

where we will feed x as input, and obtain the desired control u as output. The new system $\mathcal{R}$ 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"



(3.30)

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 [31, 80, 103] for details. ♢

COMMENT NO. 16. 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.

3.16 The brachistochrone problem truly belongs to optimal control

We have already pointed out that "brachistochrone" means "shortest time" (cf. Page 3), 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 13 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 nontrivial 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 §3.15

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

$$\left\{ \begin{array}{lcl} \dot{y}_1 & = & y_2 + \alpha_1(x - y_1)\,, \\ \dot{y}_2 & = & u - ay_1 + \alpha_2(x - y_1)\,, \\ u & = & ay_1 - y_1 - y_2\,. \end{array} \right. \tag{3.31}$$

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

$$u = ax - x - \dot{x}\,, \tag{3.32}$$

which would transform our equation (3.26) into the obviously stable dynamics given by

$$\ddot{x} + \dot{x} + x = 0\,. \tag{3.33}$$

The law (3.32) cannot, however, be implemented, because $\dot{x}$ is not directly available, so we use instead the law

$$u = ay_1 - y_1 - y_2\,, \tag{3.34}$$

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 (3.32), so x will essentially evolve according to (3.33).

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

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

$$\dot{e}_1 = e_2 + \alpha_1 e_1 \,. \tag{3.35}$$

Also,

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

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

$$\ddot{e}_1 + \dot{e}_1 + e_1 = 0 \,, \tag{3.36}$$

so e_1 and $\dot{e}_1$ decay exponentially. With this choice of the α_i it also follows—for example, because of (3.35) and (3.36)—that e_2 and $\dot{e}_2$ decay exponentially. If we now rewrite (3.31) as

$$\begin{cases} \dot{y}_1 &= y_2 + \alpha_1 e_1 \,, \\ \dot{y}_2 &= -y_1 - y_2 + \alpha_2 e_1 \,, \end{cases} \tag{3.37}$$

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. $\diamondsuit$

COMMENT NO. 17. JAN: I have written down in full detail the solution of the adaptive control problem stated as No. 3 in page 61. 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 4

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.

4.1 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.

4.1.1 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 know for at least 3,000 years. Yet, a careful analysis—to be carried out in §5.1—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].

[1] that is, cheat

But, as we shall show in §5.1, 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.

4.1.2 The classical isoperimetric problem.

The oldest curve optimization problem after the one about segments 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 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

(II) *to determine the shortest of all 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[3]. To transform the rotational symmetry argument into a rigorous proof that the solution of the IP is a circle, one would have to establish, for example, uniqueness of the solution modulo translations.)

The Greeks knew that the solution was be the circle, and realized that this needed proof. ***Zenodorus***[4] (*ca.* 200-140 BCE) showed that

[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 [93]).

[3]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.

[4]For more information on Zenodorus's work, cf. Heath [61], pp. 382-383.

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

and

(Z2) a circle has a larger area than any regular polygon of the same perimeter.

These observations essentially[5] 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 §14.1. 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 [109], 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. §4.2 in page 72 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 §3.10.

4.1.3 Light reflection.

Hero (or Heron) of Alexandria[6] 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 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

[5] 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.

[6] Exact dates unknown. For along 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.

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.

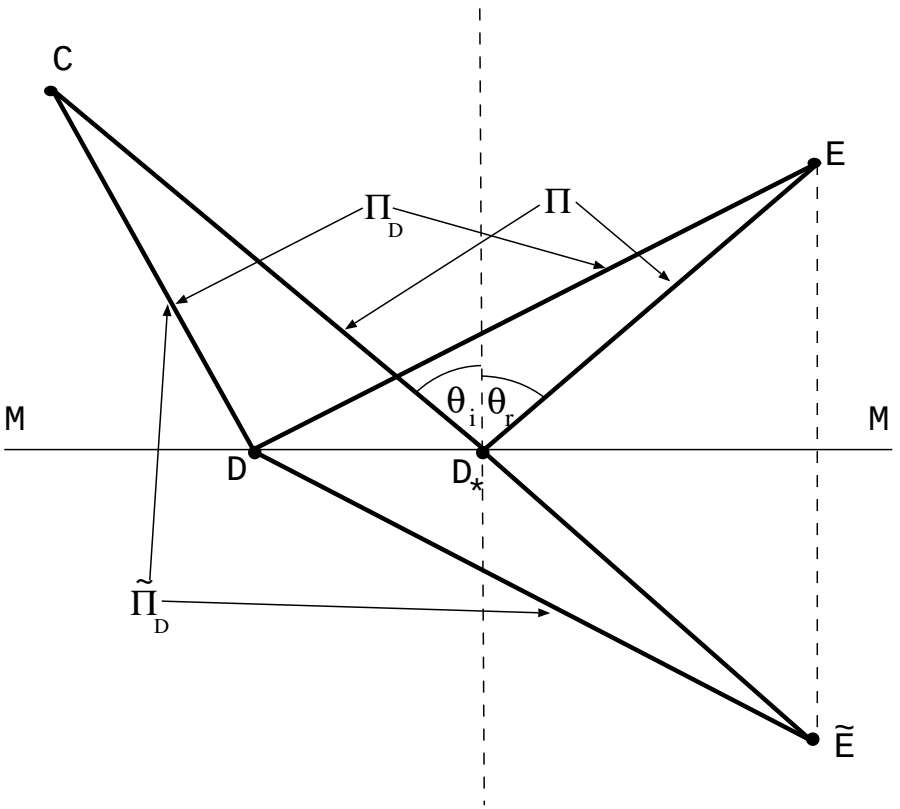


FIGURE 6. Reflection of light

4.1.4 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 105).

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

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

Therefore

$$\frac{\bar{x}}{c_+\sqrt{\bar{x}^2 + y_+^2}} = \frac{x_- - \bar{x}}{c_-\sqrt{(x_- - \bar{x})^2 + y_-^2}} = 0\,. \tag{4.1}$$

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 (4.1) says that

$$\frac{\sin\theta_i}{c_+} = \frac{\sin\theta_i}{c_-}\,. \tag{4.2}$$

4.1.5 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 §5.3 and in Chapter 13.

4.2 Euler

Leonhard Euler (1707-1793) entered the University of Basel in 1720 at the age of 13[7]. 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 [50]) 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).

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

COMMENT NO. 18. 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.

[7]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.

4.3 Lagrange

In H. Goldstine's words ([59], 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 (3.5), known today as the "Euler-Lagrange equation." This condition says[8] that the identity

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial L}{\partial q} \tag{4.3}$$

must be satisfied along the curve ξ_*.

Equation (4.3) 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:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) = \frac{\partial L}{\partial q^i}\,, \quad i = 1, \ldots, n\,. \tag{4.4}$$

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

Alternatively, we can regard Equation (4.3) 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

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

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 (4.3) as

$$\boxed{\frac{d}{dt}\Big[\frac{\partial L}{\partial u}\Big(\xi(t), \dot{\xi}(t), t\Big)\Big] = \frac{\partial L}{\partial q}\Big(\xi(t), \dot{\xi}(t), t\Big)\,,\ i = 1, \ldots, n\,,} \tag{4.5}$$

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 (4.3) 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.

4.4 Legendre

The Euler-Lagrange system (4.3)—or (4.5)—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 8. So in fact (4.3) 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

$$\boxed{\frac{\partial^2 L}{\partial \dot{q}^2}\Big(\xi_*(t), \dot{\xi}_*(t), t\Big) \geq 0 \quad (\text{i.e., } \frac{\partial^2 L}{\partial u^2}\Big(\xi(t), \dot{\xi}(t), t\Big) \geq 0)\,.} \tag{4.6}$$

With an appropriate reinterpretation, Legendre's condition (4.6) is also necessary in the vector case: all we have to do is read (4.6) as asserting that the Hessian matrix $\Big\{\frac{\partial^2 L}{\partial u^i \partial u^j}\Big(\xi_*(t), \dot{\xi}_*(t), t\Big)\Big\}_{1 \leq i,j \leq n}$ has to be nonnegative definite.

4.5 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 (4.3) and the Legendre condition (4.6), 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 (4.3) 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 (4.3). Define the "momentum" $\pi_{\xi,L}$ corresponding to L along a curve ξ by letting[9]

$$\pi_{\xi,L}(t) \stackrel{\text{def}}{=} \frac{\partial L}{\partial u}\Big(\xi(t), \dot{\xi}(t), t\Big)\,. \tag{4.7}$$

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[10]

$$\boxed{H(q, u, p, t) = \langle p, u\rangle - L(q, u, t)\,.} \tag{4.8}$$

Now suppose that a curve $t \mapsto \xi_*(t)$ is a solution of (4.3).

