VISUAL DIFFERENTIAL GEOMETRY *and* FORMS

A mathematical drama in five acts

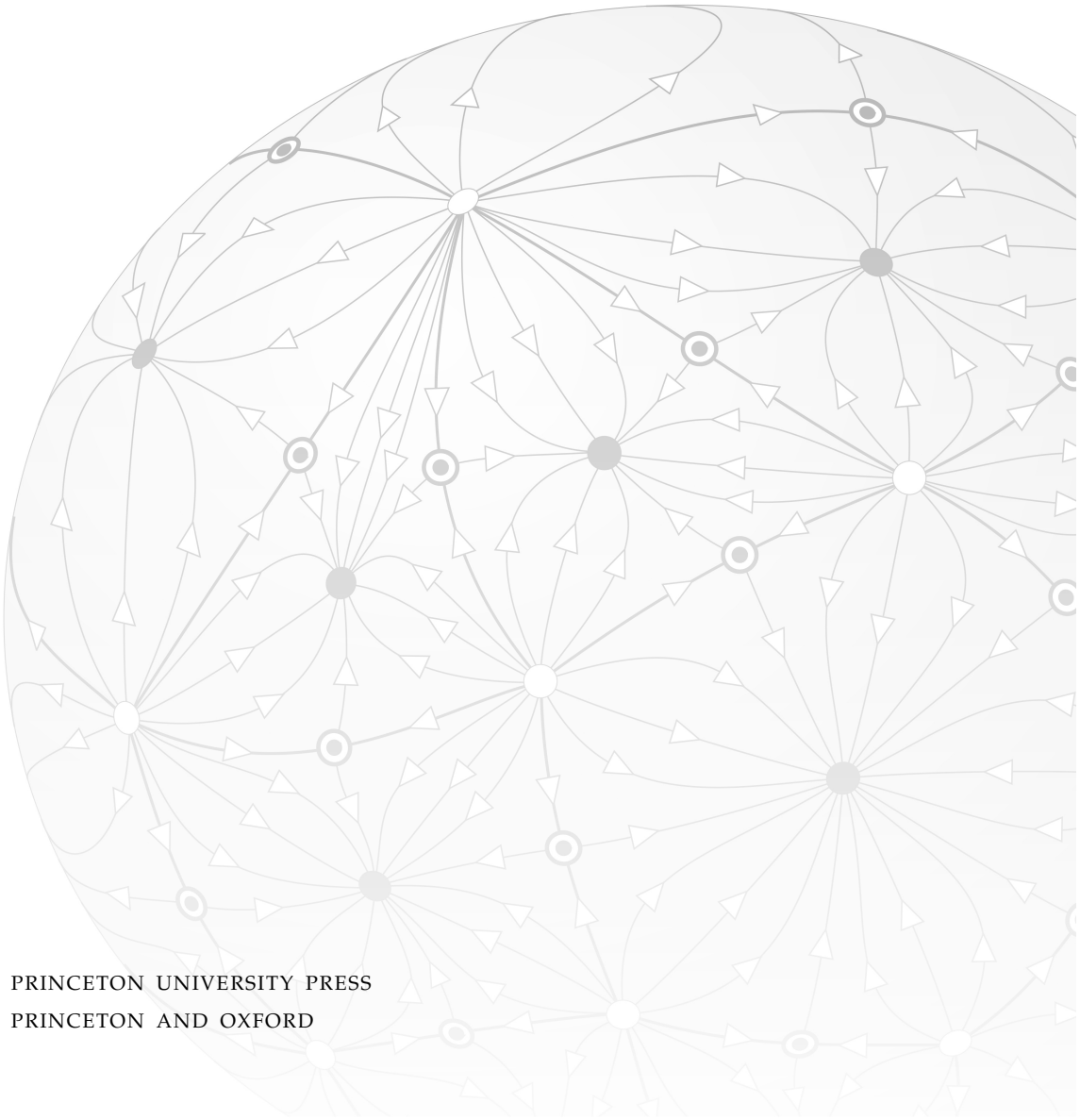
TRISTAN NEEDHAM

VISUAL DIFFERENTIAL GEOMETRY *and* FORMS

VISUAL DIFFERENTIAL GEOMETRY *and* FORMS

A mathematical drama in five acts

TRISTAN NEEDHAM



PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
6 Oxford Street, Woodstock, Oxfordshire OX20 1TR

press.princeton.edu

ISBN 9780691203690
ISBN (pbk.) 9780691203706
ISBN (ebook) 9780691219899

Library of Congress Control Number: 2021934723

British Library Cataloging-in-Publication Data is available

Editorial: Susannah Shoemaker, Kristen Hop
Production Editorial: Terri O'Prey
Text Design: Wanda España
Jacket/Cover Design: Wanda España
Production: Brigid Ackerman
Publicity: Matthew Taylor, Amy Stewart
Copyeditor: Gregory W. Zelchenko
Cover image: Stiefel vector field (figure [19.12])

This book has been composed in Palatino (text) and Euler (mathematics)

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For Roger Penrose

Contents

ACT V
Forms

Prologue

The Faustian Offer

> Algebra is the offer made by the devil to the mathematician. The devil says: *"I will give you this powerful machine, and it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvellous machine."* ... the danger to our soul is there, because when you pass over into algebraic calculation, essentially you stop thinking: you stop thinking geometrically, you stop thinking about the meaning.
>
> Sir Michael Atiyah[1]

"*Differential Geometry*" contains the word "*Geometry*."

A tautology? Well, the undergraduate who first opens up the assigned textbook on the subject may care to disagree! In place of geometry, our hapless student is instead confronted with a profusion of *formulas*, and their proofs consist of lengthy and opaque *computations*. Adding insult to injury, these computations are frequently *ugly*, involving a "debauch of indices"[2]—a phrase coined by Élie Cartan (one of the principal heroes of our drama) in 1928. If the student is honest and brave, the professor may be forced to confront an embarrassingly blunt question: "Where has the *geometry* gone?!"

Now, truth be told, most modern texts *do* in fact contain many *pictures*, usually of computer-generated curves and surfaces. But, with few exceptions, these pictures are of specific, concrete examples, which merely *illustrate* theorems whose proofs rest entirely upon symbolic manipulation. In and of themselves, these pictures *explain nothing!*

The present book has *two* distinct and equally ambitious objectives, the first of which is the subject of the first four Acts—to put the "Geometry" back into introductory "Differential Geometry." The 235 hand-drawn diagrams contained in the pages that follow are qualitatively and fundamentally of a different character than mere computer-generated examples. They are the conceptual fruits of many years of intermittent but intense effort—they are the visual embodiment of *intuitive geometric explanations of stunning geometric facts.*

The words I wrote in the Preface to VCA[3] apply equally well now: "A significant proportion of the geometric observations and arguments contained in this book are, to the best of my knowledge, new. I have not drawn attention to this in the text itself as this would have served no useful purpose: students don't need to know, and experts will know without being told. However, in cases where an idea is clearly unusual but I am aware of it having been published by someone else, I have tried to give credit where credit is due." In addition, I have attributed *exercises* that appear to be original, but that are not of my making.

On a personal note (but with a serious mathematical point to follow), the roots of the present endeavour can be traced back decades, to my youth. The story amounts to a tale of two books.

---

[1] *Mathematics in the 20th Century* (Shenitzer & Stillwell, 2002, p. 6)

[2] The full quotation begins to reveal Cartan's heroic stature: "The utility of the absolute differential calculus of Ricci and Levi-Civita must be tempered by an avoidance of excessively formal calculations, where the debauch of indices disguises an often very simple geometric reality. It is this reality that I have sought to reveal." (From the preface to Cartan 1928.)

[3] Given the frequency with which I shall have occasion to refer back to my first book, *Visual Complex Analysis* (Needham 1997), I shall adopt the compact conceit of referring to it simply as VCA.

The *first book* ignited my profound fascination with Differential Geometry and with Einstein's General Theory of Relativity. Perhaps the experience was so intense because it was my *first* love; I was 19 years old. One day, at the end of my first year of studying physics at Merton College, Oxford, I stumbled upon a colossal black book in the bowels of Blackwell's bookshop. Though I did not know it then, the 1,217-page tome was euphemistically referred to by relativity theorists as "The Bible." Perhaps it is appropriate, then, that this remarkable work altered the entire course of my life. Had I not read *Gravitation* (Misner, Thorne, and Wheeler 1973), I would never have had the opportunity[4] to study under (and become lifelong friends with) Roger Penrose, who in turn fundamentally transformed my understanding of mathematics and of physics.

In the summer of 1982, having been intrigued by the mathematical glimpses contained in Westfall's (1980) excellent biography of Newton, I made an intense study of Newton's (1687) masterpiece, *Philosophæ Naturalis Principia Mathematica*, usually referred to simply as the *Principia*. This was the *second book* that fundamentally altered my life. While V. I. Arnol'd[5] and S. Chandrasekhar (1995) sought to lay bare the remarkable nature of Newton's *results* in the *Principia*, the present book instead arose out of a fascination with Newton's *methods*.

As we have discussed elsewhere,[6] Newtonian scholars have painstakingly dismantled the pernicious myth[7] that the results in the 1687 *Principia* were first derived by Newton using his original 1665 version of the calculus, and only later recast into the geometrical form that we find in the finished work.

Instead, it is now understood that by the mid-1670s, having studied Apollonius, Pappus, and Huygens, in particular, the mature Newton became disenchanted with the form in which he had originally discovered the calculus in his youth—which is different again from the Leibnizian form we all learn in college today—and had instead embraced purely geometrical methods.

Thus it came to pass that by the 1680s Newton's algebraic infatuation with power series gave way to a new form of calculus—what he called the "synthetic method of fluxions"[8]—in which the geometry of the Ancients was transmogrified and reanimated by its application to shrinking geometrical figures in their moment of vanishing. *This* is the potent but nonalgorithmic form of calculus that we find in full flower in his great *Principia* of 1687.

Just as I did in VCA, I now wish to take full advantage of Newton's approach throughout this book. Let me therefore immediately spell it out, and in significantly greater detail than I did in VCA, in the vain hope that this second book may inspire more mathematicians and physicists to adopt Newton's intuitive (yet rigorous[9]) methods than did my first.

If two quantities A and B depend on a small quantity ϵ, and their ratio approaches unity as ϵ approaches zero, then we shall avoid the more cumbersome language of limits by following Newton's precedent in the *Principia*, saying simply that, "A is ultimately equal to B." Also, as we did in earlier works (Needham 1993, 2014), we shall employ the symbol $\asymp$ to denote this concept of ultimate equality.[10] In short,

$$\text{“}A \text{ is ultimately equal to } B\text{”} \iff A \asymp B \iff \lim_{\epsilon \to 0} \frac{A}{B} = 1.$$

---

[4] Years later I was privileged to meet with Wheeler several times, and to correspond with him, so I was finally able to thank him directly for the impact that his *Gravitation* had had upon my life.

[5] See Arnol'd and Vasil'ev (1991); Arnol'd (1990).

[6] See Needham (1993), the Preface to VCA, and Needham (2014).

[7] Sadly, this myth originated with Newton himself, in the heat of his bitter priority battle with Leibniz over the discovery of the calculus. See Arnol'd (1990), Bloye and Huggett (2011), de Gandt (1995), Guicciardini (1999), Newton (1687, p. 123), and Westfall (1980).

[8] See Guicciardini (2009, Ch. 9).

[9] Fine print to follow!

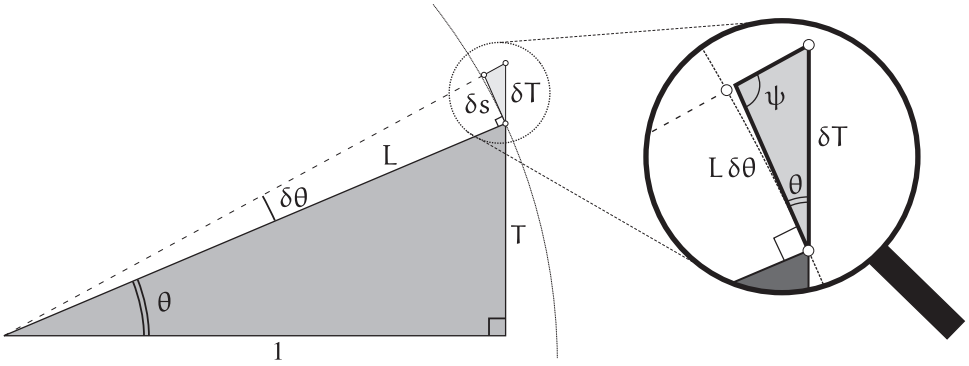
[10] This notation was subsequently adopted by the Nobel physicist, Subrahmanyan Chandrasekhar (see Chandrasekhar 1995, p. 44).

It follows [exercise] from the theorems on limits that ultimate equality is an equivalence relation, and that it also inherits additional properties of ordinary equality, e.g., $X \asymp Y \;\&\; P \asymp Q \Rightarrow X \cdot P \asymp Y \cdot Q$, and $A \asymp B \cdot C \Leftrightarrow (A/B) \asymp C$.

Before we begin to apply this idea in earnest, we also note that the jurisdiction of ultimate equality can be extended naturally to things other than numbers, enabling one to say, for example, that two triangles are "ultimately similar," meaning that their angles are ultimately equal.

Having grasped Newton's method, I immediately tried my own hand at using it to simplify my teaching of introductory calculus, only later realizing how I might apply it to Complex Analysis (in VCA), and now to Differential Geometry. Though I might choose any number of simple, illustrative examples (*see* Needham 1993 for more), I will reuse the specific one I gave in the preface to VCA, and for one simple reason: *this* time I will use the "$\asymp$"–notation to *present* the argument rigorously, whereas in VCA I did not. Indeed, this example may be viewed as a recipe for transforming most of VCA's "explanations" into "proofs,"[11] merely by sprinkling on the requisite $\asymp$ s.

Let us show that if $T = \tan\theta$, then $\frac{dT}{d\theta} = 1 + T^2$. See figure below. If we increase θ by a small (ultimately vanishing) amount $\delta\theta$, then T will increase by the length of the vertical hypotenuse δT of the small triangle, in which the other two sides of this triangle have been constructed to lie in the directions $(\theta + \delta\theta)$ and $(\theta + \frac{\pi}{2})$, as illustrated. To obtain the result, we first observe that in the limit that $\delta\theta$ vanishes, the small triangle with hypotenuse δT is ultimately similar to the large triangle with hypotenuse L, because $\psi \asymp \frac{\pi}{2}$. Next, as we see in the magnifying glass, the side δs adjacent to θ in the small triangle is ultimately equal to the illustrated arc of the circle with radius L, so $\delta s \asymp L\,\delta\theta$. Thus,



$$\frac{dT}{L\,d\theta} \asymp \frac{\delta T}{L\,\delta\theta} \asymp \frac{\delta T}{\delta s} \asymp \frac{L}{1} \quad \Longrightarrow \quad \frac{dT}{d\theta} = L^2 = 1 + T^2.$$

So far as I know, Newton never wrote down this specific example, but compare the illuminating directness of his *style*[12] of geometrical reasoning with the unilluminating computations we teach our students today, more than three centuries later! As Newton himself put it,[13] the geometric method is to be preferred by virtue of the "clarity and brevity of the reasoning involved and because of the simplicity of the conclusions and the illustrations required." Indeed, Newton went even further, resolving that *only* the synthetic method was "worthy of public utterance."

---

[11] I was already using the $\asymp$ notation (both privately and in print) at the time of writing VCA, and, in hindsight, it was a mistake that I did not employ it in that work; this led some to suppose that the arguments presented in VCA were less rigorous than they actually were (and remain).

[12] The best ambassador for Newton's approach will be you yourself. We therefore suggest that you *immediately* try your *own* hand at Newtonian reasoning, by doing Exercises 1, 2, 3, and 4, on page 24.

[13] See Guicciardini (2009, p. 231)

Newton himself did not employ *any* symbol to represent his concept of "ultimate equality." Instead, his devotion to the geometrical *method* of the Ancients spilled over into emulating their *mode* of expression, causing him to write out the words "ultimately have the ratio of equality," every single time the concept was invoked in a proof. As Newton (1687, p. 124) explained, the *Principia* is "written in words at length in the manner of the Ancients." Even when Newton claimed that two ratios were ultimately equal, he insisted on expressing *each ratio* in words. As a result, I myself was quite unable to follow Newton's reasoning without first transcribing and summarizing each of his paragraphs into "modern" form (which was in fact already quite common in 1687). Indeed, back in 1982, this was the catalyst for my private introduction and use of the symbol, $\asymp$.

It is my view that Newton's choice *not* to introduce a symbol for "ultimate equality" was a tragically consequential error for the development of mathematics. As Leibniz's symbolic calculus swept the world, Newton's more penetrating geometrical method fell by the wayside. In the intervening centuries only a handful of people ever sought to repair this damage and revive Newton's approach, the most notable and distinguished recent champion having been V. I. Arnol'd[14] (1937–2010).

Had Newton shed the trappings of this ancient mode of exposition and instead employed some symbol (*any* symbol!) in place of the words "ultimately equal," his dense, paragraph-length proofs in the *Principia* might have been reduced to a few succinct lines, and his mode of thought might still be widely employed today. Both VCA and this book are attempts to demonstrate, very concretely, the continuing relevance and vitality of Newton's geometrical approach, in areas of mathematics whose discovery lay a century in the future at the time of his death in 1727.

Allow me to insert some fine print concerning my use of the words "rigour" and "proof." Yes, my explicit use of Newtonian ultimate equalities in this work represents a quantum jump in rigour, as compared to my exposition in VCA, but there will be some mathematicians who will object (with justification!) that even this increase in rigour is insufficient, and that *none* of the "proofs" in this work are worthy of that title, including the one just given: I did not actually prove that the side of the triangle is ultimately equal to the arc of the circle.

I can offer no *logical* defence, but will merely repeat the words I wrote in the Preface of VCA, more than two decades ago: "My book will no doubt be flawed in many ways of which I am not yet aware, but there is one 'sin' that I have intentionally committed, and for which I shall not repent: many of the arguments are not rigorous, at least as they stand. This is a serious crime if one believes that our mathematical theories are merely elaborate mental constructs, precariously hoisted aloft. Then rigour becomes the nerve-racking balancing act that prevents the entire structure from crashing down around us. But suppose one believes, as I do, that our mathematical theories are attempting to capture aspects of a robust Platonic world that is not of our making. I would then contend that an initial lack of rigour is a small price to pay if it allows the reader to see into this world more directly and pleasurably than would otherwise be possible." So, to preemptively address my critics, let me therefore concede, from the outset, that when I claim that an assertion is "proved," it may be read as, *"proved beyond a reasonable doubt"!*[15]

Separate and apart from the issue of rigour is the sad fact that in rethinking so much classical mathematics I have almost certainly made mistakes: The blame for all such errors is mine, and mine alone. But please do not blame my geometrical tools for such poor craftsmanship—I am *equally* capable of making mistakes when performing symbolic computations! Corrections will be received with gratitude at VDGF.correction@gmail.com.

The book can be fully understood without giving a second thought to the complete arc of the unfolding drama, told as it is in five Acts. That said, I think that plot matters, and that the book's unorthodox structure and title are fitting, for the following reasons. First, I have sought

---

[14] See, for example, Arnol'd (1990).

[15] Upon reading these words, a strongly supportive member of the Editorial Board of Princeton University Press suggested to my editor that in place of "Q.E.D.," I conclude each of my proofs with the letters, "P.B.R.D."!

to present the ideas as dramatically as I myself see them, not only in terms of their historical development,[16] but also (more importantly) in terms of the cascading, interconnected flow of the ideas themselves, and their startling implications for the rest of mathematics and for physics. Second, more by instinct than design, the role of each of the five Acts does indeed follow (more or less) the classical structure of a Shakespearean drama; in particular, the anticipated "Climax" is indeed Act III: "Curvature." It was in fact years after I had begun work on the book that one day it suddenly became clear to me that what I had been composing all along had been *a mathematical drama in five acts.* That very day I "corrected" the title of the work, and correspondingly changed its five former "Parts" into "Acts":

- Act I: The Nature of Space
- Act II: The Metric
- Act III: Curvature
- Act IV: Parallel Transport
- Act V: Forms

The first four Acts fulfill the promise of a self-contained, *geometrical* introduction to Differential Geometry. Act IV is the true mathematical powerhouse that finally makes it possible to provide *geometric proofs* of many of the assertions made in the first three Acts.

Several aspects of the *subject matter* are as unorthodox (in a first course) as the geometrical methods by which they are treated. Here we shall describe only the three most important examples.

First, the climax *within* the climax of Act III is the *Global Gauss–Bonnet Theorem*—a remarkable link between local geometry and global topology. While the inclusion of this topic is standard, our treatment of it is not. Indeed, we celebrate its centrality and fundamental importance with an extravagant display of mathematical fireworks: we devote *five* chapters to it, offering up *four* quite distinct proofs, each one shedding new light on the result, and on the nature of Differential Geometry itself.

Second, the transition (usually in graduate school) from 2-dimensional surfaces to n-dimensional spaces (called "manifolds") is often confusing and intimidating for students. Chapter 29—the second longest chapter of the book—seeks to bridge this gap by focusing (initially) on the curvature of **3**-dimensional manifolds, which can be *visualized*; yet we frame the discussion so as to apply to *any* number of dimensions. We use this approach to provide an intuitive, geometrical, yet technically complete, introduction to the famous *Riemann tensor*, which measures the curvature of an n-dimensional manifold.

Third, having committed to a full treatment of the Riemann tensor, we felt it would have been *immoral* to have hidden from the reader its single greatest triumph in the arena of the natural world. We therefore conclude Act IV with a prolonged, *geometrical* introduction to Einstein's glorious *General Theory of Relativity*, which explains gravity as the curvature impressed upon 4-dimensional spacetime by matter and energy. This is the third longest chapter of the book. Not only does it treat (in complete geometrical detail) the famous *Gravitational Field Equation* (which Einstein discovered in 1915) but it also explains some of the most recent and exciting discoveries regarding its implications for black holes, gravitational waves, and cosmology!

Now let us turn to Act V, which is quite different in character from the four Acts that precede it, for it seeks to accomplish a *second* objective of the work, one that is quite distinct from the first, but no less ambitious.

Even the most rabid geometrical zealot must concede that Atiyah's diabolical machine (described in the opening quotation) is a *necessary* evil; but if we *must* calculate, let us at least

[16] As I did in VCA, I *strongly* recommend Stillwell's (2010) masterpiece, *Mathematics and Its History*, as a companion to this book, for it provides deeply insightful and detailed analysis of many historical developments that we can only touch on here.

do so gracefully! Fortunately, starting in 1900, Élie Cartan developed a powerful and elegant new method of *computation*, initially to investigate Lie Groups, but later to provide a new approach to Differential Geometry.

Cartan's discovery is called the "Exterior Calculus," and the objects it studies and differentiates and integrates are called "Differential Forms," here abbreviated simply to *Forms*. We shall ultimately follow Cartan's lead, illustrating his method's power and elegance in the final chapter of Act V—the longest chapter of the book—*reproving symbolically results that were proven geometrically in the first four Acts.* But Forms will carry us *beyond* what was possible in the first four Acts: in particular, they will provide a beautifully efficient method of calculating the Riemann tensor of an n-manifold, via its *curvature 2-forms*.

First, however, we shall fully develop Cartan's ideas in their own right, providing a self-contained introduction to Forms that is *completely independent* of the first four Acts. Lest there be any confusion, we repeat, *the first six chapters*—out of seven—*of Act V make no reference whatsoever to Differential Geometry!* We have done this because Forms find fruitful applications across diverse areas of mathematics, physics, and other disciplines. *Our aim is to make Forms accessible to the widest possible range of readers, even if their primary interest is* not *Differential Geometry.*

To that end, we have sought to treat Forms much more intuitively and *geometrically* than is customary. That said, the reader should be under no illusions: the principal purpose of Act V is to construct, at the *undergraduate* level, the "Devil's machine"—a remarkably powerful method of *computation*.

The immense power of these Forms is reminiscent of the complex numbers: a tiny drop goes in, and an ocean pours out—Cartan's Forms explain vastly more than was asked of them by their discoverer, a sure sign that he had hit upon *Platonic* Forms!

To give just one example, Forms unify and clarify *all* of Vector Calculus, in a way that would be a *revelation* to undergraduates, if only they were permitted to see it. Indeed, Green's Theorem, Gauss's Theorem, and Stokes's Theorem are merely different manifestations of a *single* theorem about Forms that is simpler than any of these special cases! Despite the indisputable importance of Differential Forms across mathematics and physics, most *undergraduates* will leave college without ever having seen them, and I have long considered this a scandal. Only a precious handful[17] of undergraduate textbooks (on either Vector Calculus *or* Differential Geometry) even mention their existence, and they are instead relegated to graduate school.

This lamentable state of affairs is now well into its second century, and I see no signs of an impending sea change. In response, Act V seeks not to curse the dark, but rather to light a candle,[18] striving to convince the reader that Cartan's Forms (and their underlying "tensors") are as *simple* as they are beautiful, and that they (and the name Cartan!) deserve to become a standard part of the *undergraduate* curriculum. *This* is the brazenly ambitious goal of Act V. After drowning the reader in *pure* Geometry for the first four Acts, we hope that the computational aspect of this final Act may serve as a suitably cathartic dénouement!

Before we close, let us simply list some housekeeping details:

- First, I have made no attempt write this book as a classroom textbook. While I hope that some brave souls may nevertheless choose to use it for that purpose—as some previously did with VCA—my primary goal has been to communicate a majestic and powerful subject to the reader as honestly and as lucidly as I am able, regardless of whether that reader is a tender neophyte, or a hardened expert.

---

[17] See *Further Reading*, at the end of this book.

[18] Ours is certainly not the first such candle to be lit. Indeed, just as our work was nearing completion, Fortney (2018) published an entire book devoted to this same goal. However, Fortney's work does not include any discussion of Differential Geometry, and, at 461 pages, Fortney's book is considerably longer than the 100-page introduction to Forms contained in Act V of this book.

- My selection of topics may seem eclectic at times: for example, why is no attention paid to the fascinating and important topic of minimal surfaces? Frequently, as in this case, it is for one (or both) of the following two reasons: (1) our focus is on intrinsic geometry, *not* extrinsic[19] geometry; (2) an excellent literature already exists on the subject; in such cases, I have tried to provide useful pointers in the *Further Reading* section at the end of the book.
- *Equations* are numbered with (ROUND) brackets, while *figures* are numbered with [SQUARE] brackets.
- Bold italics are used to highlight the ***definition of a new term***.
- For ease of reference when flipping through the book, noteworthy results are `framed`, while doubly remarkable facts are `double-framed`. In the entire work, only a handful of results are *so* fundamental that they are *triple*-framed; we hope the reader will enjoy finding them, like Easter eggs.
- I have tried to make you, the reader, into an active participant in developing the ideas. For example, as an argument progresses, I have frequently and deliberately placed a pair of logical stepping stones sufficiently far apart that you may need to pause and stretch slightly to pass from one to the next. Such places are marked "[exercise]"; they often require nothing more than a simple calculation or a moment of reflection.
- Last, we encourage the reader to take full advantage of the *Index*; its creation was a painful labour of love!

We bring this Prologue to a close with a broader philosophical objective of the work, one that transcends the specific mathematics we shall seek to explain.

One of the rights [*sic*] of passage from mathematical adolescence to adulthood is the ability to distinguish *true miracles* from *false miracles.* Mathematics itself is replete with the former, but examples of the latter also abound: "I can't believe all those ugly terms cancelled and left me such a beautifully simple answer!"; or, "I can't believe that this complicated expression has such a simple meaning!"

Rather than congratulating oneself in such a circumstance, one should instead hang one's head in shame. For if all those ugly terms cancelled, *they should never have been there in the first place!* And if that complicated expression has a wonderfully simple meaning, *it should never have been that complicated in the first place!*

In my own case, I am embarrassed to confess that mathematical puberty lasted well into my 20s, and I only *started* to grow up once I became a graduate student, thanks to the marvellous twin influences of Penrose and of my close friend George Burnett-Stuart, a fellow advisee of Penrose.

The Platonic Forms of mathematical reality are always perfectly beautiful and they are always perfectly simple; transient impressions to the contrary are manifestations of our own imperfection. My hope is that this book may help nudge the reader towards humility in the face of this perfection, just as my two friends first nudged me down this path, so many years ago amidst the surreal, Escher-like spires of Oxford.

T. N.

Mill Valley, California
Newtonmas, 2019

[19] The meanings of "intrinsic" and "extrinsic" are explained in Section 1.4.

Acknowledgements

Roger Penrose transformed my understanding of mathematics and of physics. From the very first paper of his I ever read, when I was 20 years old, the *perfection*, beauty, and almost musical counterpoint of his ideas elicited in me a profound aesthetic exhilaration that I can only liken to the experience of listening to the opening of Bach's Cantata 101 or Beethoven's *Grosse Fuge*.

From the time that I was his student, Roger's ability to unravel the deepest of mysteries through *geometry* left an indelible mark—it instilled in me a lifelong, unshakable *faith* that a geometric explanation must always exist. (My study of Newton's *Principia* later served to *deepen* my belief in the universality of the geometrical approach.) Without that faith, this work could not exist, for it sometimes took me many *years* of groping before I discovered the geometric explanation of a particular mathematical phenomenon.

To be able to count myself amongst Roger's friends has been a great joy and a high honour for 40 years, and my dedication of this imperfect work to Roger can scarcely repay the intellectual debt that I owe to him, but it is the best that I can do.

In order to properly introduce the next person to whom I owe thanks, I am forced to reveal a somewhat shabby detail about myself: when I first came to America from England in 1989, I smoked two packs of cigarettes per day! In 1995 I was finally able to quit: it was the hardest thing I had ever done, and I would likely have failed, had it not been for the invention of the nicotine patch.

Perhaps five years later, in response to the 1997 publication of my *Visual Complex Analysis* (VCA), I received a "fan letter" from a *medical* researcher at Duke University; he planned to visit the Bay Area and asked if we could meet. With some trepidation, I agreed. My visitor turned out to be Professor Jed Rose, *inventor* of my (saviour) nicotine patch! Jed had started out in mathematics and physics, and had never lost his love of those disciplines, but shrewdly calculated that he could have a greater impact if he directed his energies to medical research, instead. I'm so glad that he did!

Once I began work on VDGF (*Visual Differential Geometry and Forms*, this book) in 2011, Jed became my most enthusiastic supporter, demonstrating great generosity in using funds from his medical inventions to buy out some of my teaching, thereby greatly assisting my research for VDGF. Every time Jed visited me and my family in California during the nine years of work on the book, my spirits were lifted by Jed's relentlessly upbeat personality and his belief in the importance of what I was trying to accomplish. And, as the manuscript slowly evolved, Jed offered a remarkably large number of detailed and helpful suggestions and corrections; the finished work is significantly better as a result of his helpful observations. Thus, as you see, Jed helped me in three linearly independent directions, and I cannot thank him enough. And as if all this were not enough, what started out as a purely intellectual relationship, subsequently blossomed into a very warm and close friendship between our two families.

The next key person I would like to thank is Professor Thomas Banchoff, the distinguished geometer of Brown University. During the writing of this book, I managed to arrange for Tom to come to USF as a visiting scholar in two separate years, for one semester each. Tom was extremely generous to me during both of those semesters, offering to read my evolving manuscript and giving me extremely valuable feedback. Each week he would read the latest installment of the manuscript, hot off the presses, and then he and I would meet in his office each Friday afternoon, and go over his detailed corrections and suggestions, written in red pen in the margins, line by

line. Although this partnership sadly ended when the book was only half done, I have adopted essentially all of his helpful suggestions and corrections, and I am immensely grateful to him for sharing his deep geometrical wisdom and expertise with me.

I wish to express my sincere gratitude to Dr. Wei Liu[20] (a physicist specializing in optics research, whom I hope to eventually meet, some fine day). In 2019 he wrote to me to express his appreciation of VCA, and he enclosed a research paper of his[21] that cited my treatment of the Poincaré–Hopf Theorem. This paper totally opened my eyes to how physicists continue to make wonderful use of a beautiful result of Hopf that seems to have completely evaporated from all *mathematical* textbooks. The result is the subject of Section 19.6.4, and it says this: The Poincaré–Hopf Theorem not only applies to vector fields, but also to Hopf's *line fields*,[22] which greatly generalize vector fields and which can have singular points with *fractional* indices. Witness the examples shown in [19.14] on page 212. At my request, Dr. Liu then further assisted me by pointing out many other applications that physicists have made of Hopf's ideas. I have in turn shared his kind guidance with you, dear reader, in the *Further Reading* section at the end of the book.

In addition to the principal players above, I have received all manner of advice, support, and suggestions from colleagues and friends, near and far.

My beloved brother Guy is an anchor whose love and faith in me is too often taken for granted, but it should not be!

Stanley Nel and Paul Zeitz, my friends of 30 years, have always believed in me more than I have believed in myself, and their encouragement has meant a great deal to me over the many years it has taken to create this book, first struggling to discover the needed geometrical insights, then writing and *drawing* the book.

Douglas Hofstadter—whose *Gödel, Escher, Bach* transfixed me (and millions more) as an undergraduate—has honoured me with his support for more than 20 years. First, he has repeatedly and forcefully promoted VCA, both in print and in interviews. Second, he read and provided very valuable feedback on Needham (2014), which was ultimately incorporated into this book.

Dr. Ed Catmull—co-founder with Steve Jobs of Pixar, and later president of both Pixar and Walt Disney Animation Studios—wrote me a very flattering email about VCA back in 1999. At first I was convinced that one of my the University of San Francisco (USF) maths-pals was playing a practical joke on me, but the email was real. Ed invited me to visit the Pixar campus (still in Point Richmond back in those days), gave me a guided tour of the studios, took me out to lunch, and offered me a job! (I will leave it to the reader to assign a numerical value to my stupidity in turning that offer down.) Though my contact with Ed has been sporadic, he has been a faithful bolsterer of VCA (praising it in interviews he has done), he also wrote me a letter of recommendation for a grant, and he has been very supportive of my effort to create VDGF. I deeply appreciate Ed's encouragement over the years, and I thank him for the extremely kind words on the back of this book.

Professor Frank Morgan (whom I know by reputation only) was originally approached by Princeton University Press to provide an *anonymous* review of the manuscript of VDGF. But when he submitted his review to my editor, he *also* sent it to me directly, under his own name. I am very grateful that he chose to do this, as it now allows me to thank him publicly for his concrete suggestions and corrections. Furthermore, I was especially grateful for the tremendous boost his report gave to my *morale* at the time. Finally, I offer him my sincerest thanks for being willing to share his remarkably generous assessment of VDGF on the back of this book.

---

[20] College for Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, Hunan, P. R. China.

[21] Chen et al. (2020b).

[22] This is the modern terminology; Hopf (1956) originally called them *fields of line elements*.

I am likewise also grateful for the constructive criticism, suggestions, and corrections I have received from all of the truly anonymous reviewers—I have tried to incorporate all of their improvements, and I'm sorry I cannot thank each of them by name.

I thank The M. C. Escher Company for permission to reproduce two modifications of *Circle Limit I*: [5.11] and [5.12], the latter being an explicit mathematical transformation, carried out by John Stillwell, and used with his kind permission. Note that M. C. Escher's *Circle Limit I* is © 2020, The M. C. Escher Company, The Netherlands. All rights reserved. www.mcescher.com.

Finally, I am very grateful to Professor Henry Segerman for supplying me with the image of his *Topology Joke* [18.8], page 191, and for granting me permission to reproduce it here.

This is my second book, and it is also my last book. I therefore wish to not only thank all those who directly helped me create VDGF, listed above, but also those who influenced and supported me much earlier in my life. Some of these people were so seamlessly integrated into the fabric of my existence that they became invisible, and, shamefully, I failed to thank them properly in VCA; now is my last chance to put things right.

First amongst these is Anthony Levy, my oldest friend, from our undergraduate days together at Merton College, Oxford. Inexplicably, Anthony (or Tony, as I knew him then) believed in me long before there was any evidence to support such a belief, and that continued belief in me has buoyed me up repeatedly during periods of mathematical self-doubt over the decades. And, beyond the world of pure intellect, Anthony's sage advice and love have helped me navigate some of the most fraught episodes of my life.

Also from those undergraduate Merton days, I will always be grateful to my two physics tutors, Dr. Michael Baker (1930–2017) and Dr. Michael G. Bowler, who not only taught me a great deal of physics themselves, but who also went out of their way to arrange for me to have more advanced, individual tuition on General Relativity and on spinors, from two remarkable Fellows of Merton, Dr. Brian D. Bramson and Dr. Richard S. Ward. In particular, Dr. Bramson's enthusiasm for science was utterly contagious, and it was he who gave me my very first exposure to the (revelatory!) work of Penrose, and who pushed me to apply to undertake a DPhil in Penrose's "Relativity Group."

Moving forward to my graduate student days studying under Penrose, I wish to thank, once again, my friend George Burnett-Stuart, a fellow advisee of Penrose. George and I shared a small house together on Great Clarendon St. for several years as we carried out our doctoral work, and in the course of our endless discussions of music, physics, and mathematics, George helped me to refine both my conception of the nature of mathematics, and of what constitutes an acceptable explanation within that subject. For better or for worse, George bears great responsibility for the mathematician that I am today.

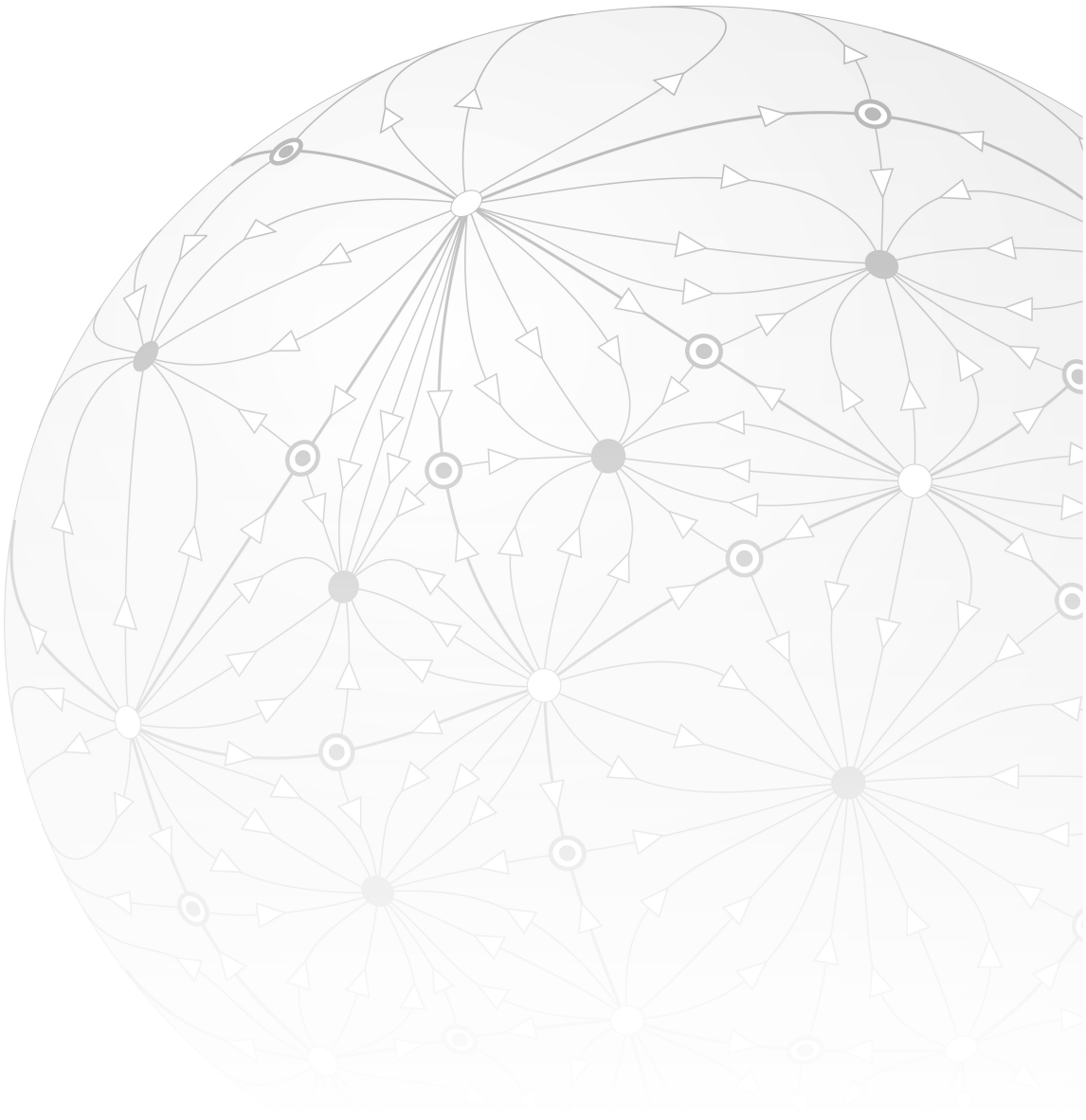
Moving forward again, to my life at USF, I am grateful to John Stillwell—who just retired—and his wonderful wife, Elaine, for 20 happy and productive years as colleagues and friends. While I have picked his brains many times during the course of writing this book, my greatest debt is to his many *writings*. Indeed, our relationship began with a "fan letter" I wrote to him in reaction to the first edition of his magnificent magnum opus, *Mathematics and Its History*. A few years later, while serving as Associate Dean for Sciences, I was able to lure John away from Australia, creating a professorship for him at USF. Both VCA and VDGF owe a great deal to John's holistic grasp of the entire expanse of mathematics, and his ability to use this perspective to give back *meaning* to mathematical ideas, all of which he has so generously shared with the world through his many wonderful books.

In 1996, I concluded the Acknowledgements of VCA with these words: "Lastly, I thank my dearest wife Mary. During the writing of this book she allowed me to pretend that science was the most important thing in life; now that the book is over, she is my daily proof that there is something even more important." Today, more than two decades later, my love for Mary has only grown more profound, but I now have two *more* daily proofs than I had before!

In 1999, Mary and I were blessed with twins: Faith and Hope have been our dazzling beacons of pride and joy ever since. I'm sorry that VDGF has hung like a dark cloud over the life of my family for almost half of my daughters's lives, and that it has robbed us of so much time together. Yet it is the love of these three souls that has given my life meaning and purpose, and has sustained me throughout the long struggle that created this work.

ACT I

The Nature of Space



Chapter 1

Euclidean and Non-Euclidean Geometry

1.1 Euclidean and Hyperbolic Geometry

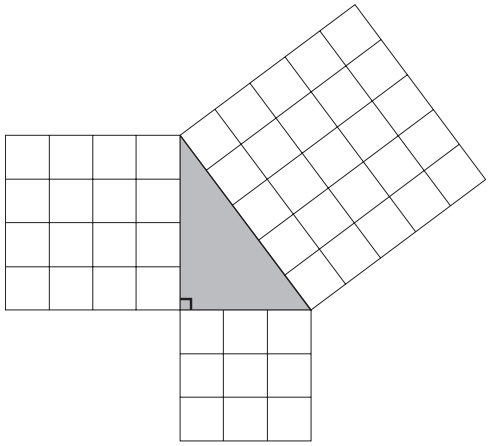
Differential Geometry is the application of calculus to the geometry of space that is *curved*. But to understand space that is curved we shall first try to understand space that is *flat*.

We inhabit a natural world pervaded by curved objects, and if a child asks us the meaning of the word "flat," we are most likely to answer in terms of the *absence* of curvature: a smooth surface *without* any bumps or hollows. Nevertheless, the very earliest mathematicians seem to have been drawn to the singular simplicity and uniformity of the flat plane, and they were rewarded with the discovery of startlingly beautiful facts about geometric figures constructed within it. With the benefit of enormous hindsight, some of these facts can be seen to *characterize* the plane's flatness.

One of the earliest and most profound such facts to be discovered was Pythagoras's Theorem. Surely the ancients must have been awed, as any sensitive person must remain today, that a seemingly unalloyed fact about *numbers*,

$$3^2 + 4^2 = 5^2,$$

in fact has *geometrical* meaning, as seen in [1.1].[1]



[1.1] *Pythagoras's Theorem: the geometry of* $3^2 + 4^2 = 5^2$.

While Pythagoras himself lived in Greece around 500 BCE, the theorem bearing his name was discovered much earlier, in various places around the world. The earliest known example of such knowledge is recorded in the Babylonian clay tablet (catalogued as "Plimpton 322") shown in [1.2], which was unearthed in what is now Iraq, and which dates from about 1800 BCE.

The tablet lists ***Pythagorean triples***:[2] integers (a, b, h) such that h is the hypotenuse of a right triangle with sides a and b, and therefore $a^2 + b^2 = h^2$. Some of these ancient examples are impressively large, and it seems clear that they did not stumble upon them, but rather possessed a mathematical process for generating solutions. For example, the fourth row of the tablet records the fact that $13500^2 + 12709^2 = 18541^2$.

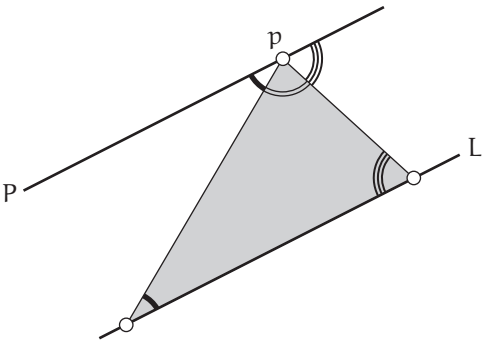
The deeper knowledge that underlay these ancient results is not known,[3] but to find the first evidence of the "modern," logical, deductive approach to mathematics we must jump 1200 years into the future of the clay tablet. Scholars believe that it was Thales of Miletus (around 600 BCE)

[1] We repeat what was said in the Prologue: *equations* are labelled with parentheses (*round* brackets)—(...), while *figures* are labelled with *square* brackets—[...].

[2] In fact the tablet only records *two* members (a, h) of the triples (a, b, h).

[3] In the seventeenth century, Fermat and Newton reconstructed and generalized a *geometrical* method of generating the general solution, due to Diophantus. See Exercise 5.

[1.2] *Plimpton 322: A clay tablet of Pythagorean triples from 1800 BCE.*



[1.3] *Euclid's Parallel Axiom:* P *is the* unique *parallel to* L *through* p*, and the angle sum of a triangle is* π.

who first pioneered the idea of deducing new results from previously established ones, the logical chain beginning at a handful of clearly articulated assumptions, or ***axioms***.