It is then clear that

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

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

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

Therefore, along our curve ξ_*, the equality

$$\frac{d\xi_*}{dt}(t) = \frac{\partial H}{\partial p}\Big(\xi_*(t), \dot{\xi}_*(t), \pi(t), t\Big) \tag{4.9}$$

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

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

It follows that (4.5) says that

$$\frac{dp}{dt}(t) = -\frac{\partial H}{\partial q}\Big(\xi_*(t), \dot{\xi}_*(t), \pi(t), t\Big)\,. \tag{4.10}$$

Finally,

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

so (4.7) says:

$$\frac{\partial H}{\partial u}\Big(\xi_*(t), \dot{\xi}_*(t), \pi(t), t\Big) = 0\,. \tag{4.11}$$

The system of equations (4.9), (4.10), (4.11), usually written more concisely as

$$\boxed{\frac{dq}{dt} = \frac{\partial H}{\partial p}\,, \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial q}\,, \qquad \frac{\partial H}{\partial u} = 0\,,} \tag{4.12}$$

is ***exactly equivalent*** to (4.3), provided that H is defined as in (4.8).

We will call the function H the "control Hamiltonian," and refer to (4.12) as the ***control Hamiltonian form*** of the Euler-Lagrange equations. In our view, Formula (4.8) is the definition that Hamilton should have given for the Hamiltonian, and Equations (4.12) are *"Hamilton's equations as he should have written them."*

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

$$\frac{dq}{dt} = \frac{\partial \mathcal{H}}{\partial p}\,, \quad \frac{dp}{dt} = -\frac{\partial \mathcal{H}}{\partial q}\,, \tag{4.13}$$

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

$$\mathcal{H}(q,p,t) = \langle p, \dot{q}\rangle - L(q,\dot{q},t)\,, \tag{4.14}$$

which resembles (4.8) 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

$$p = \frac{\partial L}{\partial \dot{q}}(q,\dot{q},t)\,. \tag{4.15}$$

It is easy to see that

If the map $(q, \dot{q}, t) \mapsto (q, p, t)$ **defined by (4.15) can be inverted,** *i.e., if we can "solve (4.15) for* $\dot{q}$ *as a function of* q, p, t*," then (4.13) is equivalent to (4.12).*

Indeed, it is clear that

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

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

$$\frac{\partial \mathcal{H}}{\partial q} = \frac{\partial H}{\partial q} + \frac{\partial H}{\partial u} \cdot \frac{\partial u}{\partial q}\,. \tag{4.16}$$

Since

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

we see that

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

along solutions of (4.12), and then the first equation of (4.13) holds as well. Similarly, the second equation of (4.13) also holds. The converse is also easily proved.

4.6 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 (4.13)-(4.14)-(4.15), 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 (4.13)-(4.14)-(4.15) only makes sense when the transformation (4.15) 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 (4.6). Since $H(q,u,p,t)$ is equal to $-L(q,u,t)$ plus a linear function of u, (4.6) is completely equivalent to

$$\frac{\partial^2 H}{\partial u^2}\Big(\xi_*(t),\dot{\xi}_*(t),\pi(t),t\Big)\le 0\,,\quad \text{i.e.,}\quad \frac{\partial^2 H}{\partial \dot{q}^2}\Big(\xi_*(t),\dot{\xi}_*(t),\pi(t),t\Big)\le 0\,. \tag{4.17}$$

Now let us write (4.17) side by side with the third equation of (4.12):

$$\boxed{\frac{\partial H}{\partial u}=0 \quad \text{and} \quad \frac{\partial^2 H}{\partial u^2}\le 0}\,, \tag{4.18}$$

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 has a maximum as a function of u.***

We state this formally as a conjecture.

CONJECTURE M1: *Besides (4.12) (or the equivalent form (4.3)), 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 (4.18) and been led by it to Conjecture M1.

3. On the other hand, it is only by using the Hamiltonian of (4.8), 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 $\mathcal{H}$, which isn't even a function of u. *Only the use of the "control" Hamiltonian leads naturally to Conjecture M1.*

Remark 4.6.1 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 $\mathcal{H}$. Indeed, the Legendre condition is equivalent, in to his case, to the statement that

$$\text{the matrix } \left\{\frac{\partial^2 \mathcal{H}}{\partial p_i \partial p_j}(\xi_*(t), \pi_{\xi_*,L}(t), t)\right\}_{1\le i,j\le n} \text{ is positive definite.} \tag{4.19}$$

(The proof of (4.19) is essentially contained in our discussion of Carathéodory's Hamiltonian formulation of Weierstrass' condition, cf. §4.9. Indeed, we show there that

$$\frac{\partial \mathcal{H}}{\partial p_i}(q,p,t) = \varphi_i(q,p,t)\,. \tag{4.20}$$

Moreover,

$$\sum_{\ell=1}^{n} \frac{\partial \chi_j}{\partial \dot q_\ell}(q, \varphi(q,p,t), t)\frac{\partial \varphi_\ell}{\partial p_i}(q,p,t) = \delta_{j,i}\,, \tag{4.21}$$

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

$$\chi_j(q,\dot q,t) = \frac{\partial L}{\partial \dot q_j}(q,\dot q,t)\,. \tag{4.22}$$

It follows from (4.20), (4.21) and (4.22) that

$$\sum_{\ell=1}^{n} \frac{\partial^2 L}{\partial \dot q_\ell \partial \dot q_j}(q, \varphi(q,p,t), t)\frac{\partial^2 \mathcal{H}}{\partial p_i \partial p_\ell}(q,p,t) = \delta_{j,i}\,. \tag{4.23}$$

So the matrix

$$M = \Big\{ \frac{\partial^2 \mathcal{H}}{\partial p_i \partial p_\ell}(\xi_*(t), \pi_{\xi_*,L}(t), t) \Big\}_{1 \le i,j \le n}$$

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, (4.20) implies that

$$\nabla_p \mathcal{H}(\xi_*(t), \pi_{\xi_*,L}(t), t) = \dot{\xi}_*(t)\,,$$

so $\pi_{\xi_*,L}(t)$ is *not* in general a critical point of the function $p \mapsto \mathcal{H}(\xi_*(t), p, t)$. $\diamondsuit$

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.

4.7 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 [127] by J. L. Synge. For a discussion of what this idea becomes in the optimal control setting, we will have to wait until Chapter 7. For the specific case of the brachistochrone, we will show in 13.12 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
>
> $$V = \int v\Big(x, y, z, \frac{dx}{du}, \frac{dy}{du}, \frac{dz}{du}\Big) du\,. \tag{4.24}$$
>
> 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 7 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 7, 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.

4.8 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 4.8.1, we have a good reason for using s rather than t as the "time" variable in the expression for I.

Remark 4.8.1 In a sense, one can always assume (W) "without loss of generality." Indeed, suppose we are interested in a completely arbitrary problem (3.5). 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 (3.5). 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. ♢

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 (4.13), 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"

$$\mathcal{E}(q,\bar{u},u) = L(q,u) - \frac{\partial L}{\partial u}(q,\bar{u})\cdot u\,, \tag{4.25}$$

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 $\mathcal{E}$ 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

$$L(q,u) = \frac{\partial L}{\partial u}(q,u)\cdot u\,. \tag{4.26}$$

Therefore Weierstrass could equally well have written his excess function as

$$\mathcal{E}(q,\bar{u},u) = L(q,u) - \frac{\partial L}{\partial u}(q,\bar{u})\cdot u - \left(L(q,\bar{u}) - \frac{\partial L}{\partial u}(q,\bar{u})\cdot\bar{u}\right), \tag{4.27}$$

or as

$$\mathcal{E}(q,\bar{u},u,t) = L(q,u,t) - \frac{\partial L}{\partial u}(q,\bar{u},t)\cdot u - \left(L(q,\bar{u},t) - \frac{\partial L}{\partial u}(q,\bar{u},t)\cdot\bar{u}\right), \tag{4.28}$$

if he had allowed his Lagrangians to depend on time.