Leaping forward again, 300 years beyond Thales, we find one of the most perfect exemplars of this new approach in Euclid's *Elements*, dating from 300 BCE. This work sought to bring order, clarity, and rigour to geometry by deducing everything from just five simple axioms, the fifth and last of which dealt with parallel lines.

Defining two lines to be ***parallel*** if they do not meet, Euclid's Fifth Axiom[4] is illustrated in [1.3]:

> ***Parallel Axiom.*** *Through any point* p *not on the line* L *there exists precisely one line* P *that is parallel to* L.

But the character of this axiom was more complex and less immediate than that of the first four, and mathematicians began a long struggle to dispense with it as an assumption, instead seeking to show that it must be a logical *consequence* of the first four axioms.

This tension went unresolved for the next *2000 years*. As the centuries passed, many attempts were made to prove the Parallel Axiom, and the number and intensity of these efforts reached a crescendo in the 1700s, but all met with failure.

Yet along the way useful *equivalents* of the axiom emerged. For example: *There exist similar triangles of different sizes* (Wallis in 1663; see Stillwell (2010)). But the very first equivalent was already present in Euclid, and it is the one still taught to every school child: *the angles in a triangle add up to two right angles.* See [1.3].

The explanation of these failures only emerged around 1830. Completing a journey that had begun 4000 years earlier, Nikolai Lobachevsky and János Bolyai independently announced the

[4]Euclid did not state his axiom in this form, but it is logically equivalent.

discovery of an entirely new form of geometry (now called ***Hyperbolic Geometry***) taking place in a new kind of plane (now called the ***hyperbolic plane***). In this Geometry the first four Euclidean axioms hold, but the parallel axiom does *not*. Instead, the following is true:

> ***Hyperbolic Axiom.*** *There are at least two parallel lines through* P *that do not meet* L. (1.1)

These pioneers explored the logical consequences of this axiom, and by purely abstract reasoning were led to a host of fascinating results within a rich new geometry that was bizarrely different from that of Euclid.

[1.4] *Johann Heinrich Lambert (1728–1777).*

Many others before them, perhaps most notably Saccheri (in 1733; see Stillwell 2010) and Lambert (in 1766; see Stillwell 2010), had discovered some of these consequences of (1.1), but their aim in exploring these consequences had been to find a *contradiction*, which they believed would finally prove that Euclidean Geometry to be the One True Geometry.

Certainly Saccheri believed he had found a clear contradiction when he published "Euclid Freed of Every Flaw." But Lambert is a much more perplexing case, and he is perhaps an unsung hero in this story. His results penetrated so deeply into this new geometry that it seems impossible that he did not at times believe in the reality of what he was doing. Regardless of his motivation and beliefs[5], Lambert (shown in [1.4]) was certainly the first to discover a remarkable fact[6] about the angle sum of a triangle under axiom (1.1), and his result will be central to much that follows in Act II.

Nevertheless, Lobachevsky and Bolyai richly deserve their fame for having been the first to recognize and fully embrace the idea that they had discovered an entirely new, consistent, non-Euclidean Geometry. But what this new geometry really *meant*, and what it might be useful for, even they could not say.[7]

Remarkably and surprisingly, it was the *Differential Geometry of curved surfaces* that ultimately resolved these questions. As we shall explain, in 1868 the Italian mathematician Eugenio Beltrami finally succeeded in giving Hyperbolic Geometry a concrete interpretation, setting it upon a firm and intuitive foundation from which it has since grown and flourished. Sadly, neither Lobachevsky nor Bolyai lived to see this: they died in 1856 and 1860, respectively.

This non-Euclidean Geometry had in fact already manifested itself in various branches of mathematics throughout history, but always in disguise. Henri Poincaré (beginning around 1882) was the first not only to strip it of its camouflage, but also to recognize and exploit its power

---

[5] I thank Roger Penrose for making me see that Lambert deserves greater credit than he is usually granted. Penrose did so by means of the following analogy: "Should we not give credit to Einstein for the cosmological constant, even if he introduced it for the wrong reasons? And should we blame him for later retracting it, calling it the "greatest blunder of my life"? Or what about General Relativity itself, which Einstein seemed to become less and less convinced was the right theory (needing to be replaced by some kind of non-singular unified field theory) as time went on?" [Private communication.]

[6] If you cannot wait, it's (1.8).

[7] Lobachevsky did in fact put this geometry to use to evaluate previously unknown integrals, but (at least in hindsight) this particular application must be viewed as relatively minor.

in such diverse areas as Complex Analysis, Differential Equations, Number Theory, and Topology. Its continued vitality and centrality in the mathematics of the 20th and twenty-first centuries is demonstrated by Thurston's work on 3-manifolds, Wiles's proof of ***Fermat's Last Theorem***, and Perelman's proof of the ***Poincaré Conjecture*** (as a special case of Thurston's ***Geometrization Conjecture***), to name but three examples.

In Act II we shall describe Beltrami's breakthrough, as well as the nature of Hyperbolic Geometry, but for now we wish to explore a different, simpler kind of non-Euclidean Geometry, one that was already known to the Ancients.

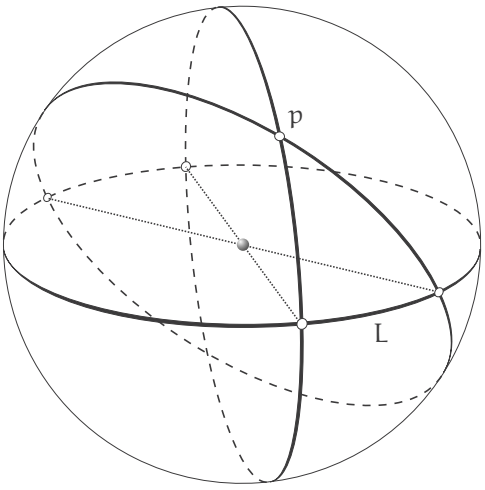
1.2 Spherical Geometry

To construct a non-Euclidean Geometry we must deny the existence of a unique parallel. The Hyperbolic Axiom assumes two or more parallels, but there is one other logical possibility—*no* parallels:

> ***Spherical Axiom.*** *There are* no *lines through* p *that are parallel to* L *:* every *line meets* L. (1.2)

Thus there are actually *two* non-Euclidean[8] geometries: spherical and hyperbolic.

As the name suggests, Spherical Geometry can be realized on the surface of a sphere—denoted $\mathbb{S}^2$ in the case of the *unit* sphere—which we may picture as the surface of the Earth. On this sphere, what should be the analogue of a "straight line" connecting two points on the surface? Answer: the shortest route between them! But if you wish to sail or fly from London to New York, for example, what *is* the shortest route?



[1.5] *The great circles of* $\mathbb{S}^2$ *intersect in pairs of antipodal points.*

The answer, already known to the ancient mariners, is that the shortest route is an arc of a ***great circle***, such as the equator, obtained by cutting the sphere with a plane passing through its centre. In [1.5] we have chosen L to be the equator, and it is clear that (1.2) is satisfied: every line through p meets L in a pair of ***antipodal*** (i.e., diametrically opposite) points.

In the plane, the shortest route is also the *straightest* route, and in fact the same is true on the sphere: in a precise sense to be discussed later, the great circle trajectory bends neither to the right nor to the left as it traverses the spherical surface.

There are other ways of constructing the great circles on the Earth that do not require thinking about planes passing through the completely inaccessible centre of the Earth. For example, on a globe you may map out your great circle journey by holding down one end of a piece of string on London and pulling the string tightly over the surface so that the other end is on New York. The taut string has

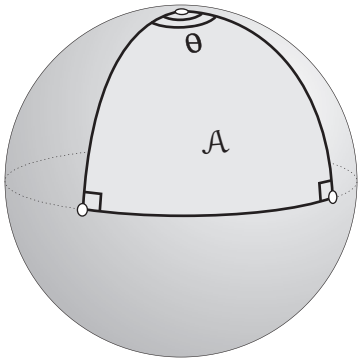
[8]Nevertheless, the reader should be aware that in modern usage "non-Euclidean Geometry" is usually synonymous with "Hyperbolic Geometry."

automatically found the shortest, straightest route—the shorter[9] of the two arcs into which the great circle through the two cities is divided by those cities.

With the analogue of straight lines now found, we can "do geometry" within this spherical surface. For example, given three points on the surface of the Earth, we can connect them together with arcs of great circles to obtain a "triangle." Figure [1.6] illustrates this in the case where one vertex is located at the north pole, and the other two are on the equator.

But if this non-Euclidean Spherical Geometry was already used by ancient mariners to navigate the oceans, and by astronomers to map the spherical night sky, what then was so shocking and new about the non-Euclidean geometry of Lobachevsky and Bolyai?

The answer is that this Spherical Geometry was merely considered to be inherited from the *Euclidean* Geometry of the 3-dimensional space in which the sphere resides. No thought was given in those times to the sphere's internal 2-dimensional geometry as representing an alternative to Euclid's plane. Not only did it violate Euclid's fifth axiom, it also violated a much more basic one (Euclid's first axiom) that we can always draw a unique straight line connecting two points, for this fails when the points are antipodal.



[1.6] *A particularly simple "triangle" on the sphere.*

On the other hand, the Hyperbolic Geometry of Lobachevsky and Bolyai was a much more serious affront to Euclidean Geometry, containing familiar lines of infinite length, yet flaunting multiple parallels, ludicrous angle sums, and many other seemingly nonsensical results. Yet the 21-year-old Bolyai was confident and exuberant in his discoveries, writing to his father, *"From nothing I have created another entirely new world."*

We end with a tale of tragedy. Bolyai's father was a friend of Gauss, and sent him what János had achieved. By this time Gauss had himself made some important discoveries in this area, but had kept them secret. In any case, János had seen further than Gauss. A kind word in public from Gauss, the most famous mathematician in the world, would have assured the young mathematician a bright future. But Nature and nurture sometimes conspire to pour extraordinary mathematical gifts into a vessel marred by very ordinary human flaws, and Gauss's reaction to Bolyai's marvellous discoveries was mean-spirited and self-serving in the extreme.

First, Gauss kept Bolyai in suspense for six months, then he replied as follows:

> Now something about the work of your son. You will probably be shocked for a moment when I begin by saying that *I cannot praise it*, but I cannot do anything else, since to praise it would be to praise myself. The whole content of the paper, the path that your son has taken, and the results to which he has been led, agree almost everywhere with my own meditations, which have occupied me in part for 30–35 years.

Gauss did however "thank" Bolyai's son for having "saved him the trouble"[10] of having to write down theorems he had known for decades.

János Bolyai never recovered from the surgical blow delivered by Gauss, and he abandoned mathematics for the rest of his life.[11]

---

[9] If the two points are antipodal, such as the north and south poles, then the two arcs are the *same* length. Furthermore, the great circle itself is no longer unique: *every* meridian is a great circle connecting the poles.

[10] Gauss had previously denigrated Abel's discovery of elliptic functions in precisely the same manner; see Stillwell (2010, p. 236).

[11] If this depresses you, turn your thoughts to the uplifting counterweight of Leonhard Euler. An intellectual volcano erupting with wildly original thoughts (some of which we shall meet later) he was also a kind and generous spirit. We cite one, parallel

1.3 The Angular Excess of a Spherical Triangle

As we have said, the parallel axiom is equivalent to the fact that the angles in a triangle sum to π. It follows that both the spherical axiom and the hyperbolic axiom must lead to geometries in which the angles do *not* sum to π. To quantify this departure from Euclidean Geometry, we introduce the ***angular excess***, defined to be the amount $\mathcal{E}$ by which the angle sum exceeds π:

$$\mathcal{E} \equiv (\text{angle sum of triangle}) - \pi.$$

For example, for the triangle shown in [1.6], $\mathcal{E} = (\theta + \frac{\pi}{2} + \frac{\pi}{2}) - \pi = \theta$.

A crucial insight now arises if we compare the triangle's angular excess with its area $\mathcal{A}$. Let the radius of the sphere be R. Since the triangle occupies a fraction $(\theta/2\pi)$ of the northern hemisphere, $\mathcal{A} = (\theta/2\pi)\, 2\pi R^2 = \theta\, R^2$. Thus,

$$\mathcal{E} = \frac{1}{R^2}\,\mathcal{A}. \tag{1.3}$$

[1.7] *Thomas Harriot (1560–1621).*

In 1603 the English mathematician Thomas Harriot (see [1.7]) made the remarkable discovery[12] that this relationship holds for *any* triangle Δ on the sphere; see [1.8a]. Harriot's elementary but ingenious argument[13] goes as follows.

Prolonging the great-circle sides of Δ divides the surface of the sphere into eight triangles, the four triangles labelled Δ, Δ_α, Δ_β, Δ_γ each being paired with a congruent antipodal triangle. This is clearer in [1.8b]. Since the area of the sphere is $4\pi R^2$, we deduce that

$$\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\alpha) + \mathcal{A}(\Delta_\beta) + \mathcal{A}(\Delta_\gamma) = 2\pi R^2. \tag{1.4}$$

On the other hand, it is clear in [1.8b] that Δ and Δ_α together form a wedge whose area is a fraction $(\alpha/2\pi)$ of the area of the sphere:

$$\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\alpha) \;=\; 2\alpha R^2.$$

Similarly,

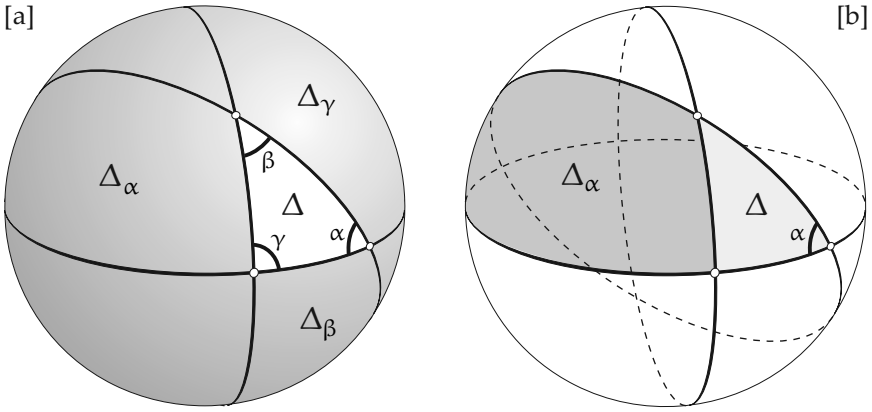
$$\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\beta) \;=\; 2\beta R^2,$$
$$\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\gamma) \;=\; 2\gamma R^2.$$

---

example. When the then-unknown 19-year-old Lagrange wrote to him with overlapping discoveries in the calculus of variations, Euler wrote back: "... I deduced this myself. However, I decided to conceal this until you publish your results, since in no way do I want to take away from you any of the glory that you deserve." See Gindikin (2007, p. 216). Incidentally, Euler also personally intervened to rescue Lambert's career!

[12] This discovery is most often attributed to Girard, who rediscovered it about 25 years later.

[13] This argument was later rediscovered by Euler in 1781.



[1.8] *Harriot's Theorem (1603):* $\mathcal{E}(\Delta) = \mathcal{A}(\Delta)/R^2$.

Adding these last three equations, we find that

$$3\mathcal{A}(\Delta) + \mathcal{A}(\Delta_\alpha) + \mathcal{A}(\Delta_\beta) + \mathcal{A}(\Delta_\gamma) = 2(\alpha+\beta+\gamma)R^2. \tag{1.5}$$

Finally, subtracting (1.4) from (1.5), we find that

$$\mathcal{A}(\Delta) = R^2(\alpha+\beta+\gamma-\pi) = R^2\,\mathcal{E}(\Delta),$$

thereby proving (1.3).

1.4 Intrinsic and Extrinsic Geometry of Curved Surfaces

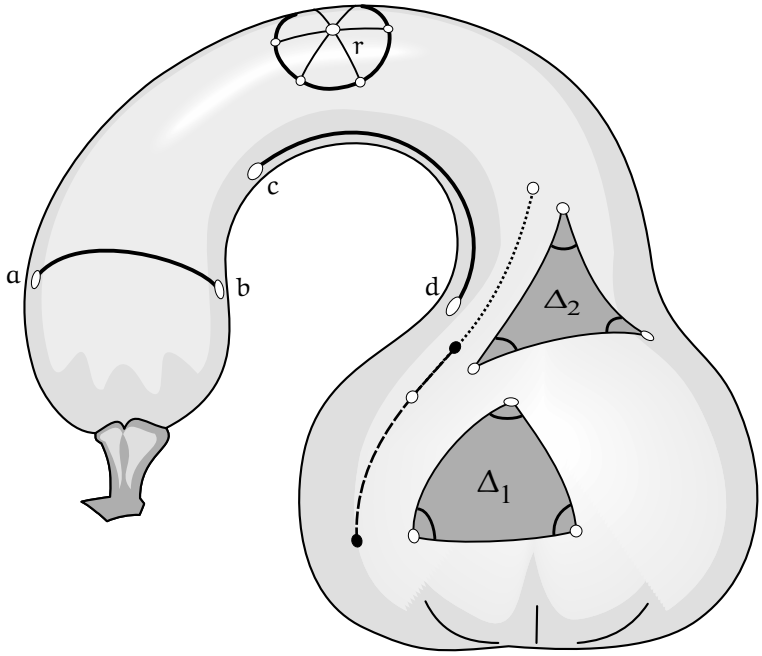
The mathematics associated with this stretched-string construction of a "straight line" will be explored in depth later in the book. For now we merely observe that the construction can be applied equally well to a *non*spherical surface, such as the crookneck squash shown in [1.9].

Just as on the sphere, we stretch a string over the surface, thereby finding the shortest, straightest route between two points, such as a and b. Provided that the string can slide around on the surface easily, the tension in the string ensures that the resulting path is as short as possible. Note that in the case of cd, we must imagine that the string runs over the *inside* of the surface.

In order to deal with all possible pairs of points in a uniform way, it is therefore best to imagine the surface as made up of two thinly separated layers, with the string trapped between them. On the other hand, this is only useful for thought experiments, not actual experiments. We shall overcome this obstacle shortly by providing a *practical* method of constructing these straightest curves on the surface of a physical object, even if the patch of surface bends the wrong way for a string to be stretched tightly over the outside of the object.

These shortest paths on a curved surface are the equivalent of straight lines in the plane, and they will play a crucial role throughout this book—they are called ***geodesics***. Thus, to use this new word, we may say that geodesics in the plane are straight lines, and geodesics on the sphere are great circles.

But even on the sphere the *length-minimizing* definition of "geodesic" is provisional, because we see that nonantipodal points are connected by *two* arcs of the great circle passing through them: the short one (which *is* the shortest route) and the long one. Yet the long arc is every bit as much a geodesic as the short one. There is the additional complication on the sphere that antipodal points



[1.9] *The* **intrinsic geometry** *of the surface of a crookneck squash:* **geodesics** *are the equivalents of straight lines, and triangles formed out of them may possess an angular excess of either sign, depending on how the surface bends:* $\mathcal{E}(\Delta_1) > 0$ *and* $\mathcal{E}(\Delta_2) < 0$.

are connected by *multiple* geodesics, and this *nonuniqueness* occurs on more general surfaces, too. What *is* true is that any two points that are *sufficiently close together* can be joined by a unique geodesic segment that is the shortest route between them.

Just as a line segment in the plane can be extended indefinitely in either direction by laying down overlapping segments, so too can a geodesic segment be extended on a curved surface, and this extension is unique. For example, in [1.9] we have extended the dashed geodesic segment connecting the black dots, by laying down the overlapping dotted segment between the white points.

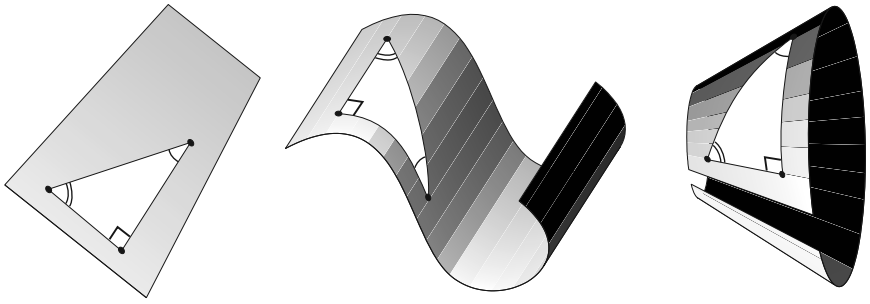
Because of the subtleties associated with the length-minimizing characterization of geodesics, before long we will provide an alternative, purely *local* characterization of geodesics, based on their straightness.

With these caveats in place, it is now clear how we should define distance within a surface such as [1.9]: the distance between two sufficiently close points a and b is the length of the geodesic segment connecting them.

Figure [1.9] shows how we may then define, for example, a "circle of radius r and centre c" as the locus of points at distance r from c. To construct this ***geodesic circle*** we may take a piece of string of length r, hold one end fixed at c, then (keeping the string taut) drag the other end round on the surface. But just as the angles in a triangle no longer sum to π, so now the circumference of a circle no longer is equal to $2\pi r$. In fact you should be able to convince yourself that for the illustrated circle the circumference is *less* than $2\pi r$.

Given three points on the surface, we may join them with geodesics to form a ***geodesic triangle***; [1.9] shows two such triangles, Δ_1 and Δ_2:

- Looking at the angles in Δ_1, it seems clear that they sum to *more* than π, so $\mathcal{E}(\Delta_1) > 0$, like a triangle in Spherical Geometry.



[1.10] *Bending a piece of paper changes the extrinsic geometry, but not the intrinsic geometry.*

- On the other hand, it is equally clear that the angles of Δ_2 sum to *less* than π: $\mathcal{E}(\Delta_2) < 0$, and (as we shall explain) this opposite behaviour is in fact exhibited by triangles in *Hyperbolic Geometry*. Note also that if we construct a circle in this saddle-shaped part of the surface, the circumference is now *greater* than $2\pi r$.

The concept of a geodesic belongs to the so-called ***intrinsic geometry*** of the surface—a fundamentally new view of geometry, introduced by Gauss (1827). It means the geometry that is knowable to tiny, ant-like, intelligent (but 2-dimensional!) creatures living *within* the surface. As we have discussed, these creatures can, for example, define a geodesic "straight line" connecting two nearby points as the shortest route within their world (the surface) connecting the two points. From there they can go on to define triangles, and so on. Defined in this way, it is clear that the intrinsic geometry is unaltered when the surface is bent (as a piece of paper can be) into quite different shapes in space, as long as distances *within* the surface are not stretched or distorted in any way. To the ant-like creatures within the surface, such changes are utterly undetectable.

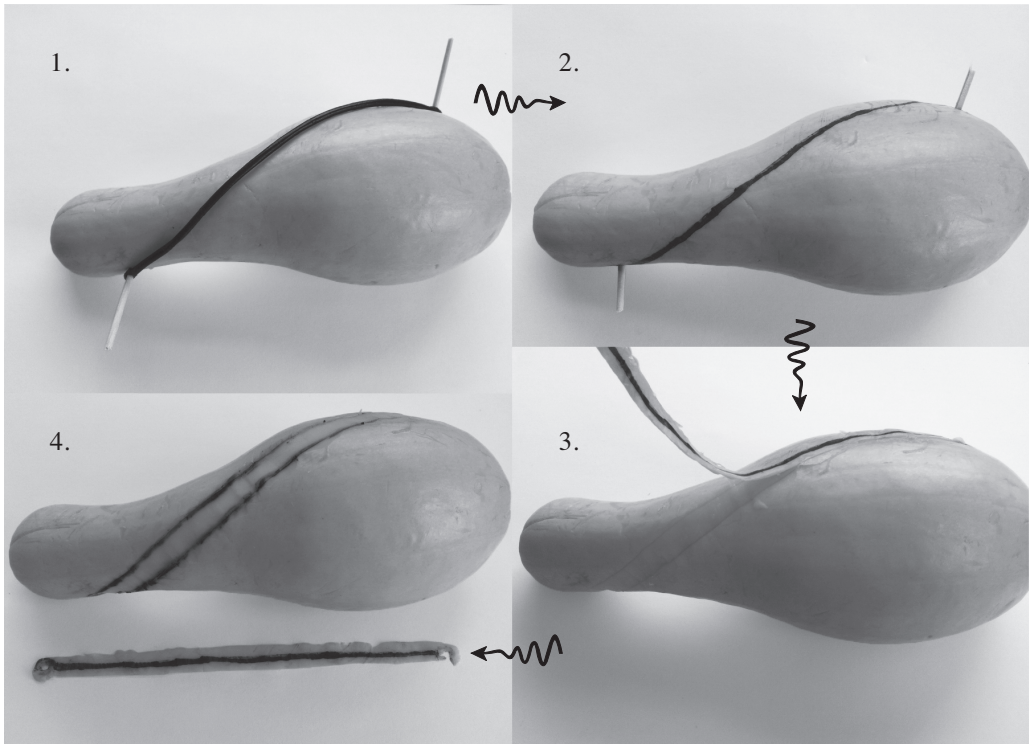
Under such a bending, the so-called ***extrinsic geometry*** (how the surface sits in space) most certainly does change. See [1.10]. On the left is a flat piece of paper on which we have drawn a triangle Δ with angles $(\pi/2)$, $(\pi/6)$, and $(\pi/3)$. Of course $\mathcal{E}(\Delta) = 0$. Clearly we can bend such a flat piece of paper into either of the two (extrinsically) curved surfaces on the right.[14] However, *intrinsically* these surfaces have undergone no change at all—they are both as flat as a pancake! The illustrated triangles on these surfaces (into which Δ is carried by our stretch-free bending of the paper) are identical to the ones that intelligent ants would construct using geodesics, and in both cases $\mathcal{E} = 0$: geometry on these surfaces is Euclidean.

Even if we take a patch of a surface that is intrinsically *curved*, so that a triangle within it has $\mathcal{E} \neq 0$, it too can generally be bent somewhat without stretching or tearing it, thereby altering its extrinsic geometry while leaving its intrinsic geometry unaltered. For example, cut a ping pong ball in half and gently squeeze the rim of one of the hemispheres, distorting that circular rim into an oval (but not an oval lying in a single plane).

1.5 Constructing Geodesics via Their Straightness

We have already alluded to the fact that geodesics on a surface have at least *two* characteristics in common with lines in the plane: (1) they provide the *shortest* route between two points that are not too far apart *and* (2) they provide the "straightest" route between these points. In this section we seek to clarify what we mean by "straightness," leading to a very simple and *practical* method of constructing geodesics on a physical surface.

[14] But note that we must first trim the edges of the rectangle to bend it into the shape on the far right.



[1.11] *On the curved surface of a fruit or vegetable, peel a narrow strip surrounding a geodesic, then lay it flat on the table. You will obtain a* straight line *in the plane!*

Most texts on Differential Geometry pay scant attention to such practical matters, and it is perhaps for this reason that the construction we shall describe is surprisingly little known in the literature.[15] In sharp contrast, in this book we *urge* you to explore the ideas by all means possible: theoretical contemplation, drawing, computer experiments, and (especially!) physical experiments with actual surfaces. Your local fruit and vegetable shop can supply your laboratory with many interesting shapes, such as the yellow summer squash shown in [1.11].

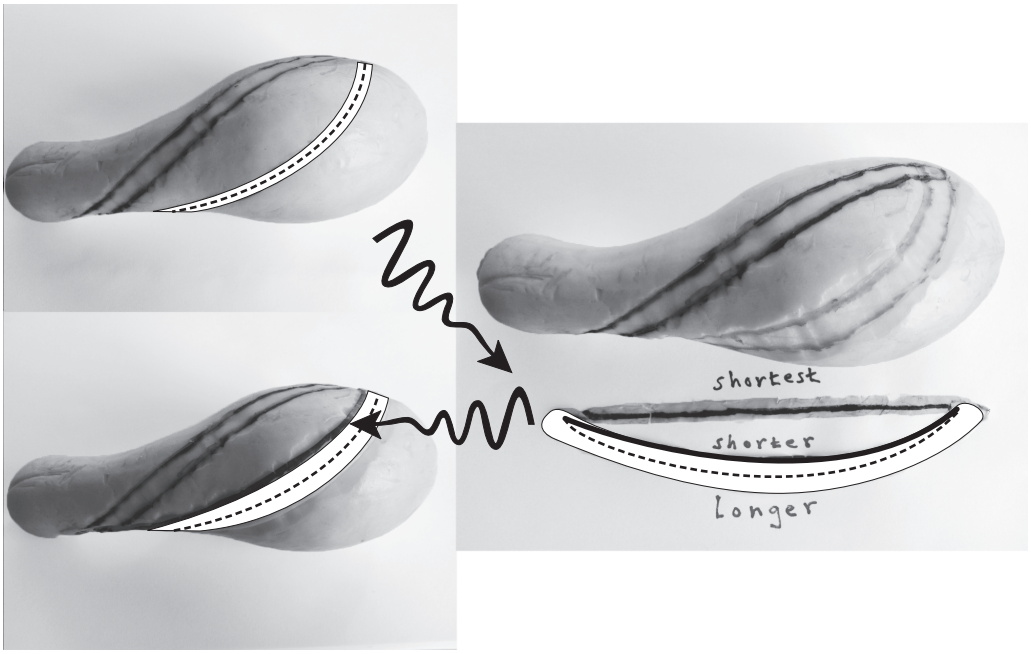
We can now use this vegetable to reveal the hidden straightness of geodesics via an experiment that we hope you will repeat for yourself:

1. On a fruit or vegetable, construct a geodesic by stretching a string over its curved surface.
2. Use a pen to trace the path of the string, then remove the string.
3. Make shallow incisions on either side of (and close to) the inked path, then use a vegetable peeler or small knife to remove the narrow strip of peel between the two cuts.
4. Lay the strip of peel flat on the table, and witness the marvellous fact that the geodesic within the peeled strip has become a *straight line* in the plane!

But why?!

To understand this, first let us be clear that although the strip is free to bend in the direction perpendicular to the surface (i.e., perpendicular to itself), it is *rigid* if we try to bend it sideways, tangent to the surface. Now let us employ proof by contradiction, and imagine what would happen if such a peeled geodesic did *not* yield a straight line when laid flat on the table. It is both a

[15]One of the rare exceptions is Henderson (1998), which we strongly recommend to you; for more details, see the *Further Reading* section at the end of this book.



[1.12] *Suppose that the illustrated dotted path is a geodesic such that a narrow (white) strip surrounding it does* not *become a straight line when laid flat in the plane. But in that case we can shrink the dotted path in the plane (towards the shortest, straight-line route in the plane) thereby producing the solid path. But if we then reattach the strip to the surface, this solid path is still shorter than the original dotted path, which was supposed to be the shortest path within the surface—a contradiction!*

drawback and an advantage of conducting such physical experiments that they will simply not *permit* us to construct something that is impossible, as is required in our desired mathematical proof by contradiction. Nevertheless, let us *suppose* that there exists a geodesic path, such as the dotted one shown on top left of [1.12], that when peeled and laid flat on the table (on the right) does *not* become a straight line.

The shortest route between the ends of this dotted (nonstraight) plane curve is the straight line connecting them. (As illustrated, this is the path of the *true* geodesic we already found using the string—but pretend you don't know that for now!) Thus we may shorten the dotted curve by deforming it slightly towards this straight, shortest route, yielding the solid path along the edge of the peeled strip. Therefore, after reattaching the strip to the surface (bottom left) the solid curve provides a shorter route over the surface than the dotted one, which we had supposed to be the shortest: a contradiction! Thus we have proved our previous assertion:

> *If a narrow strip surrounding a segment* G *of a geodesic is cut out of a surface and laid flat in the plane, then* G *becomes a segment of a* straight line. (1.6)

We are now very close to the promised simple and practical construction of geodesics. Look again at step 3 of [1.11], where we peeled off the strip of surface. But imagine now that we are *reattaching* the strip to the surface, instead. Ignore the history of how we got to this point: what are we actually doing right now in this reattachment process? We have picked up a narrow straight strip (of three-dimensional peel—but mathematically idealized as a two-dimensional strip) and we have unrolled it back onto the surface into the shallow channel from which we cut it. But here

is the crucial observation: this shallow channel need not exist—the *surface* decides where the strip must go as we unroll it!

Thus, as a kind of time-reversed converse of (1.6), we obtain a remarkably simple and practical method[16] of constructing geodesics on a physical surface:

> *To construct a geodesic on a surface, emanating from a point* p *in direction* **v**, *stick one end of a length of narrow sticky tape down at* p *and unroll it onto the surface, starting in the direction* **v**. (1.7)

(Note, however, that this does *not* provide a construction of the geodesic connecting p to a specified *target* point q.)

If this construction seems too simple to be true, please try it on any curved surface you have to hand. You can check that the sticky tape[17] is indeed tracing out a geodesic by stretching a string over the surface between two points on the tape: the string will follow the same path as the tape. But note that, as a promised bonus, this new tape construction works on *any* part of a surface, even where the surface is concave towards you, so that the stretched-string construction breaks down.

Of course all of this is a concrete manifestation of a mathematical idealization. A totally flat narrow strip of tape of nonzero width *cannot*[18] be made to fit perfectly on a genuinely curved surface, but its centre line *can* be made to rest on the surface, while the rest of the tape is tangent to the surface.

1.6 The Nature of Space

Let us return to the history of the discovery of non-Euclidean Geometry, and take our first look at how these two new geometries differ from Euclid's.

As we have said, Euclidean Geometry, is characterized by the vanishing of $\mathcal{E}(\Delta)$. Note that, unlike the original formulation of the parallel axiom, *this statement can be checked against experiment*: construct a triangle, measure its angles, and see if they add up to π. Gauss may have been the first person to ever conceive of the possibility that physical space might not be Euclidean, and he even attempted the above experiment, using three mountain tops as the vertices of his triangle, and using light rays for its edges.

Within the accuracy permitted by his equipment, he found $\mathcal{E} = 0$. Quite correctly, Gauss did not conclude that physical space is definitely Euclidean in structure, but rather that if it is *not* Euclidean then its deviation from Euclidean Geometry is extremely small. But he did go so far as to say (see Rosenfeld 1988, p. 215) that he wished that this non-Euclidean Geometry might apply to the real world. In Act IV we shall see that this was a prophetic statement.

---

[16] This important fact is surprisingly hard to find in the literature. After we (re)discovered it, more than 30 years ago, we began searching, and the earliest mention of the underlying idea we could find at that time was in Aleksandrov (1969, p. 99), albeit in a less practical form: he imagined pressing a flexible metal ruler down onto the surface. Later, the basic idea also appeared in Koenderink (1990), Casey (1996), and Henderson (1998). *However*, we have since learned that the essential idea (though not in our current, *practical* form) goes all the way back to Levi-Civita, more than a century ago! See the footnote on page 236.

[17] We recommend using masking tape (aka painter's tape) because it comes in bright colours, and once a strip has been created, it can be detached and reattached repeatedly, with ease. A simple way to create narrow strips (from the usually wide roll of tape) is to stick a length of tape down onto a kitchen cutting board, then use a sharp knife to cut down its length, creating strips as narrow as you please.

[18] This is a consequence of a fundamental theorem we shall meet later, called the *Theorema Egregium*.

Although Gauss had bragged to friends that he had anticipated the Hyperbolic Geometry of Lobachevsky and Bolyai by decades, even he had unknowingly been scooped on some of its central results.

In 1766 (eleven years before Gauss was born) Lambert rediscovered Harriot's result on the sphere and then broke totally new ground in pursuing the analogous consequences of the Hyperbolic Axiom (1.1). First, he found that a triangle in Hyperbolic Geometry (if such a thing even existed) would behave *oppositely* to one in Spherical Geometry:

- In Spherical Geometry the angle sum of a triangle is greater than π: $\mathcal{E} > 0$.
- In Hyperbolic Geometry the angle sum of a triangle is less than π: $\mathcal{E} < 0$.

Thus a hyperbolic triangle behaves like a triangle drawn on a saddle-shaped piece of surface, like Δ_2 in [1.9]. Later we shall see that this is no accident.

Furthermore, Lambert discovered the crucial fact that $\mathcal{E}(\Delta)$ is again simply proportional to $\mathcal{A}(\Delta)$:

In both Spherical and Hyperbolic Geometry,

$$\mathcal{E}(\Delta) = \mathcal{K}\,\mathcal{A}(\Delta), \tag{1.8}$$

where $\mathcal{K}$ is a constant that is positive in Spherical Geometry, and negative in Hyperbolic Geometry.

Several interesting observations can be made in connection with this result:

- Although there are no qualitative differences between them, there are nevertheless *infinitely many* different Spherical Geometries, depending on the value of the positive constant $\mathcal{K}$. Likewise, each negative value of $\mathcal{K}$ yields a different Hyperbolic Geometry.
- Since the angle sum of a triangle cannot be negative, $\mathcal{E} \geqslant -\pi$. Thus in Hyperbolic Geometry ($\mathcal{K} < 0$) we have the strange and surprising result that *no triangle can have an area greater than* $|\pi/\mathcal{K}|$.
- From (1.8) we deduce that two triangles of different size cannot have the same angles. In other words, in non-Euclidean Geometry, *similar triangles do not exist!* (This accords with Wallis's 1663 discovery that the existence of similar triangles is equivalent to the Parallel Axiom.)
- Closely related to the previous point is the fact that in non-Euclidean Geometry there exists an *absolute unit of length*. (Gauss himself found it to be an exciting possibility that this purely mathematical fact might be realized in the physical world.) For example, in Spherical Geometry we could define this absolute unit of length to be the side of *the* equilateral triangle having, for instance, angle sum 1.01π. Similarly, in Hyperbolic Geometry we could define it to be the side of *the* equilateral triangle having angle sum 0.99π.
- A somewhat more natural way of defining the absolute unit of length is in terms of the constant $\mathcal{K}$. Since the radian measure of angle is defined as a ratio of lengths, $\mathcal{E}$ is a pure number. On the other hand, the area $\mathcal{A}$ has units of (length)^2. It follows that $\mathcal{K}$ must have units of $1/(\text{length})^2$, and so there exists a length R such that $\mathcal{K}$ can be written as follows: $\mathcal{K} = +(1/\mathsf{R}^2)$ in Spherical Geometry; $\mathcal{K} = -(1/\mathsf{R}^2)$ in Hyperbolic Geometry. Of course in Spherical Geometry we already know that the length R occurring in the formula $\mathcal{K} = +(1/\mathsf{R}^2)$ is simply the

radius of the sphere. Later we will see that this length R occurring in the formula $\mathcal{K} = -(1/R^2)$ can be given an equally intuitive and *concrete* interpretation in Hyperbolic Geometry.

- The smaller the triangle, the harder it is to distinguish it from a Euclidean triangle: only when the linear dimensions are a significant fraction of R will the differences become discernable. For example, we humans are small compared to the radius of the Earth, so if we find ourselves in a boat in the middle of a lake, its surface appears to be a Euclidean plane, whereas in reality it is part of a sphere. This Euclidean illusion for small figures is the reason that Gauss chose the largest possible triangle to conduct his light-ray experiment, thereby increasing his chances of detecting any small curvature that might be present in the space through which the light rays travelled.

Chapter 2

Gaussian Curvature

2.1 Introduction

The proportionality constant,

$$\mathcal{K} = +\frac{1}{R^2},$$

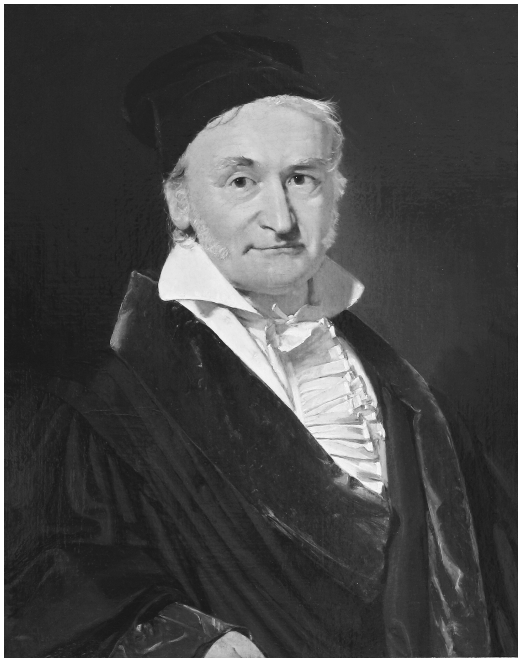
that enters into Spherical Geometry via Harriot's result (1.3), is called the ***Gaussian curvature***[1] of the sphere. The smaller the radius R, the more tightly curved is the surface of the sphere, and the greater the value of the Gaussian curvature $\mathcal{K}$.

Likewise, in Hyperbolic Geometry the negative constant

$$\mathcal{K} = -\frac{1}{R^2}$$

occurring in (1.8) is *again* called the Gaussian curvature, for reasons we shall explain shortly.

This intrinsic[2] concept $\mathcal{K}$ was announced by Gauss (after a decade of private investigation) in his revolutionary "General Investigations of Curved Surfaces,"[3] published in 1827.



[2.1] *Carl Friedrich Gauss (1777–1855).*

As we now explain, Gauss introduced this concept to measure the *curvature at each point* of a general, irregular surface such as that depicted in [1.9]. This one idea of curvature will dominate all that is to come. According to Harriot's and Lambert's results (1.8),

$$\mathcal{K} = \frac{\mathcal{E}(\Delta)}{\mathcal{A}(\Delta)} = \textit{angular excess per unit area}.$$

In both Spherical and Hyperbolic Geometry this interpretation holds for a triangle Δ of any size and any location. But on a more general surface such as in [1.9] this definition makes no sense, for even the *sign* of $\mathcal{E}$ varies between triangles, such as Δ_1 and Δ_2, that reside in different parts of the surface.

---

[1] Also called the ***Gauss curvature***, ***intrinsic curvature***, ***total curvature***, or just plain ***curvature***.

[2] As we shall discuss later, Olinde Rodrigues had arrived at and published the same concept as early as 1815, but from an *extrinsic* point of view. Gauss was not aware that he had been anticipated in this way.

[3] Gauss (1827).

[2.2] *The Gaussian curvature $\mathcal{K}(\mathrm{p})$ at a point is the angular excess per unit area as a geodesic triangle shrinks to that point. In this example, $\mathcal{K}(\mathrm{p}) > 0$ and $\mathcal{K}(\mathrm{q}) < 0$.*

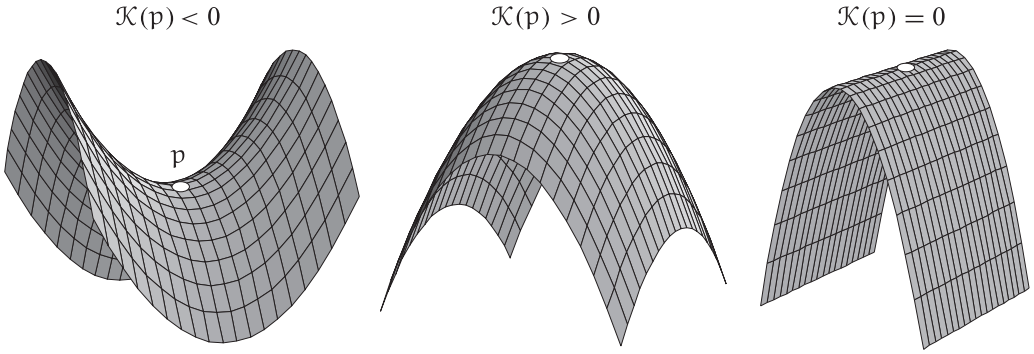
To define the Gaussian curvature *at* a point p on such a surface we now imagine a small geodesic triangle Δ_{p} containing p, and then allow the triangle to shrink down towards p.

Using the construction of geodesics discovered in the previous section, [2.2] depicts such a sequence of shrinking triangles, converging towards a point on the surface of an inflatable swimming pool ring, the mathematical name for which is a ***torus***. We now define the Gaussian curvature $\mathcal{K}(\mathrm{p})$ at p to be the limit as this triangle shrinks down towards p:

$$\mathcal{K}(\mathrm{p}) = \lim_{\Delta_{\mathrm{p}} \to \mathrm{p}} \frac{\mathcal{E}(\Delta_{\mathrm{p}})}{\mathcal{A}(\Delta_{\mathrm{p}})} = \textit{angular excess per unit area at } \mathrm{p}. \tag{2.1}$$

At this stage it is *not* meant to be obvious to you that this limit exists, independently of the shape of the triangle and the precise manner in which it shrinks; this will be proved later. As our drama unfolds we shall discover several other ways[4] of interpreting the Gaussian curvature and of calculating its value in concrete cases.

[4]For a mathematical concept to be truly fundamental it must lie at the intersection of different branches of mathematics. Thus it is to be expected that each of these branches will provide a seemingly distinct yet equally natural way of looking at one and the same concept.



[2.3] *The Gaussian curvature $\mathcal{K}$ is the local angular excess per unit area: its sign is negative if the surface looks like a saddle, positive if it's like a hill, and it vanishes if it's like a curled piece of paper.*

The definition in (2.1) extends beyond triangles. If we replace Δ_p with a small n-gon then (see Ex. 10),

$$\mathcal{E}(n\text{-gon}) \equiv [\text{angle sum}] - (n-2)\pi, \tag{2.2}$$

and the interpretation of curvature in (2.1) as *angular excess per unit area* applies without change.

Inspection of the inflatable pool ring in [2.2] should make it clear that $\mathcal{K}(p) > 0$ at every point p on the outer half, where the immediate neighbourhood of p resembles a hill, whereas $\mathcal{K}(q) < 0$ at every point q on the inner half, where the immediate neighbourhood of q resembles a saddle. Figure [2.3] summarizes this.

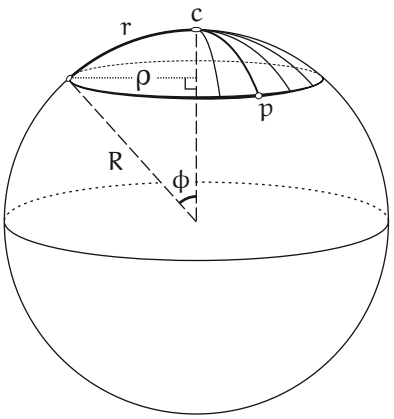
2.2 The Circumference and Area of a Circle

But why is $\mathcal{K}(p)$ so important? Yes, clearly it controls small triangles to some extent, but there is so much more to geometry than just triangles! The answer is that while we may have chosen to define $\mathcal{K}(p)$ (for the moment) in terms of small triangles, we will gradually discover that the curvature has an iron grip over *every* aspect of geometry within the surface. Let us give just two examples for now.