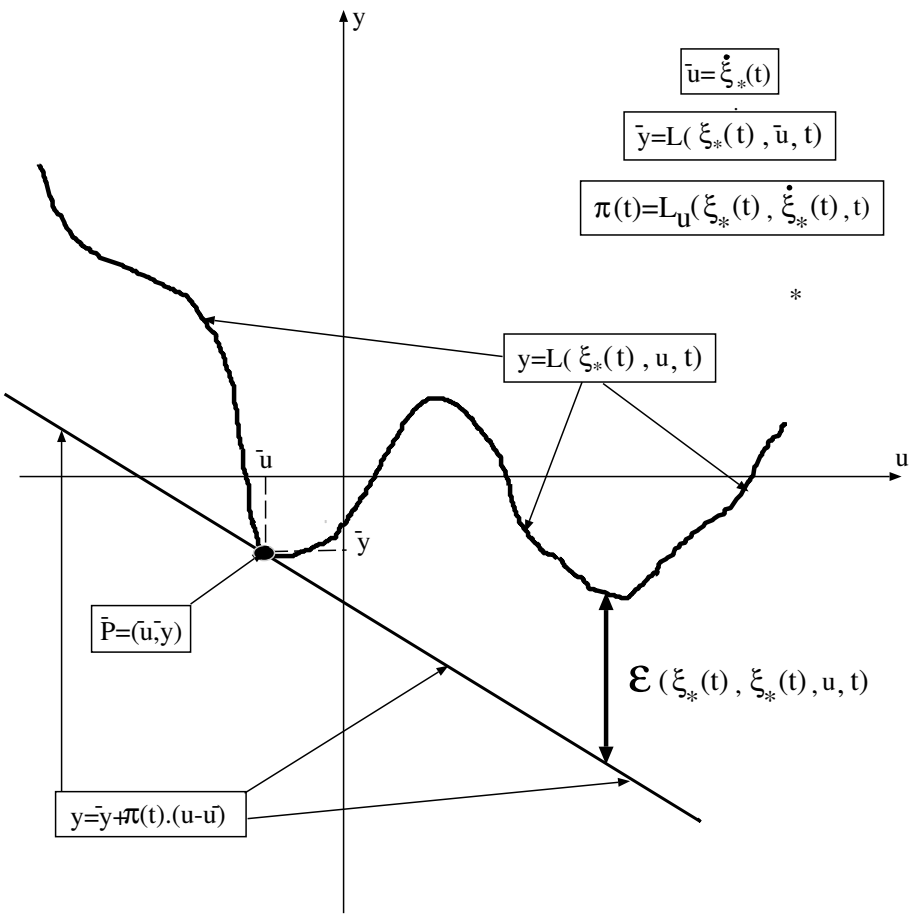


FIGURE 7. The Weierstrass side condition

Remark 4.8.2 Formula (4.27) 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

$$\mathcal{E}(q, \bar{u}, u) = y_{graph}(u) - y_{plane}(u) \,.$$

In other words (cf. Figure 7):

> ***the number*** $\mathcal{E}(q, \bar{u}, u)$ ***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.* $\diamondsuit$

Using

$$\pi(t) = \frac{\partial L}{\partial u}(\xi_*(t), \dot{\xi}_*(t), t) \,,$$

as in (4.7), we see that

$$\mathcal{E}(\xi_*(t), \dot{\xi}_*(t), u, t) = \Big(L(\xi_*(t), u, t) - \langle \pi(t), u \rangle\Big) - \Big(L(\xi_*(t), \dot{\xi}_*(t)) - \langle \pi(t), \dot{\xi}_*(t) \rangle\Big) \,, \tag{4.29}$$

which the reader will immediately recognize as saying that

$$\mathcal{E}(\xi_*(t), \dot{\xi}_*(t), u, t) = H(\xi_*(t), \dot{\xi}_*(t), \pi(t), t) - H(\xi_*(t), u, \pi(t), t)\,, \quad (4.30)$$

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. 19. 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 (4.30)). 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 4.8.1, 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 $\mathcal{E}$, then one also ends up with a statement valid for all problems, but the expression for $\mathcal{E}$ is now much more complicated, since one has to use (4.27) and cannot substitute for it the simpler expression (4.25).

Moreover, (MAX) can be considerably simplified. Indeed,

the requirement that $p(t)$ be defined via (4.7) 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 (4.7). Moreover, the vanishing of $\frac{\partial H}{\partial u}(q_*(t), \dot{q}_*(t), p(t), t)$ is also one of the conditions of (4.12). So we can just drop (4.7) altogether, and state (4.12) and (MAX) together:

(NCO) *If a curve $t \mapsto q_*(t)$ is a solution of the minimization problem (3.5), 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:*

$$\dot{q}_*(t) = \frac{\partial H}{\partial p}(q_*(t), \dot{q}_*(t), p(t), t) \tag{4.31}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial q}(q_*(t), \dot{q}_*(t), p(t), t) \tag{4.32}$$

$$H(q_*(t), \dot{q}_*(t), p(t), t) = \max_u H(q_*(t), u, p(t), t) \tag{4.33}$$

Remark 4.8.3 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 (4.32) 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. 20. 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.

4.9 Carathéodory gets even closer, but misses too

C. Carathéodory (1873-1950) published his work on the calculus of variations in a 1935 book, cf. [32]. His discussion of the excess function, and of how Weierstrass' condition can be expressed in Hamiltonian form, is worth quoting in detail[11], 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

$$\begin{aligned}
\mathcal{E}(t, x, \dot{x}, x') &= L(t, x, x') - L(t, x, \dot{x}) - \frac{\partial L}{\partial \dot{x}}(t, x, \dot{x}) \cdot (x' - \dot{x}) \\
&= L(t, x, x') - L(t, x, \dot{x}) - y \cdot (x' - \dot{x}) \\
&= y \cdot \dot{x} - L(t, x, \dot{x}) - \Big(y \cdot x' - L(t, x, x')\Big) \\
&= y \cdot \dot{x} - L(t, x, \dot{x}) - \Big(y' \cdot x' - L(t, x, x')\Big) - (y - y') \cdot x' \\
&= H - H' - (y - y') \cdot x' \,.
\end{aligned}$$

Since

$$\frac{\partial H'}{\partial y'_j} = \frac{\partial}{\partial y'_j}\Big(\sum_{i=1}^n y'_i x'_i - L(t, x, x')\Big)$$

[11] We follow [32], pages 210-212.

$$
\begin{aligned}
&= x'_j + \sum_{i=1}^{n} y'_i \frac{\partial \varphi_i}{\partial y'_j}(t,x,y') - \frac{\partial}{\partial y'_j}\Big(L(t,x,\varphi(t,x,y'))\Big) \\
&= x'_j + \sum_{i=1}^{n} y'_i \frac{\partial \varphi_i}{\partial y'_j}(t,x,y') - \sum_{i=1}^{n} \frac{\partial L}{\partial x'_i}(t,x,x') \frac{\partial \varphi_i}{\partial y'_j}(t,x,x') \\
&= x'_j + \sum_{i=1}^{n} y'_i \frac{\partial \varphi_i}{\partial y'_j}(t,x,y') - \sum_{i=1}^{n} y'_i \frac{\varphi_i}{\partial y'_j}(t,x,x') \\
&= x'_j \,,
\end{aligned}
$$

Carathéodory concluded that "the equation

$$
\begin{aligned}
\mathcal{E}(t,x,\dot{x},x') &= L' - L - L_{\dot{x}_j}(x'_j - \dot{x}_j) \\
&= H - H' - H'_{y'_j}(y_j - y'_j)
\end{aligned}
$$

therefore holds, by which the $\mathcal{E}$ function is represented in the desired form."

4.10 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.*[12] ***The answer will be given below.*** ♢

[12] *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. [42], 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 ([96]), 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. In particular, you should avoid following that book's unjustifiably critical opinion about Pontryagin. (In p. 47 of Adair's translation, [96] 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.)

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 (4.33) does this automatically. So the addition of the new Weierstrass condition to the three equations of (4.12) 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 (4.18), and almost certainly by Weierstrass, if only Hamilton's equations had been written in the form (4.8), (4.12).

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 (4.18), 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 (4.13) of Hamilton's equations—which involves no functions of u and therefore no u-derivatives—the fact remains that in order to get to (4.13) we first have to solve (4.15), 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 (4.33) 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 (4.33) no longer implies (4.15). So we see that ***(4.33) is more fundamental than (4.15)***.

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 (4.15), 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 (4.15). 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 (4.15) follows from (4.33). Otherwise, p is a "new," "independent" object.

Since Weierstrass was working under assumptions that imply (4.15), 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 $\frac{\partial L}{\partial u}(q_*(t), \dot{q}_*(t), t)$.

4.11 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 4.11.1 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

$$x(0) = 1\,, \quad \text{and} \quad x(L) = 1\,, \tag{4.34}$$

and minimizes the integral $\int_0^L |\dot{x}(t)|\, dt$ among all the functions on $[0, L]$ satisfying the constraints (4.34). 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 (4.31) and (4.32) 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 4.11.2 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

$$\xi(0) = (0, 0)\,, \quad \text{and} \quad \xi(1) = (1, 2)\,, \tag{4.35}$$

and minimizes the integral

$$J(\xi) = \int_0^1 \Big(|\dot{\xi}_1(t)| + |\dot{\xi}_2(t)| \Big)\, dt \tag{4.36}$$

in the set Ξ of all the Lipschitz functions $[0,1] \ni t \mapsto \xi(t) \in \mathbb{R}^2$ satisfying the constraints (4.35). 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

$$1 = \xi_1(1) - \xi_1(0) = \int_0^1 \dot{\xi}_1(t)\, dt \leq \int_0^1 |\dot{\xi}_1(t)|\, dt \,, \tag{4.37}$$

and a similar argument shows that

$$2 \leq \int_0^1 |\dot{\xi}_2(t)|\, dt \,. \tag{4.38}$$

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

$$H(x_1, x_2, u_1, u_2, p_1, p_2, t) = p_1 u_1 + p_2 u_2 - |u_1| - |u_2| \,,$$

(so in particular H does not depend on x and t). Then (4.31) and (4.32) 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

$$\xi_1(0) = 0\,, \quad \xi_1(1) = 1\,, \quad \text{and} \quad \xi_1 \text{ is monotonically nondecreasing}\,. \tag{4.39}$$

Similarly, ξ_2 is an arbitrary Lipschitz function such that

$$\xi_2(0) = 0\,, \quad \xi_2(1) = 2\,, \quad \text{and} \quad \xi_2 \text{ is monotonically nondecreasing}\,. \tag{4.40}$$

Conversely, *every Lipschitz curve* $[0,1] \ni t \mapsto \xi(t) = (\xi_1(t), \xi_2(t)) \in \mathbb{R}^2$ *such that (4.39) and (4.40) hold is a minimizer.* (This is easy to prove. It suffices to observe that under those conditions the inequalities occurring in (4.37) and (4.38) 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 4.11.3 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

$$x(0) = 1\,, \quad x(L) = 1\,, \text{ and } |\dot{x}(t)| \leq 1 \text{ for all } t \in [0, L]\,, \tag{4.41}$$

and minimizes the integral $\int_0^L x(t)^2 dt$ among all the functions satisfying the constraints (4.41). 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 (4.31) and (4.32) 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, (4.33) 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 (4.33) does not directly determine the value of u when $p = 0$, since in this case every value of u satisfies (4.33), 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. $\diamondsuit$

Examples 4.11.1 and 4.11.3 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.

4.12 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 4.12.1 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 (4.31) and (4.32) 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 4.12.2 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

$$U = \{(u_1,u_2) \in \mathbb{R}^2 : u_1^2 + u_2^2 \leq 1\}\,,\ \ L(q_1,q_1,u_1,u_2,t) = q_1 q_2\,.$$

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

$$\dot{\pi}_2(t) = \xi_1(t)\,. \tag{4.42}$$

On the other hand, the Hamiltonian maximization condition says that for almost all t the function

$$U \ni (u_1,u_2) \mapsto \pi_1(t)u_1 + \pi_2(t)u_2 - \xi_1(t)\xi_2(t) \tag{4.43}$$

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 (4.43) 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

$$\pi_2(t) \equiv 0\,. \tag{4.44}$$

Clearly, (4.44) contradicts (4.42). Therefore π does not exist, and (NCO) is *not* satisfied by the solution. ♢

Once again, as in Example 4.13.1, we seem to be in trouble. But, as opposed to what happened in Example 4.13.1, this time there is no mistake. The problem of Example 4.12.1 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,"[13] and write the Hamiltonian as

$$H(q, u, p, p_0, t) = \langle p, f(q, u, t)\rangle - p_0 L(q, u, t)\,. \tag{4.45}$$

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.

4.13 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

[13] The need for the abnormal multiplier had already been noticed by Bolza in 1913, cf. [15].

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.* ♢

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 (4.8)?

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*

$$H(q, u, p, t) = \langle p, f(q, u, t)\rangle - L(q, u, t)\,. \tag{4.46}$$

As before, we can test this with a simple toy problem where the answer can easily be found directly.

Example 4.13.1 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 4.11.3.) 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 $(4 + t_0 + h + t_0 h + \frac{t_0^2 - h^2}{2}, 1 + t_0 - h)$ 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 4.8.1. 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 4.8.1 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[14] 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 (4.8) and (4.12).*

Remark 4.13.2 Notice that what we did in Example 4.13.1—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 (4.15) for $\dot{q}$ as a function of q, p, and t. This 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[15] in 1621, but he did not publish it at the time. The law only become and only in 1703 did

[14] We say "essentially" for a good reason. We will get to it soon.

[15] 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

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[16] and Desargues intervened in the dispute, Descartes eventually admitted that Fermat was right, and wrote

> . . . ... 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 [12]. We quote from Struik [109], p. 393':

---

decimal places. He studied navigation, and described the curve on the sphere —called the ***loxodrome***—that intersects the meridians at equal angles.

[16]The father of Blaise Pascal

> 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 [83], 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 [84],

> 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 [109], p. 279.]

Subsequently, in 1684, Leibniz published a much longer paper [84], 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 [109], 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 12.

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.

([96], 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 $\square$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.
>
> ([96], pages 46-47)

Chapter 5

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*

$$\boxed{\begin{array}{ll} \text{Given} & Q,\, U,\, f,\, L,\, a,\, b,\, \bar{q},\, \hat{q},\, \mathcal{C}\,, \\ \text{minimize} & I = \int_a^b L(\xi(t), \eta(t), t)dt\,, \\ \text{subject to} & \xi(a) = \bar{q}\,,\ \xi(b) = \hat{q}\,, \\ & (\xi, \eta) \in \mathcal{C}\,, \\ \text{and} & \dot{\xi}(t) = f(\xi(t), \eta(t), t) \text{ for } a \leq t \leq b\,, \end{array}} \tag{5.1}$$

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 $(\xi_*, \eta_*) \in \mathcal{C}$ 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 (3.10), a necessary condition for (ξ_*, η_*) to

[1] It is because Conjecture M3 is false unless this adjustment is made that we used the word "essentially" in Footnote 14.

solve the minimization 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*

$$\begin{aligned}
&(NT) \quad (\pi(t), \pi_0) \neq (0,0) \text{ for all } t \in [a,b];\\
&(HS) \quad \dot{\xi}_*(t) = \tfrac{\partial H}{\partial p}(\Xi_*(t)) \text{ and } \quad -\dot{\pi}(t) = \tfrac{\partial H}{\partial q}(\Xi_*(t)) \text{ for } t \in [a,b];\\
&(MC) \quad H(\Xi_*(t)) = \max_{u \in U} H(\xi_*(t), u, \pi(t), \pi_0, t) \text{ for } t \in [a,b],
\end{aligned}$$

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*

$$H(q,u,p,p_0,t) = \langle p, f(q,u,t)\rangle - p_0 L(q,u,t)\,. \tag{5.2}$$

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 (4.31) and (4.32), with the new H, and (MC) is a restatement of (4.33). 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 (3.10)—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.

[2]For differential geometers: p is really a ***field of covectors*** along q_*.

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.

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 4.11.1, 4.11.3 and 4.13.1 one can always either exclude the possibility that $p_0 = 0$ or show that it leads to nothing new. Indeed, in Examples 4.11.1 and 4.11.3 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 4.13.1 (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 4.13.1, 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 4.13.1, 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 4.12.1, 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 4.11.1, 4.11.3, 4.13.1 and 4.12.1.

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 4.13.1. One has to transform the given problem to a fixed-time problem by means of a time reparametrization. What was done in Example 4.13.1 works in general, so we will not repeat the argument here.

Statement (MP') applies in particular to *minimum-time problems*, i.e., problems where $L \equiv 1$.

5.1 Example I: the straight-line segment

COMMENT NO. 21. 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.

5.2 Example II: the isoperimetric problem

.

As explained in §4.1.2, 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 $t_1, t_2 \in [a, b]$, $t_1 \leq t_2$ 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 $\{\xi(t) : a \leq t \leq b\}$, 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 $\mathrm{Ins}(\xi)$ and $\mathrm{Out}(\xi)$ precisely. Indeed, if r is large enough (such that $\|\xi(t)\| \leq r$ 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 $\mathrm{Ins}(\xi)$ and $\mathrm{Out}(\xi)$ to be, respectively, the bounded and unbounded components $\mathbb{R}^2 \backslash |\xi|$.

Now that we know what "$\mathrm{Ins}(\xi)$" means, we can use use the well-known "Green's formula"

$$\mathrm{Area}\Big(\,\mathrm{Ins}(\xi)\,\Big) = |A(\xi)|\,, \qquad \text{where} \qquad A(\xi) = \frac{1}{2}\int_\xi (x\,dy - y\,dx) \tag{5.3}$$

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, (5.3) 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 (5.3).