In [1.9] we indicated how a "circle of radius r" centred at c could be defined by taking the end p of a geodesic segment cp of fixed length r and swinging it fully around c. Let us calculate the circumference $C(r)$ of such a circle constructed on the sphere of radius R.

Referring to [2.4], we see that

$$\rho = R\sin\phi \quad \text{and} \quad \phi = \frac{r}{R} \quad \Longrightarrow \quad C(r) = 2\pi R\sin(r/R). \tag{2.3}$$



[2.4] *A circle of radius r on a sphere of radius R has circumference $C(r)$, given by $C(r) = 2\pi R\sin(r/R)$.*

Just as the curvature governs the departure of the angle sum of a triangle from the Euclidean prediction of π, so too does it govern the departure of $C(r)$ from the Euclidean prediction of $2\pi r$. To see this, recall the power series for sine:

$$\sin\phi = \phi - \frac{1}{3!}\phi^3 + \frac{1}{5!}\phi^5 + \cdots .$$

Thus, as ϕ vanishes,

$$\phi - \sin\phi \asymp \frac{1}{6}\phi^3.$$

(We remind the reader that here, $\asymp$ denotes Newton's concept of ***ultimate equality***, as introduced in the Prologue.) It follows from (2.3) that as r shrinks to zero,

$$2\pi r - C(r) = 2\pi R[(r/R) - \sin(r/R)] \asymp \frac{\pi r^3}{3R^2}.$$

In other words, the inhabitants of $\mathbb{S}^2$ can now determine the curvature of their world by examining the circumference of a small circle, just as easily as they previously could by examining the angles of a small triangle:

$$\boxed{\mathcal{K} \asymp \frac{3}{\pi}\left[\frac{2\pi r - C(r)}{r^3}\right].} \tag{2.4}$$

Remarkably, as we will be able to show much later, in Act IV, this formula continues to correctly measure the Gaussian curvature on a *general* surface! (Note that the power of r in the denominator could have been anticipated: we know that $\mathcal{K}$ has dimensions of $1/(\text{length})^2$, and circumference is a length, so we require $(\text{length})^3$ in the denominator.)

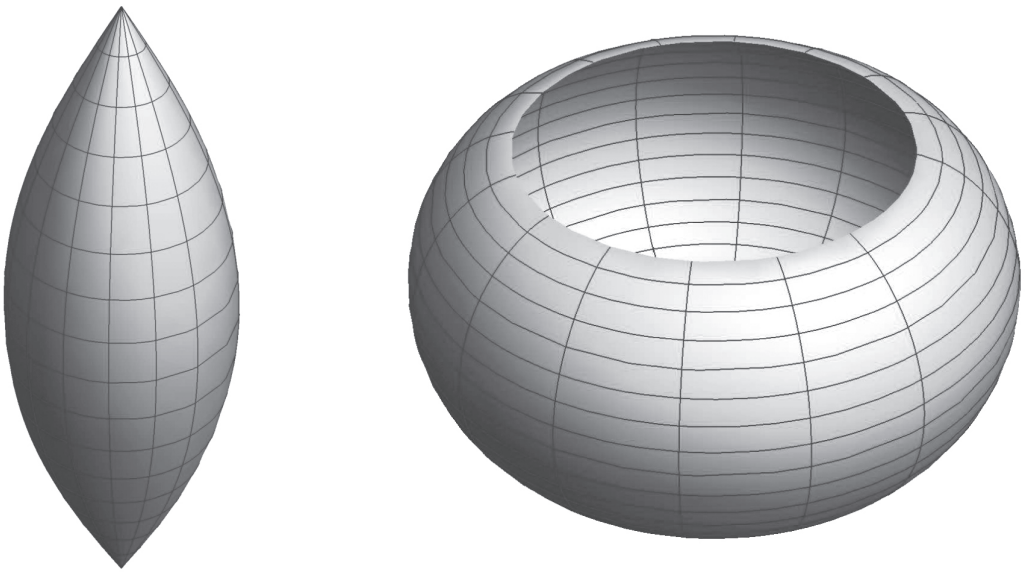
Continuing with this example, let us instead examine the *area* $\mathcal{A}(r)$ of the polar cap bounded by this circle. Again it is the curvature that governs how the area departs from the Euclidean prediction of πr^2. With the assistance of the formula for the polar cap (see Ex. 10, p. 85) it is not hard to verify [exercise] that, in fact,

$$\boxed{\mathcal{K} \asymp \frac{12}{\pi}\left[\frac{\pi r^2 - \mathcal{A}(r)}{r^4}\right].} \tag{2.5}$$

And *again* this formula turns out to be universal! (Again, the same reasoning as above explains the fourth power in the denominator.)

While we are not yet in a position to prove the universality of (2.4) and (2.5), we can at least see that they do indeed yield the correct *sign* at each point of a variably curved surface, such as that shown in [1.9]. For if the immediate vicinity of a point on such a surface is positively curved, then it is hill-shaped near that point (as it everywhere in the region of [1.9] containing Δ_1). Thus both the circumference and area of a small circle centred there will indeed be squeezed by the curvature and be *less* than they would have been in a flat Euclidean plane. Thus, both the preceding formulas yield $\mathcal{K} > 0$, as they should.

On the other hand, if the surface is saddle-shaped near the point, the opposite happens. Recall that we pointed out in [1.9] that a circle drawn in the saddle-shaped part of the surface (where Δ_2 is located) will have $C(r) > 2\pi r$. To grasp this, stand up and hold one arm out at right angles to your body. If you spin around on your heels, the tip of your hand will trace out a horizontal circle. Now repeat this pirouette, but this time wave your arm up and down as you turn; clearly the tip of hand has travelled *further* than before. But this waving up and down is just what happens when we trace out a circle on a saddle-shaped surface, and therefore both of the preceding formulas yield $\mathcal{K} < 0$, as they should.



[2.5] *Nonspherical surfaces of revolution exist that possess constant positive curvature, but these necessarily have either spikes or edges.*

We have said that curvature has an "iron grip" on geometry, but just how absolute is this control? For example, if we know that a patch of surface has constant positive curvature $\mathcal{K} = (1/R^2)$, must it in fact be a portion of a sphere of radius R? Well, take a ping pong ball and cut it in half—now flex one of the hemispheres slightly. Clearly we have obtained a new *non*spherical patch of surface, but since we have not stretched distances within the surface, geodesics and angles are unchanged, and therefore according to the definition (2.3) the curvature $\mathcal{K}$ has not changed. Thus we certainly can obtain at least patches of surface of constant curvature that are not extrinsically spherical, although they all have identical intrinsic geometry.

Figure [2.5] illustrates that even if we restrict attention just to surfaces of revolution, the sphere is not the only one of constant positive curvature. In fact there is an entire family of such surfaces, with the sphere representing the boundary case between the two illustrated types (see Ex. 22). Though they hardly look like spheres, an intelligent ant living on either of these surfaces would never know that she wasn't living on a sphere. Well, that's almost true: eventually she might discover sharp creases or spikes at which the surface is not smooth, or else she might run into an edge where the surface ends. In 1899 Heinrich Liebmann proved[5] that if a surface of constant positive curvature does not suffer from these defects then it can *only* be a sphere.

Ignoring such superficial extrinsic differences, can two surfaces have *constant* positive curvature $\mathcal{K} = (1/R^2)$ and yet have genuinely different *intrinsic* geometries? More explicitly, if our intelligent ant were suddenly transported from one surface to the other, could she devise an experiment to discover that her world had changed? In 1839 Minding (one of Gauss's few students) discovered the answer: "no!" In other words, Minding found[6] that if two surfaces have *constant* positive curvature $\mathcal{K} = (1/R^2)$ then their intrinsic geometries are locally *identical* to that of the sphere of radius R.

We have discussed the fact that the inner rim of the pool ring in [2.2] has negative curvature, but it is not *constant* negative curvature. Indeed, if C is the circle of contact between the ring and the ground, separating the inner and outer halves, then it's clear that the negative curvature $\mathcal{K}(q)$

[5]The proof will have to wait till Section 38.11.

[6]The proof will have to wait till Act IV (Exercise 7, p. 336).

must die away to zero as q approaches C, switching to positive if q crosses over C to the outer half. (This is investigated in Act II: Ex. 23, page 89.)

[2.6] *The* **pseudosphere** *of base radius* R *has constant negative curvature* $\mathcal{K} = -(1/\mathrm{R}^2)$.

In fact there do exist surfaces that have constant negative curvature. Eugenio Beltrami (whom we shall meet properly shortly) called all such surfaces ***pseudospherical***, in honour of the simplest exemplar, the ***pseudosphere***,[7] shown in [2.6]. (The precise construction of this surface will be explained in Act II, but it is the surface of revolution generated by a curve called the ***tractrix***, which was first investigated by Newton, in 1676.) If R is the radius of the circular base of the pseudosphere then (as we shall prove later) the constant negative value of the curvature over the entire surface is $\mathcal{K} = -(1/\mathrm{R}^2)$.

While the name of this surface is too established to be changed, it is perhaps unfortunate. As you see, it is certainly not closed like the sphere. In fact we shall prove later in the book that no closed pseudospherical surfaces can exist. Further, we see that while the pseudosphere extends upwards indefinitely, it has a circular edge at its base. It turns out to be impossible to extend the pseudosphere beyond this edge while maintaining its constant negative curvature. This is no accident: in 1901 David Hilbert proved that if a surface of constant negative curvature is to be embedded within ordinary Euclidean 3-space then it must *necessarily* have an edge beyond which it cannot be extended.

Minding's result applies here too: if two patches of surfaces have constant negative curvature $\mathcal{K} = -(1/\mathrm{R}^2)$ then their intrinsic geometry is *identical* to that of the pseudosphere of radius R.

To sum up, if a surface has constant curvature $\mathcal{K}$ (positive or negative) then this single number determines the intrinsic geometry of the surface *completely*.

But what of more typical surfaces, within which the curvature varies? While the control of the curvature remains very great, it is no longer absolute: it is possible for two surfaces to have different intrinsic geometries while still having identical curvatures at corresponding points. (A concrete example is provided in Ex. 19, p. 224.)

2.3 The Local Gauss–Bonnet Theorem

Recall Harriot's 1603 result (1.8) on the sphere: *the angular excess of a triangle is the curvature times the area*, which we may think of as the total amount of curvature residing within the triangle.

The ***Local***[8] ***Gauss–Bonnet Theorem***, as originally stated by Gauss in the *Disquisitiones Generales* of 1827, is a stunning generalization of this result to a geodesic[9] triangle Δ on a *general* curved

[7] This surface, also known as the ***tractroid***, was first investigated by Huygens in 1693; see Stillwell (2010, p. 345).

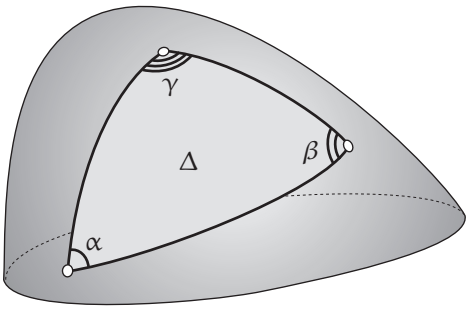
[8] Here the word "local" does not signify infinitesimal, but instead distinguishes this result from a subsequent "global" version that applies to an *entire*, closed surface.

[9] In 1865 Bonnet generalized the formula to nongeodesic triangles, hence the name of the theorem. The most general version of the theorem will not be proved until the end of Act IV (Ex. 6, p. 336).

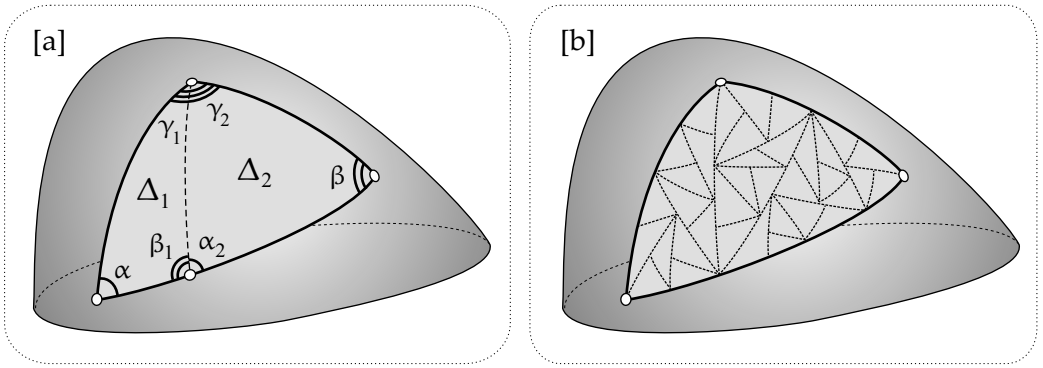
surface of variable curvature, illustrated in [2.7]. It says that the angular excess of such a triangle is simply the *total curvature* inside it:

$$\boxed{\mathcal{E}(\Delta) = \alpha + \beta + \gamma - \pi = \iint_{\Delta} \mathcal{K}\, d\mathcal{A}.} \quad (2.6)$$

In the case of the sphere, $\mathcal{K} = 1/R^2$, so (2.6) yields Harriot's Formula (1.3) as a very special case.



[2.7] *A general geodesic triangle on a general surface.*



[2.8] [a] *Angular excess is additive:* $\mathcal{E}(\Delta) = \mathcal{E}(\Delta_1) + \mathcal{E}(\Delta_2)$. **[b]** *This persists if we continue subdividing triangles:* $\mathcal{E}(\Delta) = \sum \mathcal{E}(\Delta_i)$.

To see why this is so, first recall our original definition of curvature, (2.1). As the triangle Δ_p shrinks towards a point p on $\mathcal{S}$,

$$\mathcal{E}(\Delta_p) \asymp \mathcal{K}(p)\,\mathcal{A}(\Delta_p). \quad (2.7)$$

The key fact is that the angular excess is ***additive***.

In [2.8a] a geodesic segment (dashed) has been drawn from one vertex of Δ to an arbitrary point on the opposite edge, thereby splitting Δ into two geodesic subtriangles, Δ_1 and Δ_2. Observing that $\beta_1 + \alpha_2 = \pi$, we find that

$$\mathcal{E}(\Delta_1) + \mathcal{E}(\Delta_2) = [\alpha + \beta_1 + \gamma_1 - \pi] + [\alpha_2 + \beta + \gamma_2 - \pi] = \alpha + \beta + \gamma_1 + \gamma_2 - \pi,$$

and therefore

$$\boxed{\mathcal{E}(\Delta) = \mathcal{E}(\Delta_1) + \mathcal{E}(\Delta_2).}$$

These subtriangles may then be subdivided in their turn, and so on and so forth, yielding [2.8b], and the additive property ensures that $\mathcal{E}(\Delta) = \sum \mathcal{E}(\Delta_i)$. As the subdivision becomes finer and finer, the curvature varies less and less within each Δ_i, approaching the constant value $\mathcal{K}_i$, and in this limit (2.7) yields $\mathcal{E}(\Delta) \asymp \sum \mathcal{K}_i \mathcal{A}_i$, and so we arrive at the Local Gauss–Bonnet Theorem, (2.6).

Chapter 3

Exercises for Prologue and Act I

Prologue: Newtonian Ultimate Equality ($\asymp$)

1. (This is a model of how an "ultimate equality"—see Prologue—becomes an equality.) Sketch a cube of side x, hence of volume $\mathcal{V} = x^3$. In the same picture, keeping one vertex fixed, sketch a slightly larger cube with side $x + \delta x$. If $\delta\mathcal{V}$ is the resulting increase in volume, use your sketch to deduce that as δx vanishes,

$$\delta\mathcal{V} \asymp 3x^2\,\delta x \quad \Longrightarrow \quad \frac{d\mathcal{V}}{dx} \asymp \frac{\delta\mathcal{V}}{\delta x} \asymp 3x^2 \quad \Longrightarrow \quad \frac{d\mathcal{V}}{dx} \asymp 3x^2.$$

But *the quantities in the final ultimate equality are independent of* δx, so they are equal:

$$(x^3)' = \frac{d\mathcal{V}}{dx} = 3x^2.$$

2. (This example is taken from Needham 1993.) Let $c = \cos\theta$ and $s = \sin\theta$. In the first quadrant of $\mathbb{R}^2$, draw a point $p = (c, s)$ on the unit circle. Now let p rotate by a small (ultimately vanishing) angle $\delta\theta$. With one vertex at p, draw the small triangle whose sides are δc and δs. By emulating the Newtonian geometrical argument in the Prologue, *instantly and simultaneously* deduce that

$$\frac{ds}{d\theta} = c \qquad \textit{and} \qquad \frac{dc}{d\theta} = -s.$$

3. (This example is taken from Needham 1993.) Let L be a general line through the point (a, b) in the first quadrant of $\mathbb{R}^2$, and let $\mathcal{A}$ be the area of the triangle bounded by the x-axis, the y-axis, and L.

(i) Use ordinary calculus to find the position of L that minimizes $\mathcal{A}$, and show that $\mathcal{A}_{\min} = 2ab$.

(ii) Use Newtonian reasoning to solve the problem *instantly*, without calculation! (*Hints:* Let $\delta\mathcal{A}$ be the change in the area resulting from a small (ultimately vanishing) rotation $\delta\theta$ of L. By *drawing* $\delta\mathcal{A}$ in the form of two triangles, and observing that each triangle is ultimately equal to a sector of a circle, write down an ultimate equality for $\delta\mathcal{A}$ in terms of $\delta\theta$. Now set $\delta\mathcal{A} = 0$.)

4. The following problem is taken from Arnol'd (1990, p. 28), which also contains the solution. Evaluate the limit

$$\lim_{x\to 0} \frac{\sin\tan x - \tan\sin x}{\sin^{-1}\tan^{-1} x - \tan^{-1}\sin^{-1} x},$$

(i) using *any* traditional method you can think of. (We would be remiss if we did not wish you the best of British luck in this endeavour! Arnol'd notes that the *only* mathematician who was ever able to solve this problem quickly was the Fields Medalist, Gerd Faltings.)

(ii) using Newtonian geometrical reasoning.

Euclidean and Non-Euclidean Geometry

5. It is not known with certainty how the Babylonians generated the Pythagorean triples in [1.2], but, 1500 years later, Euclid (around 300 BCE) was the first to state and prove the most general formulas for such triples. Half a millennium after that, Diophantus[1] (around 250 CE) was the first to use a *geometrical* construction[2] to systematically generate ***rational points*** on the unit circle, meaning points with rational coordinates. These rational points can then be used to generate Pythagorean triples, as follows.

(i) Let L be the line $y = m(x+1)$ through the point $(-1,0)$ of the unit circle $\mathcal{C}$, and let (X,Y) be the second intersection point of L and $\mathcal{C}$. Prove that

$$X = \frac{1-m^2}{1+m^2} \quad \text{and} \quad Y = \frac{2m}{1+m^2}.$$

(ii) Deduce that *if the slope* $m = (q/p)$ *is* rational, *then so are* X *and* Y*:*

$$X = \frac{p^2-q^2}{p^2+q^2} \quad \text{and} \quad Y = \frac{2pq}{p^2+q^2}.$$

(iii) Deduce that if (a,b,h) is an arbitrary Pythagorean triple, then

$$\frac{a}{h} = \frac{p^2-q^2}{p^2+q^2} \quad \text{and} \quad \frac{b}{h} = \frac{2pq}{p^2+q^2},$$

for some integers p and q.

(iv) Deduce that the most general Pythagorean triple is given by the following formulas, first stated by Euclid:

$$a = (p^2-q^2)r \quad \text{and} \quad b = 2pqr \quad \text{and} \quad h = (p^2+q^2)r,$$

where p, q, r are arbitrary integers.

6. Use (1.3) to deduce that as the size of a triangle on the sphere shrinks to nothing, it ultimately appears to the inhabitants of the sphere to be *Euclidean*, i.e., with angle sum equal to π.

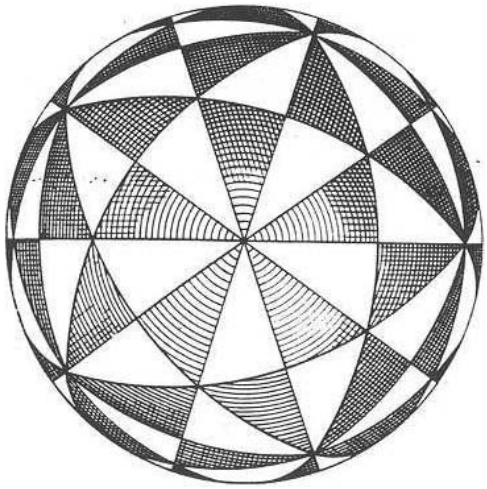
7. Let p and q be distinct, nonantipodal points on the sphere, and consider the unique great circle $\mathcal{C}$ through them. Let m_1 and m_2 be the midpoints of the two arcs into which $\mathcal{C}$ is split by p and q. Show that the locus of points that are equidistant from p and q is the great circle through m_1 and m_2 that cuts $\mathcal{C}$ orthogonally: this is the generalized "perpendicular bisector" of pq. (*Hint:* while it is mathematically immaterial, it can be *psychologically* helpful to imagine that the sphere has been rotated so that p and q lie on the equator.)

8. If the sides of a triangle on the unit sphere $\mathbb{S}^2$ are each less than π, show that the triangle is contained within a hemisphere. (*Hint:* while it is mathematically immaterial, it can be *psychologically* helpful to imagine that the sphere has been rotated so that one vertex is at the north pole.)

9. One of the characteristics of the Euclidean plane is that it possesses ***regular tessellations***: it can be ***tiled*** (completely filled, without any gaps) by regular polygons. There are precisely three such regular tessellations of the plane, using equilateral triangles, squares, or regular

[1] For the little that is known of his life, see Stillwell (2010, §3.6).

[2] This is the prototype of Newton's method of generating rational points on *cubic* curves, using chords and tangents. See Stillwell (2010, §3.5).



[3.1] Icosahedral tessellation of the sphere. *Here the image of each triangular face of the icosahedron has been further divided into six congruent triangles, by joining its centre to the midpoints of its edges. (This lovely drawing is taken from Fricke (1926).)*

hexagons. The sphere *also* admits regular tessellations. Imagine a wire-frame icosahedron inscribed in a sphere of radius R. Now imagine a point source of light at the centre of the sphere. The shadow on the sphere is shown in [3.1]. Here the image of each triangular face of the icosahedron has been further divided into six congruent triangles, by joining its centre to the midpoints of its edges.

(i) Explain why the shadows of the icosahedral edges are arcs of *great circles* on the sphere, thereby creating genuine spherical triangles.

(ii) Verify that if we instead inscribe a dodecahedron, centrally project its pentagons onto the sphere, then join their centres to the midpoints of the edges, we obtain the *same* tessellation.

(iii) Into how many congruent triangles has the sphere (of area $4\pi R^2$) been divided? So what is the area $\mathcal{A}$ of each triangle?

(iv) By observing how many like angles come together at each vertex, deduce that the angles of the triangles are $(\pi/2)$, $(\pi/3)$, and $(\pi/5)$. Hence calculate the angular excess $\mathcal{E}$ of each triangle.

(v) Verify that the previous two answers are in accord with Harriot's Theorem, (1.3).

10. (i) Prove that in Euclidean Geometry the angle sum of a quadrilateral is 2π.

(ii) If Q is a geodesic quadrilateral on the sphere of radius R, its angular excess is therefore

$$\mathcal{E}(Q) = (\text{angle sum of } Q) - 2\pi.$$

By drawing in a diagonal of Q, thereby splitting it into two geodesic triangles, deduce that (1.3) generalizes to

$$\mathcal{E}(Q) = \frac{1}{R^2}\,\mathcal{A}(Q).$$

(iii) Prove (2.2), and hence generalize (ii) to geodesic n-gons on the sphere.

11. Using the technique described in the footnote on page 14, or otherwise, manufacture narrow strips of sticky tape, ideally coloured masking tape. Then use (1.7) to conduct the following

experiments on the surface of a vase of the approximate form shown in [11.7], on page 127. If you do not own such a vase, we suggest you borrow one—this exercise is too interesting to be missed!

(i) Starting at a point on the horizontal circle of greatest radius, ρ_{max}, launch a geodesic straight up the vase, creating a ***meridian*** of the surface of revolution.

(ii) Starting at the same point, launch several more geodesics in directions that make ever-larger angles ψ with the meridian.

(iii) Note that beyond some critical angle ψ_c, the geodesics initially make their way up the vase, but then turn back and come down the vase!

(iv) As best you can, try to find the critical geodesic—the one that separates those that turn back from those that don't. By pressing a protractor against the surface at its launch point, measure its angle, ψ_c. NOTE: To find the critical geodesic, you will need to construct extra-long geodesic segments, which you may do by extending an existing segment, overlapping the new with the old, as illustrated in [1.9].

(v) Let ρ_{max} be the maximum radius of the vase (at the launch point), and let ρ_{min} be the minimum radius, at the throat of the vase. By measuring diameters and dividing by 2, find these two radii as accurately as you can. Now verify that, within experimental error,

$$\psi_c = \sin^{-1}\left[\rho_{min}/\rho_{max}\right] !$$

(This is a physical instantiation of ***Clairaut's Theorem***, which we will prove much later, in Section 11.7.4.)

Gaussian Curvature

12. ***Zero*** **Curvature.** Using the technique described in the footnote on page 14, or otherwise, manufacture narrow strips of sticky tape, ideally coloured masking tape. Then use (1.7) to conduct the following experiments.

(i) Roll up a piece of paper into a cone, and tape it so that it does not unfurl. *Start* to create a long geodesic segment originating at a point on the rim, but try to guess its form *before* you stick your strip down onto the surface. Starting at the same point on the rim, repeat this construction, launching new geodesics in different directions.

(ii) Next, construct a geodesic triangle, and use a protractor to verify that $\mathcal{E} = 0$. (This is true of *all* geodesic triangles on this surface, proving that $\mathcal{K} = 0$.)

(iii) Finally, cut open the cone (along a generator) and press the paper flat again, and observe the Euclidean form of your tape constructions.

13. ***Positive*** **Curvature.** Stick a toothpick into the north pole of any approximately spherical fruit (of radius R), such as a melon. Tie one end of a piece of string or dental floss to the toothpick and stretch the string tightly over the surface to a point about halfway to the equator. Holding a pen at the end of this geodesic segment of length r, drag the pen around (keeping the string taut) to create a circle of latitude of length $C(r)$. Stick toothpicks into this circle at 16 evenly spaced points. Now wrap a piece of string or dental floss around the outside of the approximately circular ring of toothpicks, and gently pull it tight, so that it follows the circle of latitude *on* the surface of the fruit. With a pen, carefully mark the string at the start and end of the circle. Unwrap it and stretch it along a ruler to measure the length $C(r)$ between the two marks.

(i) Treating (2.4) as an approximate equality (instead of an exact ultimate equality), estimate $\mathcal{K}$. From these *intrinsic* measurements, estimate the *extrinsic* radius R of the fruit. Compare your answer with a direct measurement of R.

(ii) Continuing from (i), suppose that your measurements of r and $C(r)$ are perfect. Use the third term of the (decreasing and alternating) Taylor expansion of $\sin(r/R)$ to show that an upper bound on the percentage error in $\mathcal{K}$ that results from applying (2.4) in the manner of part (i)—i.e., *without* taking the limit implied by the ultimate equality—is given by

$$\left|\frac{\Delta\mathcal{K}}{\mathcal{K}}\right| < 5\left[\frac{r}{R}\right]^2 \%$$

Deduce that even for a circle as large as the one you constructed, the error cannot be larger than approximately 3% !

(iii) Use the result of (ii) to deduce a formula for the upper bound of the percentage error in R.

14. ***Negative*** **Curvature.** Using the technique described in the footnote on page 14, or otherwise, manufacture narrow strips of sticky tape, ideally coloured masking tape. Then use (1.7) to conduct the following experiments.

(i) By following the instructions that accompany [5.3], page 53, construct your own, personal pseudosphere out of discs of radius R—the more cones, the better; the bigger, the better!

(ii) Starting at a point on the circular base of radius R, launch geodesics in various directions, and try to predict their course *before* you lay the tape down onto the surface. When a strip of tape runs out, continue the geodesic by simply *overlapping* a new strip with the old one, as illustrated in [1.9]. With the sole exception of the meridian geodesic that heads straight up the surface—tracing a ***tractrix*** generator of the surface of revolution—note that every geodesic initially heads up the pseudosphere but then turns around and comes back down the pseudosphere, ultimately returning to the base circle.

(iii) Construct a right-angled geodesic triangle, Δ, measure its angles, and hence estimate its angular excess, $\mathcal{E}(\Delta)$. Estimate (as best you can) its areas, $\mathcal{A}(\Delta)$. Hence estimate the (constant) curvature $\mathcal{K}$ of your pseudosphere, using

$$\mathcal{K} = \frac{\mathcal{E}(\Delta)}{\mathcal{A}(\Delta)}.$$

(iv) The larger the triangle, the larger (i.e., more negative) the value of $\mathcal{E}(\Delta)$, making its measurement easier and more accurate. But the tradeoff is that it becomes harder to accurately estimate the area $\mathcal{A}(\Delta)$. To overcome this difficulty, do the following. Make narrow strips of the same kind that you use to generate geodesics, but create them all with the *same* (accurately measured) width, W, perhaps $1/4$ inch. Now fill your Δ with these strips, cutting them off when they hit an edge. Remove the strips and lay them end-to-end on a flat surface, and measure the total length, L. Then $\mathcal{A}(\Delta) \asymp LW$.

(v) Repeat (iii) with several more triangles, but no longer restrict them to be right-angled, because (iv) now makes it easy to measure $\mathcal{A}(\Delta)$ for any shape of triangle. Verify that (within experimental error) all triangles yield the *same* value of $\mathcal{K}$.

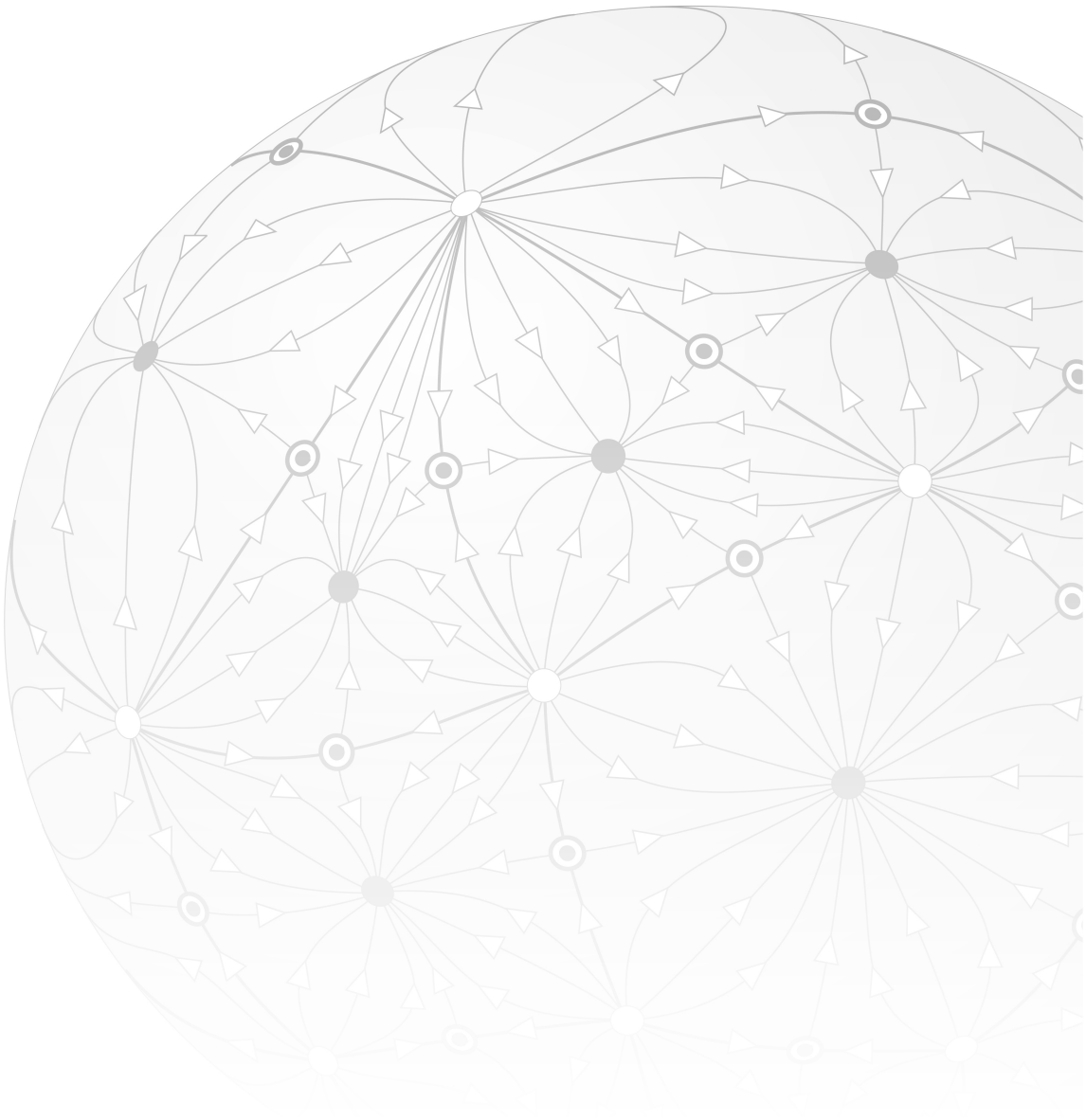
(vi) Assuming that

$$\mathcal{K} = -\frac{1}{R^2},$$

estimate R, and compare this to the actual radius of the discs you used to construct your pseudosphere.

ACT II

The Metric



Chapter 4

Mapping Surfaces: The Metric

4.1 Introduction

The perfect extrinsic symmetry of the sphere has the advantage of making it obvious that its intrinsic geometry is likewise uniform. By contrast, in [2.5] it's certainly not clear that a flexible but unstretchable shape fitted to the surface can be freely slid about and rotated on the surface, all the while remaining fully in contact with the surface. But in fact this must be so, for the above discussion shows that the actual shape of a surface in space is a distraction: these surfaces are intrinsically indistinguishable from the sphere (at least locally).

From this point of view it would be better to have a more abstract model that captured the essence of all possible surfaces having the same intrinsic geometry. By the "essence" we mean knowledge of the distance between any two points, for this and this alone determines the intrinsic geometry. In fact—and this was Gauss's fundamental insight into Differential Geometry—it is sufficient to have a rule for the *infinitesimal* distance between neighbouring points. This rule is called the ***metric***.[1] Given this metric, we may determine the length of any curve as an infinite sum (i.e., integral) of the infinitesimal segments into which it may be divided. Consequently, we may also identify the geodesics of the geometry as the shortest routes from one point to another, and we can likewise determine angles.

This leads to the following strategy for capturing the essence of any curved surface $\mathcal{S}$ (not necessarily one of constant curvature). To avoid the distraction of the shape of the surface in space, we draw a (cartographic) ***map*** of $\mathcal{S}$ on a flat piece of paper. That is, we set up a one-to-one correspondence between points $\widehat{z}$ on $\mathcal{S}$ and points z on the plane. Of course in the case of the spherical Earth and the spherical night sky, mariners and astronomers have been devising such geographical and celestial maps for thousands of years.

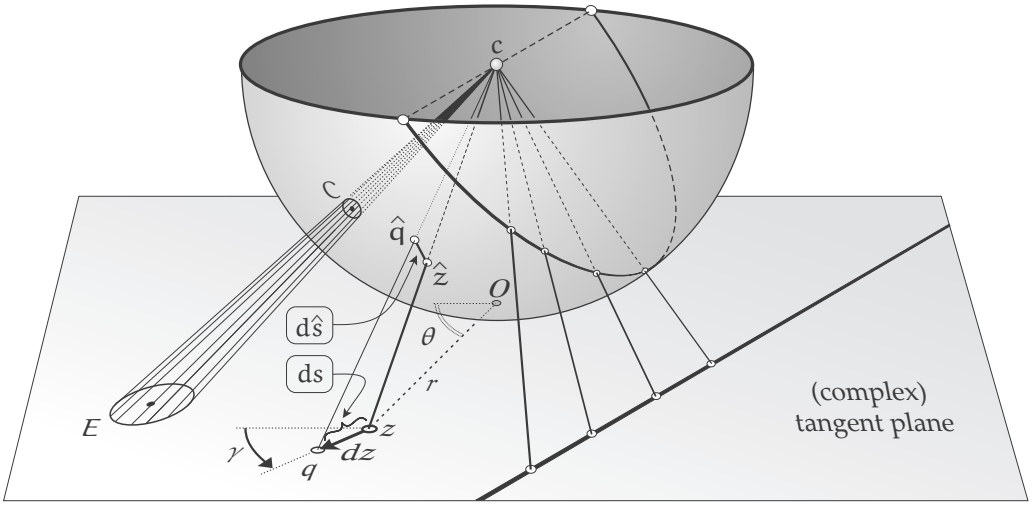
IMPORTANT NOTE ON TERMINOLOGY: In most of mathematics, "map" is synonymous with "mapping" or "function." However, when used as a noun, we shall always use "map" in the current sense of a *cartographic* map. When we mean "mapping" or "function," we shall *say* "mapping" or "function"! That said, we will retain some of the traditional meaning when it comes to the *verb*: e.g., the coordinate functions map this curve on the surface to that curve in the map.

In general, such a map of a curved surface cannot be created without introducing some kind of distortion: if you peel part of an orange and try to press the patch of peel flat on the table, it will tear. Euler was the first to prove (in 1775) the mathematical impossibility of a perfect map of the Earth, in which all "straight lines" (geodesics) on the surface become straight lines in the maps, and all terrestrial distances are represented by proportional distances on the map.

The preceding discussion explains this impossibility of a perfect map. A triangle on the Earth's surface has $\mathcal{E} \neq 0$, but if this triangle could be pressed flat without altering distances then its image in the map would be a Euclidean triangle with $\mathcal{E} = 0$: a contradiction.

We will eventually discover that there exist deep and mysterious connections between non-Euclidean Geometry and *complex numbers*. Let us therefore, from the very outset, imagine that the flat piece of paper on which we draw our map is in fact the *complex plane*, $\mathbb{C}$.

[1] Another common name, particularly in older works, is the ***First Fundamental Form***.



[4.1] *The central projection of the sphere maps geodesics to straight lines and circles to ellipses. The* metric *tells us the local scale factor relating map distance* $\mathrm{d}s$ *and surface distance* $\mathrm{d}\,\hat{s}$.

Now consider the distance $\delta\,\hat{s}$ separating two neighbouring points $\hat{z}$ and $\hat{q}$ on $\mathcal{S}$. The points $\hat{z}$ and $\hat{q}$ will then be represented by the complex numbers $z = r\,e^{i\theta}$ and $q = z + \delta z$ in this complex plane, separated by (Euclidean) distance $\delta s = |\delta z|$. For a concrete example of such a map (to be explained in a moment) look ahead to [4.1]. Once we have a rule for calculating the actual separation $\delta\,\hat{s}$ on $\mathcal{S}$ from the apparent separation δs in the map, then (in principle) we know everything there is to know about the intrinsic geometry of $\mathcal{S}$.

The rule giving $\delta\,\hat{s}$ in terms of δz is called the ***metric***. In general, $\delta\,\hat{s}$ depends on the direction of δz as well as its length δs: writing $\delta z = e^{i\gamma}\,\delta s$, so $\delta\,\hat{s} \asymp \Lambda(z,\gamma)\,\delta s$, in which we remind the reader that $\asymp$ denotes Newton's concept of ***ultimate equality***, which was introduced in the Prologue. This relation is more traditionally expressed using the notation of infinitesimals, as,

$$\boxed{\mathrm{d}\,\hat{s} = \Lambda(z,\gamma)\,\mathrm{d}s.} \tag{4.1}$$

According to this formula, $\Lambda(z,\gamma)$ is the amount by which we must expand the apparent separation $\mathrm{d}s$ in the map—located at z, and in the direction γ—to obtain the true separation $\mathrm{d}\,\hat{s}$ on the surface $\mathcal{S}$.

4.2 The Projective Map of the Sphere

Figure [4.1] illustrates the meaning of (4.1) for one particular method of drawing a map of the southern hemisphere, called ***central projection***. Imagine the hemisphere to be a glass bowl resting on the complex plane at the origin 0, and imagine a light source at the centre of the sphere, c. Then a light ray passing through $\hat{z}$ on the hemisphere goes on to hit a point z in $\mathbb{C}$. The resulting plane map is the so-called ***projective map*** (or ***projective model***) of the southern hemisphere.

If we draw a circle C on the hemisphere then the rays passing through it form a cone in space, and hence they hit $\mathbb{C}$ in a perfect ellipse E. This is a very special and unusual property of this particular method of drawing a map. However, it turns out[2] that if C is an intrinsically defined circle on a *general* surface $\mathcal{S}$ (such as in [1.9]) then as its radius shrinks its image E in a *general* map

[2]Because if the mapping is differentiable then its local effect is a linear transformation.

will ultimately be an ellipse, also. Returning to the perfect ellipses of the central projection, it's clear that E's major axis will be radial, pointed directly away from the bowl's point of contact with the plane. In other words, if you imagine dz rotating about z, then $\Lambda(z, \gamma)$ achieves its minimum and maximum values at $\gamma = \theta$ and $\gamma = \theta + (\pi/2)$, respectively.

How we choose to draw a map of a surface depends on which features we wish to accurately or ***faithfully*** represent. For example, [4.1] illustrates that the projective map faithfully represents lines: a straight line in the map represents a great-circle geodesic on the sphere. But, the price that we pay for preserving the concept of lines is that angles are not faithfully represented: the angle at which two curves meet on the sphere is not (in general) the angle at which they meet on the map.

That said, there are in fact two orthogonal families of curves on the sphere that map to orthogonal curves in the plane: they are the circles of longitude and latitude. A circle of latitude (i.e., a horizontal cross section of the hemisphere) maps to an origin-centred circle in the plane, and a (semi-)circle of longitude (i.e., a vertical cross section of the hemisphere through its centre) maps to a line through the origin. As claimed, these circles and lines do indeed meet at right angles. We now use this fact to derive a formula for the metric of the sphere in terms of polar coordinates (r, θ) in the projective map. It is not hard to accomplish this by calculation [exercise], but we shall instead use the Newtonian form of geometric reasoning that was introduced in the Prologue, and which we shall employ throughout this work.

See [4.2]. A small rotation of $\delta\theta$ in the plane rotates z a distance $r\,\delta\theta$ round the circle of radius r and rotates $\widehat{z}$ by $\delta\widehat{s_1}$ round a horizontal circle of latitude. If z then moves radially outward by δr then $\widehat{z}$ moves north by $\delta\widehat{s_2}$ along a vertical circle of longitude. By Pythagoras's Theorem, $\delta\,\widehat{s}^2 \asymp \delta\,\widehat{s_1}^2 + \delta\,\widehat{s_2}^2$, and we shall now find each of these terms separately. (Recall from the Prologue that $\asymp$ is our symbol for Newton's concept of "ultimate equality.")

Let $H = cz$ denote the distance in $\mathbb{R}^3$ from the centre c of the sphere to the complex number z, as illustrated in [4.2]. Since $\widehat{z}$ is at distance R from c, and since the rotation causes $\widehat{z}$ and z to move in proportion to their distances from the centre of the sphere:

$$\frac{d\widehat{s_1}}{r\,d\theta} \asymp \frac{\delta\widehat{s_1}}{r\,\delta\theta} = \frac{R}{H}.$$

Next, imagine H to be a string of fixed length, attached to c. If we swing its free end at z upward a distance ϵ in the plane perpendicular to the complex plane, then the northward motion $\delta\widehat{s_2}$ of $\widehat{z}$ will be in the same proportion as before. Furthermore, we see that the small vertical grey triangle with edges ϵ and δr is ultimately similar to the large triangle 0cz with sides R and H. Thus,

$$\frac{\delta\widehat{s_2}}{\epsilon} \asymp \frac{R}{H} \quad \text{and} \quad \frac{\epsilon}{\delta r} \asymp \frac{R}{H} \quad \Longrightarrow \quad \frac{d\widehat{s_2}}{dr} = \frac{R^2}{H^2}$$

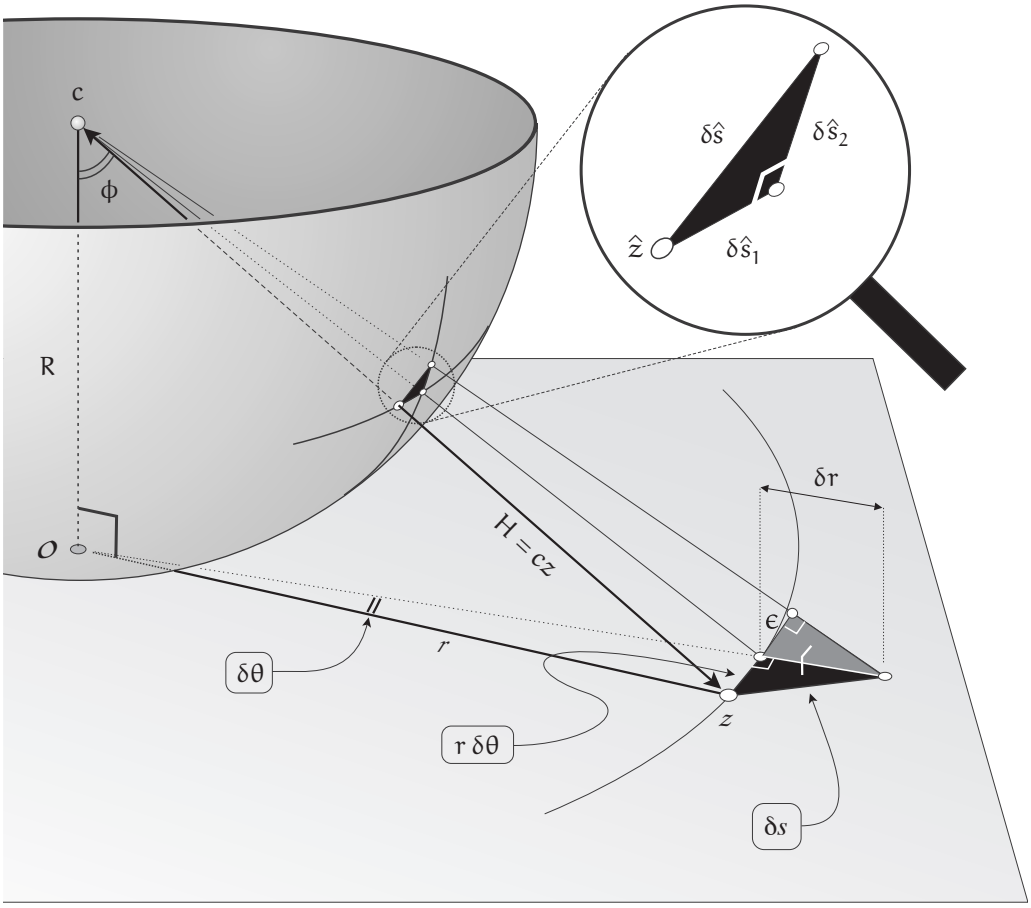
Finally, by the Pythagoras's Theorem again, $H^2 = R^2 + r^2$. So, reverting to the industry-standard, infinitesimal-based notation, the metric describing the true distance on the sphere in terms of the coordinates in the projective map is given by

$$d\,\widehat{s}^2 = \frac{1}{1 + (r/R)^2}\left[\frac{dr^2}{1 + (r/R)^2} + r^2\,d\theta^2\right]. \tag{4.2}$$

In geography, the angle of latitude ϕ is usually defined to be zero at the equator, but we choose to instead measure it from a pole (as shown in both [4.2] and [2.4]). Returning to [4.1], we can now quantify the amount of distortion induced by the map, by looking at the shape of the small ellipse E. As C shrinks, you should now be able to confirm [exercise] that

$$\left(\frac{\text{major axis of E}}{\text{minor axis of E}}\right) \asymp \sec\phi, \tag{4.3}$$

so as C moves north towards the rim ($\phi = \frac{\pi}{2}$), the distortion of shapes increases without limit.