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,

[3]Actually, even assigning a precise meaning to (5.3) 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 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 (5.3) 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:

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 §3.8. As explained in §3.10, 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" $\dot{x}^2 + \dot{y}^2 = 1$.

5.2.1 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 5.2.1 *If (IP.V.a) has a solution γ_* such that $\gamma_* \in \Gamma_L$, then γ_* is a solution of (IP.V).* ♢

5.2.2 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

$$\dot{m}_x(t) = -\frac{\partial H}{\partial x}(\gamma_*(t), \mu(t), \dot{\gamma}(t)), \tag{5.4}$$

$$\dot{m}_y(t) = -\frac{\partial H}{\partial y}(\gamma_*(t), \mu(t), \dot{\gamma}(t)), \tag{5.5}$$

and the Hamiltonian maximization condition

$$H(\gamma_*(t), \mu(t), \dot{\gamma}_*(t)) = \max\left\{H(\gamma_*(t), \mu(t), w) : w \in \mathbb{R}^2, \|w\| \leq 1\right\} \tag{5.6}$$

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

$$\mathbb{R}^6 \ni (x, y, p_x, p_y, u, v) \mapsto H(x, y, p_x, p_y, u, v) \stackrel{\text{def}}{=} p_x u + p_y v + xv - yu\,. \quad (5.7)$$

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

$$\dot{m}_x(t) = -\dot{y}(t)\,, \quad (5.8)$$
$$\dot{m}_y(t) = \dot{x}(t)\,. \quad (5.9)$$

Define

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

Then

$$\dot{\alpha}(t) = -2\dot{y}(t)\,, \quad (5.10)$$
$$\dot{\beta}(t) = 2\dot{x}(t)\,, \quad (5.11)$$

and

$$H(\gamma_*(t), \mu(t), u, v) = \alpha(t)u + \beta(t)v\ . \quad (5.12)$$

Therefore, if we define

$$\nu(t) = \sqrt{\alpha(t)^2 + \beta(t)^2}\,, \quad (5.13)$$

(A.8) says that

$$\left.\begin{array}{l} \dot{x}(t) = \frac{\alpha(t)}{\nu(t)}\,, \\ \\ \dot{y}(t) = \frac{\beta(t)}{\nu(t)}\,, \end{array}\right\} \quad \text{if} \quad \nu(t) \neq 0\,. \quad (5.14)$$

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

$$\nu(t)\dot{x}(t) = \alpha(t) \qquad \text{and} \qquad \nu(t)\dot{y}(t) = \beta(t)\,. \quad (5.15)$$

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 (A.12) and (A.13) imply that $\dot{x}(t) \equiv \dot{y}(t) \equiv 0$, so γ_* is a constant curve and $A(\gamma_*) = 0$.

It is clear that there do exist curves $\gamma \in \tilde{\Gamma}_L$ for which $A(\gamma) > 0$. Therefore the fact that $A(\gamma_*) = 0$ contradicts the assumption that γ_* is a solution of (IP"). So we can exclude the possibility that $\bar{\nu} = 0$.

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

$$\|\dot{\gamma}(t)\| = 1 \qquad \text{for all } t \in [0, L]\,. \tag{5.16}$$

Write

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

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

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

Therefore

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

This implies that

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

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

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

we find that

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

Then

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

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

$$\xi(t) = -\sin\Big(\frac{2(t - t_0)}{\bar{\nu}}\Big) \qquad \text{and} \qquad \eta(t) = \cos\Big(\frac{2(t - t_0)}{\bar{\nu}}\Big)\,. \tag{5.17}$$

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

$$\xi(t) = -\sin\Big(\frac{2\pi k(t - t_0)}{L}\Big) \qquad \text{and} \qquad \eta(t) = \cos\Big(\frac{2\pi k(t - t_0)}{L}\Big)\,. \tag{5.18}$$

We then get, by integration,

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

where

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

The above formulas define a family $\Big\{\gamma^{t_0,k}\Big\}_{t_0\in\mathbb{R},k\in\mathbb{Z}}$ of curves belonging to $\tilde{\Gamma}_L$, depending on the real parameter t_0 and the nonzero integer parameter k. If we let $\bar{\gamma}^{t_0,k}$ be the point with coordinates $\frac{-L}{2\pi k}\cos\Big(\frac{2\pi k t_0}{L}\Big)$, $\frac{L}{2\pi k}\sin\Big(\frac{2\pi k t_0}{L}\Big)$, then the curves $\gamma^{t_0,k}$ satisfy

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

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 5.2.2 *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* $\frac{L^2}{4\pi}$. ♢

5.2.3 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 5.2.3 $\tilde{\Gamma}_L$ *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 5.2.4 *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

$$\lim_{j\to\infty} \Big(A(\gamma_j) - A(\gamma_\infty) \Big) = 0\,. \tag{5.19}$$

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 $|\tilde{I}_j| \leq \varepsilon$ whenever $j \geq J$. Let h be a function of class C^1 on $[0, L]$ such that

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

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

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

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

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

It then follows that

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

completing our proof. ♢

It then follows from Propositions 3 and 4 that

Proposition 5.2.5 *The maximization problem (IP") has a solution.* ♢

5.2.4 The final result

By combining Propositions 14.1.2 and 14.2.3, we get

Theorem 5.2.6 *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 $\frac{L^2}{4\pi}$.* ♢

5.2.5 How to make the solution of the isoperimetric problem rigorous

COMMENT NO. 22. 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. 23. 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

$$e^{2\pi i \tilde{\psi}(t)} = \psi(e^{2\pi i t}) \tag{5.20}$$

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 (A.37) holds for all $t \in \mathbb{R}$. Moreover,

$$e^{g(t)} = e^{f(\tilde{\psi}(t))} = \eta(e^{2\pi i \tilde{\psi}(t)}) = \eta\Big(\psi(e^{2\pi i t})\Big) = (\eta \circ \psi)(e^{2\pi i t}) = \xi(e^{2\pi i t}) \,.$$

Therefore $\nu(\xi,p)=\frac{g(1)-g(0)}{2\pi i}=\frac{f(\tilde{\psi}(1))-f(\tilde{\psi}(0))}{2\pi i}=\frac{f(a+1)-f(a)}{2\pi i}=\nu(\eta,p).$)

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 $[t_j,t_{j+1}]\ni t\mapsto\xi(e^{2\pi it})$ 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 $A_j=\xi(e^{2\pi it_j})$ 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.

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 $\mathbb{R}^2 \backslash |\xi|$. We refer to $\mathrm{Ins}(\xi)$ as *the region enclosed by the simple loop* ξ. We use $A(\xi)$ to denote the area of $\mathrm{Ins}(\xi)$.

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 5.2.7 *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. The winding number $\nu(\xi, p)$ clearly depends continuously on p as long as p varies on $\mathbb{C} \backslash |\xi|$. (Let $p \in \mathbb{C} \backslash |\xi|$, and let $\{p_j\}_{j=1}^{\infty}$ be a sequence in $\mathbb{C} \backslash |\xi|$ converging to p. Let $f : \mathbb{R} \mapsto \mathbb{C}$ be a continuous map with the property that $e^{f(t)} = \xi(e^{2\pi i t}) - p \stackrel{\mathrm{def}}{=} h(t)$ for all $t \in \mathbb{R}$. Let $h_j(t) = \xi(e^{2\pi i t}) - p_j$. Then $h_j \to h$ as $j \to \infty$, uniformly on compact sets, so $h_j \to h$ uniformly, since the h_j are 1-periodic. Since $h(t) \neq 0$ for all t, the functions $g_j \stackrel{\mathrm{def}}{=} \frac{h_j}{h}$ 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

$$e^{f_j(t)} = e^{f(t)}.e^{\lambda(g_j(t))} = h(t).g_j(t) = h_j(t) = \xi(e^{2\pi i t}) - p_j\,.$$

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 $\mathbb{C}\backslash|\xi| \ni p \mapsto \nu(\xi, p) \in \mathbb{Z}$ 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 $q_j = \xi(e^{2\pi i t_j})$, $t_j \in [\bar{t} - \frac{1}{2}, \bar{t} + \frac{1}{2}]$. Also, let $q_j = r_j e^{2\pi i\varphi(\tau_j)}$, 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 $0 < \delta < \bar{r} = \operatorname{dist}(0, |\xi|)$. 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.

$$\Omega_j = \mathrm{Ins}(\omega_j)\,. \tag{5.21}$$

To prove (5.21, 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 $z \in \mathbb{S}^1$.

5.3 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

$$x_c \;=\; \frac{1}{\Lambda}\int_a^b x\, ds = \frac{1}{\Lambda}\int_a^b x\sqrt{1+y'(x)^2}\, dx\,,$$

$$y_c \;=\; \frac{1}{\Lambda}\int_a^b y\,ds = \frac{1}{\Lambda}\int_a^b y(x)\sqrt{1+y'(x)^2}\,dx\,,$$

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 $\mathcal{Y}$ 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 $\mathcal{Y}$ is nonempty if and only if

$$\Lambda \geq \Delta = \sqrt{(b-a)^2+(\beta-\alpha)^2}\,, \tag{5.22}$$

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 $y(\cdot) \in \mathcal{Y}$ 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. 24. 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 $y_* \in \mathcal{Y}$ 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. 25. 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

$$\pi_y'(x) = \frac{\partial L}{\partial y}(\xi_*(x), \xi_*'(x)) = \pi_0 v_*(x) = \pi_0 \sqrt{1 + u_*(x)^2}\,, \tag{5.23}$$

$$\pi_s'(x) = \frac{\partial L}{\partial s}(\xi_*(x), \xi_*'(x)) = 0 \tag{5.24}$$

for a.e. $x \in [a, b]$, $(\pi_y(x), \pi_s(x), \pi_0) \neq (0, 0, 0)$ for all x, and

$$\begin{aligned}
&\pi_y(x)u_*(x) + \pi_s(x)v_*(x) - \pi_0 y_*(x)v_*(x) \\
&\quad = \pi(x) \cdot \xi_*'(x) - \pi_0 L(\xi_*(x), \xi_*'(x)) \\
&\quad = \max\Big\{\, \pi_y(x)u + \pi_s(x)v - \pi_0 L(\xi_*(x), w) : w \in U \,\Big\} \\
&\quad = \max\Big\{\, \pi_y(x)u + \pi_s(x)v - \pi_0 y_*(x)v : (u, v) \in U \,\Big\}
\end{aligned}$$

for almost all $x \in [a, b]$. It then follows that

$$\begin{aligned}
&\pi_y(x)u_*(x) + (\pi_s(x) - \pi_0 y_*(x))\sqrt{1 + u_*(x)^2} \\
&\qquad = \max\Big\{\, \pi_y(x)u + (\pi_s(x) - \pi_0 y_*(x))\sqrt{1 + u^2} : u \in \mathbb{R} \,\Big\}
\end{aligned} \tag{5.25}$$

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 (5.24) 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 (5.23) and (5.24) 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 (5.25) 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 (5.25) is $\bar{\pi}_y + \frac{\bar{\pi}_s u}{\sqrt{1+u^2}}$, which can only vanish if

$$\frac{u}{\sqrt{1+u^2}} = -\frac{\bar{\pi}_y}{\bar{\pi}_s}\,. \tag{5.26}$$

The function $\mathbb{R} \ni u \mapsto \frac{u}{\sqrt{1+u^2}}$ is strictly increasing, so the solution u of (5.26) 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 (5.23) 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 (5.25) says that $u = u_*(x)$ maximizes the function

$$u \to (\bar{\pi}_y + s_*(x))u - Y(x)\sqrt{1+u^2}\,, \tag{5.27}$$

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

$$Y(x)^2 = S(x)^2 + c\,, \tag{5.28}$$

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, (5.28) 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 (5.28) 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 (5.27) 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 (5.27) 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

$$\kappa^2\tau'(x)^2\cosh^2(\tau(x)) = 1 + \kappa^2\tau'(x)^2\sinh^2(\tau(x))\,. \tag{5.29}$$

Since $\cosh^2 r = 1 + \sinh^2 r$ for all r, (5.29) 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

$$y_*(x) = \gamma + \kappa \cosh\left(\frac{x - x_0}{\kappa}\right), \tag{5.30}$$

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 (5.30) to conclude that

$$\begin{aligned}
\alpha &= y_*(a) = \gamma + \kappa \cosh \mathcal{A}\,, \\
\beta &= y_*(b) = \gamma + \kappa \cosh \mathcal{B}\,, \\
\frac{\beta - \alpha}{\kappa} &= \cosh \mathcal{B} - \cosh \mathcal{A} \\
&= 2 \sinh \frac{\mathcal{B} + \mathcal{A}}{2} \sinh \frac{\mathcal{B} - \mathcal{A}}{2} \\
&= 2 \sinh \mathcal{C} \sinh \mathcal{D}\,,
\end{aligned}$$

where

$$\mathcal{A} = \frac{a - x_0}{\kappa}\,, \qquad \mathcal{B} = \frac{b - x_0}{\kappa}\,, \qquad \mathcal{C} = \frac{a + b - 2x_0}{2\kappa}\,, \qquad \mathcal{D} = \frac{b - a}{2\kappa}\,.$$

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

$$\Lambda = \int_a^b \sqrt{1 + y_*'(x)^2}\, dx = \kappa(\sinh \mathcal{B} - \sinh \dashv)\,,$$

so

$$\begin{aligned}
\frac{\Lambda}{\kappa} &= \sinh \mathcal{B} - \sinh \mathcal{A} \\
&= 2 \sinh \frac{\mathcal{B} - \mathcal{A}}{2} \cosh \frac{\mathcal{B} + \mathcal{A}}{2} \\
&= 2 \sinh \mathcal{D} \cosh \mathcal{C}\,.
\end{aligned}$$

Then

$$\frac{\Lambda^2}{\kappa^2} = 4 \sinh^2 \mathcal{D} \cosh^2 \mathcal{C}\,,$$

and

$$\frac{(\beta - \alpha)^2}{\kappa^2} = 4 \sinh^2 \mathcal{C} \sinh^2 \mathcal{D}\,,$$

so

$$\frac{\Lambda^2 - (\beta - \alpha)^2}{\kappa^2} = 4 \sinh^2 \mathcal{D}\,,$$

and then

$$2\kappa \sinh \mathcal{D} = \sqrt{\Lambda^2 - (\beta - \alpha)^2}\,,$$

so

$$\frac{\sinh \mathcal{D}}{\mathcal{D}} = \frac{\sqrt{\Lambda^2 - (\beta - \alpha)^2}}{b - a}\,. \tag{5.31}$$

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 (5.31) is > 1, so there exists a unique positive solution $\mathcal{D}$ of (5.31). Then κ is determined by

$$\kappa = \frac{b - a}{2\mathcal{D}}\,. \tag{5.32}$$

In addition,

$$\tanh \mathcal{C} = \frac{\beta - \alpha}{\Lambda}\,, \tag{5.33}$$

and the assumption that $\Lambda > \Delta$ implies that the right-hand side of (5.33) belongs to the interval $]-1, 1[$, so there exists a unique real solution $\mathcal{C}$ of (5.33). Then x_0 is determined by

$$x_0 = \frac{a + b}{2} - \kappa \mathcal{C} = \frac{a + b}{2} - \frac{(b - a)\mathcal{C}}{2\mathcal{D}}\,. \tag{5.34}$$

Finally, we determine γ by

$$\gamma = \alpha - \kappa \cosh \mathcal{A} = \alpha - \frac{b - a}{2\mathcal{D}} \cosh(\mathcal{C} - \mathcal{D})\,. \tag{5.35}$$

This leads to the formula

$$y_*(x) = \alpha + \frac{b - a}{2\mathcal{D}} \left(\cosh\left(\mathcal{C} + \frac{\mathcal{D}(2x - a - b)}{b - a} \right) - \cosh(\mathcal{C} - \mathcal{D} \right), \tag{5.36}$$

where $\mathcal{C}$ and $\mathcal{D}$ are determined by (5.31) and (5.33).

All this was done under the assumption that $\Lambda > \Delta$. Let us now examine the limit of the right-hand side of (5.36) as $\Lambda \downarrow \Delta$. For this purpose, we rewrite (5.36) as

$$\begin{aligned} y_*(x) &= \alpha + \frac{b - a}{\mathcal{D}} \left(\sinh\left(\mathcal{C} + \frac{\mathcal{D}(x - b)}{b - a} \right) \sinh\left(\frac{\mathcal{D}(x - a)}{b - a} \right) \right) \\ &= \alpha + (x - a) \frac{\sinh\left(\frac{\mathcal{D}(x - a)}{b - a} \right)}{\frac{\mathcal{D}(x - a)}{b - a}} \sinh\left(\mathcal{C} + \frac{\mathcal{D}(x - b)}{b - a} \right), \end{aligned}$$

and observe that $\mathcal{D} \to 0$ and $\mathcal{C} \to \mathcal{C}_0$ as $\Lambda \downarrow \Delta$, where $\mathcal{C}_0$ is characterized by

$$\tanh \mathcal{C}_0 = \frac{\beta - \alpha}{\Delta} \,. \tag{5.37}$$

It follows that, as $\Lambda \downarrow \Delta$, $y_*(x)$ converges to

$$\bar{y}_*(x) = \alpha + (\sinh \mathcal{C}_0)(x - a) \,. \tag{5.38}$$

Equation (5.37) easily implies $\sinh \mathcal{C}_0 = \frac{\beta - \alpha}{b - a}$, so the final expression for $\bar{y}_*(x)$ is

$$\bar{y}_*(x) = \alpha + \frac{\beta - \alpha}{b - a} \cdot (x - a) \,, \tag{5.39}$$

which is, as expected, the unique member of $\mathcal{Y}$ for the limiting case $\Lambda = \Delta$.