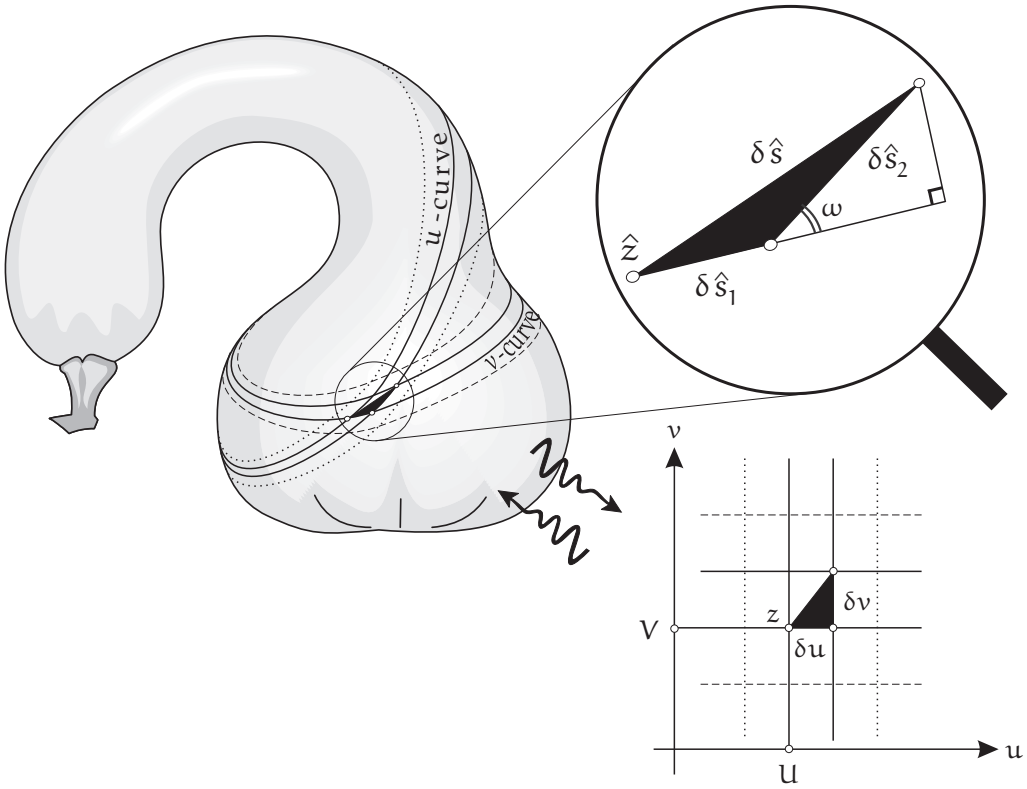
[4.2] *A small rotation of $\delta\theta$ in the plane moves z by $r\,\delta\theta$ and it moves $\hat{z}$ by $\delta\hat{s}_1$ round a horizontal circle of latitude. If z then moves radially outward by δr then $\hat{z}$ moves north by $\delta\hat{s}_2$ along a vertical circle of longitude.*

4.3 The Metric of a General Surface

Given the critical importance of maps in navigation, mathematicians have for centuries explored many different methods of drawing maps of the Earth, and we shall meet an especially important one shortly; other maps are explored in the exercises at the end of this Act. For now we wish merely to point out that each such map has a different metric formula associated with it, despite the fact that they all describe the *same* intrinsic geometry.

For example, the most common way of describing the location of a place on Earth is to supply its longitude θ and latitude ϕ. Suppose we use these two angles to draw a very straightforward kind of map in the plane, with θ along the horizontal axis, and ϕ along the vertical axis. That is, if a particular house has longitude θ and latitude ϕ then in the flat map we represent it by the point with Cartesian coordinates (θ, ϕ). With the assistance of [2.4], you should now easily find [exercise] that the metric formula telling us the true distance on the sphere corresponding to neighbouring points in the map is

$$d\hat{s}^2 = R^2\left[\sin^2\phi\, d\theta^2 + d\phi^2\right]. \tag{4.4}$$



[4.3] *On a general surface, draw two parameterized families of curves: "u-curves" (*$u = \text{const.}$*) and "v-curves." A point is then specified by the particular pair of curves* (U, V) *that intersect there, yielding a point* $z = U + iV$ *in the map. A small horizontal movement* δu *in the map ultimately produces a proportional movement* $\delta \hat{s}_1 \asymp A\,\delta u$ *on the surface, along a* v*-curve.*

This certainly *looks* very different from our previous formula, (4.2), but we know that it in fact describes exactly the same intrinsic geometry.

Even if we simply take our surface to be a flat plane, an infinite variety of metric formulas are possible. For example, if we use regular Cartesian coordinates then $d\hat{s}^2 = dx^2 + dy^2$, whereas if we use polar coordinates then $d\hat{s}^2 = dr^2 + r^2 d\theta^2$.

Now let us investigate the form and meaning of the most general metric formula for a general surface. Figure [4.3] illustrates how we may draw a general map of a patch of such a surface $\mathcal{S}$. For each point $\hat{z}$ in this patch, our first aim is to assign a pair of coordinates (u, v) so that it can be represented by the complex number $z = u + iv$ in the flat map.

To get started, we simply draw, quite arbitrarily, a family of curves covering the patch, so that one (and only one) curve from our family passes though each point $\hat{z}$ on the patch. We now label each of these curves with a unique number u, and agree to call these curves, with their assigned u-values, the ***u-curves***. The labelling can be done quite arbitrarily, *except* that we shall insist that the numbering vary *smoothly*, in the sense that as we move over the surface across the u-curves, the u-values should change at a definite *rate*, i.e., differentiably—more on this in a moment.

To complete the coordinate system, we now draw a second family of curves, again quite arbitrarily, except that they should cross the u-curves and never coincide with them. Now label these new curves with v-values, and, again, the v-values should vary differentiably, at a definite *rate* (in the same manner as before); call these the ***v-curves***. Thus, as illustrated, the point $\hat{z}$ can be labelled by the unique u-curve (say $u = U$) and v-curve (say $v = V$) that intersect there. So, in

the map, $\widehat{z}$ can now be represented as $z = U + iV$. Thus, as illustrated, u-curves are represented in the map by vertical lines, while v-curves are mapped to horizontal lines.

Now that we have the coordinates on at least some patch of the surface, the task is to find the metric formula that gives the distance between neighbouring points. Suppose that in the map we make a small movement away from z along $\delta z = \delta u + i\,\delta v$. By virtue of the differentiable assignment of u-values to the u-curves, the small change δu produces—and we now come to the *definition* of ***differentiable***—a small movement $\delta\widehat{s_1}$ on the surface (along a v-curve) that is ultimately *proportional* to δu, and we define A to be this proportionality constant at the point:

$$A \equiv \frac{\partial \widehat{s_1}}{\partial u} \asymp \frac{\delta \widehat{s_1}}{\delta u}.$$

This is important, so we reiterate: A *is the local scale factor in the horizontal direction of the* (u, v)*-map*, the factor by which we must stretch a small horizontal distance in the map to obtain the true distance on the surface.

There is also a way to visualize this without even looking at the map. If in [4.3] we imagine that the u-curves have been drawn at small fixed increments of ϵ ($u = U$, $u = U + \epsilon$, $u = U + 2\epsilon, \ldots$) then A can also be visualized as being inversely proportional to the crowding or density of the u-curves. For the greater the crowding, the greater the change δu resulting from a given movement $\delta\widehat{s_1}$ on the surface.

In exactly the same way, the small change δv (while keeping u constant) produces a movement $\delta\widehat{s_2}$ on the surface that is ultimately proportional to it, enabling us to define B to be the local scale factor in the vertical direction of the map:

$$B \equiv \frac{\partial \widehat{s_2}}{\partial v} \asymp \frac{\delta \widehat{s_2}}{\delta v}.$$

Finally, as illustrated, we define ω to be the angle between the u-curves and the v-curves. Obviously this angle is not a constant: just as with the scale factors A and B, the angle ω is a function of position.

We can now apply Pythagoras's Theorem to the right triangle shown in the magnifying glass of [4.3]:

$$\delta\widehat{s}^2 \asymp (\delta\widehat{s_1} + \delta\widehat{s_2}\,\cos\omega)^2 + (\delta\widehat{s_2}\,\sin\omega)^2 \tag{4.5}$$

$$\asymp (A\,\delta u + B\,\delta v\,\cos\omega)^2 + (B\,\delta v\,\sin\omega)^2. \tag{4.6}$$

After simplifying, and reverting to more standard, infinitesimal-based notation, we find that

The general metric formula for a general surface is

$$d\widehat{s}^2 = A^2\,du^2 + B^2\,dv^2 + 2F\,du\,dv, \quad \textit{where} \quad F = AB\,\cos\omega. \tag{4.7}$$

Assuming that you will eventually look at other books, we should immediately issue a NOTATIONAL WARNING. In his original masterpiece of 1827, Gauss made the decision to write[3] the metric as

$$d\widehat{s}^2 = E\,du^2 + G\,dv^2 + 2F\,du\,dv,$$

[3] In fact Gauss wrote p and q in place of u and v, which later became the standard notation. We have chosen not to mind our historical p's and q's.

In the subsequent centuries, almost every[4] research paper and textbook on Differential Geometry has slavishly perpetuated this E, F, G-notation. And yet we have seen that it is $\sqrt{E} = A$ and $\sqrt{G} = B$ that have the simple geometric interpretation given above, and thus it should come as no surprise that it is they (*not* E and G) that manifest themselves in many important formulas. The consequence has been that a literature has arisen that is needlessly cluttered with square roots. We shall therefore continue to employ the notation A and B (in place of $\sqrt{E}$ and $\sqrt{G}$) throughout the book, so we stress that when you look elsewhere you must translate using this

$$\text{NOTATIONAL DICTIONARY:} \qquad E \equiv A^2, \qquad G \equiv B^2, \qquad F \equiv AB\,\cos\omega. \tag{4.8}$$

Next, the general metric formula can be simplified as follows. It should be clear that once we have drawn the family of u-curves we may then draw the family of *orthogonal trajectories*, and then *choose* these as the v-curves. Since this construction insists that $\omega = (\pi/2)$, it follows that $F = 0$, and hence,

Locally, we may always construct an **orthogonal** (u, v)*-coordinate system on a general surface; the metric then takes the form,*

$$d\hat{s}^2 = A^2\,du^2 + B^2\,dv^2. \tag{4.9}$$

Note, however, that it is generally *not* possible to cover the *entire* surface with a single (u, v)-coordinate system, even if we do *not* insist on orthogonal coordinates. The problem that arises is that two u-curves (and/or v-curves) may be forced to intersect, in which case the intersection point would have to be assigned two *different* u-values. In fact, as will eventually see in Chapter 19, such problems are in fact *inevitable* on *every closed surface*, other than the doughnut.

For example, on the surface of the Earth, suppose we choose our u-curves to be circles of latitude (though not necessarily with $u =$ latitude). Then the orthogonal trajectories (i.e., the v-curves) are *necessarily* the circles of longitude: great circles intersecting at the north and south poles. Thus each pole must be assigned infinitely many v-values.

4.4 The Metric Curvature Formula

Now suppose we are simply handed a metric *formula* (4.9), without any direct geometric knowledge of the surface $\mathcal{S}$ it describes, nor the geometric meaning of the coordinates u and v, themselves. What does this surface actually look like? As far as the intrinsic geometry of $\mathcal{S}$ is concerned, this formula tells us *everything*, but only in principle. How can we actually extract useful information from the formula?

If we knew the curvature $\mathcal{K}$ at each point then (via [2.3]) we would possess a clear understanding of the nature and shape of $\mathcal{S}$. But since the metric knows everything about the intrinsic geometry, it must (in particular) contain this information about the curvature. Thus the existence of a formula for $\mathcal{K}$ is assured, and, by virtue of the symmetry of the metric formula, it is also clear that the formula for $\mathcal{K}$ must be symmetrical under the simultaneous exchanges $u \leftrightarrow v$ and $A \leftrightarrow B$.

[4] There are only a few exceptions, but their high pedigree perhaps lends a modicum of respectability to our choice of notation. For example, Hopf (1956, p. 92) occasionally wrote $E = e^2$ and $G = g^2$, while Blaschke (1929, p. 162) occasionally employed precisely our A, B-notation.

Nevertheless, it is remarkable how *beautiful*[5] the formula for K turns out to be:

$$\boxed{K = -\frac{1}{AB}\left(\partial_v\left[\frac{\partial_v A}{B}\right] + \partial_u\left[\frac{\partial_u B}{A}\right]\right).} \tag{4.10}$$

The road will be long, but we shall ultimately derive this formula with the simplicity and grace it deserves, first geometrically in Act IV (Chapter 27), and finally by calculation (using Cartan's Differential Forms) in Act V (Section 38.8.2). For now, though, we treat it as hyper-advanced technology delivered to us from the future, like a phaser weapon from *Star Trek*'s twenty-third century: we may fire it at targets without beginning to understand (for now) how it works.[6]

For example, the Euclidean metric $d\hat{s}^2 = dr^2 + r^2 d\theta^2$ has $u = r$, $v = \theta$, $A = 1$, and $B = r$, so

$$K = -\frac{1}{r}\left(\partial_\theta\left[\frac{\partial_\theta 1}{r}\right] + \partial_r\left[\frac{\partial_r r}{1}\right]\right) = -\frac{1}{r}\left(\partial_r 1\right) = 0,$$

as it should.

Next, let's aim at the spherical metric (4.4). In this case, $u = \theta$, $v = \phi$, $A = R\sin\phi$, and $B = R$, so

$$K = -\frac{1}{R^2\sin\phi}\left(\partial_\phi\left[\frac{\partial_\phi R\sin\phi}{R}\right] + \partial_\theta\left[\frac{\partial_\theta R}{R\sin\phi}\right]\right) = -\frac{\partial_\phi\cos\phi}{R^2\sin\phi} = +\frac{1}{R^2},$$

as it should.

Although the calculation is longer, we encourage you to try your own hand at applying this formula to the projective metric formula (4.2) for the sphere; this too should yield $K = +(1/R^2)$.

Before moving on, we note another result we shall need later. The metric tells us how to convert small distances in the map into distances on the surface. But how should we convert *areas*? Well, in [4.3] a small rectangle in the map has area $\delta u\,\delta v$, and on the surface the corresponding parallelogram has an area that is ultimately equal to $(A\,\delta u)(B\,\delta v\sin\omega)$. Thus, using (4.7), the formula for the infinitesimal element of area dA on the surface is

$$\boxed{dA = \sqrt{(AB)^2 - F^2}\,du\,dv.} \tag{4.11}$$

If we specialize (as we usually shall) to an *orthogonal coordinate system*, so that $F = 0$, then this formula simplifies to

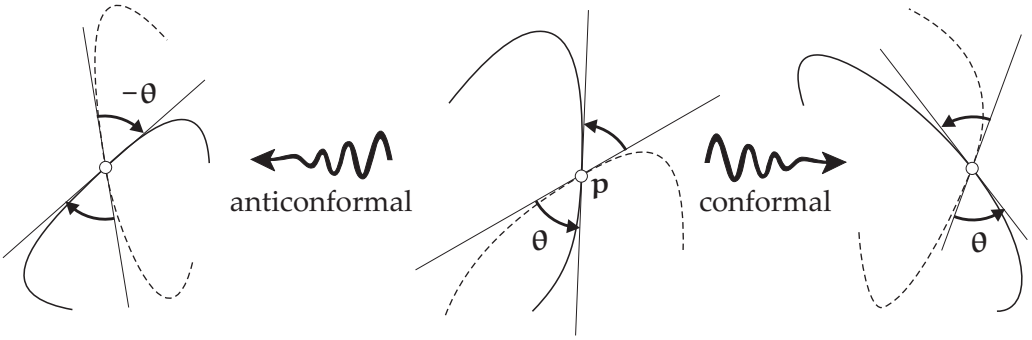
$$\boxed{dA = AB\,du\,dv.} \tag{4.12}$$

4.5 Conformal Maps

While the projective map of the sphere enjoys the advantage of preserving straight lines, for almost all purposes it is much better to sacrifice straight lines in favour of preserving *angles*. A map that

[5] Again, most texts retain Gauss's original E, G-notation, so the beauty of this remarkable formula is traditionally marred by the appearance of *five* distracting and unnecessary square roots!

[6] This approach is not without risk: witness the tragic twentieth-century use of Dr. McCoy's stolen phaser in *The City on the Edge of Forever*!



[4.4] *In the centre, two curves intersect at* p*, and the angle between them (from dashed to solid) is defined to be the angle* θ *between their tangents. A* conformal *mapping (right) preserves both the magnitude and sense of the angle, whereas an* anticonformal *mapping (left) preserves the magnitude but reverses the sense.*

preserves the magnitude and ***sense***[7] of angles is called ***conformal***; if it preserves the magnitude but reverses the sense, it is called ***anticonformal***.

When we speak of the angle between two curves, we mean the angle between their tangents. See [4.4].

In terms of the metric formula (4.1), a map is conformal if and only if the expansion factor Λ does not depend on the direction γ of the infinitesimal vector dz emanating from z:

$$\textbf{\textit{Conformal map}} \iff d\hat{s} = \Lambda(z)\, ds. \tag{4.13}$$

The great advantage of such a map is that

> *An infinitesimal shape on the surface* $\mathcal{S}$ *is represented in a* **conformal map** *by a* similar *shape that differs from the original only in size: the linear dimensions of the shape on* $\mathcal{S}$ *are just* Λ *times bigger than the linear dimensions of the shape in the map.*

Indeed, eighteenth-century mathematicians used the expression ***similar in the small***, in place of the modern term ***conformal***.

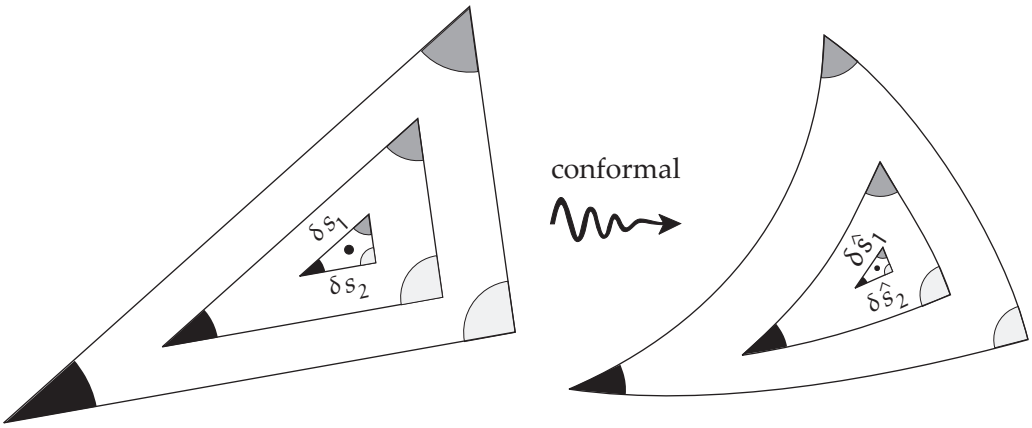
It is clear that (4.13) implies conformality, and to see that the converse is true (as claimed) consider [4.5]. On the left we see a triangle shrinking down towards a point in a *conformal* map of a surface. As it does so, the corresponding curvilinear triangle on the surface shrinks towards a rectilinear triangle that—to use the language introduced in the Prologue—is ***ultimately similar***:

$$\frac{\delta\hat{s_1}}{\delta s_1} \asymp \frac{\delta\hat{s_2}}{\delta s_2} \asymp \Lambda,$$

for some Λ, independent of the directions of the triangle sides δs_1 and δs_2. Thus, we see that conformality implies (4.13).

In discussing the general metric formula (4.7), it was clear that we could always specialize to an orthogonal coordinate system of u-curves and v-curves on the surface, corresponding to orthogonal vertical and horizontal lines in the map. But at this stage the expansion factors A and B

[7] Anticlockwise $(+)$ or clockwise $(-)$.



[4.5] *As a triangle shrinks in the map (left) its conformal image on the surface (right) is* **ultimately similar***:* $\frac{\delta\hat{s}_1}{\delta s_1} \asymp \frac{\delta\hat{s}_2}{\delta s_2}$.

in these directions were different, so an infinitesimal circle in the map would be stretched into an ellipse on the surface, and angles would be changed, in general.

We are now contemplating an even greater specialization in which the expansion factor is the *same* in all directions, so that $A = B = \Lambda$, and infinitesimal circles are mapped to infinitesimal *circles*, and *angles are preserved*. In this case, the (u, v)-coordinates are called ***conformal coordinates*** (or ***isothermal coordinates***), and (4.9) reduces to being a simple multiple of the Euclidean metric:

$$\boxed{d\hat{s}^2 = \Lambda^2\,[du^2 + dv^2].} \tag{4.14}$$

This is such a strong restriction that one might fear that such maps might not even exist. But as Gauss discovered in 1822, it is *always* possible to draw such a map of a general surface, at least locally. Remarkably, the proof (see Ex. 8) depends on *complex numbers*—indeed, the deep connection between Complex Analysis and conformal maps is the subject of the next section.

The already elegant curvature formula (4.10) now becomes even *more* elegant.[8] Recall that the second-order ***Laplacian***[9] differential operator ∇^2 is defined by

$$\boxed{\nabla^2 \equiv \partial_u^2 + \partial_v^2.} \tag{4.15}$$

[8] According to Dombrowski (1979, p. 128), this was the very *first* curvature formula that Gauss discovered, recorded in his private notes, dated 13th of December, 1822; see Gauß (1973, p. 381). Only in 1825 did he find a formula in more general (nonconformal) orthogonal coordinates, but, again, he kept this secret. Finally, in 1827, he found the completely general formula in a nonorthogonal coordinate system, and *only* this horrendously complicated and ugly formula appears in print in the final masterpiece (Gauss 1827). We note that Gauss was positively *proud* of his deliberate and utterly *deplorable* (in our view) habit of concealing his motivations and his path of discovery, declaring, "No architect leaves the scaffolding in place after completing the building."

[9] While physicists tend to denote this operator ∇^2, mathematicians instead tend to write it as Δ. We have avoided the latter notation because we often use Δ to denote a triangle.

Since now $A = B = \Lambda$, we easily find [exercise] that (4.10) simplifies to

$$\boxed{\mathcal{K} = -\frac{\nabla^2 \ln \Lambda}{\Lambda^2}.} \tag{4.16}$$

4.6 Some Visual Complex Analysis

Even if we take the surface $\mathcal{S}$ to be the plane, so that we are dealing with conformal mappings of the plane to itself, this turns out to be a deep and rich field of study. Remarkably, such conformal mappings are inextricably intertwined with the complex numbers, as we shall briefly explain in this section. (Additional concrete examples will appear later in this Act.) For a *full* exploration of how the spectacular results of Complex Analysis emerge from this geometric foundation, we shall immodestly recommend our first book, *Visual Complex Analysis* (VCA).

Not only does every surface $\mathcal{S}$ possess a conformal map, it possesses an *infinite variety* of conformal maps! We begin by noting that there is nothing unique about the specific u-curves and orthogonal v-curves that yield a conformal (u, v)-coordinate system with metric (4.14):

$$d\hat{s} = \Lambda(u, v)\, ds.$$

The real magic resides within the conformal *mapping* $F : \mathbb{C} \to \mathcal{S}$ itself. Given such a conformal mapping F, we may create an infinite variety of conformal $(\tilde{u}, \tilde{v})$-coordinates on $\mathcal{S}$ simply by rotating, expanding, and translating our (u, v)-coordinate grid in $\mathbb{C}$, yielding (by application of F) completely new $\tilde{u}$-curves and orthogonal $\tilde{v}$-curves on $\mathcal{S}$. And this new, orthogonal $(\tilde{u}, \tilde{v})$-coordinate system on $\mathcal{S}$ is just as conformal as the original one was.

Let us introduce some notation to explain this fully. As is customary in Complex Analysis, we shall think of $z = u + iv$ as living in one copy of $\mathbb{C}$, and its image $\tilde{z} = \tilde{u} + i\tilde{v}$ under a complex function $z \mapsto \tilde{z} = f(z)$ as living in a separate copy of $\mathbb{C}$:

$$z \mapsto \tilde{z} = \tilde{u} + i\tilde{v} = f(z) = f(u + iv).$$

In the example just discussed, we have $f(z) = ae^{i\tau}z + w$, which does an expansion by the (real) factor a, a rotation through angle τ, followed by a translation by the (complex) number w.

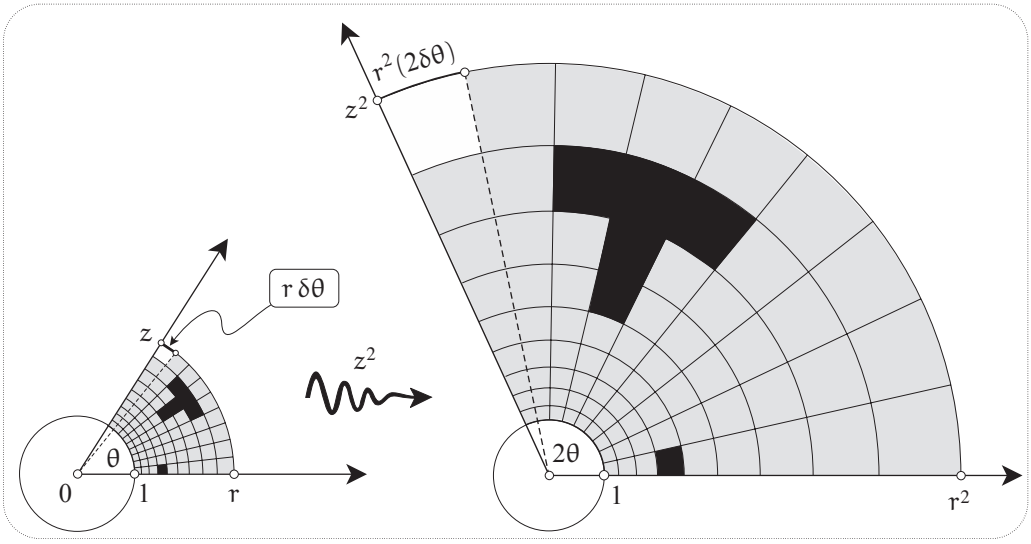
If we compose this mapping f with the mapping F then we obtain the new conformal mapping $\tilde{F} \equiv F \circ f$ from $\mathbb{C}$ to $\mathcal{S}$. If z moves along a small complex number δz then its image under the first mapping $z \mapsto \tilde{z} = f(z)$ will move along what we shall call δz's "image" $\delta\tilde{z}$, emanating from $\tilde{z}$, and clearly

$$\delta\tilde{z} = ae^{i\tau}\delta z\,. \tag{4.17}$$

Thus the length $\delta s = |\delta z|$ is stretched by factor a, so that $\delta\tilde{s} = |\delta\tilde{z}| = a\,\delta s$. Next, under the second mapping F (up to the surface $\mathcal{S}$), the length of $\delta\tilde{s}$ will be stretched by the conformal metric factor $\Lambda(\tilde{z})$. Thus the net magnification of δs under the new mapping $\tilde{F}$ is given by the *product* of these two expansion factors:

$$d\hat{s} = \widetilde{\Lambda}(z)\, ds, \qquad \text{where} \qquad \widetilde{\Lambda}(z) = a\,\Lambda(\tilde{z}).$$

But even this freedom barely scratches the surface, and to explain why we must simply quote one fundamental fact from Complex Analysis; see VCA for details. *Every* familiar, useful real function $f(x)$ you have ever studied (such as x^m, e^x, and $\sin x$) has a unique generalization $f(z)$ to complex number inputs z. Picturing this, as before, as a mapping from one $\mathbb{C}$ to another $\mathbb{C}$, the miracle of Complex Analysis is that *all* of these naturally occurring mappings $f(z)$ are automagically *conformal*!



[4.6] *The mapping* $z \mapsto f(z) = z^2$ *(in common with* all *powers of* z*) is* conformal*, so the fine grid of "squares" on the left is mapped to approximate squares on the right, and these ultimately become perfect squares in the limit that the size of the squares shrinks to zero.*

For example, [4.6] illustrates the effect of

$$z \mapsto \widetilde{z} = f(z) = z^2 = \left[r\,e^{i\theta}\right]^2 = r^2\,e^{i2\theta},$$

which squares the length, and doubles the angle, of every point. As we see, the small "squares" on the left are sent to *similar* "squares" on the right. Of course both sets of "squares" only become *actual* squares in the limit that they shrink to zero. Likewise, the smaller the black T-shape on the left, the more perfectly similar is the image T-shape on the right.

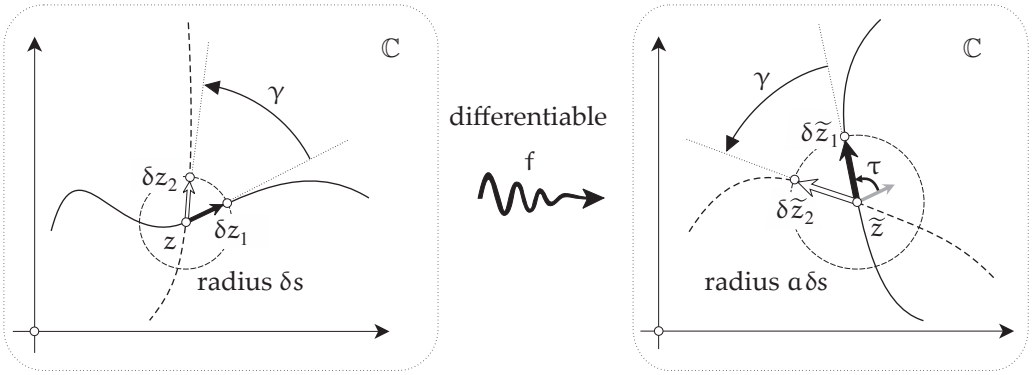
To see how remarkable this is, suppose you were to randomly write down two real functions $\widetilde{u}$ and $\widetilde{v}$ (of the real variables u and v), and then perform a shotgun wedding between these two functions to amalgamate them into a single complex mapping, $f(u,v) = \widetilde{u} + i\widetilde{v}$. Then there is essentially *no chance* that f will be conformal, and, as we shall now see, this also means that there is no chance that $f'(z)$ exists!

Let us begin to redo the earlier analysis, but now replacing the linear function $f(z)$ with a *general* mapping *such that the derivative* $f'(z)$ *exists*. Such functions are called ***analytic***. As we have just indicated, such analytic mappings are extraordinarily rare, and *yet* they include essentially *all* the useful functions that naturally arise out of mathematics and physics.

The analogue of (4.17) is now

$$\boxed{\delta\widetilde{z} \asymp f'(z)\,\delta z = a e^{i\tau}\,\delta z\,.} \tag{4.18}$$

The main difference is that now both the expansion factor $a(z)$ and the angle of rotation $\tau(z)$ both depend on the location z, rather than being constant throughout $\mathbb{C}$. For example, in [4.6] we can see that the black square adjacent to the real axis undergoes a smaller expansion than does the white square in the corner of the grid, so $a(z)$ is smaller at the black square than it is at the white square. Likewise, it is clear that the black square undergoes *no* rotation, so $\tau = 0$ there, but the white square clearly *does* need to be rotated to obtain its image.



[4.7] The Amplitwist. *The local effect of a differentiable complex mapping* $f(z)$ *is an* **amplitwist** *(***amplification** *and* **twist***) described by the complex number* $f'(z) = \text{amplitwist}$ *of* $f(z) = (\text{amplification})\, e^{i\,(\text{twist})} = \alpha e^{i\tau}$. *Any pair of tiny complex numbers* $\delta z_{1,2}$ *emanating from* z *are* **amplitwisted** equally *to produce a pair* $\delta\widetilde{z}_{1,2} \asymp \alpha e^{i\tau}\,\delta z_{1,2}$ *emanating from* $\widetilde{z} = f(z)$*, and thus the angle* γ *between curves is preserved:* f *is* conformal!

Equation (4.18) says that every tiny complex arrow δz emanating from z undergoes the *same* expansion α and rotation τ to obtain its image arrow $\delta\widetilde{z}$ emanating from $\widetilde{z} = f(z)$. See [4.7]. As illustrated, the angle between a pair of δz emanating from z will be the same as the angle between the image pair of $\delta\widetilde{z}$ at $\widetilde{z} = f(z)$, so it follows that differentiable complex mappings are automatically *conformal!*

We see that the derivative $f'(z)$ describes the local behaviour of the conformal mapping, and in VCA we introduced (*nonstandard*) terminology to describe this geometrically. We call the local expansion factor α the ***amplification***; we call the local angle of rotation τ the ***twist***; and we call the combined amplification and twist (needed to transform the original shape into its image) the ***amplitwist***. In summary,

$$\boxed{f'(z) = \text{amplitwist of } f \text{ at } z = (\text{amplification})\, e^{i[\text{twist}]} = \alpha\, e^{i\tau}.} \tag{4.19}$$

Before we turn to conformal metric formulas for surfaces in space, let us return to $f(z) = z^2$ and show how we may use [4.6] to geometrically deduce its amplitwist.

Focus attention on the illustrated white square, with one corner at $z = re^{i\theta}$. Since the radial edge through z at angle θ is mapped to a radial edge through z^2 at angle 2θ, this edge has undergone a rotation of θ, so

$$\tau = \text{twist} = \theta.$$

To find the amplification α, consider the highlighted outer edge of the square (ultimately equal to the arc of the circle though z, connecting the white dots). If this subtends angle $\delta\theta$ at the origin, then its length is ultimately $r\,\delta\theta$. Since angles are doubled by the mapping, the image arc will subtend angle $2\delta\theta$, and since it now lies on a circle of radius r^2, the length of this image edge is ultimately $r^2(2\delta\theta)$. Thus,

$$(\text{image edge}) \asymp 2r\,(\text{original edge}) \implies \alpha = \text{amplification} = 2r.$$

We conclude that,

$$(z^2)' = \text{amplitwist of } z^2 = (\text{amplification})\, e^{i[\text{twist}]} = 2r\, e^{i\theta} = 2z,$$

a result that looks formally identical to the real result, $(x^2)' = 2x$, but which now means so much more.

It is straightforward [exercise] to generalize this geometric argument to deduce that $(z^m)' = m\,z^{m-1}$. The amplitwists of other important mappings can likewise be deduced purely geometrically; see VCA for details.

Now let us return to our main interest: conformal coordinates on a surface. We can now replace our simple linear function with our enormously richer class of differentiable (i.e., conformal) mappings, $f(z)$. Once again defining $\widetilde{F} \equiv F \circ f$ from $\mathbb{C}$ to $\mathcal{S}$, the new metric formula is

$$d\widehat{s} = \widetilde{\Lambda}(z)\, ds, \qquad \text{where} \qquad \widetilde{\Lambda}(z) = (\text{amplification})\, \Lambda(\widetilde{z}) = |f'(z)|\, \Lambda(\widetilde{z}).$$

The existence of a conformal mapping $F : \mathbb{C} \mapsto \mathcal{S}$ allows us to take any analytic mapping $f : \mathbb{C} \mapsto \mathbb{C}$, and *transfer* it to $\mathcal{S}$, so that it becomes a conformal mapping of $\mathcal{S}$ to itself. To see this, consider the effect of

$$\widehat{f} \equiv F \circ f \circ F^{-1}, \tag{4.20}$$

acting on $\mathcal{S}$. First F^{-1} conformally maps $\mathcal{S}$ to $\mathbb{C}$, then f conformally maps $\mathbb{C}$ to $\mathbb{C}$, and finally F conformally maps $\mathbb{C}$ back to $\mathcal{S}$. Since each of the three mappings preserves angles, so do does their composition, so $\widehat{f} : \mathcal{S} \mapsto \mathcal{S}$ is indeed conformal.

In the next section we shall meet an extremely important example of a conformal mapping F in the case where $\mathcal{S} = S^2$. Later, we will use this F, via (4.20), to express the rotations of S^2 as complex functions (given by (6.10)). These rotations are not merely conformal, they are the (orientation-preserving) ***isometries*** of the sphere.

4.7 The Conformal Stereographic Map of the Sphere

Hipparchus (c. 150 BCE) may have been the first to construct a conformal[10] map of the sphere using the method illustrated in [4.8], which is called ***stereographic projection***. Certainly, by 125 CE Ptolemy (who is instead usually credited as the discoverer) was using it to plot the positions of heavenly bodies on the ***celestial sphere***.

The construction is similar to central projection, except now we imagine a point source of light at the north pole N, and we project onto a plane that passes through the equator, instead of touching the south pole.[11] Then a light ray passing through $\widehat{z}$ on Σ goes on to hit a point z in $\mathbb{C}$, which we call the ***stereographic image*** of $\widehat{z}$. Since this gives us a one-to-one correspondence between points in $\mathbb{C}$ and points on Σ, let us also say that $\widehat{z}$ is the stereographic image of z. No confusion should arise from this, the context making it clear whether we are mapping $\mathbb{C}$ to Σ, or vice versa.

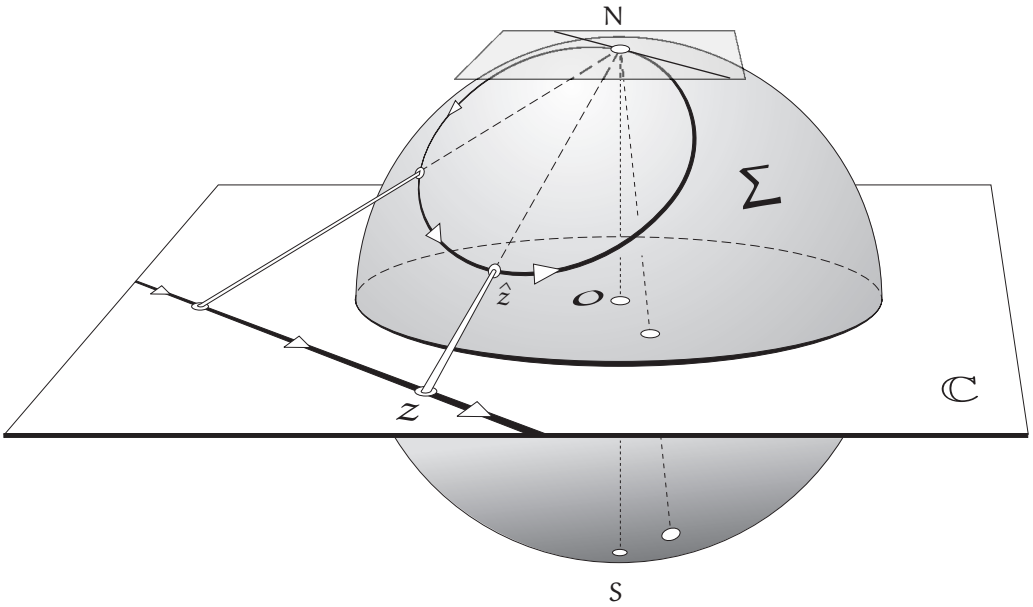
Note the following immediate facts: (i) the southern hemisphere of Σ is mapped to the interior in $\mathbb{C}$ of the circle $|z| = R$, and in particular the south pole S is mapped to the origin 0 of $\mathbb{C}$; (ii) each point on the equator is mapped to itself; (iii) in $\mathbb{C}$, the exterior of the circle $|z| = R$ is mapped to the northern hemisphere of Σ, *except* that N is not the image of any finite point in the plane. However, it is clear that as z moves further and further away from the origin (in any direction), $\widehat{z}$ moves closer and closer to N. In Complex Analysis, and later in this Act, one takes Σ to be the *unit* sphere, and once its points are stereographically labelled with complex numbers, it is called the ***Riemann sphere***. The point N is then the concrete manifestation of the ***point at infinity*** of the so-called ***extended complex plane***.

Figure [4.8] illustrates the fact that

---

[10] The conformality of the construction in [4.8] is *not* meant to be immediately obvious; it will be explained shortly, in [4.9].

[11] In some older texts the plane *is* in fact taken to touch the south pole; this [exercise] alters the map only by a constant scale factor of 2.



[4.8] Stereographic Projection: *light from* N *shines through the glass sphere* Σ*, casting shadows onto* ℂ*. The shadow of a circle through* N *is a line in* ℂ *that is parallel to the circle's tangent at* N.

> *The stereographic image of a line in the plane is a circle on* Σ *that passes through* N *in a direction parallel to the original line.* (4.21)

To see this, first observe that as z moves along the line shown in [4.8], the line connecting N to z sweeps out part of a plane through N. Thus $\hat{z}$ moves along the intersection of this plane with Σ, which is a circle passing through N. Next, note that the tangent plane to Σ at N is parallel to ℂ. But if we slice through two parallel planes with a third plane, we obtain two parallel lines of intersection. Therefore, the tangent at N to the circle is parallel to the original line, as claimed.

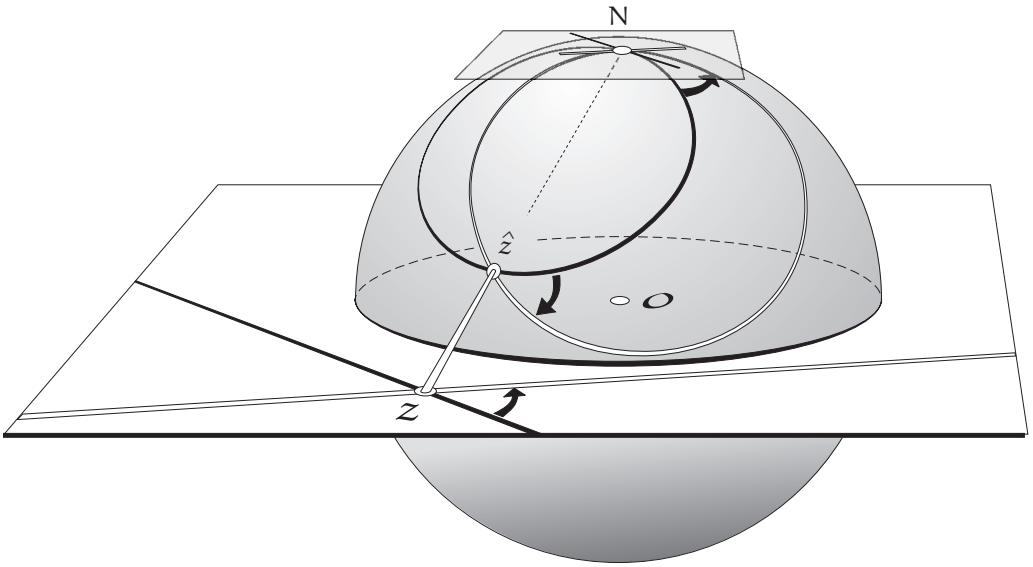
From this last fact it follows that *stereographic projection preserves angles.* Consider [4.9], which shows two lines intersecting at z, together with their circular, stereographic projections, passing through N.

First note that the magnitude of the angle of intersection between the circles is the *same* at their two intersection points, $\hat{z}$ and N. This is because the figure has *mirror symmetry* in the plane passing through the centre of the sphere and the centres of the two circles.[12] Since the tangents to the circles at N are parallel to the original lines in the plane, it follows that the illustrated angles at z and $\hat{z}$ are of equal *magnitude*. But before we can say that stereographic projection is "conformal," we must assign a ***sense*** to the angle on the sphere.

According to our convention, the illustrated angle at z (from the black curve to the white one) is *positive*; i.e., it is counterclockwise when viewed from above the plane. From the perspective from which we have drawn [4.9], the angle at $\hat{z}$ is negative (clockwise). However, if we were looking at this angle from *inside* the sphere then it would be positive. Thus,

> *If we define the sense of an angle on* Σ *by its appearance to an observer inside* Σ*, then stereographic projection is* conformal.

[12] This will become crystal clear if you draw for yourself any two intersecting circles on an orange.



[4.9] Stereographic Projection Is Conformal. *As the line rotates around* z*, the tangent at* N *of its circular image rotates with it, so the angles at* z *and* N *are equal. But, by symmetry, the angle at* $\hat{z}$ *equals the angle at* N*, so this equals the angle at* z*, proving that stereographic projection is conformal.*

HISTORICAL NOTE: It is remarkable that although stereographic projection had been well known since Ptolemy first put it to practical use around 125 CE, its beautiful and fundamentally important *conformality* was not discovered for another 1500 years! This was first done around 1590 by Thomas Harriot[13]—yes, the *same* Thomas Harriot who in 1603 discovered the fundamental formula (1.3) linking angular excess and area on the sphere!