COMMENT NO. 26. Must put in remarks on how in the "singular case" $\Lambda = \Delta$ the catenary problem gives rise to an abnormal extremal.

Chapter 6

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

$$L = -mc||\dot{q}|| + \frac{\rho}{c}A(q) \cdot \dot{q}\,. \tag{6.1}$$

Here

1. the variable q takes values in 4-dimensional Minkowski space-time Q,
2. $||\ldots||$ is length with respect to the Lorentz-Minkowski metric, so that

$$||v|| = \sqrt{\langle v, v\rangle}\,, \tag{6.2}$$

where $\langle \cdot, \cdot \rangle$ is the Minkowski inner product, given, with respect to any Lorentz frame, by

$$\langle v, w \rangle = \sum_{i=0}^{3} g_i v_i w_i \,, \text{ with } g_0 = c^2 \,, \; g_1 = g_2 = g_3 = -1 \,, \tag{6.3}$$

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

$$w \cdot v = w_0 v_0 + w_1 v_1 + w_2 v_2 + w_3 v_3 \,. \tag{6.4}$$

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 6.0.1 Formula (6.1) 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

$$\alpha mc ||\dot{q}|| = \alpha mc \sqrt{c^2 - ||\dot{q}_\#||^2} \,, \tag{6.5}$$

where $||\dot{q}_\#||$ is now the ordinary Euclidean length of the 3-vector $\dot{q}_\#$. Therefore

$$\alpha mc ||\dot{q}|| = \alpha mc^2 \sqrt{1 - \frac{||\dot{q}_\#||^2}{c^2}} \,, \tag{6.6}$$

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 [58], 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 [79].

On the other hand, a truly succesful approach was proposed by Tulczyjew in 1977 in [129] (cf. also [130]), 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 (4.15) 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

$$H = p \cdot u - L = p \cdot u + mc||u|| - \frac{\rho}{c} A(q) \cdot u \,, \tag{6.7}$$

so that $H = \pi \cdot u + mc||u||$, where $\pi = p - \frac{\rho}{c}A(q)$.

Then $\frac{\partial H}{\partial u} = 0$ yields

$$\pi_0 = -mc^3 \frac{u_0}{||u||}\,, \quad \pi_i = mc\frac{u_i}{||u||} \quad \text{for} \quad i = 1, 2, 3\,, \tag{6.8}$$

while the other Hamilton equations say that

$$\frac{dq_i}{ds} = u_i\,, \quad \frac{dp_i}{ds} = \frac{\rho}{c}\sum_j \frac{\partial A_j}{\partial q_i}(q)\, u_j\,. \tag{6.9}$$

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 (6.8), (6.9) is invariant under arbitrary reparametrizations of the curve. Notice also that the 4-momentum covector π satisfies

$$||\pi||^2 = m^2c^2\,, \tag{6.10}$$

where $||\pi||$ is the norm $\sqrt{\langle \pi, \pi\rangle}$ of $||\pi||$ arising from the inner product

$$\langle p, p'\rangle = \frac{p_0 p_0'}{c^2} - \sum_{i=1}^{3} p_i p_i' \tag{6.11}$$

of covectors. This in particular implies the famous formula

$$E = mc^2 \tag{6.12}$$

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 (6.9) we get

$$\frac{d\pi_i}{ds} = \frac{\rho}{c}\sum_j \left(\frac{\partial A_j}{\partial q_i}(q) - \frac{\partial A_i}{\partial q_j}(q)\right) u_j\,, \tag{6.13}$$

i.e.,

$$\frac{d\pi_i}{ds} = \frac{\rho}{c}\sum_j F_{ij}(q) u_j\,, \tag{6.14}$$

where

$$F_{ij} = \frac{\partial A_j}{\partial q_i} - \frac{\partial A_i}{\partial q_j}\,. \tag{6.15}$$

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

$$\pi_i = \frac{mv_i}{\sqrt{1 - \frac{||v||^2}{c^2}}}\,, \tag{6.16}$$

showing that π_1, π_2, π_3 are the components of the relativistic momentum of our particle. Equation (6.13) 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

$$\text{Force} = \rho\Big(E + \frac{1}{c}(v \times B)\Big)\,, \tag{6.17}$$

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 7

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 §4.7, page 81, one defines $V(A', A)$, for two points A', A, to be the minimum of the values of the integral (4.24) 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" $\mathcal{T} \subseteq Q$, then we can also consider the value function $q \mapsto V_{\mathcal{T}}(q)$, defined by letting

$$V_{\mathcal{T}}(q) = \inf\{V(q, q') : q' \in \mathcal{T}\}\,. \tag{7.1}$$

It is clear that $V_{\mathcal{T}}$ 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 $V_{\mathcal{T}}$) the central object of the method of "dy-

namic programming," one of the theory's main tools.

According to this method, one can determine the value function $V_{\mathcal{T}}$ 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

$$\inf\{\langle \nabla V(q), f(q,u)\rangle + L(q,u) : u \in U\} = 0\,. \tag{7.2}$$

and it can be proved, under appropriate technical conditions, that the function $V_{\mathcal{T}}$ satisfies (7.2) and is the unique solution that satisfies a certain boundary condition.

However, $V_{\mathcal{T}}$ in general fails to be continuously differentiable, so (7.2) has to be interpreted in a weak sense, for example as a "viscosity solution" (cf. [52] for details). On the other hand, at those points q where $V_{\mathcal{T}}$ is differentiable and the infimum of (7.2) 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

$$\dot{x} = \frac{u}{1+y^2}\,,\ \ \dot{y} = v\,, \tag{7.3}$$

with control constraint $|u| + |v| \le 1$, 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

$$V(x,y) = |x| + |y|\,. \tag{7.4}$$

Notice that V is not continuously differentiable. However, V is continuously differentiable on the set $S = \{(x,y) : xy \neq 0\}$, and

$$\nabla V(x,y) = (\mathrm{sgn}(x), \mathrm{sgn}(y)) \ \text{ for } \ (x,y) \in S\,. \tag{7.5}$$

The HJB equation says

$$\inf\Big\{\frac{u.\mathrm{sgn}(x)}{1+y^2} + v.\mathrm{sgn}(y) : |u| + |v| \le 1\Big\} = -1\,, \tag{7.6}$$

from which we get $v = -\text{sgn}(y)$ 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.

7.1 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

$$p(t) = \frac{\partial L}{\partial u}(q_*(t), \dot{q}_*(t), t)\,, \tag{7.7}$$

(cf. (4.7)), 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 (7.7), i.e., by

$$p(t) = \frac{\partial L}{\partial u}(q_*(t), \dot{q}_*(t))\,, \tag{7.8}$$

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 (7.7), 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 (7.7) anyhow, it would appear that nothing is gained by substituting (A) for (7.7). This appearance, however, is quite deceptive, as was explained in Chapter 4. We saw there that *the gain achieved by using (A) instead of (7.7) is in fact enormous,* because (A) implies (7.7) 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 (7.7) is not. (We have seen examples of this in Chapter 4. In Example 4.11.1 L fails to be differentiable with respect to u at $u = \dot{q}(t)$, so (7.7) does not make sense. In Example 4.11.3 there is an inequality constraint on $\dot{q}(t)$, so (A) does not imply (7.7) 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 (7.7) as a characterization of "the momentum along a curve." The substitution of (A) for (7.7) has a cost, however: while (7.7) 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

[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.

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.