It follows from conformality that the metric takes the form (4.13). Thus a very small circle of radius $\delta\hat{s}$ on the sphere is ultimately mapped to a circle of radius δs in the plane, where

$$\delta\hat{s} \asymp \Lambda\, \delta s.$$

See [4.10]. Our task now is to find Λ.

Since Λ is independent of the direction of radius $\delta\hat{s}$ emanating from $\hat{z}$, we are free to choose this direction so as to make the analysis of the geometry as simple as possible. Thus, as illustrated, let us choose the direction to be horizontal, along a circle of latitude.

Under the stereographic projection, this circle of latitude undergoes a uniform expansion to yield an origin-centred circle through z, with δs pointing along it. As $\hat{z}$ rotates around its circle of latitude, z rotates with it, and their motions are in proportion to their distances from N. Thus,

$$\frac{\delta\hat{s}}{\delta s} = \frac{N\hat{z}}{Nz}.$$

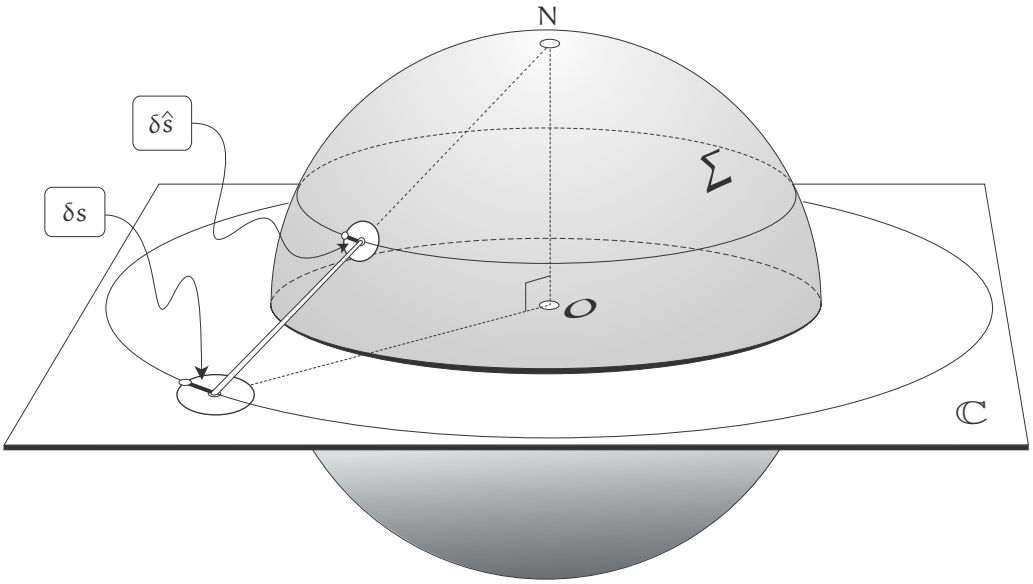
Now consider [4.11], which shows the vertical cross section of [4.10] taken through N, $\hat{z}$, z. The triangle $N\hat{z}S$ is similar to N0z, so

$$\frac{N\hat{z}}{2R} = \frac{R}{Nz}.$$

Combining this with the previous equation, we deduce that

$$\frac{\delta\hat{s}}{\delta s} = \frac{2R^2}{[Nz]^2}.$$

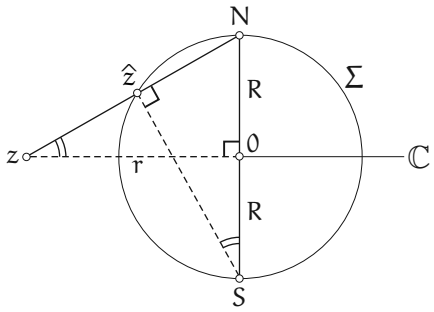
[13]See Stillwell (2010, §16.2). For a short sketch of Harriot's life, see Stillwell (2010, §17.7).



[4.10] *Conformality implies that a very small circle of radius $\delta\hat{s}$ on the sphere is ultimately mapped to a circle of radius δs in the plane. To find the metric, we must find the ratio of their radii, which we do by choosing $\delta\hat{s}$ along the illustrated circle of latitude.*

Finally, writing $r = |z|$, and applying Pythagoras's Theorem to the triangle N0z, we obtain $[Nz]^2 = R^2 + r^2$, so the conformal metric for the stereographic map is

$$d\hat{s} = \frac{2}{1 + (r/R)^2}\, ds. \tag{4.22}$$



[4.11] *The triangle $N\hat{z}S$ is similar to $N0z$.*

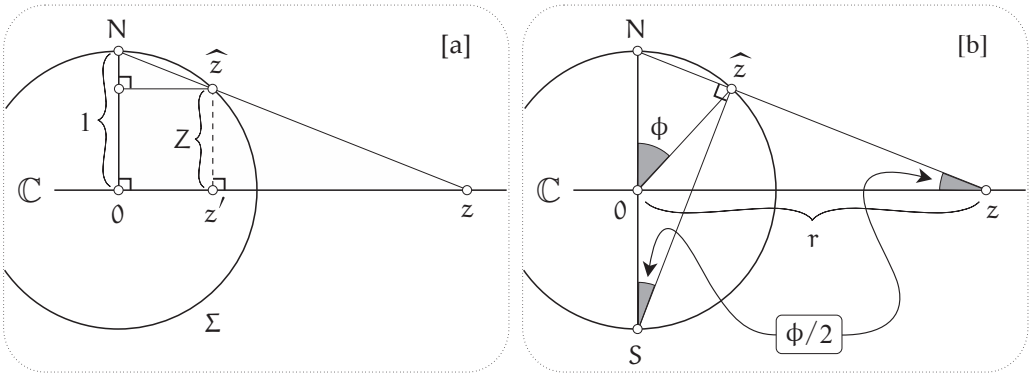
Here is our first opportunity (there will be others) to try out the conformal curvature formula (4.16); of course we should find that $\mathcal{K} = (1/R^2)$. We suggest you confirm this in two ways. First, write $r^2 = x^2 + y^2$ and use our original Cartesian form of the Laplacian operator, (4.15). Second, use the fact that in polar coordinates the Laplacian instead takes the form

$$\nabla^2 = \partial_r^2 + \frac{1}{r}\partial_r + \frac{1}{r^2}\partial_\theta^2. \tag{4.23}$$

4.8 Stereographic Formulas

In this section we derive explicit formulas connecting the coordinates of a point $\hat{z}$ on Σ and its stereographic projection z in $\mathbb{C}$. To simplify matters, *we now restrict ourselves to the standard case* $R = 1$.

To begin with, let us describe z with Cartesian coordinates: $z = x + iy$. Similarly, let (X, Y, Z) be the Cartesian coordinates of $\hat{z}$ on Σ; here the X- and Y-axes are chosen to coincide with the x- and



[4.12] [a] *By similar triangles,* $|z|/|z'| = 1/(1-Z)$. **[b]** $r = \cot(\phi/2)$.

y-axes of $\mathbb{C}$, so that the positive Z-axis passes through N. To make yourself comfortable with these coordinates, check the following facts: the equation of Σ is $X^2 + Y^2 + Z^2 = 1$, the coordinates of N are $(0,0,1)$, and similarly $S = (0,0,-1)$, $1 = (1,0,0)$, and $i = (0,1,0)$.

Now let us find the formula for the stereographic projection $z = x + iy$ of the point $\hat{z}$ on Σ in terms of the coordinates (X,Y,Z) of $\hat{z}$. Let $z' = X + iY$ be the foot of the perpendicular from $\hat{z}$ to $\mathbb{C}$. Clearly, the desired point z is in the same direction as z', so

$$z = \frac{|z|}{|z'|}\, z'.$$

Now look at [4.12a], which shows the vertical cross section of Σ and $\mathbb{C}$ taken through N and $\hat{z}$; note that this vertical plane necessarily also contains z' and z. From the similarity of the illustrated right triangles with hypotenuses $N\hat{z}$ and Nz, we immediately deduce [exercise] that

$$\frac{|z|}{|z'|} = \frac{1}{1-Z}\ ,$$

and so we obtain our first stereographic formula:

$$x + iy = \frac{X + iY}{1 - Z}. \tag{4.24}$$

For later use, let us now also invert this formula to find the coordinates of $\hat{z}$ in terms of those of z. Since [exercise]

$$|z|^2 = \frac{1+Z}{1-Z}\ ,$$

we obtain [exercise]

$$\boxed{X + iY = \frac{2z}{1+|z|^2} = \frac{2x + i2y}{1 + x^2 + y^2}, \quad \text{and} \quad Z = \frac{|z|^2 - 1}{|z|^2 + 1}.} \tag{4.25}$$

Although it is often useful to describe the points of Σ with the three coordinates (X,Y,Z), this is certainly unnatural, for the sphere is intrinsically 2-dimensional. If we instead describe $\hat{z}$ with the more natural (2-dimensional) spherical polar coordinates (ϕ, θ) then we obtain a particularly neat stereographic formula.

First recall that θ measures the angle around the Z-axis, with $\theta = 0$ being assigned to the vertical half-plane through the positive X-axis. Thus for a point z in $\mathbb{C}$, the angle θ is simply the usual angle from the positive real axis to z. The definition[14] of ϕ is illustrated in [4.12b]—it is the angle subtended at the centre of Σ by the points N and $\widehat{z}$; for example, the equator corresponds to $\phi = (\pi/2)$. By convention, $0 \leqslant \phi \leqslant \pi$.

If z is the stereographic projection of the point $\widehat{z}$ having coordinates (ϕ, θ), then clearly $z = r\,e^{i\theta}$, and so it only remains to find r as a function of ϕ. From [4.12b] it is clear [exercise] that the triangles $N\,\widehat{z}\,S$ and $N0\,z$ are similar, and because the angle $\angle NS\,\widehat{z} = (\phi/2)$, it follows [exercise] that $r = \cot(\phi/2)$. Thus our new stereographic formula is

$$\boxed{z = \cot(\phi/2)\, e^{i\theta}.} \tag{4.26}$$

In Exercise 33 we show how this formula may be used to establish a beautiful alternative interpretation of stereographic projection, due to Sir Roger Penrose.

We will now illustrate this formula with an application that we shall need shortly: the relationship between the complex numbers representing ***antipodal points***. Recall that this means two points on a sphere that are diametrically opposite each other, such as the north and south poles. Let us show that

If $\widehat{p}$ and $\widehat{q}$ are antipodal points of Σ, then their stereographic projections p and q are related by the following formula:

$$q = -(1/\overline{p}). \tag{4.27}$$

Note that the relationship between p and q is actually symmetrical (as clearly it should be): $p = -(1/\overline{q})$. To verify (4.27), first check for yourself that if $\widehat{p}$ has coordinates (ϕ, θ) then $\widehat{q}$ has coordinates $(\pi - \phi, \pi + \theta)$. Thus,

$$q = \cot\left[\frac{\pi}{2} - \frac{\phi}{2}\right] e^{i(\pi+\theta)} = -\frac{1}{\cot(\phi/2)}\, e^{i\theta} = -\frac{1}{\cot(\phi/2)\, e^{-i\theta}} = -\frac{1}{\overline{p}},$$

as was to be shown.

For an elementary geometric proof of (4.27), see Exercise 7.

4.9 Stereographic Preservation of Circles

This section is devoted to proving a single fact that is beautiful, surprising, and critically important:

Stereographic projection preserves circles! (4.28)

That is, not only does it send infinitesimal circles to infinitesimal circles (as all conformal maps must do) but it sends a finite circle of any size and location on the sphere to a perfect circle in

[14] This is the American convention; in my native England the roles of θ and ϕ are the reverse of those stated here. For a lovely example of the same diagram, but labelled per British convention, see Penrose and Rindler (1984, Vol. 1, p. 12).

the plane, although the centre on the sphere is *not* mapped to the centre in the plane. Note that if the circle passes close to N then its image will be extremely large, and in the limit that it passes through N this very large circle becomes the line shown in [4.8].

There exists a beautiful, completely conceptual, geometric explanation of (4.28)—see VCA, page 142—but in our current haste we must instead settle for a calculation.

Every circle on the unit sphere Σ is the intersection of Σ with a plane whose distance from O is less than one:

$$lX + mY + nZ = k, \qquad \text{where} \qquad l^2 + m^2 + n^2 \geqslant k^2.$$

Substituting (4.25) into the equation of this plane we find [exercise] that the circle of intersection on Σ stereographically projects to a curve in C whose equation is

$$2lx + 2my + n(x^2 + y^2 - 1) = k(x^2 + y^2 + 1).$$

If $k = n$ then [exercise] the circle on Σ passes through N and its image is a line (as it should be!) with equation $lx + my = n$. If $k \neq n$ then we may complete squares [exercise] to rewrite the equation as

$$\left[x - \frac{l}{k-n}\right]^2 + \left[y - \frac{m}{k-n}\right]^2 = \frac{l^2 + m^2 + n^2 - k^2}{(k-n)^2},$$

which is indeed a circle with

$$\text{centre} = \left(\frac{l}{k-n}, \frac{m}{k-n}\right) \qquad \text{and} \qquad \text{radius} = \frac{\sqrt{l^2 + m^2 + n^2 - k^2}}{|k-n|}.$$

(Comment: This is an excellent example of the seductive but corrupting power of calculation. We invoked the "the Devil's machine"—see Prologue—and in just a couple of lines its work was complete and the result was proved. Yet here we stand, bereft of any understanding of *why* the result is true!)

From the preservation of circles we readily deduce from [1.5] that geodesics on the sphere (the great circles) appear in the map as circles that intersect the equatorial circle at opposite ends of a diameter.

Chapter 5

The Pseudosphere and the Hyperbolic Plane

5.1 Beltrami's Insight

While the long investigation of parallel lines reached a climax around 1830 with the discovery of Hyperbolic Geometry by Lobachevsky and Bolyai, a quite different parallel line—pun intended!—of enquiry into Differential Geometry also reached a climax at almost the same time, with Gauss's differential-geometric discoveries of 1827.

Just as lines on the sphere that are initially parallel[1] must eventually intersect, so too did these two parallel lines of thought collide, and in a powerful and fruitful way.

In 1868 the Italian geometer Eugenio Beltrami (see [5.1]) recognized that a connection might exist between two results from these seemingly unrelated realms of thought. On the one hand, he knew of Lambert's result (1.8)—later rediscovered by Gauss, Lobachevsky, and Bolyai—that in Hyperbolic Geometry the angular excess of a triangle is a fixed *negative* multiple of its area. On the other hand, he also knew of the Local Gauss–Bonnet Theorem.

Beltrami had the insight that if one could find a surface of *constant negative curvature* $\mathcal{K} = -(1/\mathsf{R}^2)$, then, by virtue of (2.6), geodesic triangles constructed within it would automatically obey the central law of Hyperbolic Geometry:

$$\mathcal{E}(\Delta) = -\frac{1}{\mathsf{R}^2}\,\mathcal{A}(\Delta).$$

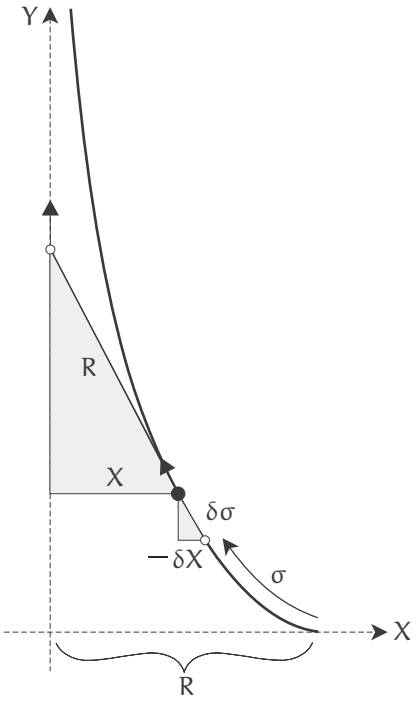
[5.1] *Eugenio Beltrami (1835–1900).*

Up to this point, the bizarre Hyperbolic Geometry of Lobachevsky and Bolyai had languished in obscurity for close to 40 years, vilified by some, but ignored by most. Now, at last, Beltrami had an idea that could put it on a secure and intuitive foundation. Perhaps Hyperbolic Geometry simply *was* the intrinsic geometry of a surface of constant negative curvature! A 2000-year-old struggle was about to come to an end.

[1] Think of two neighbouring points on the equator, and the meridians passing through them.

5.2 The Tractrix and the Pseudosphere

Beltrami already knew that the pseudosphere shown in [2.6] was indeed a surface of constant negative curvature $\mathcal{K} = -(1/R^2)$, where R is the radius of its circular bottom rim. (We shall prove this fact within this section.) Thus, more concretely, his insight was that the local geometry within this surface obeys the laws of the abstract non-Euclidean Geometry of Lobachevsky and Bolyai. But this abstract Hyperbolic Geometry is understood to take place in an infinite ***hyperbolic plane*** that is exactly like the Euclidean plane, obeying the first four of Euclid's axioms, *but* with lines that obey the Hyperbolic Axiom (1.1), instead of Euclid's Parallel Axiom.



[5.2] The Tractrix. *A weight is attached to a string of length* R*, laid out along the* X*-axis. If the free end is moved up the* Y*-axis, the weight is dragged along the illustrated curve, called the* **tractrix**.

While the constant negative curvature of the pseudosphere ensures that it faithfully embodies local consequences of this axiom, the pseudosphere will not do as a model of the *entire* hyperbolic plane, because it departs from the Euclidean plane in two unacceptable ways: (1) the pseudosphere is akin to a cylinder instead of a plane and (2) a line segment cannot be extended indefinitely in both directions: we hit the rim. (As we noted earlier, Hilbert discovered in 1901 that such a rim is an *essential* feature of *all* surfaces of constant negative curvature—not intrinsically, but by virtue of trying to force the surfaces to fit inside ordinary Euclidean 3-space.)

Beltrami recognized these obstacles, and in the next section we will see how he overcame them both, in one fell swoop, by constructing a conformal map of the pseudosphere. For now, though, we turn to the construction of the pseudosphere itself.

Try the following experiment. Take a small heavy object, such as a paperweight, and attach a length of string to it. Now place the object on a table and drag it by moving the free end of the string along the edge of the table. You will see that the object moves along a curve like that in [5.2], where the Y-axis represents the edge of the table. This curve is called the ***tractrix***, and the Y-axis (which the curve approaches asymptotically) is called the ***axis***. The tractrix was first investigated by Newton, in 1676.

If the length of the string is R, then it follows that the tractrix has the following geometric property: *the segment of the tangent from the point of contact to the* Y*-axis has constant length* R. Newton identified this as the defining property of the tractrix.

Returning to [5.2], let σ represent arc length along the tractrix, with $\sigma = 0$ corresponding to the starting position $X = R$ of the object we are dragging. Just as the object is about to pass through (X, Y), let δX denote the very small change in X that occurs while the object moves a distance $\delta\sigma$ along the tractrix. From the ultimate similarity of the illustrated triangles, we deduce that

$$\frac{-dX}{d\sigma} = \frac{X}{R},$$

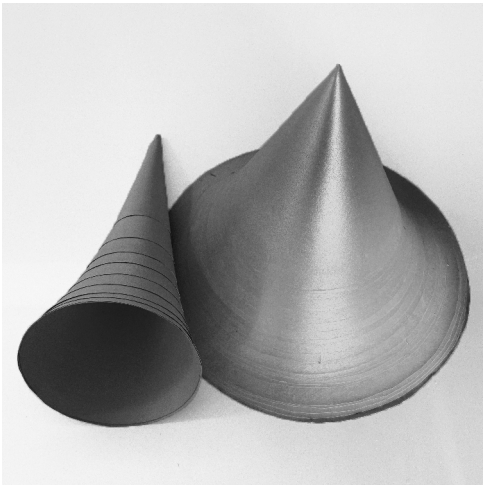
and therefore

$$\boxed{X = R\, e^{-\sigma/R}.} \tag{5.1}$$

The ***pseudosphere*** of radius R, [2.6], may now be simultaneously defined and constructed as the surface obtained by rotating the tractrix about its axis. Remarkably, this surface was investigated as early as 1693 (by Christiaan Huygens), two centuries prior to its catalytic role in the acceptance of Hyperbolic Geometry, and the constancy of its curvature was already known to Gauss's student Minding as early as 1839.

To demonstrate this constancy of the curvature, let us find the metric of the pseudosphere. We choose very natural orthogonal coordinate curves on the surface, as follows. See [5.4a] (but please ignore [5.4b] for now). To specify a point on the pseudosphere we need only say (i) which tractrix generator it lies on and (ii) how far along this tractrix it lies. To answer (i) we specify the angle x around the axis of the pseudosphere, and to answer (ii) we specify the distance σ travelled up the tractrix, starting at the base.

Thus the curves $x = const.$ are the tractrix generators of the pseudosphere (note that these are clearly geodesics[2]), and the curves $\sigma = const.$ are circular cross sections of the pseudosphere (note that these are clearly *not* geodesics). Since the radius of such a circle is the same thing as the X-coordinate in [5.2], it follows from (5.1) that



[5.3] *The author's personal pseudophere, built out of cones, themselves built out of discs of radius* R*. The top half has been removed to make the construction easier to see.*

> *The radius* X *of the circle* $\sigma = \text{const.}$ *passing through the point* (x, σ) *on the pseudosphere is given by* $X = R\,e^{-\sigma/R}$.

An increase dx therefore rotates the point through the arc $X\,dx$, as illustrated. Thus, the metric is

$$ds^2 = X^2\,dx^2 + d\sigma^2 = (R\,e^{-\sigma/R})^2\,dx^2 + d\sigma^2. \tag{5.2}$$

Finally, we may now apply the curvature formula (4.10) to this metric to obtain

$$\mathcal{K} = -\frac{1}{R\,e^{-\sigma/R}}\partial_\sigma\left[\frac{\partial_\sigma(R\,e^{-\sigma/R})}{1}\right] = -\frac{1}{R^2},$$

thereby confirming the crucial fact that Beltrami needed to interpret Hyperbolic Geometry:

> *The pseudosphere has constant negative curvature* $\mathcal{K} = -(1/R^2)$*, where* R *is its base radius.* (5.3)

This proposition carries such great mathematical and historical interest that in the course of this book we shall attempt to understand it as directly and geometrically as possible. Indeed, in

[2] By symmetry, the meridians of *any* surface of revolution are geodesics.

later Acts we shall offer *two* geometric proofs: one using extrinsic geometry (Act III), and one using intrinsic geometry (Act IV).

Before moving on, we can think of no better way to develop a feel (literally!) for the geometry of the pseudosphere than to *build your own!* To see the idea behind the construction, imagine rotating [2.6] to create the pseudosphere. In this process, no matter the position of the dragged weight, *the rotating string always traces out a cone (tangent to the pseudosphere) of fixed slant length* R.

Therefore, take as many sheets of paper as you can cut through with your scissors, and staple them together along three edges, a few staples per edge. Find the largest bowl or plate that will fit inside the paper, and trace its circular edge. Now cut along this circle to produce identical discs; repeat this step till you have at least 20 discs—the more, the better! Cut a small wedge from the first disc and tape the edges together to create a very shallow cone. Take the next disc and repeat with a slightly larger wedge[3] to create a slightly taller cone, but still with the same slant length. Place this new cone on top the previous cone and repeat, and repeat, ... Behold your own personal pseudosphere!

5.3 A Conformal Map of the Pseudosphere

As a first step towards creating a map of an infinite hyperbolic plane akin to the Euclidean plane, we now construct a conformal map of the pseudosphere in $\mathbb{C}$. In our map, let us choose the angle x as our horizontal axis, so that the tractrix generators of the pseudosphere are represented by vertical lines. See [5.4b]. Thus a point on the pseudosphere with coordinates (x, σ) will be represented in the map by a point with Cartesian coordinates (x, y), which we will think of as the complex number $z = x + iy$.

If our map were not required to be special in any way, then we could simply choose $y = y(x, \sigma)$ to be an arbitrary function of x and σ. But now let us insist that our map be *conformal*. Thus an infinitesimal[4] triangle on the pseudosphere is mapped to a *similar* infinitesimal triangle in the map, and more generally it follows that any small shape on the pseudosphere looks the same (only bigger or smaller) in the map. Having decided upon such a conformal map, it turns out there is (virtually) no freedom in the choice of the y-coordinate. Let us see why. First, the tractrix generators $x = const.$ are orthogonal to the circular cross sections $\sigma = const.$, so the same must be true of their images in our conformal map. Thus the image of $\sigma = const.$ must be represented by a horizontal line $y = const.$, and from this we deduce that $y = y(\sigma)$ must be a function solely of σ.

Second, on the pseudosphere consider the arc of the circle $\sigma = const.$ (of radius X) connecting the points (x, σ) and $(x + dx, \sigma)$. By the definition of x, these points subtend angle dx at the centre of the circle, so their separation on the pseudosphere is $X\,dx$, as illustrated. In the map, these two points have the same height and are separated by distance dx. Thus in passing from the pseudosphere to the map, this particular line segment is shrunk by factor X.

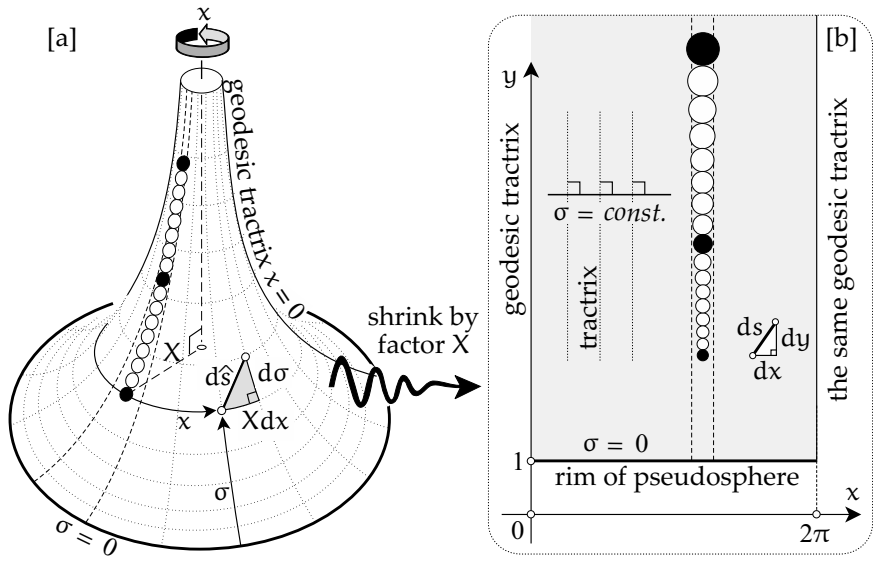
However, since the map is conformal, an infinitesimal line segment emanating from (x, σ) in *any* direction must be multiplied by the *same* factor $(1/X) = \frac{1}{R}e^{\sigma/R}$. In other words, the metric is

$$\boxed{d\hat{s} = X\,ds.}$$

Third, consider the uppermost black disc on the pseudosphere shown in [5.4a]. Think of this disc as infinitesimal, say of diameter ϵ. In the map, it will be represented by *another disc*, whose diameter (ϵ/X) may be interpreted more vividly as the angular width of the original disc as seen by an observer at the same height on the pseudosphere's axis. Now suppose we repeatedly

[3] In practice you may find it easier simply to cut a radial slit, then *overlap* the paper to create the cone.

[4] Here, for once, we shall prefer the intuitive abbreviation "infinitesimal," instead of spelling out the rigorous "ultimate equalities."



[5.4] *A conformal map of the pseudosphere. First, choose the* x*-coordinate to be the angle round the axis. Then conformality dictates the* y*-coordinate via the similarity of the illustrated infinitesimal triangles:* $\frac{dy}{d\sigma} = \frac{1}{X}$.

translate the original disc towards the pseudosphere's rim, moving it a distance ϵ each time. Figure [5.4a] illustrates the resulting chain of touching, congruent discs. As the disc moves down the pseudosphere, it recedes from the axis, and its angular width as seen from the axis therefore diminishes. Thus the image disc in the map appears to gradually shrink as it moves downward, and the equal distances 8ϵ between the successive black discs certainly do not appear equal in the map.

Having developed a feel for how the map works, let's actually calculate the y-coordinate corresponding to the point (x, σ) on the pseudosphere. From the above observations (or directly from the requirement that the illustrated triangles be similar) we deduce that

$$\frac{dy}{d\sigma} = \frac{1}{X} = \frac{1}{R} e^{\sigma/R} \qquad \Longrightarrow \qquad y = e^{\sigma/R} + \textit{const.}$$

The standard choice of this constant is 0, so that

$$y = e^{\sigma/R} = (R/X). \tag{5.4}$$

Thus the entire pseudosphere is represented in the map by the shaded region lying above the line $y = 1$ (which itself represents the pseudosphere's rim), and the metric associated with the map is

$$\boxed{d\hat{s} = \frac{R\,ds}{y} = \frac{R\sqrt{dx^2 + dy^2}}{y}.} \tag{5.5}$$

For future use, also note that an infinitesimal rectangle in the map with sides dx and dy represents a similar infinitesimal rectangle on the pseudosphere with sides $(R\,dx/y)$ and $(R\,dy/y)$. Thus the apparent area $dx\,dy$ in the map is related to the true area dA on the pseudosphere by

$$\boxed{dA = \frac{R^2\,dx\,dy}{y^2}.} \tag{5.6}$$

(Of course this is just a special case of (4.12), with $A = B = (R/y)$.)

5.4 The Beltrami–Poincaré Half-Plane

We now have a conformal map of the *cylinder-like, rimmed*, pseudosphere: $\{(x, y) : 0 \leqslant x < 2\pi,\ y \geqslant 1\}$. To instead create a map of an infinite hyperbolic plane, Beltrami knew that he must remove both of these adjectives. Note that while different choices of R yield quantitatively different geometries, they are all qualitatively the same, so there is no harm in making a specific choice:

> *In essentially all books and papers on Hyperbolic Geometry, the specific choice* $R = 1$ *is made, so that* $\mathcal{K} = -1$. *In this section we too shall make this conventional choice.*

If one wishes to return from this specific case to the general case, one need only insert the appropriate power of R into the special case ($R = 1$) formulas. For example, if you are dealing with area, then you must multiply by R^2.

To remove the "cylinder-like" adjective, imagine painting a wall with a standard cylindrical paint roller (of unit radius). After one revolution you have painted a strip of wall of width 2π, and every point on the surface of the roller has been mapped to a unique point within this strip of flat wall. To paint the entire wall, you can simply keep on rolling! Now imagine that our paint roller instead takes the form of a pseudosphere. To make it fit onto the flat wall you must first stretch out its surface, according to the metric (5.5), but then, just as before, you can keep on rolling (let's say horizontally). If a particle moves along a horizontal line on the wall, for example, the corresponding particle on the pseudosphere goes round and round the horizontal circle $\sigma = const.$ The "cylinder-like" adjective has been successfully removed[5] and we now have map $\{(x, y) : -\infty < x < \infty,\ y \geqslant 1\}$.

Our second problem, the pseudosphere's rim, is solved by the conformal map with equal ease. On the left of [5.5] is the image of a particle moving down the pseudosphere along a tractrix. Of course on the pseudosphere the journey is rudely interrupted at some point $\hat{p}$ on the rim ($\sigma = 0$), corresponding to a point p on the line $y = 1$. But in the map this point p is just like any other, and there is absolutely nothing preventing us from continuing all the way down to the point q on $y = 0$, with the true distances $d\hat{s}$ continuing to be given by the ***standardized hyperbolic metric***,

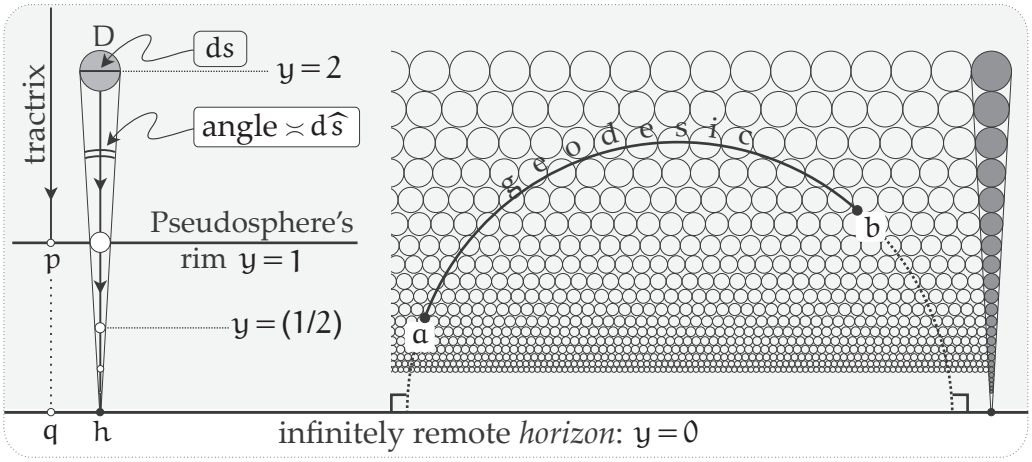
$$d\hat{s} = \frac{ds}{y}. \tag{5.7}$$

Why stop at q? The answer is that the particle will never even get that far, because q is infinitely far from p! Consider the small disc D of diameter ds on the line $y = 2$ shown on the left of [5.5]. Its true size on the pseudosphere is $d\hat{s} = \frac{ds}{y}$, and this is ultimately equal [exercise] to the illustrated angle it subtends at the point h directly below it on the line $y = 0$. Now imagine D moving down the pseudosphere at steady speed. Its apparent size in the map must shrink so that its subtends a constant angle at h. In the map it hits $y = 1 \ldots$ and keeps on going!

Assuming it took one unit of time to go from $y = 2$ to $y = 1$, then in the next unit of time it will reach $y = (1/2)$, then $y = (1/4)$, ..., for these points are all separated by the same hyperbolic distance:

$$\ln 2 = \int_1^2 \frac{dy}{y} = \int_{1/2}^1 \frac{dy}{y} = \int_{1/4}^{1/2} \frac{dy}{y} = \cdots.$$

[5] Stillwell (1996) notes that this was perhaps the first appearance in mathematics of what topologists now call a ***universal cover***.



[5.5] *The hyperbolic diameter of the disc* D *is the (Euclidean) angle it subtends on the horizon. Thus, as it moves downward, its image in the map shrinks. The discs on the right are all the same size, and a geodesic* a b *therefore passes through the smallest number of them.*

Thus, viewed within the map, the motion *slows down* and each successive unit of time only halves the distance from $y = 0$, and therefore D will never reach it. (An appropriate name for this phenomenon might be "Zeno's Revenge"!)

At last, we now possess a concrete model of

> ***The Hyperbolic Plane*** $\mathbb{H}^2$*: the entire shaded half-plane* $y > 0$*, with metric* $d\hat{s} = \frac{ds}{y}$. (5.8)

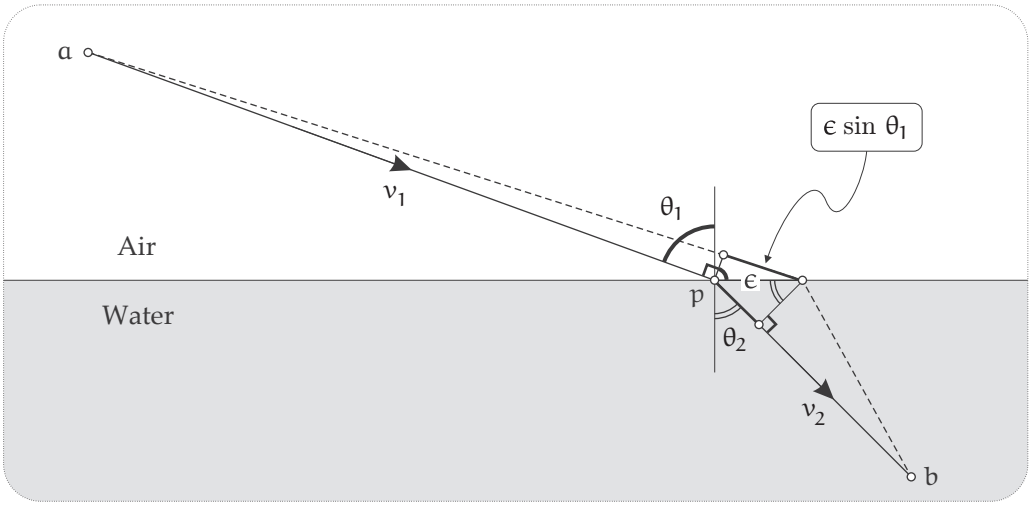
The points on the real axis $y = 0$ are infinitely far from ordinary points and are not (strictly speaking) considered part of the hyperbolic plane. They are called ***ideal points***, or ***points at infinity***. The complete line $y = 0$ of points at infinity is called the ***horizon***.

Although Beltrami discovered this map in 1868 (anticipating Poincaré by 14 years) it is now universally known as the ***Poincaré half-plane***. But in an attempt to restore some semblance of historical balance, we shall doggedly refer to this map as the ***Beltrami–Poincaré half-plane***.

Let us attempt to make the metric of this map more vivid. On the far right of [5.5] is a vertical string of touching circles of equal hyperbolic size ϵ, as in [5.4]. To the left of this we have filled part of the hyperbolic plane with such circles, *all of equal hyperbolic diameter* ϵ. Thus the hyperbolic length of any curve is ultimately equal to the number of circles it intercepts, multiplied by ϵ. This makes it clear that the shortest route from a to b is the one that intercepts the smallest number of circles, and therefore has the approximate shape shown.

If you followed our earlier advice and built your own model pseudosphere, you can also check the shape of a geodesic by stretching a string over its surface between two points at similar heights. To investigate geodesics over regions where you cannot stretch a string, such as along the tractrix generators themselves, you may instead employ the tape construction—(1.7), page 14—which works *everywhere*.

Our next task is to confirm the lovely fact illustrated in [5.5]: the exact form of such a geodesic is a perfect *semicircle that meets the horizon at right angles*. The only geodesics that are not of this form are the vertical half-lines (extending the tractrix generators), but even these may be viewed as a limiting case in which the radius of the semicircle tends to infinity.



[5.6] *Snell's Law: In order for light to take equal time to travel along the two neighbouring routes, the extra time in the air must be exactly compensated by the reduced time in the water:* $\sin\theta_1/v_1 = \sin\theta_2/v_2$.

5.5 Using Optics to Find the Geodesics

In order to explain the semicircular form of geodesics in the Beltrami–Poincaré half-plane model of Hyperbolic Geometry, we shall draw on ideas from physics, specifically from *optics*.[6] Our inspiration derives from a law called ***Fermat's Principle,***[7] discovered in 1662:

Light travels from one place to another in the least amount of time. (5.9)

We shall begin our brief excursion into physics by using Newtonian reasoning[8] to see how Fermat's Principle allows us to determine geometrically the abrupt bending of light (called ***refraction***) that occurs when it passes from air into water (for example). This is why [exercise] your spoon appears to bend when you stick it into a cup of tea.

In [5.6], a ray of light heads out from a in direction θ_1 to the vertical, travelling at speed v_1 through the air, hitting the water at p. It is then refracted into direction θ_2, travelling through the water at reduced speed v_2, finally arriving at b. As early as 130 CE, Ptolemy conducted such experiments and compiled a fairly accurate table of the pair of angles, θ_1 and θ_2. But the precise mathematical relationship between the two angles eluded Ptolemy, as it would continue to elude scientists for centuries to come.

Finally, in 1621 the Dutch mathematician Willebrord Snell van Royen (1591–1626) discovered the correct law, now universally known as ***Snell's Law***:[9]

---

[6] This approach does not seem to be widely known. I thank Sergei Tabachnikov for pointing out to me that it was previously published by Gindikin (2007, p. 324). The essential idea of using Fermat's principle to find a path that minimizes some quantity goes back to Johann Bernoulli's solution of the Brachistochrone Problem in 1697.

[7] First, this is the same Pierre de Fermat (1601–1665) famous for discoveries in number theory, including Fermat's Last Theorem. Second, Feynman discovered that there is a beautiful quantum-mechanical explanation of this principle, a masterful account of which can be found in Feynman (1985).

[8] We were honoured to realize that here we merely retraced the path taken by Feynman: see Feynman et al. (1963, Vol. 1, 26-3). Fermat himself first gave an analytic proof and later a geometric one, but neither is as elegant as the present Newtonian argument; both of Fermat's proofs can be found in Mahoney (1994, pp. 399–401).

[9] As usual, the history is far more complex than the name suggests. The same Thomas Harriot that we met earlier also discovered "Snell's Law," 20 years before Snell, but as with most of his discoveries Harriot was secretive, though this result he did communicate to Kepler. But even Harriot had been beaten to the result—by 600 years! The Islamic mathematician and physicist Ibn Sahl published it in 984 CE, and even used it to design sophisticated anaclastic lenses.

$$\sin\theta_1 = n\sin\theta_2, \quad \text{where } n = \textit{const.} \tag{5.10}$$

The value of n (called the ***index of refraction***) depends on the materials on either side of the interface, but for air/water $n \approx 1.33$.

It should be clear, at least qualitatively, that Fermat's Principle requires the light to bend. For if it were to travel in a straight line from a to b then it would waste valuable time travelling relatively slowly through the water, instead of travelling quickly through the air. Quantitatively, the amount of bending needed to minimize the flight time will occur when the derivative of the time (with respect to the position of p) is equal to zero.

Put geometrically, if the position of p minimizes the time, then an infinitesimal displacement ϵ of p should produce no change (to first order in ϵ) in the time. But, as we see in [5.6], this displacement causes the light to travel an additional distance through the air that is ultimately equal to $\epsilon\sin\theta_1$, increasing the flight time by an amount that is ultimately equal to $(\epsilon\sin\theta_1)/v_1$. On the other hand, by the same reasoning, the time *saved* in the water is ultimately equal to $(\epsilon\sin\theta_2)/v_2$. For the net time change to vanish, these two individual time changes must be equal. Thus, by cancelling ϵ,

$$\frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2}. \tag{5.11}$$

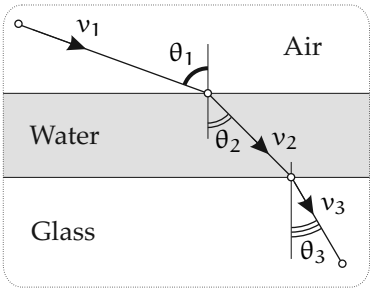
This not only proves Snell's Law (5.10), but it also makes a physical prediction, confirmed by direct experiment: the index of refraction is the ratio of the speed of light in the two materials, $n = (v_1/v_2)$.

Now suppose that our water sits at the bottom of a drinking glass. See [5.7]. How will the light bend when it gets to the bottom of the water and enters the glass base, where its speed is v_3? Applying the same law as before,

$$\frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2} = \frac{\sin\theta_3}{v_3}.$$

Thus, more generally, if we have many thin horizontal strips of material, the speed of light being v_i within each, then the entire journey of the light through the layers is governed by the law

$$\frac{\sin\theta_i}{v_i} = \textit{const.} = k.$$

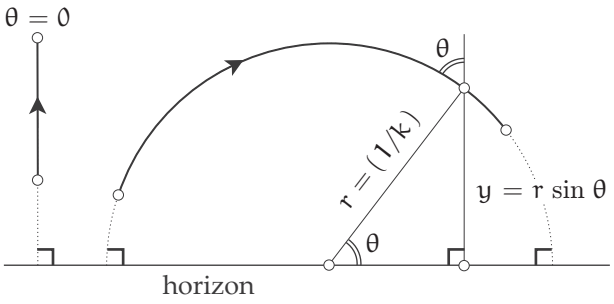


[5.7] *Snell's Law applied to multiple layers of material.*

Finally, let us go one step further and imagine the passage of light through a block of nonuniform material whose density is the same on each horizontal slice $y = \textit{const.}$, but which varies *continuously* as y varies, the speed of light at height y being $v(y)$, and the angle of the light there being $\theta(y)$. Then the Generalized Snell's Law is

$$\frac{\sin\theta(y)}{v(y)} = k. \tag{5.12}$$

All very interesting, you say, but what has this to do with proving the semicircular form of geodesics in the hyperbolic plane?! Well, suppose points $\widehat{a}$ and $\widehat{b}$ on the pseudosphere correspond to the points a and b in $\mathbb{H}^2$, [5.5]. Imagine a particle that travels along different routes over the surface of the pseudosphere from $\widehat{a}$ and $\widehat{b}$, but always at the same constant speed, say, 1. As the particle follows one of these paths, its image in the hyperbolic plane will follow a corresponding path from a to b, but *not* with uniform speed: as we saw in [5.5], if a particle moves down the pseudosphere at constant speed, its image in the hyperbolic plane *slows down*.



[5.8] *The hyperbolic geodesics must satisfy the Generalized Snell's Law* $(\sin\theta/y) = k$*, and are therefore semicircles and half-lines orthogonal to the horizon.*

The next crucial observation is that because the map is *conformal*, the slowing down only depends on the *location* of the particle, not on the direction in which it moves. For suppose we simultaneously launch from $\hat{a}$ a multitude of unit-speed particles in all directions. After infinitesimal time ϵ they will form a circle of radius ϵ centred at $\hat{a}$ on the pseudosphere, while in the hyperbolic plane (5.8) tells us they will also form a circle (centred at a) but of radius ϵy, where y is the height of a. In other words, the speed of the particles emanating from a is $v(y) = y$: the closer to the horizon we are, the slower the particles move.

Of course the time taken for each journey $\hat{a}\hat{b}$ on the pseudosphere will be the same as the time taken along ab in the hyperbolic plane. But on the pseudosphere, the geodesic route, being the shortest, will also be the path of least time, and therefore the geodesic in the hyperbolic plane is *also* the quickest route from a to b: *geodesic motion in the hyperbolic plane automatically obeys Fermat's Principle*, and its shape is therefore dictated by the Generalized Snell's Law! Substituting $v(y) = y$ into (5.12), suddenly the answer becomes clear in [5.8]:

> *Geodesics in the Beltrami–Poincaré half-plane model of* $\mathbb{H}^2$ *satisfy* $(\sin\theta/y) = k$*. If* $k \neq 0$ *this is a semicircle of radius* $r = (1/k)$ *centred on the horizon. If* $k = 0$ *then it a vertical half-line* $\theta = 0$*.* (5.13)

Later, in Section 11.7.5, we will provide a *second* physical explanation of this important fact, based on *angular momentum*!

5.6 The Angle of Parallelism

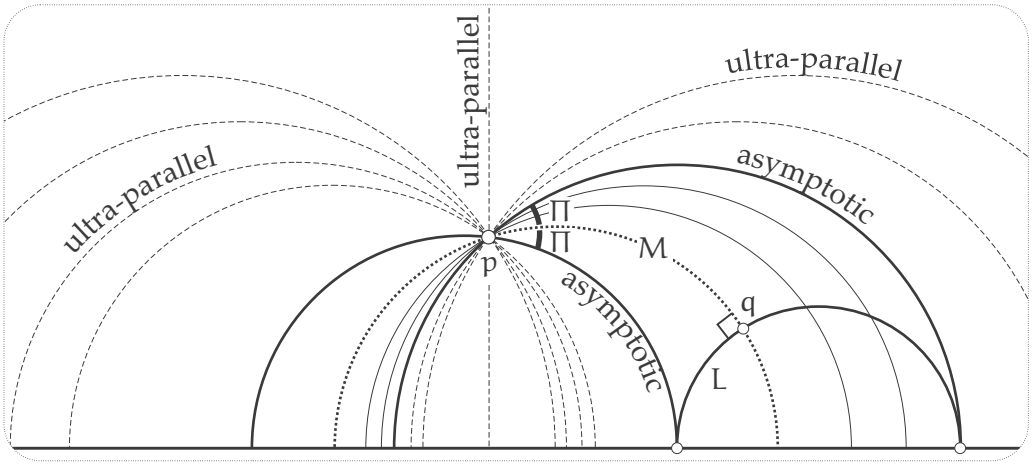
Now that we grasp the form of geodesics in the hyperbolic plane, we can return to our starting point and visually confirm in [5.9] the truth of the Hyperbolic Axiom (1.1). There are indeed infinitely many lines (shown dashed) through p that do not meet the line L. Such lines are said to be ***ultra-parallel*** to L.

Separating the ultra-parallels from the lines that do intersect L, we see that there are precisely two lines that fail to meet L anywhere within the hyperbolic plane proper, but that do meet it on the horizon. These two lines are called ***asymptotic***.[10]

As in Euclidean Geometry, the figure makes it clear that there is precisely one line M (dotted) passing through p that cuts L at right angles (say at q). The existence of M makes it possible to define the distance of a point p from a line L in the usual way, namely, as the (hyperbolic, $d\hat{s}$) length D of the segment pq of M.

In fact, as illustrated, M bisects the angle at p contained by the asymptotic lines, though for the moment this is far from obvious. The angle between M and either asymptotic line is called the ***angle of parallelism***, and is usually denoted Π. As one rotates the line M about p, its intersection

[10] Another commonly used name is ***parallel***.



[5.9] *Explicit confirmation of the Hyperbolic Axiom: there are infinitely many lines* (ultra-parallels) *(shown dashed) through* p *that fail to meet the given line* L.

point on L moves off towards infinity, and Π tells you how far one can rotate M before it starts missing L entirely.

As Lobachevsky and Bolyai both discovered, there exists a remarkable relationship between this angle Π and the hyperbolic distance D of p from L:

$$\tan(\Pi/2) = e^{-D}. \tag{5.14}$$

This is usually called the ***Bolyai–Lobachevsky Formula***.

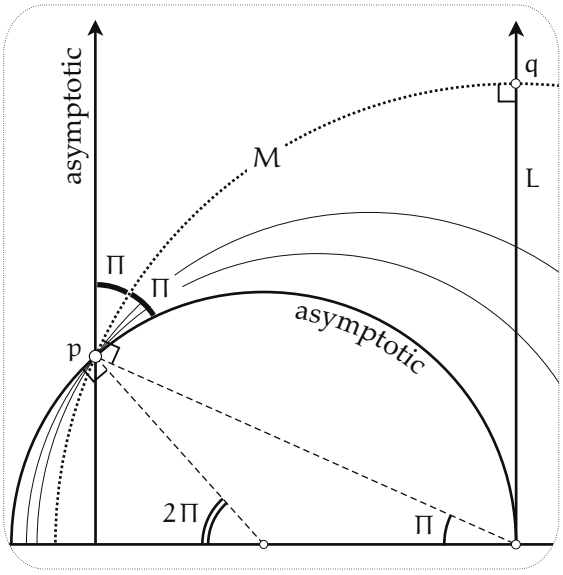
So, if p is close to L then $\Pi \approx (\pi/2)$ and the Euclidean result almost holds: about half the rays (half-lines) emanating from p eventually hit L. Of course in Euclidean Geometry exactly half the rays hit L, no matter how far p is from L. But in the hyperbolic plane we see—qualitatively in [5.9] and quantitatively in (5.14)—that as p recedes from L, the proportion of rays emanating from p that hit L shrinks to zero!

It is essential to realize that from the point of view of microscopic inhabitants of the pseudosphere, or the true hyperbolic plane, there is no way to distinguish between geodesics—every straight line is like every other. Thus, intrinsically, geodesics that are represented in our map by vertical half-lines are completely indistinguishable from geodesics represented by semicircles.

But what about the fact that the semicircles have two ends on the horizon, whereas the vertical lines appear to only have one? The answer is that, in addition to the points on the horizon, there is one more point at infinity, and all the vertical lines meet there. According to (5.8), as we move upward along two neighbouring, vertical lines, the distance between them dies away as $(1/y)$, and they converge to a single point at infinity; this is particularly vivid on the pseudosphere.

Having stressed that the two manifestations of lines in the map are mathematically identical, we now do an about-face and also stress that *psychologically* they are *not* identical. That is, standing outside this non-Euclidean world and looking in via our map, we may find it easier to *see* that some mathematical relation holds if we look at the simpler (but less typical) case where a line is vertical rather than semicircular.

What gives this idea real power is the existence of rigid motions of $\mathbb{H}^2$, such as rotations about a point. Such distance-preserving motions are called ***isometries***, and they are the subject of the next chapter. For now, we note that the existence of isometries makes it possible to rigidly



[5.10] *The same geometry as [5.9], but shown after rotating the hyperbolic plane about* p *until* L *appears vertical. This simplified picture can be used to derive the Bolyai–Lobachevsky Formula, (5.14).*

move any hyperbolic figure involving a semicircular line so that it appears in the map as a vertical line.

For example, returning to [5.9], we may rotate $\mathbb{H}^2$ about p until L becomes vertical, in which case the diagram now takes on the simpler form shown in [5.10]. An immediate payoff is that now we *can* see that M bisects the angle between the asymptotics at p: simply verify all the marked angles [exercise]. But this means that M must indeed have bisected the angle when L was in general position in [5.9], *before* we did the rotation.

By the same token, the simpler geometry of this new picture also makes it much easier to confirm the truth of the Bolyai–Lobachevsky Formula, (5.14); for details, see VCA, p. 306.

5.7 The Beltrami–Poincaré Disc

The Beltrami–Poincaré upper half-plane with metric $d\hat{s} = ds/y$ is merely one way of depicting the abstract hyperbolic plane $\mathbb{H}^2$. Several alternative models exist.[11] While these different models are, by definition, all intrinsically identical, they are not *psychologically* identical: a particular fact or formula can be very hard to see in one model, yet be transparent in another. Facility in switching between models is therefore a valuable skill when trying to come to grips with the wonders of Hyperbolic Geometry.

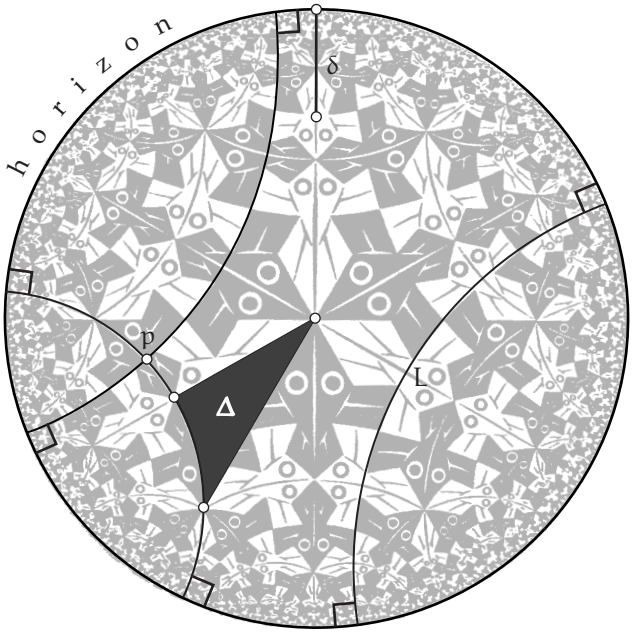
For now we wish only to illustrate one particularly useful and famous model. The new model is drawn in the unit disc; see [5.11]. Like the upper half-plane, this is a conformal model, and geodesics are again represented as arcs of circles that meet the horizon at right angles, but now the infinitely distant horizon is the boundary of this disc, the unit circle. If r is distance from the centre of the disc, the new metric formula is (see Ex. 25)

$$d\hat{s} = \frac{2}{1-r^2}\, ds. \tag{5.15}$$

For further details we refer you to VCA or to Stillwell (2010). You can at least confirm that this is indeed $\mathbb{H}^2$ by using the conformal curvature formula (4.16) to verify [exercise] that this surface has constant negative curvature $\mathcal{K} = -1$.

It was Beltrami who first discovered this model, announcing it in the same 1868 paper as the half-plane model; see Stillwell (1996). Again, Poincaré rediscovered this model 14 years later, and again it became universally known as the "Poincaré disc." However, as before, we shall steadfastly

[11] See VCA or Stillwell (2010) for a full account.



[5.11] *The Beltrami–Poincaré disc model of the hyperbolic plane. The background is Escher's* Circle Limit I*; superimposed are hyperbolic lines, which are diameters and circular arcs orthogonal to the infinitely distant boundary circle (the horizon). Clearly the Hyperbolic Axiom (1.1) is satisfied, and* $\mathcal{E}(\Delta) < 0$*. M. C. Escher's* Circle Limit I *© 2020 The M. C. Escher Company-The Netherlands. All rights reserved. www.mcescher.com.*

insist upon giving credit to both by calling this the ***Beltrami–Poincaré disc***; less contentiously, it is also called the ***conformal disc model***.

In 1958 the famous British geometer H.S.M. Coxeter (1907–2003) introduced the Dutch artist M. C. Escher (1898–1972) to the conformal disc model of $\mathbb{H}^2$. This led Escher to create his famous *Circle Limit* series of woodcuts,[12] the first of which is reproduced (deliberately faintly) in [5.11], but with hyperbolic lines prominently superimposed. (The idea for this diagram came directly from Penrose (2005, Fig. 2.12).) We see, for example, that there are indeed infinitely many lines through p that fail to meet L, in accord with the Hyperbolic Axiom (1.1). Diameters of the circle are also hyperbolic lines, so the illustrated triangle Δ is a genuine hyperbolic triangle. Note that it is visually apparent (and easily provable) that $\mathcal{E}(\Delta) < 0$, as it should be.

As you stare at this, try to imagine yourself as one of the fish. You are exactly the same size and shape as every other fish, and you can swim in a straight line forever without ever seeing any change in your surroundings or in your fellow fish. But looking into the map from outside, the compression of distances makes you appear to be swerving along a circular path and to be shrinking as you go. In fact if $\delta \equiv (1 - r)$ is the illustrated *Euclidean* distance of fish from the horizon, and we look at a point close to the edge of the map, we find that (5.15) implies [exercise] that the *(apparent size of fish)* $\propto \delta$.

In Exercise 25 you will see that there is in fact a simple *conformal* transformation that relates this new disc model to our previous conformal half-plane model, thereby explaining the conformality of the new disc model. Computer application of this transformation to Escher's original art [5.11] yields [5.12], a picture that Escher himself never created, but which he surely would have appreciated. (This image is reproduced, with permission, from Stillwell (2005, p. 195).)

[12] One can find online a short video of Coxeter himself discussing the mathematics of these Escher constructions.

[5.12] *Escher's* Circle Limit I *transformed (by Professor John Stillwell) from its original conformal disc model [5.11] to the conformal half-plane model. M. C. Escher's* Circle Limit I

Let us pause, catch our breath, and look back at how far we have come. The story[13] with which we began this book has arrived at a happy conclusion, an ending of sorts. For 2000 years Euclidean Geometry had been vexed by confusion and doubt surrounding the Parallel Axiom. Now Beltrami's concrete vindication of Hyperbolic Geometry as a legitimate alternative had at last brought about clarity and cathartic release from two millennia of mathematical tension. What a splendid place to end Act II!

Or not . . .

[13]For the complete saga, see Gray (1989).

Chapter 6

Isometries and Complex Numbers

6.1 Introduction

Great mathematical ideas not only put past mysteries to rest, they also reveal *new* ones: tunnels at the end of the light! It is to such glimpses of strange new connections with other areas mathematics and physics that we now turn.

The first of these new mysterious connections is between the three geometries of constant curvature (Euclidean, Spherical, and Hyperbolic) and the *complex numbers*.

We have already asked you to imagine our maps of curved surfaces as being drawn in the complex plane. However, the astute reader will have noticed that, up to this point, we have made but slight use of the essential structure of the complex numbers. That is about change. First, however, we need to make some more general observations about the very concept of an isometry.

Isometries necessarily preserve the magnitude of every angle, and the ones that also preserve the *sense* of the angle (clockwise versus counterclockwise) are called ***direct***, while those that reverse the sense are called ***opposite***. Thus a direct isometry is a very special kind of conformal mapping, while an opposite isometry is a very special kind of anticonformal mapping. For example, in the plane, a rotation is a direct isometry, whereas a reflection in a line is an opposite isometry.

Next we observe that, under the operation of composition,

> *The complete set of isometries (both direct and opposite) of a given surface $\mathcal{S}$ automatically has the structure of a* **group**, $\mathcal{G}(\mathcal{S})$.

To confirm this, let $e = (\text{do nothing})$ and let a, b, c be any three isometries of $\mathcal{S}$. Then the ***group axioms***[1] are satisfied:

- Since e clearly preserves distances, $e \in \mathcal{G}(\mathcal{S})$, and since $a \circ e = a = e \circ a$, we deduce that e is the ***group identity***.
- If we apply a and then apply b (both of which preserve distances) then the net transformation also preserves distances: $b \circ a \in \mathcal{G}(\mathcal{S})$.
- Since the transformation a preserves distances, its inverse does too: $a^{-1} \in \mathcal{G}(\mathcal{S})$.
- Composition of transformations (not necessarily isometries) is associative: $(a \circ b) \circ c = a \circ (b \circ c)$.

Note that (under composition) direct and opposite isometries behave like $(+)$ and $(-)$ under multiplication: $(+)(+) = (+)$, $(+)(-) = (-)$, and $(-)(-) = (+)$. It follows that

> *The direct isometries form a* **subgroup** $\mathcal{G}_+(\mathcal{S})$ *of the full group* $\mathcal{G}(\mathcal{S})$.

[1] These define the mathematical concept of a ***group***. Even if you have not met this concept before, you can still follow along by simply accepting the following axioms as the definition.

On the other hand, the opposite isometries do not form a group at all. But they do belong to the full group $\mathcal{G}(\mathcal{S})$, so how are they related to $\mathcal{G}_+(\mathcal{S})$?

Let ξ be a specific, fixed opposite isometry; then ξ^{-1} is too. Let ζ be a general opposite isometry—think of this varying over all possible opposite isometries. Then $\xi^{-1} \circ \zeta \in \mathcal{G}_+(\mathcal{S}) \Rightarrow \zeta \in \xi \circ \mathcal{G}_+(\mathcal{S})$. By the same token, $\zeta \in \mathcal{G}_+(\mathcal{S}) \circ \xi$. Thus,

If ξ is any opposite isometry, the complete set of opposite isometries is $\xi \circ \mathcal{G}_+(\mathcal{S}) = \mathcal{G}_+(\mathcal{S}) \circ \xi$, and the full symmetry group is therefore

$$\mathcal{G}(\mathcal{S}) = \mathcal{G}_+(\mathcal{S}) \cup [\xi \circ \mathcal{G}_+(\mathcal{S})] = \mathcal{G}_+(\mathcal{S}) \cup [\mathcal{G}_+(\mathcal{S}) \circ \xi] \,. \tag{6.1}$$

Does every surface $\mathcal{S}$ possess a nontrivial group $\mathcal{G}(\mathcal{S})$ of isometries? No, for an isometry must also preserve the *curvature*. Suppose an isometry carries a very small (ultimately vanishing) triangle Δ at p to a congruent triangle Δ' at p'. Then

$$\mathcal{K}(p) \asymp \frac{\mathcal{E}(\Delta)}{\mathcal{A}(\Delta)} = \frac{\mathcal{E}(\Delta')}{\mathcal{A}(\Delta')} \asymp \mathcal{K}(p').$$

Thus an irregular surface like the squash depicted in [1.9] will not have any isometries.[2]

On the other hand, it is not necessary for the curvature to be constant over $\mathcal{S}$ for isometries to exist.[3] For example, consider any surface of revolution. By its very construction, this surface of (typically) nonconstant curvature does admit a group of isometries. Rotations about the surface's axis are direct isometries, and reflections in planes passing through this axis are opposite isometries; additional isometries may exist as well.

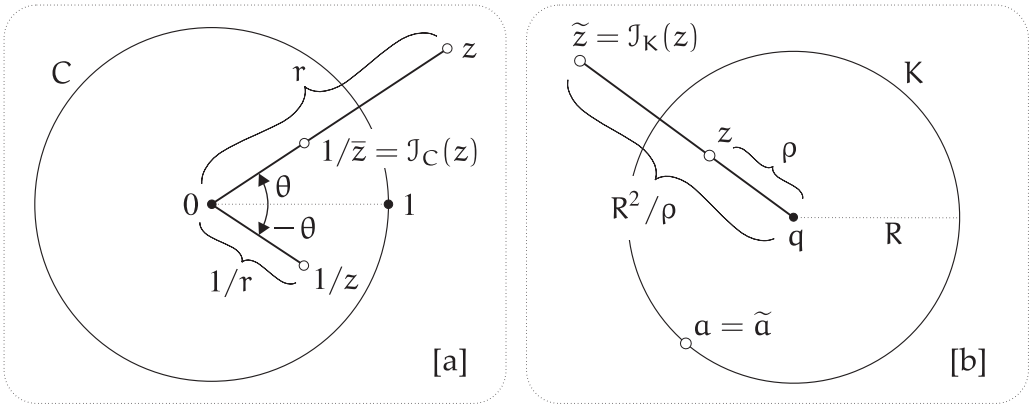
The greater the symmetry of $\mathcal{S}$, the bigger the group of isometries, and the greatest symmetry occurs in the three cases of constant curvature: $\mathcal{K} = 0$, $\mathcal{K} > 0$, and $\mathcal{K} < 0$. Extrinsically, the archetypal surfaces possessing these geometries are the Euclidean plane, the sphere, and the pseudosphere. However, the concept of an isometry belongs to *intrinsic* geometry; so, for example, the Beltrami–Poincaré half-plane map of $\mathbb{H}^2$ is in fact a much better depiction of Hyperbolic Geometry than is the pseudosphere, and it is this map (or the conformal disc model) that will be relevant in what follows.

We will refine the statement shortly, but here in brief is the startling connection between these three, maximally symmetric geometries and the complex numbers—the **Main Result**:

All three geometries of constant curvature have symmetry groups $\mathcal{G}_+(\mathcal{S})$ (of direct isometries) that are subgroups of the group of Möbius transformations of the complex plane, $z \mapsto M(z) = \frac{az+b}{cz+d}$, where a, b, c, d are complex constants. (6.2)

[2] I have not thought through precisely how to quantify the degree of irregularity needed to preclude the existence of isometries. A starting point might be to observe that if the value of $\mathcal{K}(p)$ is *uniquely* attained at p, then p must be a *fixed point* of any putative isometry—it has nowhere to go! Three such points would result in three fixed points, and I would guess this would preclude nontrivial isometries.

[3] We will be concerned with infinite, continuous sets of isometries, but it is also possible to have a *finite* set of isometries, even for a smooth surface. For example, consider the symmetries of a gaming die (a cube with its edges and corners rounded off).



[6.1] [a] Complex inversion *is the composition of geometric inversion and conjugation.* **[b] Geometric inversion** *in a general circle.*

6.2 Möbius Transformations

As the above result alone may convince you, Möbius[4] transformations are extraordinarily important in modern mathematics (and, as we shall see, in physics too). We now summarize[5] some facts that we shall need concerning these transformations.

- **Decomposition into Simpler Transformations.** Let us decompose [exercise] $z \mapsto M(z) = \frac{az+b}{cz+d}$ into the following sequence of transformations:

$$\left.\begin{array}{ll} \text{(i)} & z \mapsto z + \frac{d}{c}\text{, which is a translation;} \\ \text{(ii)} & z \mapsto (1/z)\text{, which we shall call } \textit{complex inversion}\text{;} \\ \text{(iii)} & z \mapsto -\frac{(ad-bc)}{c^2} z\text{, which is an expansion and a rotation; and} \\ \text{(iv)} & z \mapsto z + \frac{a}{c}\text{, which is another translation.} \end{array}\right\} \tag{6.3}$$

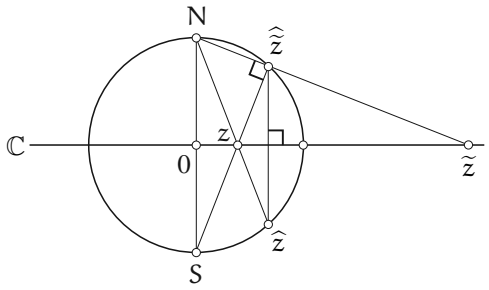
  Note that if $(ad - bc) = 0$ then $M(z)$ crushes the entire complex plane down to a single image point (a/c); in this exceptional case $M(z)$ cannot be undone, and is called ***singular***. In discussing Möbius transformations we shall therefore always assume that $M(z)$ is ***nonsingular***, meaning that it is invertible (i.e., $(ad - bc) \neq 0$).

  Of the four transformations above, only the second one (the reciprocal mapping) is unfamiliar and requires further investigation.

- **Inversion in a Circle.** This mapping $z \mapsto (1/z)$ holds the key to understanding the Möbius transformations. As we did in VCA, we shall call this reciprocal mapping ***complex inversion***. In polar coordinates, the image of $z = r\,e^{i\theta}$ under complex inversion is $1/(r\,e^{i\theta}) = (1/r)\,e^{-i\theta}$: the new length is the reciprocal of the original, and the new angle is the negative of the original. See [6.1a]. Note how a point outside the unit circle C is mapped to a point inside C, and vice versa. Figure [6.1a] also illustrates a particularly fruitful way of decomposing complex inversion into a two-stage process:
  1. Send $z = r\,e^{i\theta}$ to the point that is in the same direction as z but that has reciprocal length, namely the point $(1/r)\,e^{i\theta} = (1/\bar{z})$.

[4] Also called ***fractional linear*** or ***bilinear*** transformations.

[5] For greater depth, see Chapters 3 and 6 of VCA.



[6.2] *A vertical cross section of the Riemann sphere. Reflecting the sphere in its equatorial plane* $\mathbb{C}$ *sends* $\widehat{z} \mapsto \widehat{\widetilde{z}}$ *and* $z \mapsto \widetilde{z} = \mathcal{I}_C(z)$.

2. Apply complex conjugation (i.e., reflection in the real axis), which sends $(1/\bar{z})$ to $\overline{(1/\bar{z})} = (1/z)$.

Check for yourself that the order in which we apply these two mappings is immaterial. (This is atypical: in (6.3), for example, the order certainly does matter.)

While stage (2) is geometrically trivial, we shall see that the mapping in stage (1) is filled with surprises; it is called[6] ***geometric inversion***, or simply ***inversion***. Clearly, the unit circle C plays a special role for this mapping: the inversion interchanges the interior and exterior of C, while each point *on* C remains fixed (i.e., is mapped to itself). For this reason we write the mapping as $z \mapsto \mathcal{I}_C(z) = (1/\bar{z})$, and we call $\mathcal{I}_C$ (a little more precisely than before) "inversion in C."

This added precision in terminology is important because, as illustrated in [6.1b], there is a natural way of generalizing $\mathcal{I}_C$ to inversion in an *arbitrary* circle K (say with centre q and radius R). Clearly, this "inversion in K," written $z \mapsto \widetilde{z} = \mathcal{I}_K(z)$, should be such that the interior and exterior of K are interchanged, while each point on K remains fixed. If ρ is the distance from q to z, then *we define* $\widetilde{z} = \mathcal{I}_K(z)$ *to be the point in the same direction from* q *as* z*, and at distance* (R^2/ρ) *from* q. Check for yourself that this definition is forced on us if we imagine this figure to be [6.1a] expanded by factor R.

- **Inversion Is Equatorial Reflection of the Riemann Sphere.** If we stereographically suck the complex numbers off the plane and onto the unit sphere, thereby creating the Riemann sphere, then the effect of inversion becomes startlingly simple:

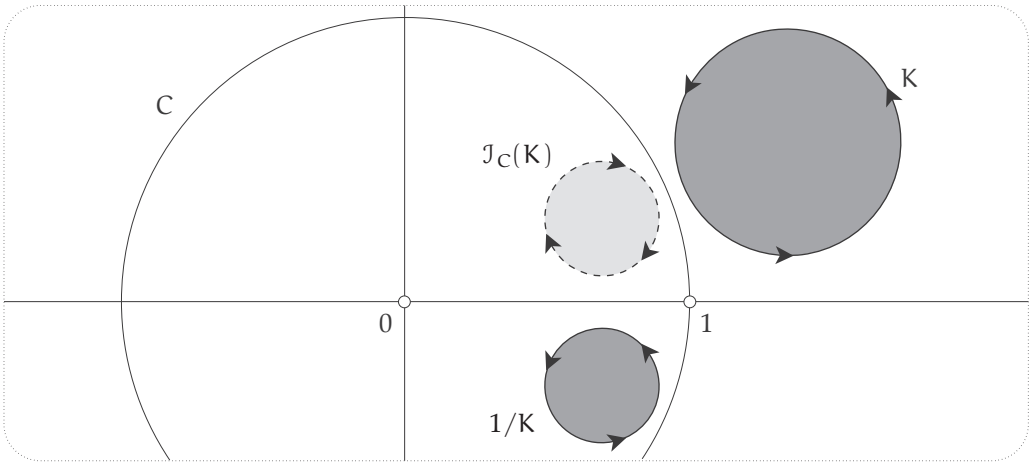
> *Inversion of* $\mathbb{C}$ *in the unit circle induces a reflection of the Riemann sphere in its equatorial plane,* $\mathbb{C}$. (6.4)

To confirm (6.4), see [6.2], which shows a vertical cross section of the Riemann sphere. The point z is stereographically projected to $\widehat{z}$, reflected across the equatorial plane to $\widehat{\widetilde{z}}$, and finally projected to $\widetilde{z}$. We see [exercise] that the triangles N0$\widetilde{z}$ and z0N are similar, and so $|\widetilde{z}|/1 = 1/|z|$. Thus, $\widetilde{z} = \mathcal{I}_C(z)$, as claimed.

Note that this figure also shows that inversion in the unit circle is equivalent to stereographic projection from the south pole, followed by stereographic projection from the north pole (or vice versa).

---

[6] Another common and appropriate name is ***reflection*** in a circle (for reasons that will become apparent shortly). In older works it is often called "transformation by reciprocal radii."



[6.3] *Complex inversion in the unit circle* C *sends the oriented circle* K *to* $(1/K)$*, and the shaded region to the left of the direction of travel along* K *is mapped to the shaded region to the left of the direction of travel along* $(1/K)$*.*

- **Inversion Preserves Circles.** By virtue of (4.28), a circle K in $\mathbb{C}$ projects to a circle on Σ, and the anticonformal reflection of the sphere in its equatorial plane (inversion) maps this to another circle on Σ, which finally projects back to a circle $\mathcal{I}_C(K)$ in $\mathbb{C}$. See [6.3].

  If we instead begin with a line, we get a circle on Σ passing through N (recall [4.21]) which reflects to a circle through S, which projects back to a circle in $\mathbb{C}$ passing through 0. Conversely, because inversion *swaps* points, a circle K through 0 is sent to a line $\mathcal{I}_C(K)$. See [6.4].

  We may think of the second result as a limiting case of the first, a line being a limiting form of a circle. Indeed, on the Riemann sphere a line is literally a circle that happens to pass through the north pole. With this unified language in place, we may summarize by saying,

  Inversion is anticonformal and maps circles to circles. (6.5)

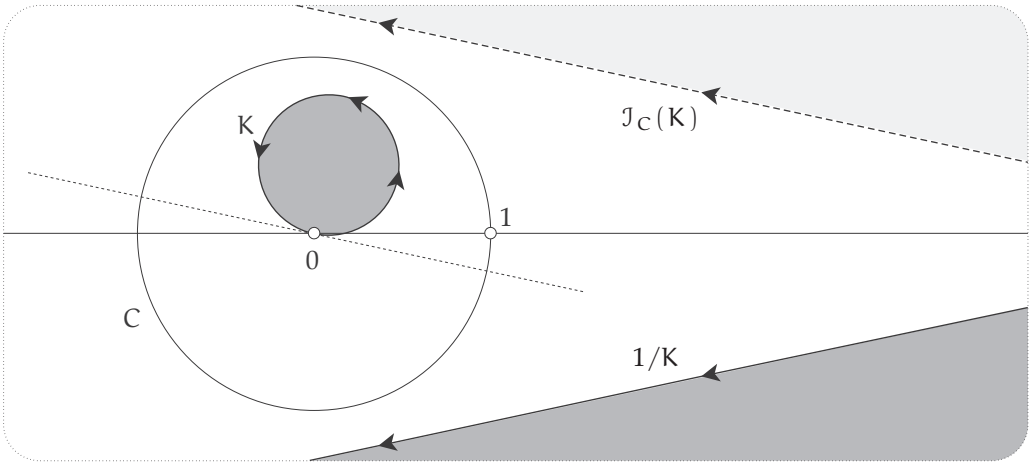
- **Complex Inversion is a Rotation of the Riemann Sphere.** If we now follow this inversion by conjugation, $z \mapsto \bar{z}$, then the net effect is complex inversion. But the effect of conjugation on the Riemann sphere is another reflection, this time in the vertical plane passing through the real axis. As you may easily check (perhaps with the assistance of an orange) the net effect of these two reflections in perpendicular planes passing through the real axis is a *rotation* about that axis:

  Complex inversion, $z \mapsto (1/z)$*, is a rotation of the Riemann sphere, through angle* π *about the real axis. It is therefore conformal and maps circles to circles.* (6.6)

  We can also provide a second proof using the stereographic formula, (4.26). First check for yourself that if $\widehat{z}$ has coordinates (ϕ, θ) then rotating it by π about the real axis carries it to a point with coordinates $(\pi - \phi, -\theta)$. So this rotated point corresponds to the complex number

$$\cot\left[\frac{\pi}{2} - \frac{\phi}{2}\right] e^{i(-\theta)} = \frac{1}{\cot(\phi/2)} e^{-i\theta} = \frac{1}{\cot(\phi/2)\, e^{i\theta}} = \frac{1}{z},$$

  as was to be shown.



[6.4] *Complex inversion in the unit circle* C *sends the oriented circle* K *to* $(1/K)$*, and the shaded region to the left of the direction of travel along* K *is mapped to the shaded region to the left of the direction of travel along* $(1/K)$.

Figure [6.3] illustrates the effect of both inversion and complex inversion on an oriented circle K. Note how the darkly shaded region to the left of the direction of travel along K (the interior of K) is mapped to the region to the left of the direction of travel along $(1/K)$. If K contains 0 then the orientation of $(1/K)$ is *reversed* by the mapping, but the rule "left $\mapsto$ left" remains in force. Check all this for yourself, both directly and by using (6.6).

Figure [6.4] illustrates the same phenomenon in the case where K passes through 0 (the south pole) and $(1/K)$ is therefore a line (a circle through the north pole).

Observe that this provides a conformal mapping between a half-plane and the interior of a disc, and this is in fact precisely how the conformal disc model [5.11] of the hyperbolic plane is constructed from the conformal half-plane model [5.12]. (The explicit Möbius transformation is given in Exercise 25, and the full explanation is given in VCA, pp. 315–317.)

- **Preservation of Angles and Circles.** As illustrated in [6.3] and [6.4], it follows immediately from (6.3) and (6.6) that, using the new generalized sense of "circle,"

> *Möbius transformations are conformal and map each oriented circle* K *to an oriented circle* $\widetilde{K}$ *in such a way that the region to the left of* K *is mapped to the region to the left of* $\widetilde{K}$. (6.7)

- **Matrix Representation.** Geometrically, the Riemann sphere allows us to think of ∞ like any other point—it's simply the north pole. From (6.6), we see that complex inversion induces a rotation that swaps the north and south poles, so the statements $0 = 1/\infty$ and $\infty = 1/0$ are literally true.

  It would be nice if we could likewise do away with the exceptional role of infinity at the *algebraic* level. To do so, we adapt an idea from Projective Geometry, describing each point on the Riemann sphere as the ratio of a *pair* of complex numbers $(\mathfrak{z}_1, \mathfrak{z}_2)$ living in $\mathbb{C}^2$: $z = (\mathfrak{z}_1/\mathfrak{z}_2)$.

  The ordered pair of complex numbers[7] $[\mathfrak{z}_1, \mathfrak{z}_2]$ are called ***projective coordinates*** or ***homogeneous coordinates*** of z. In order that this ratio be well defined we demand that $[\mathfrak{z}_1, \mathfrak{z}_2] \neq [0, 0]$.

[7] We note, sadly only in passing, that $[\mathfrak{z}_1, \mathfrak{z}_2]$ may also be viewed as the coordinate representation of a ***2-spinor***. This concept lies at the heart of huge body of fundamental, pioneering work by Sir Roger Penrose—see Penrose and Rindler (1984) for

To each ordered pair $[\mathfrak{z}_1 \text{ arbitrary}, \mathfrak{z}_2 \neq 0]$ there corresponds precisely one ordinary, finite point $z = (\mathfrak{z}_1/\mathfrak{z}_2)$, but to each point z there corresponds an infinite set of projective coordinates, $[k\mathfrak{z}_1, k\mathfrak{z}_2] = k[\mathfrak{z}_1, \mathfrak{z}_2]$, where k is an arbitrary nonzero complex number. For example, i can be represented as $[1+i, 1-i]$, as $[-1, i]$, or as $[3+2i, 2-3i]$, to give just three examples out of infinitely many.

The *point at infinity*, on the other hand, can now be represented as an unremarkable pair of the form $[\mathfrak{z}_1, 0]$.

Just as a linear transformation of $\mathbb{R}^2$ is represented by a real 2×2 matrix, so a linear transformation of $\mathbb{C}^2$ is represented by a complex 2×2 matrix:

$$\begin{bmatrix} \mathfrak{z}_1 \\ \mathfrak{z}_2 \end{bmatrix} \longmapsto \begin{bmatrix} \mathfrak{w}_1 \\ \mathfrak{w}_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \mathfrak{z}_1 \\ \mathfrak{z}_2 \end{bmatrix} = \begin{bmatrix} a\,\mathfrak{z}_1 + b\,\mathfrak{z}_2 \\ c\,\mathfrak{z}_1 + d\,\mathfrak{z}_2 \end{bmatrix}.$$

But if $[\mathfrak{z}_1, \mathfrak{z}_2]$ and $[\mathfrak{w}_1, \mathfrak{w}_2]$ are thought of as the projective coordinates in $\mathbb{C}^2$ of the point $z = (\mathfrak{z}_1/\mathfrak{z}_2)$ in $\mathbb{C}$ and its image point $w = (\mathfrak{w}_1/\mathfrak{w}_2)$, then the above linear transformation of $\mathbb{C}^2$ induces the following (nonlinear) transformation of $\mathbb{C}$:

$$z = \frac{\mathfrak{z}_1}{\mathfrak{z}_2} \quad \longmapsto \quad w = \frac{\mathfrak{w}_1}{\mathfrak{w}_2} = \frac{a\,\mathfrak{z}_1 + b\,\mathfrak{z}_2}{c\,\mathfrak{z}_1 + d\,\mathfrak{z}_2} = \frac{a\,(\mathfrak{z}_1/\mathfrak{z}_2) + b}{c\,(\mathfrak{z}_1/\mathfrak{z}_2) + d} = \frac{az+b}{cz+d}.$$

This is none other than the most general Möbius transformation!

Thus to every Möbius transformation $M(z)$ there corresponds a 2×2 matrix $[M]$,

$$M(z) = \frac{az+b}{cz+d} \quad \longleftrightarrow \quad [M] = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

and we deduce that the matrix of the composition of two Möbius transformations is the product of the corresponding matrices:

$$[M_2 \circ M_1] = [M_2]\,[M_1]. \tag{6.8}$$

Likewise, the inverse of a Möbius transformation can be found by taking the inverse of the corresponding matrix:

$$[M^{-1}] = [M]^{-1}. \tag{6.9}$$

It follows easily [exercise] that the nonsingular Möbius transformations do indeed form a group (as was implicitly claimed earlier).

Since the coefficients of the Möbius transformation are not unique, neither is the corresponding matrix: if k is any nonzero constant, then the matrix $k[M]$ corresponds to the same Möbius transformation as $[M]$. However, if $[M]$ is ***normalized*** by imposing $(ad - bc) = 1$, then there are just two possible matrices associated with a given Möbius transformation: if one is called $[M]$, the other is $-[M]$; in other words, the matrix is determined "uniquely up to sign." This apparently trivial fact turns out to have deep significance in both mathematics and physics; see Penrose and Rindler (1984, Ch. 1) and Penrose (2005, §11.3).

---

technical details—having to do with "spinorial objects," that do not return to their original state after being rotated by 2π! For an intuitive explanation of the existence of such seemingly impossible objects, see Penrose (2005, §11.3).

6.3 The Main Result

We can now state and prove a more detailed version of the Main Result, (6.2):

The symmetry groups $\mathcal{G}_+(\mathcal{S})$ of all three geometries of constant curvature are subgroups of the group of Möbius transformations.

1. *Euclidean Geometry ($\mathcal{K}=0$):*

$$E(z) = e^{i\theta}z + k.$$

   Note that $z \mapsto \bar{z}$ is an opposite isometry, being reflection in the real axis. Thus (6.1) tells us that the full group of isometries $\mathcal{G} = \{E(z)\} \cup \{E(\bar{z})\}$.
2. *Spherical Geometry ($\mathcal{K}=+1$) in the stereographic map:*

$$S(z) = \frac{az+b}{-\bar{b}z+\bar{a}}, \quad \textit{where} \quad |a|^2+|b|^2 = 1. \tag{6.10}$$

   Note that $z \mapsto \bar{z}$ is an opposite isometry, being reflection of the sphere in the vertical plane through the real axis. Thus (6.1) tells us that the full group of isometries $\mathcal{G} = \{S(z)\} \cup \{S(\bar{z})\}$.
3. *Hyperbolic Geometry ($\mathcal{K}=-1$) in the Beltrami–Poincaré half-plane map:*

$$H(z) = \frac{az+b}{cz+d}, \quad \textit{where } a, b, c, d \textit{ are } \text{real}, \textit{ and } (ad-bc) = 1. \tag{6.11}$$

   Note that $z \mapsto -\bar{z}$ is an opposite isometry, being reflection in the imaginary axis. Thus, again by (6.1), the full group of isometries $\mathcal{G} = \{H(z)\} \cup \{H(-\bar{z})\}$.

The matrix results (6.8) and (6.9) make it a simple matter [exercise] to verify that each of these three sets is indeed a group. We also note that in the spherical case the matrices are of a special kind that plays an important role in physics. They are called ***unitary***, meaning that if you take the conjugate and then transpose (the net operation being denoted $*$) then [exercise] you obtain the matrix of the inverse transformation:

$$\textbf{\textit{unitary}} \text{ means} \qquad [S][S]^* = \text{the identity matrix.}$$

The Euclidean transformation $E(z)$ is a rotation of θ followed by a translation of k, and it seems clear that this indeed the most general rigid motion of the plane. We therefore rush past this and reserve our energy for the challenge of proving the surprising results claimed for $\mathcal{K} = \pm 1$. In this work we cannot pause to do geometric justice to these lovely results; instead, we refer you to VCA. For now, let us show off the *computational* power of the metric machinery we have developed.

To do so, we need to pause and introduce a pinch of Complex Analysis. As we explained in Section 4.6, the derivative $f'(z)$ of a complex mapping $z \mapsto w = f(z)$ is defined exactly as in first-year calculus, and all the usual rules of differentiation apply without change. Thus, applying the quotient rule to the normalized (i.e., $(ad-bc)=1$) Möbius transformation, we find that

$$M(z) = \frac{az+b}{cz+d} \quad \Longrightarrow \quad M'(z) = \frac{1}{(cz+d)^2}. \tag{6.12}$$

But recall from (4.19) that the complex numbers imbue this derivative with the "amplitwist" interpretation, which is magically richer than in ordinary calculus.

We can now return to the Main Result. As Euler was the first to prove (in 1775) the rigid motions of the sphere are simply rotations, and it was Gauss (around 1819) who first recognized that they could be represented as Möbius transformations of the form (6.10).

Taking the sphere to have unit radius, we may rewrite the stereographic metric formula (4.22) in terms of the complex number z:

$$d\,\widehat{s} = \frac{2}{1+|z|^2}\,|dz|.$$

To prove that the mapping $z \mapsto w = S(z) = \frac{az+b}{-\overline{b}z+\overline{a}}$ is a direct isometry, we must therefore show that

$$\frac{2}{1+|w|^2}\,|dw| = \frac{2}{1+|z|^2}\,|dz|. \tag{6.13}$$

Direct calculation shows [exercise] that

$$1+|w|^2 = \frac{1+|z|^2}{|-\overline{b}z+\overline{a}\,|^2}\,.$$

It follows from (6.12) that

$$\left|\frac{dw}{dz}\right| = \left|S'(z)\right| = \frac{1}{|-\overline{b}z+\overline{a}\,|^2} = \frac{1+|w|^2}{1+|z|^2},$$

thereby confirming (6.13), and with it part 2 of the Main Result. (Well, this does not prove that these are the *only* direct isometries. This will be addressed by our deferred geometric analysis.)

But how could any mortal (even Gauss) have guessed the form of these Möbius transformations?! Here is a simple argument, based only on what we know so far: *If the rotation carries the point $\widehat{z}$ on the sphere to $\widehat{M(z)}$, then it must also carry the point antipodal to $\widehat{z}$ to the point antipodal to $\widehat{M(z)}$.* But we know that the stereographic images of antipodal points are related by (4.27), page 49, and so

$$M\left(-\frac{1}{\overline{z}}\right) = -\frac{1}{\overline{M(z)}}.$$

Thus,

$$\frac{a\left[-\frac{1}{\overline{z}}\right]+b}{c\left[-\frac{1}{\overline{z}}\right]+d} = -\frac{\overline{cz+d}}{\overline{az+b}} \quad\Longrightarrow\quad \frac{-b\,\overline{z}+a}{d\,\overline{z}-c} = \frac{\overline{c}\,\overline{z}+\overline{d}}{\overline{a}\,\overline{z}+\overline{b}},$$

which is clearly satisfied if $c = -\overline{b}$ and $d = \overline{a}$, which is precisely the form of (6.10)!

In Exercise 27 we give concrete examples of how to represent rotations as Möbius transformations, and we also provide a constructive proof that every rotation must be a Möbius transformation of the form (6.10).

Lastly, consider the hyperbolic plane. The metric formula (5.7) tells us that we must show that $z \mapsto w = H(z) = \frac{az+b}{cz+d}$ satisfies

$$\frac{|dw|}{\operatorname{Im} w} = \frac{|dz|}{\operatorname{Im} z}. \tag{6.14}$$

Direct calculation shows [exercise, recalling that $\overline{a} = a, \ldots$] that

$$\operatorname{Im} w = \frac{w-\overline{w}}{2i} = \frac{\operatorname{Im} z}{|cz+d|^2}.$$

It follows from (6.12) that

$$\left|\frac{dw}{dz}\right| = \left|H'(z)\right| = \frac{1}{|cz+d|^2} = \frac{\operatorname{Im} w}{\operatorname{Im} z},$$

thereby confirming (6.14), and with it the final part of the Main Result.

These hyperbolic isometries not only contain analogues of ordinary rotations and translations, but they also include a *third* type of rigid motion, called a ***limit rotation***, which has no counterpart in ordinary Euclidean Geometry. It is the limit of an ordinary rotation of $\mathbb{H}^2$ as the centre of rotation moves off to infinity, ultimately becoming a point on the horizon, $y = 0$. For details, see VCA, pp. 306–313.

6.4 Einstein's Spacetime Geometry

If mere *sub*groups of the group of Möbius transformations have such deep significance for geometry, it is natural to ask if even more remarkable powers reside within the *full* group that gave rise to them.

Indeed, the full group plays a fundamental role in at least two seemingly unrelated spheres of knowledge: Relativity Theory and *3-dimensional* Hyperbolic Geometry. In fact this is *not* a coincidence. There is a deep connection between these two subjects, but we shall not be able to explore it here, and instead direct you to the wonderful exposition of Penrose (2005, §18.4).

The first role may be called "fundamental" without hyperbole:[8]

> *The Möbius group describes the symmetries of space and time, or, more precisely, of Einstein's unification of the two*—spacetime.

Clearly it would be neither appropriate nor feasible for us to explore Einstein's Special Theory of Relativity in detail here.[9] However, for readers who have not studied this theory previously, we shall try to say enough to make sense of this particular connection with the Möbius group.[10]

The starting point of Einstein's theory is an extraordinary and bizarre experimental fact of Nature:

> *The speed of light is the same for all observers in uniform relative motion.*

As Einstein first recognized in 1905, this can *only* be possible if such observers *disagree* about measurements of space and time!

To quantify this, let us combine the time T and the 3-dimensional Cartesian coordinates (X, Y, Z) of an ***event*** $\mathfrak{E}$ into a single ***4-vector*** (T, X, Y, Z) in 4-dimensional ***spacetime***.[11] These are the space and time coordinates of event $\mathfrak{E}$ for what we shall call our *first observer*.

Of course the spatial components of the first observer's vector have no absolute significance: if a *second observer* uses the same origin but has coordinate axes rotated relative to the first, then

[8] We encourage you not to peek, but, the mathematically precise statement of this result will be given in (6.20).

[9] For a wonderfully physical account of the theory, we recommend Taylor and Wheeler (1992). For more on the *geometry* of the theory, we strongly recommend Misner, Thorne, and Wheeler (1973) and Penrose (2005).