Example 7.1.1 Consider the minimum-time optimal control problem in $\mathbb{R}^2$ discussed earlier, with equations (7.3), 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 (7.5) 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

$$H = \frac{p_x u}{1+y^2} + p_y v - p_0\,, \tag{7.9}$$

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

$$\dot{p}_x^\xi = 0\,,\ \ \dot{p}_y^\xi = -\frac{2y p_x^\xi u^\xi}{(1+y^2)^2}\,, \tag{7.10}$$

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

$$M(t) = \tilde{M}(t) - p_0\,,\ \ \text{where}\ \ \tilde{M}(t) = \max\left(\frac{|p_x^\xi(t)|}{1+y(t)^2}, |p_y^\xi(t)|\right). \tag{7.11}$$

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

$$u(t) = \mathrm{sgn}(p_x^\xi(t))\,,\ \ v(t) = 0\ \ \text{if}\ \ \frac{|p_x^\xi(t)|}{1+y(t)^2} > |p_y^\xi(t)|\,, \tag{7.12}$$

and

$$u(t) = 0\,,\ \ v(t) = \operatorname{sgn}(p_y^\xi(t)) \ \text{ if } \ \frac{|p_x^\xi(t)|}{1+y(t)^2} < |p_y^\xi(t)|\,. \tag{7.13}$$

It follows that at time $t = |y|$ the equality $|p_x^\xi(t)| = |p_y^\xi(t)|$ holds, from which it easily follows that $p_x^\xi(t) = \operatorname{sgn}(\bar{x})$ and $p_\xi^y(t) = \operatorname{sgn}(\bar{y})$.

This shows that

$$(p_x^\xi(0), p_y^\xi(0)) = -\nabla V(\bar{x}, \bar{y}) \ \text{ if } \ \bar{x} \neq 0\,,\ \bar{y} \neq 0\,. \tag{7.14}$$

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*

$$p(0) = -\nabla V(\bar{q}) \tag{7.15}$$

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. [40, 39, 52, 110] for details).

Remark 7.1.2 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 8

From principle to theorem

8.1 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 [98], 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. [109], 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 $f(x) + pi + qi^2 + ri^3 + \ldots...$, 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.

8.2 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 [107] 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).

8.3 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 (3.5) by Lagrange's method, we would set $\delta I = 0$, and then write

$$0 = \delta I = \delta \int_a^b L\, dt = \int_a^b (\delta L)\, dt = \int_a^b (M\, \delta q + N\, \delta \dot{q}) dt\,, \tag{8.1}$$

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 (4.3).

Remark 8.3.1 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

$$dy = (x + dx)^2 - x^2 = 2x\, dx + (dx)^2\,. \tag{8.2}$$

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

$$dy/dx = 2x + dx\,. \tag{8.3}$$

[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 [30], p. 194.)

They would then have to contradict themselves and arbitrarily set $dx = 0$, thereby obtaining the desired foimula

$$dy/dx = 2x\,. \tag{8.4}$$

The problem here is, of course, that $dx \neq 0$ is neded to go from (8.2) to (8.3), and $dx = 0$ to go from (8.3) to (8.4). 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 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,

[2]The Dutch Bernhard Nieuwentijt (1654-1718) was a physician-burgomaster of Purmerend, near Amsterdam.

[3]According to Cajori ([30], 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. [59], 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.

footnote 7 in page 162), without being distracted by those foundational issues that later generations would eventually settle in a completely rigorous way. ♢

8.4 The modern justification of Lagrange's derivation; the Lavrentiev phenomenon

Lagrange's derivation of (4.3), 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 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:

$$\frac{\partial}{\partial\varepsilon}\frac{\partial q}{\partial t} = \frac{\partial}{\partial t}\frac{\partial q}{\partial\varepsilon}\,. \tag{8.5}$$

(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 (8.1) is, of course, that we are differentiating the integral with respect to the parameter ε at $\varepsilon = 0$:

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \int_a^b L\Big(q_*(t) + \varepsilon r(t), \dot q_*(t) + \varepsilon\dot r(t), t\Big)\,dt$$

$$= \int_a^b \left[\frac{\partial L}{\partial q}(q_*(t), \dot{q}_*(t), t) \cdot r(t) + \frac{\partial L}{\partial \dot{q}}(q_*(t), \dot{q}_*(t), t) \cdot \dot{r}(t) \right] dt \quad (8.6)$$

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 (8.6),

(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 (8.6) 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*

$$\left\| \frac{\partial L}{\partial q}(q, u, t) \right\| + \left\| \frac{\partial L}{\partial u}(q, u, t) \right\| \leq \varphi(t) \text{ for } ||q - q_*(t)|| \leq \alpha, \ ||u - \dot{q}_*(t)|| \leq \alpha. \quad (8.7)$$

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 (8.7) 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. [81], [36] §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 (4.3) appropriately, as was already explained in Chapter 4. This may seem fairly surprising at first, since the elementary derivation of (4.3) 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 (4.3).) The true explanation of this strange fact is that Conjecture M2, in Chapter 4, is true, and the statement of Conjecture M2 contains no derivatives of L with respect to u.

8.5 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

$$\delta^2 I = \int_a^b (M\,\delta^2 q + N\,\delta^2 \dot{q} + P(\delta q)^2 + 2Q\,\delta q \delta \dot{q} + R(\delta \dot{q})^2)\,.$$

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

$$\delta^2 I = \int_a^b (P(\delta q)^2 + 2Q\,\delta q \delta \dot{q} + R(\delta \dot{q})^2)\,. \tag{8.8}$$

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 8.5.1 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 (8.8) dominates the other ones. $\diamondsuit$

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

[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.

$\int_a^b(\dot\alpha(\delta q)^2 + 2\alpha\,\delta q\,\delta\dot q)$ 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 (8.8) as

$$\delta^2 I = \int_a^b ((P+\dot\alpha)(\delta q)^2 + 2(Q+\alpha)\,\delta q\delta\dot q + R(\delta\dot q)^2)\,. \tag{8.9}$$

Now suppose we can choose α to be a solution of the differential equation

$$(P+\dot\alpha)R = (Q+\alpha)^2\,. \tag{8.10}$$

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 (8.10) need not have a solution defined on the whole interval $[a, b]$. The reader will recognize (8.10) as a Riccati differential equation.[6] It is well known that 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 (8.10) exists on J. Since the restriction to J of a solution of (3.5) 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.

[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.

[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 8.3.1 in page 156.

8.6 The needle variations approach

COMMENT NO. 27. **This section still has to be written.**

8.7 The classical version of the maximum principle

The first rigorous version of the maximum principle appears in the book [98]. This "classical" version was then improved by other authors, e.g. L. Cesari [36], M. Hestenes [67], E. B. Lee and L. Markus [82], and L. D. Berkovitz [10]. 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 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 $\mathcal{U}$ of U-valued functions on $[a, b]$, and $\bar{q}, \hat{q} \in Q$, let us use $\mathcal{C}(\mathcal{U}, \bar{q}, \hat{q})$ to denote the set of all pairs $(q(\cdot), u(\cdot))$ such that $u(\cdot) \in \mathcal{U}$, $q(\cdot)$ is a solution of (3.10) (i.e., $q(\cdot)$ is an absolutely continuous curve $[a, b] \mapsto Q$ such that (3.10) holds for almost every t), $q(a) = \bar{q}$, and $q(b) = \hat{q}$. Use $\mathcal{U}_{adm}$ 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 $\mathcal{C}(\mathcal{U}_{adm}, \bar{q}, \hat{q})$. 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 $\mathcal{U}$ of controls. (For example, we may want $\mathcal{U}$ 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 mini-

mization problem in $\mathcal{C}(\mathcal{U}, \bar{q}, \hat{q})$, where $\mathcal{U}$ is any class of controls such that *(a)* $\mathcal{U} \subseteq \mathcal{U}_{adm}$, *(b)* $\mathcal{U}$ *is closed under interval substitutions (that is, if* $u(\cdot) \in \mathcal{U}$, $v(\cdot) \in \mathcal{U}$, *and* J *is a subinterval of* $[a, b]$, *then* $u\#_J v \in \mathcal{U}$*), and (c) all constants belong to* $\mathcal{U}$. The conclusion is the same as in the classical result.

Remark 8.7.1 If $\mathcal{U} = \mathcal{U}_{adm}$, then $\mathcal{U}$ 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* $u(\cdot) \in \mathcal{U}$ *and every constant map* $v : [a, b] \mapsto U$ *there exist compact subsets* E_j *of* $[a, b]$ *such that (i)* $u\#_{E_j} v \in \mathcal{U}$ *for all* j, *and (ii)* $\text{meas}([a, b] \backslash E_j) \to 0$ *as* $j \to \infty$.

One can easily verify that (c') always holds when $\mathcal{U} = \mathcal{U}_{adm}$, and it is trivial that (c) implies (c'). ♢

Remark 8.7.2 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. 28. 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 [123, 124, 125]. ♢

Remark 8.7.3 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.