[10] According to Coxeter (1967, pp. 73–77), this connection was first recognized by H. Liebmann almost immediately, in 1905!

[11] In fact it was Hermann Minkowski who in 1908 recast Einstein's 1905 theory in these geometrical terms, and at first Einstein himself did not approve!

this second observer will ascribe different spatial coordinates $(\widetilde{X}, \widetilde{Y}, \widetilde{Z})$ to the same event $\mathfrak{E}$. Yet, if these two observers are *not in relative motion*, they will nevertheless agree on the value of $\widetilde{X}^2 + \widetilde{Y}^2 + \widetilde{Z}^2 = X^2 + Y^2 + Z^2$, for this represents the square of the distance to the point where the event happened.

In contrast to this, we are accustomed to thinking that the *time* component T *does* have an absolute significance. However, Einstein's theory—confirmed by innumerable experiments—tells us that this is wrong. *If our two (momentarily coincident) observers* are *in relative motion, they will disagree about the times at which events occur.* Furthermore, they will no longer agree about the value of $(X^2 + Y^2 + Z^2)$—this is the famous ***Lorentz contraction***.

Such effects are only discernible if the relative speed of the two observers is a significant fraction of the fantastic speed of light (which is roughly 186,000 miles per second). For example, even if the second observer shoots away from the first observer with the speed of a rifle bullet (2,000 miles per hour), then even over the period of their entire adult lifetimes (say 85 years), the total accumulated discrepancy between their two clocks would only amount to about one hundredth of a second! It is this accident of our snail-like existence relative to the speed of light that hid for thousands of years (and continues to hide, day-to-day) the truth of Einstein's discovery.

But if the second object *is* travelling at an appreciable fraction of the speed of light, the temporal and spatial distortion is dramatic, and this effect is witnessed daily in particle accelerators around the world. Under these strange circumstances, is there *any* aspect of spacetime that has absolute significance and upon which the two observers in uniform relative motion must agree?

Einstein's amazing answer is "Yes!": spacetime *does* possess an *absolute* structure that is independent of all observers. For this reason, Einstein personally disliked, and resisted for years, the name "Theory of Relativity"—it should be "Theory of the Absolute"!

Making a convenient choice of units in which the speed of light is equal to 1, Einstein[12] discovered that both observers will agree on the value of the ***spacetime interval*** $\beth$ between the the observer and the event. This interval[13] $\beth$ is defined by means of its *square*:

$$\boxed{\beth^2 \equiv T^2 - (X^2 + Y^2 + Z^2) = \widetilde{T}^2 - (\widetilde{X}^2 + \widetilde{Y}^2 + \widetilde{Z}^2).} \tag{6.15}$$

As Minkowski realized, this is the correct generalization of the concept of distance appropriate to spacetime, and the isometries/symmetries of spacetime must preserve it. But, quite unlike ordinary distance, here *the square of the interval between distinct events can be zero or even negative.*

Let us provide a more vivid interpretation of $\beth$ in the case[14] $\beth^2 > 0$. As we have said, we suppose our two observers (in uniform relative motions to one another) are momentarily coincident at one particular place and time, defining an origin event that we label $\mathfrak{O}$. (Of course for this event to be precisely defined we must imagine our observers to be pointlike, having a specific location.) Now suppose that (by accident or by design) the first observer arrives at the event $\mathfrak{E}$ just as it happens. As far as she is concerned, she has been sitting still the whole time, and both events $\mathfrak{O}$ and $\mathfrak{E}$ happened right where she sits ($X = Y = Z = 0$), and therefore the spacetime interval between $\mathfrak{O}$ and $\mathfrak{E}$ (upon which *all* observers must agree) is simply $\beth = T$:

---

[12]Just as we have previously asked Poincaré to step aside slightly, to make room for Beltrami in connection with $\mathbb{H}^2$, so now must we ask Einstein to step aside slightly for Poincaré: already in 1905 Poincaré, too, had discovered the invariance of the spacetime interval. Of course Hendrik Lorentz should be considered the third father of Special Relativity, but at least he is immortalized in the group of transformations that bear his name.

[13]Surprisingly, there is no standard symbol for this interval, but our use of ***gimel*** (the third letter of the Hebrew alphabet) seems appropriate: it resembles (visually) the English "I" of "Interval," and one connotation (culturally) of gimel is of a *bridge* connecting two points.

[14]If $\beth^2 < 0$ there is a different but equally simple interpretation; see Taylor and Wheeler (1992, Ch. 1).

If an observer in uniform motion carries a wristwatch from one event to another, the invariant spacetime interval $\beth$ between the two events is simply the elapsed time on that watch. (6.16)

Next, imagine that a spark occurs at the event $\mathfrak{O}$, so that photons (particles of light) travel out in all directions from this flash. If both observers focus their attention on a single photon, both will agree that every event along this photon's spacetime trajectory has $\beth = 0$. Such a 4-vector of vanishing "length," pointing along a particular light ray, is called a ***null*** vector.

A ***Lorentz transformation*** $\mathfrak{L}$ is a linear transformation of spacetime (a 4×4 matrix) that maps one observer's description (T, X, Y, Z) of an event to another observer's description $(\widetilde{T}, \widetilde{X}, \widetilde{Y}, \widetilde{Z})$ of the same event. Put differently, $\mathfrak{L}$ is a linear transformation that preserves the quantity $\beth^2$, upon which both observers must agree.

Now let's return to the imagined flash of light—an origin-centred sphere whose radius increases at the speed of light. After time $T = 1$ these photons will form a sphere of unit radius, which we now choose to identify with the *Riemann sphere*. Thus this sphere is now thought of as made up of points that are simultaneously labelled with spacetime coordinates $(1, X, Y, Z)$ *and* with complex numbers, assigned stereographically via (4.25) on page 48.

Substituting the projective-coordinate description $z = (\mathfrak{z}_1/\mathfrak{z}_2)$ into the stereographic formulas (4.25), we obtain,

$$X = \frac{\mathfrak{z}_1\overline{\mathfrak{z}_2} + \mathfrak{z}_2\overline{\mathfrak{z}_1}}{|\mathfrak{z}_1|^2 + |\mathfrak{z}_2|^2}, \qquad Y = \frac{\mathfrak{z}_1\overline{\mathfrak{z}_2} - \mathfrak{z}_2\overline{\mathfrak{z}_1}}{i\,(|\mathfrak{z}_1|^2 + |\mathfrak{z}_2|^2)}, \qquad \text{and} \qquad Z = \frac{|\mathfrak{z}_1|^2 - |\mathfrak{z}_2|^2}{|\mathfrak{z}_1|^2 + |\mathfrak{z}_2|^2}.$$

But a light ray may be identified by *any* null vector along it, so instead of choosing $T = 1$ let us now eliminate the denominators in the above expressions by scaling up our null vector by a factor of $(|\mathfrak{z}_1|^2 + |\mathfrak{z}_2|^2)$ (i.e., by choosing $T = |\mathfrak{z}_1|^2 + |\mathfrak{z}_2|^2$). The new null vector (T, X, Y, Z) (in the same spacetime direction as the original) is therefore given by

$$\begin{aligned} T &= |\mathfrak{z}_1|^2 + |\mathfrak{z}_2|^2, & X &= \mathfrak{z}_1\overline{\mathfrak{z}_2} + \mathfrak{z}_2\overline{\mathfrak{z}_1}, \\ Y &= -i\,(\mathfrak{z}_1\overline{\mathfrak{z}_2} - \mathfrak{z}_2\overline{\mathfrak{z}_1}), & Z &= |\mathfrak{z}_1|^2 - |\mathfrak{z}_2|^2. \end{aligned}$$

You may readily check [exercise] that these formulas may be inverted to yield

$$\begin{pmatrix} T+Z & X+iY \\ X-iY & T-Z \end{pmatrix} = 2\begin{pmatrix} \mathfrak{z}_1\overline{\mathfrak{z}_1} & \mathfrak{z}_1\overline{\mathfrak{z}_2} \\ \mathfrak{z}_2\overline{\mathfrak{z}_1} & \mathfrak{z}_2\overline{\mathfrak{z}_2} \end{pmatrix} = 2\begin{pmatrix} \mathfrak{z}_1 \\ \mathfrak{z}_2 \end{pmatrix}\begin{pmatrix} \overline{\mathfrak{z}_1} & \overline{\mathfrak{z}_2} \end{pmatrix},$$

or,

$$\boxed{\begin{pmatrix} T+Z & X+iY \\ X-iY & T-Z \end{pmatrix} = 2\begin{pmatrix} \mathfrak{z}_1 \\ \mathfrak{z}_2 \end{pmatrix}\begin{pmatrix} \mathfrak{z}_1 \\ \mathfrak{z}_2 \end{pmatrix}^*,} \tag{6.17}$$

where "$*$" again denotes the ***conjugate transpose***.

Note that the spacetime interval can now be expressed neatly as the determinant of this matrix:

$$\boxed{\beth^2 = T^2 - (X^2 + Y^2 + Z^2) = \det\begin{pmatrix} T+Z & X+iY \\ X-iY & T-Z \end{pmatrix}.} \tag{6.18}$$

This makes it easy to see [exercise] that this scaled-up spacetime vector is still null, as it should be.

Having now used stereographic projection to identify light rays with complex numbers,[15] via (6.17), let us find the effect on the flash of light effected by a Möbius[16] transformation $z \mapsto \widetilde{z} = M(z)$ of $\mathbb{C}$, or equivalently,

$$\begin{bmatrix} \mathfrak{z}1 \\ \mathfrak{z}2 \end{bmatrix} \longmapsto \begin{bmatrix} \widetilde{\mathfrak{z}1} \\ \widetilde{\mathfrak{z}2} \end{bmatrix} = [M] \begin{bmatrix} \mathfrak{z}1 \\ \mathfrak{z}2 \end{bmatrix}.$$

Substituting this into (6.17) we obtain the following *linear transformation* of the null vectors making up the flash of light:

$$\begin{pmatrix} T+Z & X+iY \\ X-iY & T-Z \end{pmatrix} \mapsto \begin{pmatrix} \widetilde{T}+\widetilde{Z} & \widetilde{X}+i\widetilde{Y} \\ \widetilde{X}-i\widetilde{Y} & \widetilde{T}-\widetilde{Z} \end{pmatrix} = [M] \begin{pmatrix} T+Z & X+iY \\ X-iY & T-Z \end{pmatrix} [M]^*. \tag{6.19}$$

Finally, imagine that this linear transformation is applied to *all* spacetime vectors (not just null ones). Because $\det[M] = 1 = \det[M]^*$, it follows from (6.18) that *this linear transformation preserves the spacetime interval*:

$$\widetilde{\mathfrak{I}}^{\,2} = \det\left\{ [M] \begin{pmatrix} T+Z & X+iY \\ X-iY & T-Z \end{pmatrix} [M]^* \right\} = \mathfrak{I}^{\,2}.$$

Thus,

> *Every Möbius transformation of $\mathbb{C}$ yields a unique Lorentz transformation of spacetime. Conversely, it can be shown (Penrose and Rindler, 1984, Ch. 1) that to every Lorentz transformation there corresponds a unique (up to sign) Möbius transformation.*

(6.20)

Even amongst professional physicists, this beautiful "miracle" is not as well known as it should be.

It turns out that every Möbius or Lorentz transformation is fundamentally equivalent to one of just four archetypes. The essential idea is to focus on the so-called ***fixed points*** of the transformation, meaning points that remain fixed in the sense that they are mapped to themselves: $M(z) = z$.

Clearly [exercise] this leads to a quadratic equation with two solutions that may be distinct or coincide. If the fixed points are distinct, one can imagine[17] them to be the north and south poles; this yields the three archetypes shown in [6.5a–c]. If the two fixed points coincide, one can imagine them both to be the north pole; this yields the fourth archetype shown in [6.5d].

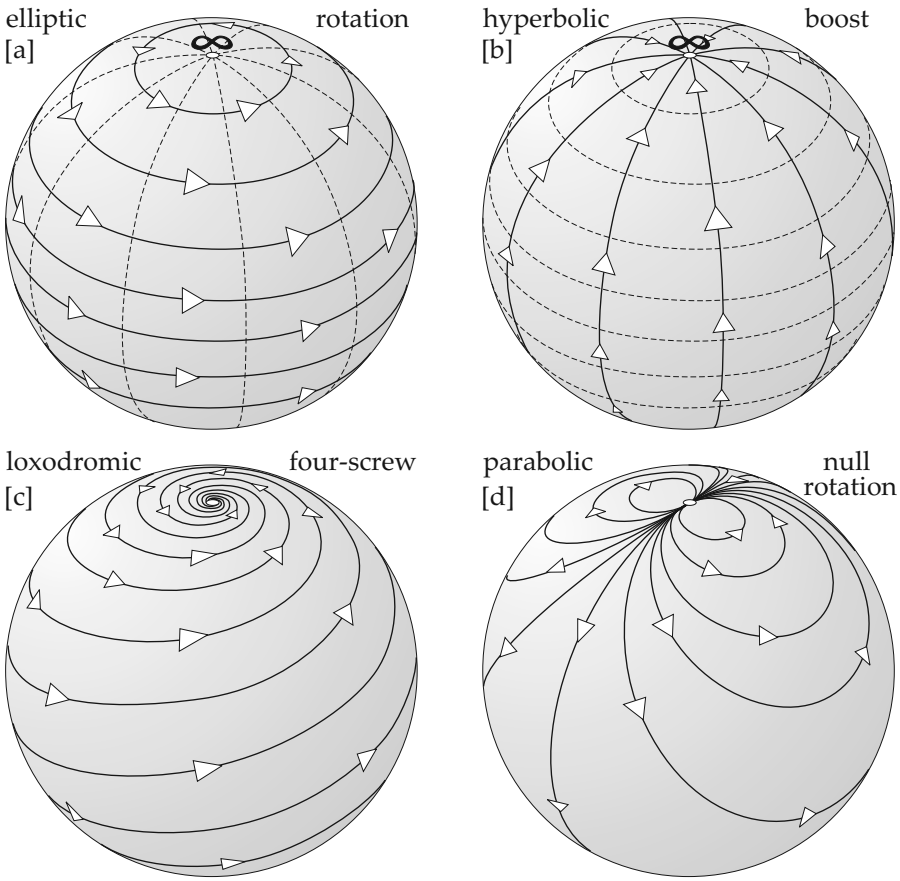
To make these four types of Lorentz transformation more vivid, imagine yourself in a spaceship in interstellar space, looking out in all directions at countless stars scattered over your ***celestial sphere***. This is an imagined unit sphere with you at its centre, a dot being marked on its surface wherever the starlight intercepts the sphere on its way to your eye. We suppose the sphere to be aligned with your spaceship, the north pole lying directly ahead of you.

---

[15] Exercise 33 explains Penrose's method of accomplishing this directly in spacetime.

[16] In the context of relativity theory, a Möbius transformation is called a ***spin transformation***, and the corresponding matrix $[M]$ is called a ***spin matrix***. See Penrose and Rindler (1984, Ch. 1) for details.

[17] For a detailed justification of this simplification in the case of Möbius transformations, see VCA, Ch. 3; in the case of Lorentz transformation, see Penrose and Rindler (1984, Ch. 1).



[6.5] Classification of Möbius and Lorentz Transformations. *Each of the four types of transformation has two names, depending on whether it is viewed as acting on* $\mathbb{C}$ *(name on the left) or on spacetime (name on the right).*

If you now fire your lateral thrusters so as to set your spaceship in a spin about its north–south axis, then the star field will rotate around you, as illustrated in [6.5a]. Unsurprisingly, this type of Lorentz transformation is called a ***rotation***. In the complex plane the corresponding Möbius transformation is also a rotation, $M(z) = e^{i\alpha}z$, and this type is called ***elliptic***.

Suppose you are no longer spinning and that you now fire your powerful main engines. Almost instantly this sends your craft hurtling at nearly light speed v directly ahead toward the north pole of your celestial sphere. As illustrated in [6.5b], you will in fact see the stars crowd *toward* your direction of travel—the exact opposite of the depiction in most science-fiction films! This type of Lorentz transformation is called a ***boost***. In the complex plane the corresponding Möbius transformation is a simple expansion by some real factor ρ, $M(z) = \rho z$, and this type is called ***hyperbolic***. In fact (see Ex. 34) the expansion factor ρ is related to the spacecraft velocity v by

$$\rho = \sqrt{\frac{1+v}{1-v}}. \tag{6.21}$$

If you fire your main engines *and* your lateral thrusters then the previous two effects combine to cause the star field to spiral toward your direction of motion, as shown in [6.5c]. This type of Lorentz transformation is called a ***four-screw***, and the corresponding Möbius transformation, $M(z) = \rho e^{i\alpha}z$, is called ***loxodromic***.

The last type of Lorentz transformation is shown in [6.5d]. It is called a ***null rotation*** and both of its fixed points are coincident at the north pole. While it is hard to give a vivid physical description of this spacetime transformation, the corresponding Möbius transformation is called ***parabolic*** and is a simple translation of $\mathbb{C}$: $M(z) = z + \tau$. This moves every complex number along a line parallel to τ. But, as we know from [4.8], page 45, these lines all stereographically project to circles through the north pole with a common tangent there that is parallel to τ, thereby explaining the form of [6.5d]. Note that movements on the sphere become smaller and smaller as the north pole is approached, leaving this as the only fixed point.

6.5 Three-Dimensional Hyperbolic Geometry

Remarkably, neither Lobachevsky nor Bolyai set out to develop a 2-dimensional non-Euclidean Geometry. Instead, from the outset, both men independently sought a hyperbolic alternative to *three*-dimensional[18] Euclidean space; hyperbolic planes then emerged naturally within this hyperbolic 3-space.

Recall that we created a conformal map of the hyperbolic plane in the upper-half-plane by taking the metric to be

$$d\hat{s} = \frac{ds}{\text{Euclidean height above the boundary}} = \frac{\sqrt{dx^2 + dy^2}}{y}. \tag{6.22}$$

To generalize this to a *three*-dimensional hyperbolic space, $\mathbb{H}^3$, consider the half-space $Z > 0$ above the horizontal (X, Y)-plane of Euclidean 3-space (X, Y, Z). We may now create a conformal model of $\mathbb{H}^3$ by again taking the metric to be[19]

$$d\hat{s} = \frac{ds}{\text{Euclidean height above the boundary}} = \frac{\sqrt{dX^2 + dY^2 + dZ^2}}{Z}. \tag{6.23}$$

Points on the (X, Y)-plane ($Z = 0$) are therefore infinitely far away, and this bounding plane is the two-dimensional horizon, which we will now think of as the complex plane $\mathbb{C}$, with coordinates $X + iY$.

It follows immediately from this construction that every vertical half-plane is a copy of $\mathbb{H}^2$. See [6.6] and compare to [5.5]. (By extension, imagine this space filled with small spheres of equal hyperbolic size, shrinking in apparent Euclidean size as the horizon is approached.)

If we imagine light travelling within such a vertical plane, symmetry dictates that it *remain* in that plane, and our previous analysis of the path taken by the light therefore applies. Thus,

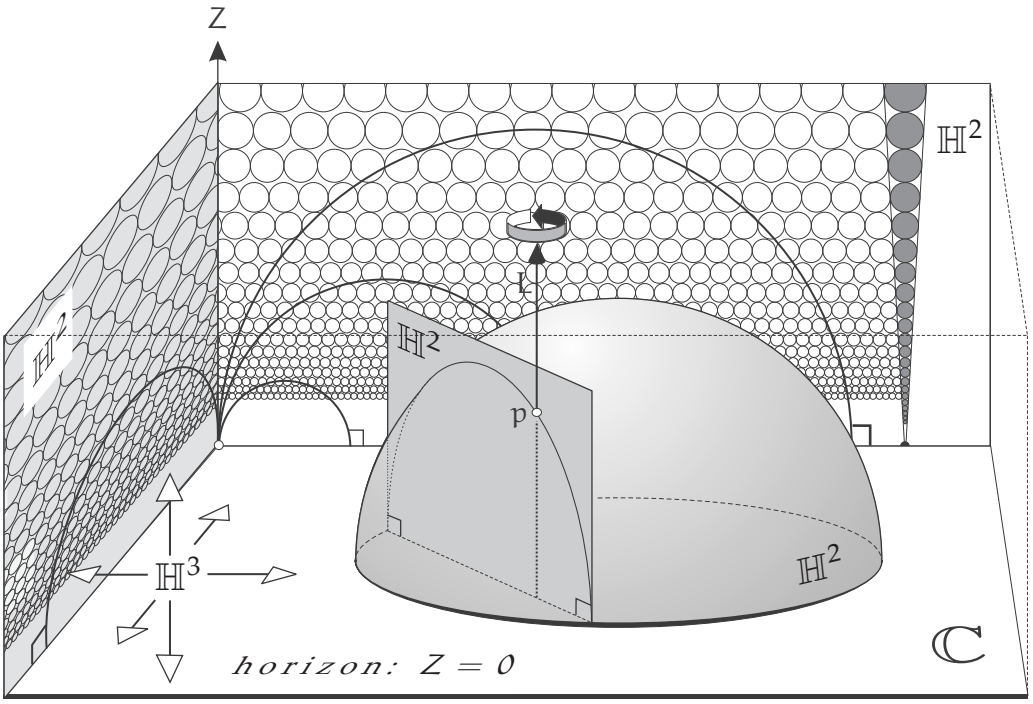
> *The geodesics of $\mathbb{H}^3$ are the vertical half-lines and the semicircles orthogonal to the horizon, $\mathbb{C}$.*

In $\mathbb{H}^2$ the vertical half-lines are exceptional geodesics, the typical ones being the semicircles orthogonal to the horizon. Likewise, in $\mathbb{H}^3$ the vertical half-planes are the exceptional hyperbolic planes; the typical planes are in fact hemispheres orthogonal to the horizon $\mathbb{C}$, such as the one shown in [6.6].

---

[18] See Stillwell (2010, p. 365).

[19] This, too, was published by Beltrami in 1868, along with a generalization to n dimensions. See Stillwell (1996).



[6.6] Hyperbolic 3-Space, $\mathbb{H}^3$. *Hyperbolic planes appear as vertical half-planes and hemispheres orthogonal to $\mathbb{C}$. The intersection of two such planes is a hyperbolic line: a vertical half-line, or a semicircle orthogonal to $\mathbb{C}$.*

To begin to see this, imagine that you live in $\mathbb{H}^3$; what would you mean by the word "plane"? Suppose I hand you an infinitesimal disc D centred at p. To extend this disc (uniquely) into an infinite "plane" H, you presumably need to prolong its diameters into infinite hyperbolic lines, i.e., semicircles meeting $\mathbb{C}$ orthogonally. If the disc is vertical this clearly yields the vertical half-plane containing D. If the disc D is *not* vertical, we claim that the construction will instead generate a hemisphere H orthogonal to $\mathbb{C}$, with D coincident with the tangent plane to H at p. Incidentally, we may easily construct the centre of this hemisphere as the intersection with $\mathbb{C}$ of the (Euclidean) normal to D through p.

In [6.6] let us take the illustrated hemisphere to be this H. Consider the illustrated intersection of H with a vertical plane through p. Clearly this is a hyperbolic line extending a diameter of D, since we have *constructed* H to have its tangent plane coincident with D at p. Now let L be the illustrated vertical line through p. If we rotate the vertical plane about L, this hyperbolic line of intersection sweeps out all of H, so this is indeed the unique extension of D into a plane. Thus,

> *The hyperbolic planes of $\mathbb{H}^3$ are the vertical half-planes and the hemispheres orthogonal to the horizon, $\mathbb{C}$.*

It is clear that a vertical half-plane is a hyperbolic plane in two senses: it is what an inhabitant of $\mathbb{H}^3$ calls a plane, *and* its internal geometry is that of $\mathbb{H}^2$. Is the same true of the hemispherical planes? That is, if we measure distance within such a hemisphere using the hyperbolic metric (6.23) of the ambient space $\mathbb{H}^3$, do we obtain an $\mathbb{H}^2$?

The answer is "Yes": each hemisphere is intrinsically an $\mathbb{H}^2$. Indeed if you live in $\mathbb{H}^3$ then every direction is like every other direction, every line is like every other line, and every plane is

like every other plane. By generalizing geometric inversion to 3-dimensional space, it is possible to provide a natural geometric explanation, by showing that there exist motions that carry a vertical hyperbolic plane into a hemispherical one, so their intrinsic geometries must be the same.[20] For now, though, we shall make do with a calculation.

If we use spherical polar coordinates (longitude and latitude) on the hemisphere of radius R, then $Z = R\cos\phi$, so (4.4) implies that the hyperbolic metric (6.23) reduces to

$$d\hat{s}^2 = \frac{\sin^2\phi\, d\theta^2 + d\phi^2}{\cos^2\phi}. \tag{6.24}$$

To see that this really is $\mathbb{H}^2$, we can either introduce new coordinates that transform this metric into the standard form (6.22) (see Ex. 28) or else we can use the intrinsic metric curvature formula (4.10) to confirm [exercise] that $\mathcal{K} = -1$.

Let us describe a couple of simple isometries of $\mathbb{H}^3$. The metric (6.23) is clearly unaltered by a horizontal Euclidean translation,

$$X + iY \longmapsto X + iY + (\text{complex constant}) \quad \text{and} \quad Z \mapsto Z.$$

Likewise, hyperbolic distance is unaltered by an origin-centred Euclidean dilation (expansion) of space,

$$(X, Y, Z) \longmapsto k(X, Y, Z) = (kX, kY, kZ),$$

where $k > 0$ is the expansion factor, for this will scale up both ds and Z by the same factor k, leaving $d\hat{s}$ unchanged.

Combining this with the translations, we see that any dilation centred on the horizon is also an isometry. The existence of these two kinds of isometries again confirms that all hemispherical planes must have the same intrinsic geometry, because any one of them can be rigidly moved to any other: translate the first so that its centre coincides with the centre of the second, then expand it till the radii are equal.

Any isometry must carry planes to planes, which means that their boundary circles in $\mathbb{C}$ must also be carried to other circles. It turns out (though we shall not be able to prove it here) that this circle-preserving property is so special that it alone suffices to completely determine[21] the complex mappings involved: *they can only be Möbius transformations!*

Conversely, *any Möbius transformation of* $\mathbb{C}$ *induces a unique transformation of all* $\mathbb{H}^3$, because a point p in $\mathbb{H}^3$ is uniquely determined as the intersection of three planes, which can be encoded as the three circles in which these hemispheres meet $\mathbb{C}$. The images of these three circles then determine three new hemispheres, meeting in the image of p. Furthermore, it can be shown[22] that this induced transformation is automatically a direct isometry of $\mathbb{H}^3$. To summarize this wonderful discovery of Poincaré (1883),

> *The group of direct isometries of hyperbolic 3-space,* $\mathbb{H}^3$*, is the Möbius group of transformations of* $\mathbb{C}$.

If you grew up in this $\mathbb{H}^3$-world, and attended $\mathbb{H}^3$-school, you would learn that angles in triangles always add up to less than two right angles, that there are infinitely many lines parallel

[20] For details, see VCA, pp. 133–136, 322–327

[21] This remarkable fact was proved, in its most general form, by Carathéodory (1937).

[22] To see how Poincaré made this discovery, see Stillwell (1996, pp. 113–122). For more on the geometry underlying the result, see VCA, pp. 322–327.

to a given line, etc. But at university you might eventually learn of a theoretical geometry that defies all your everyday experience: the angles in a triangle sum to exactly π, no matter how large the triangle! In an attempt to make intuitive sense of such bizarre phenomena, mathematicians seek to construct a model surface on which this "Euclidean" geometry actually holds true. And, remarkably, they succeed!

Returning from parable to reality, it was actually Wachter (a student of Gauss) who first discovered such a surface in 1816, called a ***horosphere***, later so named by Lobachevsky, who (along with Bolyai) independently rediscovered it. In terms of [6.6], a typical horosphere is an ordinary Euclidean sphere that rests on (touches) the horizon. Very surprisingly, if we measure distance using the hyperbolic metric (6.23), geometry within the horosphere does indeed turn out to be that of a flat Euclidean plane! See Exercises 30 and 32.

Beltrami provided a simpler way of seeing this. Using the same three-dimensional generalization of geometric inversion that allows us to swap the vertical half-planes and the hemispheres, he observed that a typical horosphere can be moved so that it becomes $Z = \mathrm{const.} = k$. Within this horizontal surface, the hyperbolic metric (6.23) then reduces to

$$d\widehat{s} = \frac{\sqrt{dX^2 + dY^2}}{k},$$

which is clearly Euclidean.

Recall that Hilbert proved that one cannot construct a full hyperbolic plane in Euclidean space. The existence of horospheres (full Euclidean planes) in hyperbolic space is therefore our first intimation of Hyperbolic Geometry's superiority, Euclidean Geometry becoming subordinate to it. In fact Spherical Geometry is subsumed as well. For it turns out that what an inhabitant of $\mathbb{H}^3$ calls a sphere (i.e., the set of points at fixed hyperbolic distance from the centre) actually appears in our model as a Euclidean sphere, but with a different centre. Furthermore, such a hyperbolic sphere has intrinsic geometry of constant positive curvature: it's as spherical as it looks! See Exercise 31.

We have repeatedly entertained the notion of intelligent creatures living in a non-Euclidean world such as $\mathbb{H}^3$; this is less fanciful than you might suppose. In 1915 Einstein discovered that the actual spacetime that *we* inhabit is *not* flat! In fact energy and matter warp the fabric of spacetime, producing a complicated pattern of curvature, both positive and negative, varying from place to place, time to time, and direction to direction.[23]

Under normal circumstance, however, the amount of curvature is so fantastically small that angle sums of triangles differ from π by an utterly undetectable amount, creating the illusion that our world obeys the laws of Euclidean Geometry. This illusion is so perfect and so compelling that it held sway for 4000 years.

Nevertheless, the subtle pattern of spacetime curvature that Einstein discovered in 1915 is extremely important, even in everyday life, and it has a name ... it is called *gravity!*

[23] Einstein's supremely beautiful theory is called ***General Relativity***, and it is the subject of Chapter 30.

Chapter 7

Exercises for Act II

Mapping Surfaces: The Metric

1. **Coordinate-Independence of $\mathcal{K}$.** Begin with the flat Euclidean metric,
$$d\hat{s}^2 = dx^2 + dy^2.$$
   (i) Now change coordinates by writing $x = f(u,v)$ and $y = g(u,v)$, where f and g are explicit, *simple* functions of your choosing, and find the new metric formula in the (u,v) coordinates.
   (ii) Apply the curvature formula (4.10) to confirm that the new metric is *still* flat, as it should be: $\mathcal{K} = 0$.
   (iii) Redo (ii) in the case where f and g are completely general.

2. **Central Projection.** Consider the central projection of the sphere shown in [4.1].
   (i) Prove that an infinitesimal circle on the sphere is distorted into an ellipse whose shape is given by (4.3).
   (ii) Explain why this distortion is symmetrical: if we instead start with an infinitesimal circle in the plane, the image on the sphere is an ellipse of the same shape.

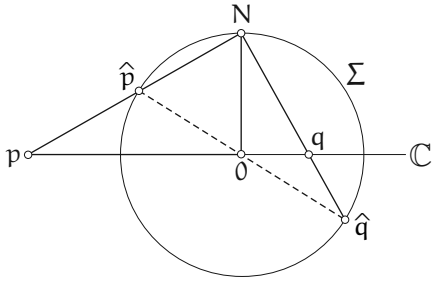
3. **Central Projection.** Rederive the central projection metric (4.2) by calculation.

4. **Central Projection.** Apply the curvature formula (4.10) to the metric (4.2) of the sphere under central projection, and confirm that $\mathcal{K} = +1/R^2$.

5. **Angular Excess of n-gon.** Show that the internal angles of a Euclidean n-gon sum to $(n-2)\pi$, thereby justifying the definition of angular excess as $\mathcal{E} \equiv [\text{Angle Sum}] - (n-2)\pi$. Now take Δ_p to be a small n-gon containing a point p on a curved surface, and divide it into triangles by joining p to the vertices. Deduce that (2.1) applies without change.

6. **Stereographic Metric.** Suppose that in [4.10] we choose $\delta\hat{s}$ due north (instead of the illustrated choice of due west). Show that this leads to the same metric (4.22), and do so in two ways:
   (i) by calculation;
   (ii) geometrically.

7. **Antipodal Points.** The figure below shows a vertical cross section of the Riemann sphere Σ through $\hat{p}$ and the antipodal point $\hat{q}$. Show that the triangles $p0N$ and $N0q$ are similar. Deduce (4.27).



8. Conformal ("Isothermal") Coordinates. Given the metric of a general surface in general (u, v)-coordinates, (4.7),

$$d\hat{s}^2 = A^2\,du^2 + B^2\,dv^2 + 2F\,du\,dv,$$

our aim is to find *conformal* (U, V)-coordinates, such that if $Z = U + iV$, then

$$d\hat{s}^2 = \Lambda^2\,(dU^2 + dV^2) = \Lambda^2\,dZ\,d\overline{Z}.$$

Let $M = \sqrt{A^2B^2 - F^2}$ be the area magnification factor in going from the (u, v)-map to the surface, so that (4.11) becomes $d\mathcal{A} = M\,du\,dv$.

(i) Verify that the metric has the following *complex* factorization:

$$d\hat{s}^2 = \left[A\,du + \frac{(F + iM)}{A}\,dv\right]\left[A\,du + \frac{(F - iM)}{A}\,dv\right].$$

(ii) Suppose that a *complex* integrating factor Ω can be found such that

$$\Omega\left[A\,du + \frac{(F + iM)}{A}\,dv\right] = dU + i\,dV = dZ.$$

Deduce that in this case (U, V) are indeed conformal coordinates on the surface:

$$d\hat{s}^2 = \Lambda^2\,(dU^2 + dV^2) = \Lambda^2\,dZ\,d\overline{Z}, \qquad \text{where} \qquad \Lambda = \frac{1}{|\Omega|}.$$

(iii) Use $df = (\partial_u f)\,du + (\partial_v f)\,dv$ to deduce that

$$\partial_u Z = \Omega A \qquad \text{and} \qquad \partial_v Z = \Omega\left[\frac{F + iM}{A}\right].$$

(iv) Deduce that

$$\big[F + iM\big]\,\partial_u Z = A^2\,\partial_v Z.$$

(v) Multiplying both sides of the previous equation by $\big[F - iM\big]$, deduce that

$$B^2\,\partial_u Z = \big[F - iM\big]\,\partial_v Z.$$

(vi) By equating real and imaginary parts, deduce that the rates of change of U can be expressed in terms of the rates of change of V (and visa versa) as follows:

$$\partial_u U = \frac{1}{M}\left[A^2\,\partial_v V - F\,\partial_u V\right] \qquad \text{and} \qquad \partial_v U = \frac{1}{M}\left[F\,\partial_v V - B^2\,\partial_u V\right];$$

$$\partial_u V = \frac{1}{M}\left[F\,\partial_u U - A^2\,\partial_v U\right] \qquad \text{and} \qquad \partial_v V = \frac{1}{M}\left[B^2\,\partial_u U - F\,\partial_v U\right].$$

(vii) Granted that $\partial_u\partial_v\Phi - \partial_v\partial_u\Phi = 0$, deduce that $\Phi = U$ and $\Phi = V$ are both solutions of the ***Beltrami–Laplace Equation***:

$$\boxed{\partial_v\left[\frac{A^2\,\partial_v\Phi - F\,\partial_u\Phi}{M}\right] + \partial_u\left[\frac{B^2\,\partial_u\Phi - F\,\partial_v\Phi}{M}\right] = 0.}$$

This equation is ***elliptic***, and the general theory of elliptic partial differential equations now guarantees that this equation does indeed have solutions, thereby confirming the existence of conformal coordinates!

(viii) Let (x, y) be a second pair of conformal coordinates, so that

$$d\hat{s}^2 = \lambda^2(dx^2 + dy^2).$$

By setting $u = x$ and $v = y$, deduce that $A = B = \lambda$, $F = 0$, and $M = \lambda^2$. From (vi), deduce that the local mapping between the $(x + iy)$-plane and the $(U + iV)$-plane is an *amplitwist*, characterized by the celebrated ***Cauchy–Riemann equations***:

$$\partial_x U = \partial_y V \qquad \text{and} \qquad \partial_y U = -\partial_x V.$$

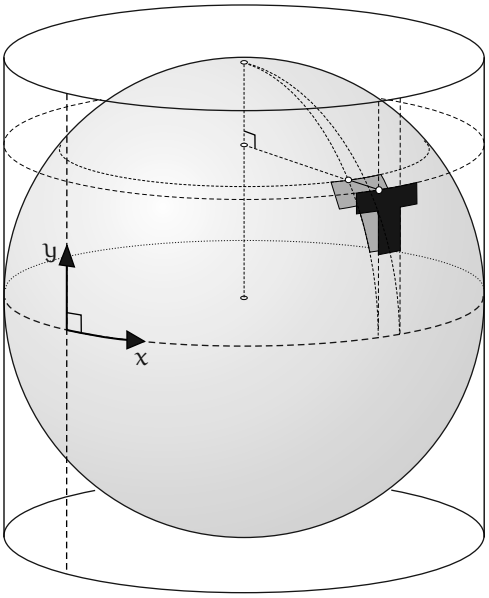
(For a complete discussion of these phenomena, see VCA.)

(ix) Deduce that in this case the Beltrami–Laplace Equation reduces to the ***Laplace Equation***:

$$\partial_x^2 \Phi + \partial_y^2 \Phi = 0.$$

9. Conformal Curvature Formula. Apply the conformal curvature formula (4.16) to the metric (4.22) of the sphere under stereographic projection, and confirm that $\mathcal{K} = +1/R^2$.

10. Archimedes–Lambert Projection. In the (x, y)-plane, consider the rectangle $\{0 \leqslant x \leqslant 2\pi R, -R \leqslant y \leqslant R\}$. Now imagine first taping together the left and right edges to create a cylinder, then slipping this over a sphere of radius R so that the cylinder touches the sphere along its equator, as shown in the figure below. We now draw a map of the sphere by projecting its



points horizontally onto the cylinder, radially outward, perpendicular to the axis of the cylinder, as shown. This is the ***Archimedes–Lambert projection***, first investigated by Archimedes

(c. 250 BCE), then rediscovered and published by Lambert about 2000 years later, in 1772. Lambert's seminal treatise was the first systematic mathematical investigation of maps, based on which properties they preserved.

(i) Show (ideally geometrically) that the metric of the sphere now takes the form

$$d\widehat{s}^2 = (R^2 - y^2)\,dx^2 + \frac{dy^2}{R^2 - y^2}.$$

(ii) Apply the curvature formula (4.10) to confirm that $\mathcal{K} = +1/R^2$.

(iii) Use (4.12) to deduce that the projection *preserves area*. For example, the two illustrated T-shapes have equal area. As Archimedes realized, this implies that the area of the entire sphere must equal that or the original rectangle: $(2\pi R)(2R) = 4\pi R^2$. Archimedes was so proud of this (and related discoveries on volumes) that he asked his friends to have this particular diagram inscribed on his tomb. Almost a century-and-a-half later, in 75 BCE, Cicero sought and found the tomb, overgrown with bushes, but with the cylinder and sphere still visible.

(iv) Use the previous part to deduce (without integration) that the area $\mathcal{A}$ of the polar cap $0 \leqslant \phi \leqslant \Phi$ is

$$\mathcal{A} = 2\pi R^2(1 - \cos\Phi).$$

11. Central Cylindrical Projection. Reconsider the sphere and cylinder in the diagram of the previous question, but now imagine that the cylinder extends infinitely upward and downward. We again project the sphere onto the cylinder, but this time by imagining a point source of light at the centre of the sphere, casting shadows onto the cylinder. This is called ***central cylindrical projection***.

(i) Sketch the projection and deduce that $y = R\cot\phi$.

(ii) Either using part (i), or directly geometrically, show that the metric now takes the form

$$d\widehat{s}^2 = \frac{R^2}{R^2 + y^2}\,dx^2 + \frac{R^4}{(R^2 + y^2)^2}\,dy^2.$$

(iii) Apply the curvature formula (4.10) to confirm that $\mathcal{K} = +1/R^2$.

12. Mercator Projection. The projections of the sphere onto the cylinder in the the previous two exercises share two properties: 1) meridians ($\theta = \mathrm{const.}$) are mapped to vertical generators of the cylinder with the same θ; 2) circles of latitude ($\phi = \mathrm{const.}$) are mapped to horizontal circular cross sections of the cylinder. In 1569 Gerardus Mercator discovered a third projection that shared these properties but that enjoyed the crucial advantage over the other two of being *conformal*.

(i) Argue geometrically that if $\delta\widehat{s}$ is a small movement along a circle of latitude on the sphere then the two required properties above imply that $\delta\widehat{s} \asymp \sin\phi\,\delta s$, where $\delta s = \delta x$ is the horizontal movement on the cylinder.

(ii) For our map to be conformal we must insist that we have the same scale factor for infinitesimal movements in *all* directions. In particular, suppose $\delta\widehat{s}$ is instead chosen along a meridian, and let $y = f(\phi)$ be the height on the cylinder of the image of (θ, ϕ) on the sphere. Deduce that

$$f'(\phi) = \frac{R}{\sin\phi}.$$

(iii) If we insist that points on the equator map to the x-axis, verify (or deduce, if you remember your integration techniques) that the ***Mercator projection*** is given by

$$y = f(\phi) = R \ln\left[\frac{1+\cos\phi}{\sin\phi}\right].$$

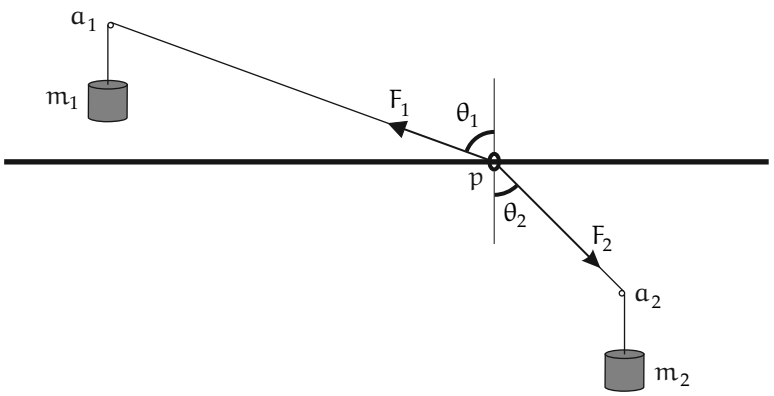
(Note: Neither logarithms nor calculus were known in Mercator's time, so *how did he do it?!* See the excellent chapter by Osserman in Hayes et al. (2004, Ch. 18).)

(iv) If you fly a plane or sail a ship on a fixed compass heading, you travel along a ***loxodrome*** (illustrated in [6.5c]), also called a ***rhumb line***. If you unroll the cylinder onto a tabletop to obtain a standard flat Mercator navigational chart, how does your loxodromic route appear in the chart? (*Hint:* No calculation required: just use conformality!)

13. An *Impossibly* Good Map. In the last few exercises we have met a map of the sphere that preserves area, and another that preserves angles. Can we have a map that preserves both? Show that no such map can exist for *any* curved surface, not just the sphere.

The Pseudosphere and the Hyperbolic Plane

14. Pólya's Mechanical Proof of Snell's Law. Pólya (1954, pp. 149–152) provided the following ingenious *mechanical* explanation of Snell's Law, and hence yet another way for us to think about the geodesics of the hyperbolic plane. The figure below is a modified form of [5.6], in which the boundary between air and water is replaced by a frictionless rod, along which slides a ring, p. Attached to this ring are two strings of fixed length. As illustrated, each string passes over a frictionless peg at a_i and is attached to a mass m_i that hangs a distance h_i below a_i.



(i) By considering the potential energy of the system, deduce that equilibrium is achieved when

$$m_1 h_1 + m_2 h_2 = \text{maximum}.$$

(ii) But for equilibrium to occur, the horizontal components of the forces F_i pulling on the ring must cancel. Deduce that

$$m_1 \sin\theta_1 = m_2 \sin\theta_2.$$

(iii) Let l_i be the length of string from a_i to p, so that $l_i + h_i = \text{const.}$. Now let us choose $m_i = (1/v_i)$, where v_1 and v_2 represent the speed of light in air and in water, respectively, as in the original optical problem. Deduce that the mechanical problem in part (i)

becomes mathematically identical to the original problem of finding the optical path of minimum time:

$$\frac{l_1}{v_1} + \frac{l_2}{v_2} = \text{minimum}.$$

(iv) From the solution of the mechanical problem in part (ii), deduce Snell's Law!

15. If the velocity of light in the (x, y)-plane is $v = 1/\sqrt{1-y}$, show that the geodesics are parabolas. Describe these parabolas.

16. **Tractrix Parameterization.** In (X, Y, Z)-space let us switch to ***cylindrical polar coordinates***, i.e., ordinary (r, θ) polar coordinates in the (X, Y)-plane, supplemented by Z. The Euclidean metric in this space is therefore

$$ds^2 = dr^2 + r^2 d\theta^2 + dZ^2.$$

In the plane $\theta = 0$, i.e., the (X, Z)-plane, consider the curve traced by a particle whose position at time t is

$$X = r = \operatorname{sech} t, \qquad Z = t - \tanh t.$$

(i) Show that this curve is the tractrix: the distance along the tangent to the Z-axis is fixed. What is this fixed distance? (Note: some interesting history and beautiful geometry underlies this particular representation of the tractrix; see Stillwell (2010, pp. 341–342).)

(ii) Deduce that if we rotate the curve about the Z-axis, we obtain the pseudosphere of unit radius. Show that the surface's metric is

$$d\hat{s}^2 = \tanh^2 t \, dt^2 + \operatorname{sech}^2 t \, d\theta^2.$$

(iii) Use (4.10) to confirm that $\mathcal{K} = -1$.

(iv) Show that if we now introduce new coordinates $x = \theta$ and $y = \cosh t$ then the metric assumes the standard form (5.7).

17. **Cylindrical-Polar Area On Sphere.** If we employ the cylindrical polar coordinates of the previous exercise, the element of area on the northern hemisphere of the unit sphere is,

$$d\mathcal{A} = \frac{r \, dr \, d\theta}{\sqrt{1-r^2}}.$$

(i) Prove this by calculation.

(ii) Prove this geometrically, using Newtonian ultimate equalities.

(iii) Integrate this to obtain the area of the complete hemisphere.

18. **The Pseudosphere Has Finite Area!** Use (5.5) and (4.12) to show that the infinite pseudosphere of radius R has *finite* area $2\pi R^2$. This was discovered by Huygens in 1693.

19. **Build Your Own Pseudosphere!** Take a stack of ten sheets of paper and staple them together, placing staples along three of the edges. Use a pair of compasses (or inverted plate or bowl) to draw the largest circle that will fit comfortably inside the top sheet. Pierce through all ten sheets in the centre of the circle. With heavy scissors, cut along the circle to obtain ten identical discs, say of radius R. Repeat this whole process to double the number of discs to 20.

(i) Cut a narrow sector out of the first disc, and tape the edges together to form a shallow cone. Repeat this process with the remaining discs, steadily increasing the angle of the sector each time, so that the cones get sharper and taller. (You may find it easier to cut a radial slit and *overlap* the paper before taping it together.) Ensure that by the end of the process you are making very narrow cones, using only a quarter disc or less.

(ii) Stack these cones in the order you made them. Explain how it is that *you have created a model of a portion of a pseudosphere of radius* R.

(iii) Use the same idea to create a disc-like piece of "hyperbolic paper," such as you would get if you could simply cut out a disc from your pseudosphere. Press it against the pseudosphere and verify that you can freely move it about and rotate it on the surface. (For detailed instructions, see Henderson (1998, p. 32).)

20. Geodesics of the Pseudosphere. Using (1.7), construct a segment of a typical geodesic on the surface of the toy pseudosphere you built in the previous exercise. Extend this segment in both directions, one strip of tape at a time. Note the surprising way the geodesic only spirals a finite distance up the pseudosphere before spiralling down again.

(i) Use the Beltrami–Poincaré upper half-plane to verify mathematically that the tractrix generators are the *only* geodesics that extend indefinitely upward.

(ii) Let L be a typical geodesic, and let α be the angle between L and the tractrix generator at the point where L hits the rim $\sigma = 0$. Show that the maximum distance $\sigma_{\max}$ that L travels up the pseudosphere is given by $\sigma_{\max} = |\ln \sin \alpha|$.

21. Conformal Curvature Formula. Show that in the case of a conformal mapping with metric (4.14) the general curvature formula (4.10) reduces to (4.16).

22. Surfaces of Revolution of Constant Curvature. Imagine a particle travelling along a curve in the (x, y)-plane at unit speed, and let its position at time t be $[x(t), y(t)]$. Now imagine rotating this plane through angle θ about the x-axis. As θ varies from 0 to 2π, the curve sweeps out a surface of revolution.

(i) Explain why $\dot{x}^2 + \dot{y}^2 = 1$, where the dot represents the time derivative.

(ii) Show geometrically that the metric of the surface is $d\hat{s}^2 = dt^2 + y^2\, d\theta^2$.

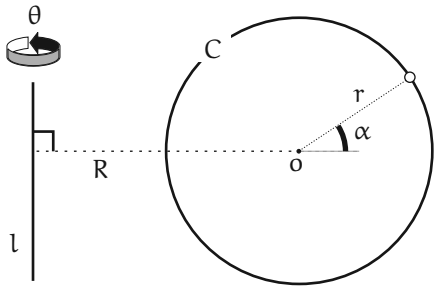
(iii) Deduce from that (4.10) that $\mathcal{K} = -\ddot{y}/y$.

(iv) Find the general solution $y(t)$ in each of the three cases of constant curvature: $\mathcal{K} = 0$, $\mathcal{K} = +1$, and $\mathcal{K} = -1$.

(v) With the assistance of part (i), calculate the velocity of the particle in each of the three cases.

(vi) Sketch some solution curves and resulting surfaces of revolution in each of the three cases; a computer may prove helpful. In particular, in the case $\mathcal{K} = +1$, confirm that one does indeed obtain surfaces like those shown in [2.5]. Likewise, verify that in the case $\mathcal{K} = -1$ one can obtain not only the pseudosphere itself, but also surfaces that look like two pseudospheres stuck together at their narrow necks.

23. Curvature of the Torus. Generate a torus T by rotating a circle C with centre o and radius r about a line l in the plane of C that is at distance $R (> r)$ from o. See the figure below.



(i) If α is the illustrated angle going round the circle, and θ is the illustrated angle of rotation about the axis of symmetry l, show geometrically that the metric of T is

$$d\hat{s}^2 = r^2 d\alpha^2 + (R + r\cos\alpha)^2 d\theta^2.$$

(ii) Use (4.10) to deduce that the curvature of the torus is

$$\mathcal{K} = \frac{1}{r(r + R\sec\alpha)}.$$

(iii) Sketch the graph of $\sec\alpha$, then of $(r + R\sec\alpha)$, and finally of $\mathcal{K}$ itself. Thereby confirm that this formula for $\mathcal{K}$ agrees with the empirical findings of Fig. [2.2].

(iv) Where on T would you expect to find $\mathcal{K} = 0$? Confirm this with the formula.

(v) We see from the formula that $\lim_{R\to\infty} \mathcal{K} = 0$. Explain this geometrically.

(vi) Use (4.12) to find the total area of T.

(vii) Check that your formula for the area agrees with the prediction of ***Pappus's Centroid Theorem***. (If you have not studied this theorem previously, first look it up in a book or on the internet.)

(viii) Show that the total curvature of the torus *vanishes*:

$$\oiint_T \mathcal{K}\, dA = 0.$$

In Act III we shall see that this is no accident!

Isometries and Complex Numbers

24. Amplitwist of z^m. Generalize the geometric argument in [4.6] to deduce that the amplitwist of z^m is given by the same formula as in real calculus: $(z^m)' = m\, z^{m-1}$; but now we can *see* why it's true!

25. Metric of the Beltrami–Poincaré Disc. We know from [6.3] that Möbius transformations can map half-planes to discs, and, as we shall be able to explain later, it turns out that a specific Möbius transformation from the upper half-plane model of $\mathbb{H}^2$ [5.12] to the disc model of $\mathbb{H}^2$ [5.11] is

$$z \mapsto w = D(z) = \frac{iz + 1}{z + i}.$$

To derive the metric (5.15) of the disc model, consider an infinitesimal vector dz emanating from z (in the half-plane) being amplitwisted to an infinitesimal vector $dw = D'(z)\, dz$ emanating from w (in the disc). By definition, the hyperbolic length $d\hat{s}$ of dw is the hyperbolic length of dz. Verify (5.15) by showing that

$$\frac{2\,|dw|}{1 - |w|^2} = \frac{|dz|}{\operatorname{Im} z} = d\hat{s}.$$

26. Curvature of the Beltrami–Poincaré Disc. Use the conformal curvature formula (4.16) to verify that the Beltrami–Poincaré disc [5.11] with metric (5.15) has constant negative curvature $\mathcal{K} = -1$.

27. Möbius Rotations of the Riemann Sphere. Our aim here is to provide a constructive, semi-geometrical proof that *every rotation of the Riemann sphere is a Möbius transformation of the type (6.10)*. Our approach coincides in its essentials with the one previously published by Wilson (2008, pp. 42–44). (Note: the corresponding group of matrices is written $SU(2)$: the "S" stands

for "special," meaning normalized, with unit determinant; the "U" stands for "unitary"; and the "2" stands for "2×2" matrices.) In the following, let R_Z^θ represent a clockwise rotation of the Riemann sphere through angle θ about the positive Z-axis of (X, Y, Z)-space, so that $R_Z^\theta(z) = e^{i\theta} z$.

(i) Find a matrix $\left[R_Z^\theta\right]$ in $SU(2)$ representing R_Z^θ.

(ii) Explain why the result (6.6) can be written

$$[R_X^\pi] = \pm \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} \in SU(2).$$

(iii) Referring to (4.24), use a sketch to show that the complex mapping corresponding to $R_X^{(\pi/2)}$ is

$$z = \frac{X + iY}{1 - Z} \longmapsto R_X^{(\pi/2)}(z) = \frac{X - iZ}{1 - Y}.$$

(iv) To confirm that $R_X^{(\pi/2)}(z)$ is in fact a *Möbius* transformation, let us suppose for now that it is, and let us try to guess its form. Use a sketch to confirm that $-i \mapsto 0$, so the numerator must be proportional to $(z + i)$. Likewise, by finding the point that rotates to the north pole, deduce the denominator. This determines the form of $R_X^{(\pi/2)}(z)$ up to a multiplicative factor; determine this factor by noting that $0 \mapsto i$. Conclude that *if* $R_X^{(\pi/2)}(z)$ is a Möbius transformation, it can only be

$$R_X^{(\pi/2)}(z) = \frac{z + i}{iz + 1}.$$

(v) Show that this *is* in fact $R_X^{(\pi/2)}(z)$ by demonstrating that it satisfies the equation in part (iii). (*Hints:* Multiply top and bottom by $(X - iZ)$, but only multiply through in the denominator. Finally, in the resulting denominator, use the fact that $X^2 + Y^2 + Z^2 = 1$.)

(vi) Deduce that $R_X^{(\pi/2)}$ can be represented in $SU(2)$ as

$$\left[R_X^{(\pi/2)}\right] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}.$$

(vii) Explain geometrically why $\left[R_X^{(\pi/2)}\right]^2 = \left[R_X^\pi\right]$, and confirm this by direct calculation.

(viii) Explain geometrically why for any angle α,

$$[R_Y^\alpha] = \left[R_X^{-(\pi/2)}\right] [R_Z^\alpha] \left[R_X^{(\pi/2)}\right],$$

and deduce that $\left[R_Y^\alpha\right] \in SU(2)$.

(ix) Finally, consider a general rotation about an arbitrary axis A. By rotating A into the (X, Y)-plane and then rotating it to the Y-axis, deduce that this too can be represented as a matrix in $SU(2)$, thereby concluding the proof.

28. Metric of $\mathbb{H}^2$ within $\mathbb{H}^3$. Let us explore different ways of expressing the metric (6.24) of a typical hyperbolic planes in $\mathbb{H}^3$, namely, a horizon-centred hemisphere.

(i) Show that if we define $u \equiv \tan \phi$, the metric (6.24) becomes

$$d\hat{s}^2 = u^2 \, d\theta^2 + \frac{du^2}{1 + u^2},$$

and use (4.10) to confirm that $\mathcal{K} = -1$.

(ii) Show that if we define a new variable ξ via $u \equiv 1/\sinh\xi$, the metric of the previous part becomes *conformal*:

$$d\hat{s}^2 = \frac{d\theta^2 + d\xi^2}{\sinh^2\xi},$$

and use (4.16) to confirm that $\mathcal{K} = -1$.

(iii) Finally, find a conformal mapping of the (θ, ξ)-plane to the (x, y)-plane such that the conformal hyperbolic metric in part (ii) takes the standard form (5.7):

$$d\hat{s}^2 = \frac{dx^2 + dy^2}{y^2}.$$

(Hint: Set $dy/y = d\xi/\sinh\xi$.)

29. Curvature of Hemispherical $\mathbb{H}^2$ in $\mathbb{H}^3$. Use the curvature formula (4.10) to verify that the horizon-centred hemispheres of $\mathbb{H}^3$ with metric (6.24) are indeed $\mathbb{H}^2$'s with constant negative curvature $\mathcal{K} = -1$.

30. Horosphere: Metric and Curvature. As discussed, a typical horosphere of $\mathbb{H}^3$ appears as a sphere resting on $\mathbb{C}$. Using the derivation of (6.24) as your inspiration, show that the metric of such a horosphere is

$$d\hat{s}^2 = \frac{\sin^2\phi\, d\theta^2 + d\phi^2}{(1+\cos\phi)^2}.$$

Apply the curvature formula (4.10) to confirm that the horosphere is intrinsically a flat Euclidean plane: $\mathcal{K} = 0$.

31. Spheres in $\mathbb{H}^3$. Consider a Euclidean sphere of radius R centred at height kR above the horizon of $\mathbb{H}^3$. Using the derivation of (6.24) as your inspiration, show that the metric of such a sphere is

$$d\hat{s}^2 = \frac{\sin^2\phi\, d\theta^2 + d\phi^2}{(k+\cos\phi)^2}.$$

Apply the curvature formula (4.10) to confirm the claim in the text that if this sphere lies entirely above the horizon (i.e., $k > 1$) then intrinsically this surface is genuinely spherical, with constant positive curvature, $\mathcal{K} = k^2 - 1$. (Note: the results of the previous two exercises follow as the special cases $k = 0$ and $k = 1$, respectively.)

32. Horosphere Metric. Let us derive the result of Exercise 30 in another way. Define a coordinate r such that

$$dr = \frac{d\phi}{1+\cos\phi}.$$

Show that the metric of the horosphere (see Ex. 30) can then be written as $d\hat{s}^2 = dr^2 + r^2\, d\theta^2$, which is just the metric of the Euclidean plane in polar coordinates.

33. Penrose's Direct Labelling of Light Rays with Complex Numbers. In the text we used the Riemann sphere to label light rays with complex numbers, via stereographic projection. Sir Roger Penrose (see Penrose and Rindler (1984, Vol. 1, p. 13)) discovered a remarkable alternative method of passing from a light ray to the associated complex number *directly*. Let the point P be one unit vertically above the origin of the (horizontal) complex plane. Now imagine that, simultaneously, P emits a flash, and $\mathbb{C}$ begins to travel straight up (in the direction $\phi = 0$) at the speed of light ($= 1$) towards P. (You may imagine that the whole of $\mathbb{C}$ flashed, creating a plane wave.) Decompose the velocity of the photon F emitted by P in the direction (θ, ϕ) into components perpendicular and parallel to $\mathbb{C}$. Hence find the time at which F hits $\mathbb{C}$. Deduce that F hits $\mathbb{C}$ at the point $z = \cot(\phi/2)\, e^{i\theta}$. Thus, *Penrose's construction is equivalent to stereographic projection!*

34. Einstein's Aberration Formula. Recall from Special Relativity that a "boost" v along the Z-axis yields the Lorentz transformation formulas,

$$\widetilde{T} = \frac{T + vZ}{\sqrt{1-v^2}}, \qquad \widetilde{X} = X, \qquad \widetilde{Y} = Y, \qquad \widetilde{Z} = \frac{Z + vT}{\sqrt{1-v^2}}.$$

(i) Show that this transformation can be rewritten as

$$\widetilde{T} + \widetilde{Z} = \rho(T + Z), \qquad \widetilde{X} = X, \qquad \widetilde{Y} = Y, \qquad \widetilde{T} - \widetilde{Z} = (1/\rho)(T - Z),$$

where

$$\rho = \sqrt{\frac{1+v}{1-v}}.$$

(ii) Deduce from (6.19) that this boost can be represented by the spin transformation,

$$\begin{bmatrix} \widetilde{\mathfrak{z}1} \\ \widetilde{\mathfrak{z}2} \end{bmatrix} = \begin{bmatrix} \sqrt{\rho} & 0 \\ 0 & 1/\sqrt{\rho} \end{bmatrix} \begin{bmatrix} \mathfrak{z}1 \\ \mathfrak{z}2 \end{bmatrix}.$$

(iii) Thereby confirm the claim made in [6.5b] and (6.21).

(iv) If (θ, ϕ) is the apparent direction of a star before you fire your spaceship engines, deduce that after you have boosted to speed v in the direction $\phi = 0$, the new direction $(\theta, \widetilde{\phi})$ of the star is given by

$$\cot\frac{\widetilde{\phi}}{2} = \rho \cot\frac{\phi}{2}.$$

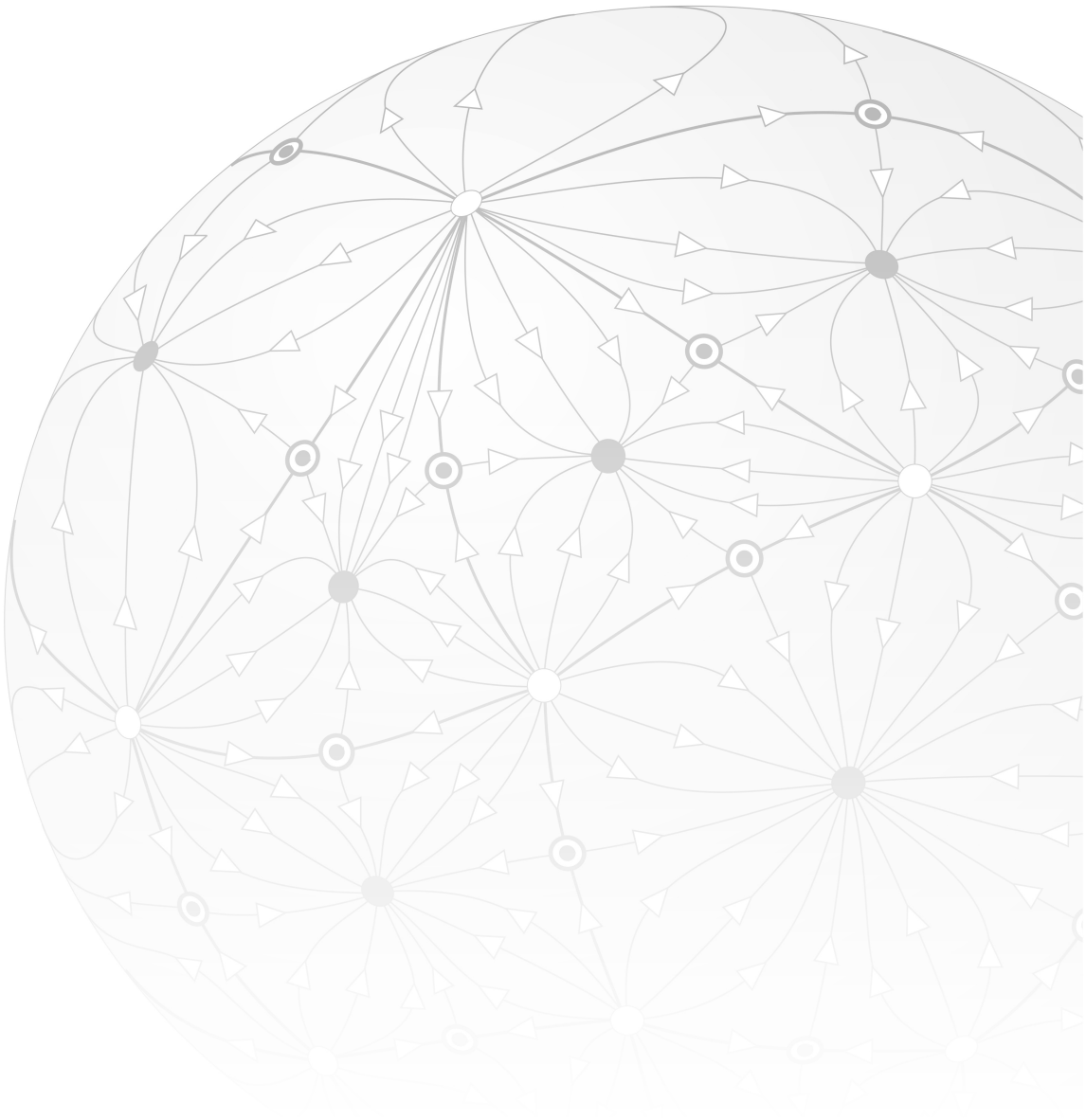
This is a more elegant and memorable form of the standard ***aberration formula***, which was discovered by Einstein in 1905.

(v) Deduce that the star field appears to crowd together towards the north pole, i.e., *towards your direction of travel*.

(vi) What happens to the star field as the speed of your spacecraft approaches the speed of light?(!)

ACT III

Curvature



Chapter 8

Curvature of Plane Curves

8.1 Introduction

The study of curvature did not begin with surfaces. Rather, as the very name suggests, curvature was born out of the study of *curves*, specifically, curves *in the plane*.

We certainly have an intuitive idea of a plane curve being straighter in one part and more curved in another, but how can we quantify this degree of curving with mathematical precision?

Throughout the first Act we focussed on the concept of *intrinsic* (Gaussian) curvature. And at the close of Act I we provided another compelling justification for this obsession with intrinsic (versus extrinsic) geometry: it is the *only* kind of geometry of spacetime that is meaningful to us as inhabitants of spacetime, and spacetime's intrinsic curvature has a fundamental importance—it is gravity.

But whereas a patch of given 2-dimensional surface could only be deformed in very limited ways while preserving its intrinsic geometry (i.e., without stretching distances within it) the same is *not* true of a 1-dimensional curve: we can bend it (without stretching it) so as to take up the form of *any* other curve of the same length. Thus the very concept of 1-dimensional intrinsic curvature simply cannot exist! Put differently, if you were a very short 1-dimensional being living within a curve, only able to measure distance back and forth along the curve as you moved along it, the *shape* of your curve would not merely be unknowable to you, it would be *meaningless*.

Thus, from the outset, it is clear that the best we can hope for is an *extrinsic* definition of curvature: how the curve sits within the plane. The big surprise—"miracle" would be a better word—is that this extrinsic concept of 1-dimensional curvature will ultimately be seen to have a direct bearing on the *intrinsic* curvature of 2-dimensional surfaces.

As we have said, a 1-dimensional being within a curve cannot sense the curvature of his world by measurements within it. But suppose we relax our notion of what is measurable and knowable for such a being. Instead of only permitting measurements of length along the curve, suppose we admit physical concepts such as mass, velocity, and force. For now, let us continue to restrict ourselves to a *plane* curve. Picture this plane curve as a frictionless wire located in outer space, with no gravitational (or other) forces at work, and suppose we launch a frictionless bead of unit mass at unit speed along the wire.[1]

By virtue of the absence of external forces, the bead will continue to travel at unit speed throughout its journey along the wire. But Newton's First Law of Motion tells us that if *no* forces acted on the bead, it would simply travel in a straight line. What is happening here is that as the bead attempts to fly in this straight line, the wire presses on it, at right angles to the direction of motion, rotating the velocity vector so as to make it continue pointing along the wire, but without changing the length of the velocity, which is the speed.

The more tightly curved a section of wire, the quicker the velocity vector must turn as it passes through it. Newton's Second Law of Motion tells us that this rate of change of the velocity vector (acceleration) is in fact equal (for a unit mass) to the force F the curved wire exerts on the

[1]Of course while this still essentially concerns a 1-dimensional curve in a 2-dimensional plane, we are now viewing all this as embedded in our physical 3-dimensional world. Indeed, we need at least three dimensions for the bead to encircle and cling to the wire.

bead. The sharper a bend in the wire, the greater the force the wire must exert to make the bead's trajectory bend.

To make this idea more visceral, recall what it's like to travel at high (constant) speed on a ride along curved track at an amusement park or fairground. You have become effectively a 1-dimensional being, carried along the path of the ride. (We will come to curves that twist through 3-space soon, but, for now, imagine you are travelling along a *horizontal* (therefore planar) stretch of track.) Even with your eyes shut, you can still *feel* the force as you go round a bend, the sides of the car pressing against your body: the more tightly curved the bend, the greater the force.

Newton was in fact the first to introduce a purely geometric definition of curvature κ, which we shall meet it in moment, and as a consequence of his definition and his laws of motion, he was indeed able to deduce that

> *If a bead of unit mass is launched at unit speed along a wire in the form of a plane curve, the* curvature κ *of that curve is the force* F*, directed perpendicularly to the curve, exerted by the wire on the bead.* (8.1)

This connection between geometry and physics (between the geometric bending of an orbit and the force required to hold the object in that orbit) became the linchpin of Newton's *magnum opus*, arguably the single most important scientific work in history, the great *Principia* of 1687.

In it, Newton used infinitesimal geometry—the "ultimate equalities" described in the Prologue—to explain the workings of the heavens, including the elliptical orbits of the planets about the Sun as the primary focus. Only now the planets are not threaded onto wires, rather their orbits are deflected from rectilinear motion by the invisible hand of gravity, reaching out across space from the Sun, diminishing in strength in proportion to the *square* of the distance: Newton's famous ***Inverse-Square Law of Gravitation***.

8.2 The Circle of Curvature

The *Principia* lay more than two decades in the future when (shortly before his Christmas Day birthday[2] in 1664) the 21-year-old Newton began to investigate what he called the "crookednesse" of plane curves, thereby introducing the concept of curvature into mathematics for the first time.

Newton identified the ***circle of curvature*** at a point p on a curve $\mathcal{C}$ as the circle that best approximates the curve in the immediate vicinity of p, just as the tangent is the *line* that does this best (see [8.1]).

He constructed the centre c (the ***centre of curvature***) of this approximating circle as the limiting position of the intersection of the normal at p with the normal at a neighbouring point q, in the limit $q \to p$. Then pc is called the ***radius of curvature***, and $\kappa \equiv (1/pc)$ is what Newton initially dubbed the "crookednesse," but later rechristened as the ***curvature***. (IMPORTANT NOTE ON NOTATION: Here "pc" denotes the ***distance between*** the points p and c.)

Next, following Newton, we measure κ by looking at how fast the curve departs from its own tangent line (see [8.2]). By definition of the curvature κ at p, the illustrated diameter has length $ps = (2/\kappa)$. Now let q be a point on $\mathcal{C}$ near to p (where $\xi = pq$) and drop a perpendicular of length $qt = \sigma$ from q to the tangent $\mathcal{T}$ at p, and finally let $\epsilon = pt$.

Since $\mathcal{T}$ is tangent to $\mathcal{C}$, $\lim_{\epsilon \to 0}(\sigma/\epsilon) = 0$, and therefore

$$\frac{\xi^2}{\epsilon^2} = \frac{\epsilon^2 + \sigma^2}{\epsilon^2} = 1 + \left[\frac{\sigma}{\epsilon}\right]^2 \asymp 1 \quad \Longrightarrow \quad \xi \asymp \epsilon.$$

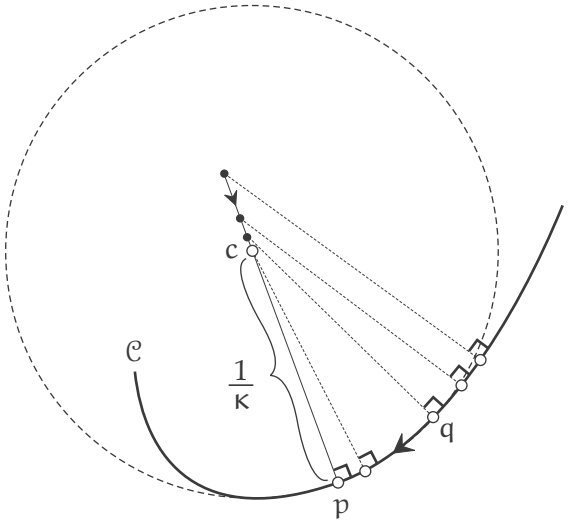
[2] In my household we (or at least *I*) refer to this event as "Newtonmas"!

IMPORTANT NOTE ON NOTATION: Here, and throughout the book, we shall employ the $\asymp$ notation, representing Newton's concept of ***ultimate equality***, as introduced and defined in the Prologue.

Also, the shaded triangle ptq is ultimately similar to the triangle sqp, so

$$\frac{\xi}{[2/\kappa]} \asymp \frac{\sigma}{\xi}.$$

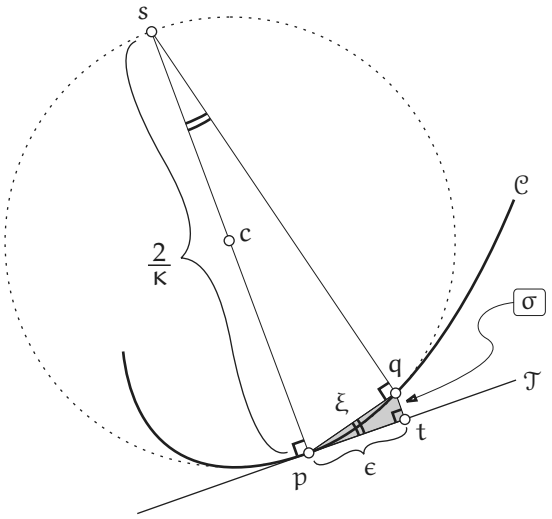
This is essentially Newton's Lemma II, from Book I of the *Principia* (Newton 1687 p. 439) (see also Brackenridge and Nauenberg (2002, p. 112)). Depending on whether it is κ or σ that needs to found in terms of the other, we may combine the previous two results to deduce that



[8.1] *The* **circle of curvature** *at* p *is the circle that best approximates the curve there. The* **curvature** κ *is defined to be the reciprocal of its radius.*

$$\kappa \asymp \frac{2\sigma}{\epsilon^2} \quad \text{or} \quad \sigma \asymp \frac{1}{2}\kappa\epsilon^2. \tag{8.2}$$

It is convenient to attach a *sign* to the curvature: κ is *positive* if $\mathcal{C}$ is concave *up*, and negative if it is concave down. For example, it follows immediately that the parabola with Cartesian equation $y = ax^2$ has curvature $\kappa = 2a$ at the origin. Likewise, the Taylor expansion of $\cos x$ tells us [exercise] that the graph $y = \cos x$ has $\kappa = -1$ at the point $(0, 1)$, and the circle of curvature is therefore the unit circle, centred at the origin.



[8.2] *The curve's departure* σ *from its tangent initially increases quadratically with distance* ϵ *(like a parabola), and the proportionality constant is one-half of the curvature.*

8.3 Newton's Curvature Formula

Suppose that the curve $\mathcal{C}$ is the graph $y = f(x)$. If the x-axis is drawn parallel to the tangent line $\mathcal{T}$, then Taylor's Theorem implies [exercise] that $\sigma \asymp (1/2)f''(x_p)\epsilon^2$, where x_p denotes the x-coordinate of p. Thus (8.2) implies

$$\boxed{\kappa = f''(x_p).}$$

Note that this formula automatically complies with our convention regarding the sign of κ.

More generally, suppose that the x-axis is now drawn in an arbitrary direction, so that $\mathcal{T}$ is inclined at angle φ to it, so that $f' = \tan\varphi$. In this general case, Newton discovered that the correct generalization of the above formula is,

$$\boxed{\kappa = \frac{f''}{\left\{1 + [f']^2\right\}^{3/2}}.} \tag{8.3}$$

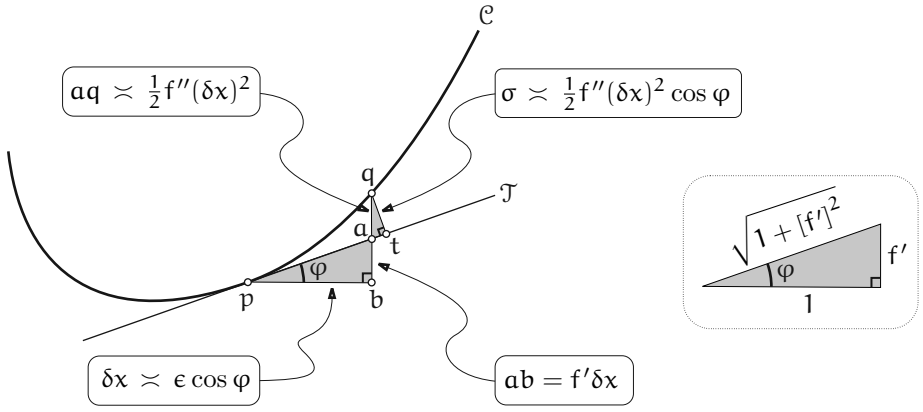
Here it is understood that the derivatives are evaluated at $x = x_p$.

Having now been exposed to several examples of the use of infinitesimal geometry, you should have little difficulty in following Newton's original proof of his formula, which is reproduced in Knoebel (2007, pp. 182–185). Here, however, we shall provide a different geometric argument that seeks to explain the complicated new denominator of (8.3) as a simple consequence of rotating the x-axis. See [8.3].

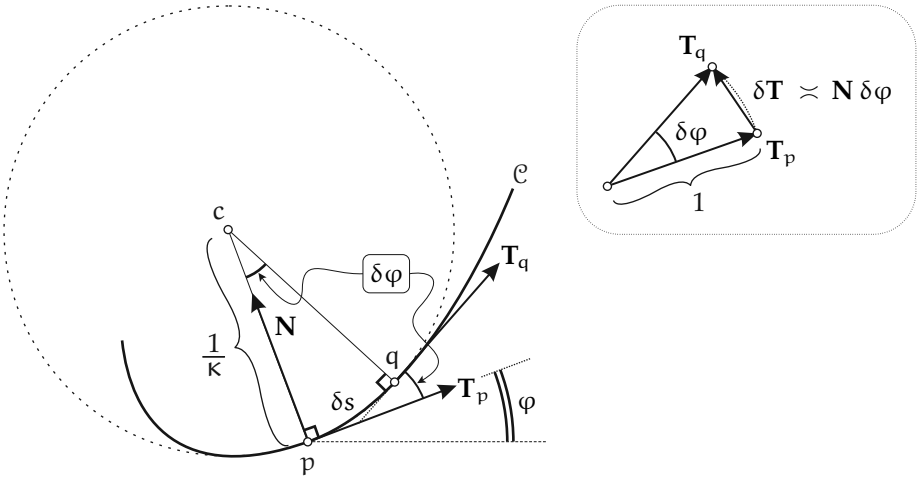
As before, here $\epsilon \equiv pt$, and $\sigma \equiv qt$, and we now suppose the x-axis to be drawn horizontally, instead of along the tangent $\mathcal{T}$. Taylor's Theorem now says that if $\delta x \equiv x_q - x_p$ then the height of q above $\mathcal{T}$ is given by $aq \asymp (1/2)f''(\delta x)^2$. Because the two shaded triangles are evidently similar, when aq is projected into the direction perpendicular to $\mathcal{T}$, we pick up one factor of $\cos\varphi$ and obtain $\sigma \asymp (1/2)f''(\delta x)^2\cos\varphi$.

Next, $pa \asymp pt = \epsilon$, and therefore $\delta x \asymp \epsilon\cos\varphi$. Thus in projecting from the tangent onto the horizontal x-axis we pick up two additional factors of $\cos\varphi$, yielding three in all:

$$\sigma \asymp \tfrac{1}{2}f''\epsilon^2\cos^3\varphi.$$



[8.3] *The simple formula $\kappa = f''$ (in the case $\varphi = 0$) must be multiplied by $\cos^3\varphi$ in the general case; as the box on the right shows, $\cos\varphi = 1/\sqrt{1+[f']^2}$.*



[8.4] *Curvature is the rate of turning of the tangent. If the unit tangent* $\mathbf{T}$ *turns through angle* $\delta\varphi$ *in moving distance* δs *along* $\mathcal{C}$*, then* $\kappa \asymp (\delta\varphi/\delta s)$*. Vectorially, in the box on the right, the rate of change of the unit tangent is* $(d\mathbf{T}/ds) \asymp (\delta\mathbf{T}/\delta s) \asymp (\mathbf{N}\,\delta\varphi/\delta s) \asymp \kappa\mathbf{N}$.

Finally, the boxed subfigure on the right of [8.3] demonstrates that

$$\cos\varphi = \frac{1}{\sqrt{1+[f']^2}},$$

and so Newton's formula (8.3) now follows immediately from (8.2).

8.4 Curvature as Rate of Turning

In 1761, about a century after Newton had introduced the concept of curvature, Kaestner published a simple alternative interpretation that would ultimately prove to be more amenable to generalization than Newton's.

> *Curvature is the rate of turning of the tangent with respect to arc length. In other words, if* φ *is the angle of the tangent, and* s *is arc length, then* $\kappa = (d\varphi/ds)$. (8.4)

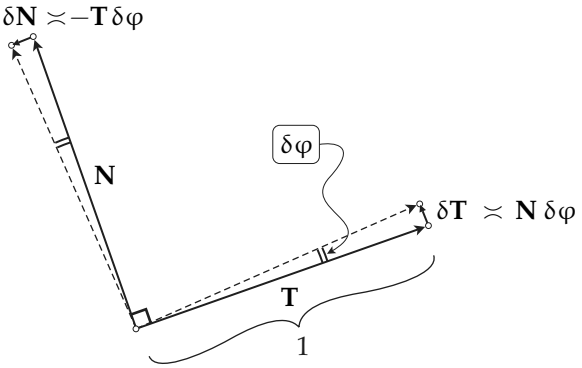
Note that this has the immediate advantage of being determined by local measurements conducted in the immediate vicinity of a tiny piece of curve: we no longer need to draw normals and follow them off into the distance until they intersect at the centre of curvature.

To confirm that (8.4) is equivalent to Newton's definition, it suffices to verify this for a perfect circle of radius ρ and curvature $\kappa = (1/\rho)$, for this circle may then be taken to be the circle of curvature in the general case. Since the rate of turning is clearly uniform in the case of a circle, consider one complete revolution: the tangent rotates 2π while traversing the full circumference of $2\pi\rho$, and therefore the rate of turning is $(2\pi)/(2\pi\rho) = \kappa$, as was to be shown.

Fig. [8.4] reproves this equivalence, and also takes us a step further. Here $\mathbf{T}$ denotes the unit tangent vector to $\mathcal{C}$, and $\mathbf{N}$ denotes the unit normal, pointing toward the centre of curvature, c. As we move a small distance δs from p to q, $\mathbf{T}$ rotates by $\delta\varphi$. Since δs is ultimately equal to the corresponding segment of the circle of curvature of radius $(1/\kappa)$ (according to Newton's original definition), $\delta s \asymp (1/\kappa)\delta\varphi$. Therefore $\kappa = (d\varphi/ds)$, reconfirming (8.4).

But now, in addition, the subfigure on the right of [8.4] allows us to determine the change $\delta\mathbf{T} \equiv (\mathbf{T}_q - \mathbf{T}_p)$ in the vector $\mathbf{T}$ itself as we pass from p to q. By drawing both unit tangents emanating from the same point, we may visualize $\delta\mathbf{T}$ as the illustrated vector connecting their tips, which will ultimately point along $\mathbf{N}$. But the length of this vector is ultimately equal to the arc of the unit circle (shown dotted) connecting these tips, so $\delta\mathbf{T} \asymp \mathbf{N}\,\delta\varphi$. Thus,

$$\frac{\delta\mathbf{T}}{\delta s} \asymp \mathbf{N}\frac{\delta\varphi}{\delta s} \quad \Longrightarrow \quad \frac{d\mathbf{T}}{ds} = \kappa\,\mathbf{N}. \tag{8.5}$$



[8.5] *As* $\mathbf{T}$ *and* $\mathbf{N}$ *rotate together, their tips begin to move along* $\mathbf{N}$ *and* $-\mathbf{T}$*, respectively, and by equal amounts.*

We can now return to our opening discussion of curvature in terms of a bead of unit mass travelling at unit speed along the curve $\mathcal{C}$. Because it travels at unit speed, the velocity vector $\mathbf{v}$ of the bead is simply $\mathbf{T}$, and also $\delta s = \delta t$. Therefore in (8.5) $(d\mathbf{T}/ds) = (d\mathbf{v}/dt)$ is in fact the *acceleration* of the bead, which in turn is the force exerted by the wire. We have thus confirmed the original claim (8.1).

Of course we may just as easily think of curvature as the rate of turning of the normal $\mathbf{N}$, instead of the tangent. Indeed [8.5] shows simply and vividly that just as the tip of $\mathbf{T}$ begins to rotate in the direction of $\mathbf{N}$, so the tip of $\mathbf{N}$ begins to rotate in the direction of $-\mathbf{T}$, and by an equal amount, so

$$\frac{d\mathbf{N}}{ds} = -\kappa\,\mathbf{T}.$$

Indeed, this use of the normal in place of the tangent will prove to be of crucial importance when we redirect our attention from curves back to surfaces, for in the latter case there is no such thing as "the" tangent, but there is still a unique *normal* vector, perpendicular to the tangent plane of the surface. How this normal vector varies in the immediate vicinity of a point on the surface will indeed tell us the Gaussian curvature at that point.

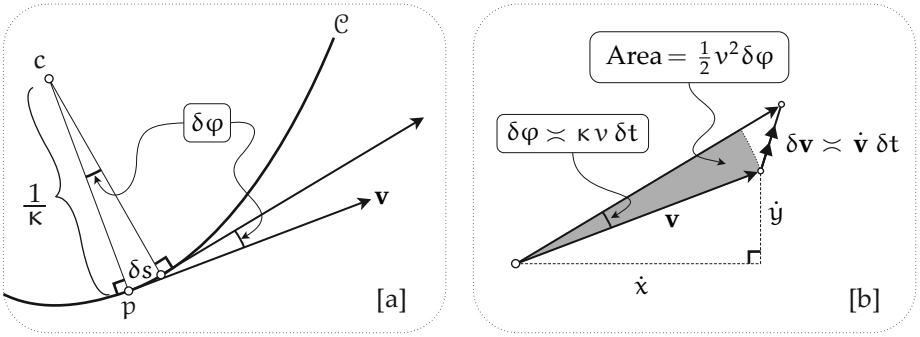
(As an aside, we note once again the distinction between geometric understanding and blind calculation, which is trivially simple in the case above. Let φ denote the angle that the tangent makes with the horizontal x-axis, so that

$$\mathbf{T} = \begin{bmatrix} \cos\varphi \\ \sin\varphi \end{bmatrix} \qquad \text{and} \qquad \mathbf{N} = \begin{bmatrix} -\sin\varphi \\ \cos\varphi \end{bmatrix}.$$

Let a prime denote differentiation with respect to arc length, so that $\kappa = \varphi'$. Then calculation immediately yields $\mathbf{T}' = \varphi'\,\mathbf{N}$ and $\mathbf{N}' = -\varphi'\,\mathbf{T}$.)

What if the bead is instead travelling at an arbitrary (but constant) speed v and has arbitrary mass m? In that case, by definition, $\delta s \asymp v\,\delta t$, and $\mathbf{v} = v\mathbf{T}$. Thus (8.5) becomes the following generalized version of the famous result for the force needed to hold an object in a circular orbit:

$$\mathbf{F} = m\frac{d\mathbf{v}}{dt} = \kappa m v^2\,\mathbf{N}.$$



[8.6] [a] *The velocity* $\mathbf{v}$ *at* p*, and a moment* δt *later, after travelling* $\delta s \asymp \nu\,\delta t$ *along* $\mathcal{C}$*.* **[b]** *Both velocity vectors are drawn emanating from the same point, so the vector connecting their tips is* $\delta\mathbf{v} \asymp \dot{\mathbf{v}}\,\delta t$*. Looking at the area of the triangle in two different ways yields formula (8.6) for* κ*.*

Newton first began to glimpse this as early as 1665,[3] as he struggled to understand the force that held the moon in its orbit.

This new interpretation of curvature (as rate of turning) allows us to deal with curves that cannot be expressed as the graph of a function, and which are therefore beyond the reach of Newton's original formula (8.3). Let $\mathcal{C}$ be the orbit of a particle whose position at time t is $(x[t], y[t])$, so that the velocity is

$$\mathbf{v} = \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix}, \quad \text{and therefore} \qquad \tan\varphi = \frac{\dot{y}}{\dot{x}},$$

where, following Newton, the dot[4] denotes differentiation with respect to time. Note that we are no longer assuming that the speed is constant: the length of $\mathbf{v}$ may vary.

By differentiating both sides of the previous equation, $\tan\varphi = (\dot{y}/\dot{x})$, and using the chain rule, it is not hard to obtain [exercise] a more general formula (also discovered by Newton):

$$\kappa = \frac{\dot{x}\,\ddot{y} - \dot{y}\,\ddot{x}}{\left[\dot{x}^2 + \dot{y}^2\right]^{3/2}}. \tag{8.6}$$

However, let us try to understand this formula in more direct, geometric terms. See [8.6a], which shows the velocity $\mathbf{v}$ at p and the velocity a moment δt later, after which the particle has travelled $\delta s \asymp \nu\,\delta t$ along $\mathcal{C}$. In [8.6b], both velocity vectors have been drawn emanating from the same point, so that the vector connecting their tips is the change in velocity, $\delta\mathbf{v} \asymp \dot{\mathbf{v}}\,\delta t$.

Thus the area of the shaded sector of the circle of radius $\nu = \sqrt{\dot{x}^2 + \dot{y}^2}$, subtending angle $\delta\varphi$, is given by

$$(\text{area of shaded sector}) = \tfrac{1}{2}\nu^2\delta\varphi \asymp \tfrac{1}{2}\nu^2\kappa\,\delta s \asymp \tfrac{1}{2}\nu^3\kappa\,\delta t.$$

But this area is ultimately equal to the area of the illustrated triangle with edge vectors

$$\mathbf{v} = \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} \quad \text{and} \qquad \delta\mathbf{v} \asymp \begin{bmatrix} \ddot{x} \\ \ddot{y} \end{bmatrix} \delta t,$$

[3] See Westfall (1980, pp. 148–150).

[4] Newton's *perfectly* minimalist notation is sometimes frowned upon, as the dot can be hard to see—in *this* work we have attempted to address that concern by simply *enlarging* the dot!

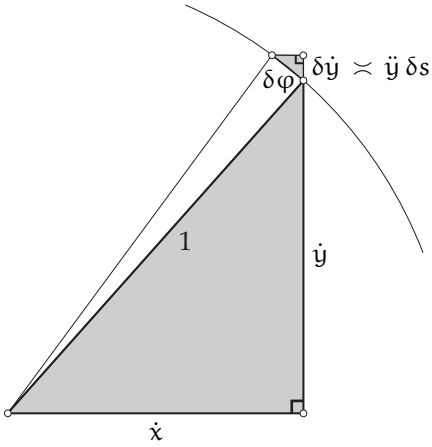
and by an elementary result (which itself may be proved geometrically [exercise]), this area is half the determinant of the matrix with these two columns:

$$\text{(area of triangle bounded by velocity vectors)} \asymp \tfrac{1}{2}(\dot{x}\ddot{y} - \dot{y}\ddot{x})\,\delta t.$$

Thus,

$$\tfrac{1}{2}v^3\kappa\,\delta t \asymp \tfrac{1}{2}(\dot{x}\,\ddot{y} - \dot{y}\,\ddot{x})\,\delta t,$$

thereby proving (8.6).



[8.7] *If the particle travels at unit speed then in a moment $\delta t = \delta s$ the tip of* $\mathbf{v}$ *traces an arc $\delta\varphi$ of the unit circle, and its height increases by $\delta\dot{y} \asymp \ddot{y}\,\delta s$. The ultimate similarity of the two shaded triangles yields (8.7).*

Note that if the particle's horizontal speed is fixed at $\dot{x} = 1$, so that its orbit is $(t, y[t])$, then we recover Newton's original formula (8.3) as a special case of (8.6). Indeed, this is precisely the manner in which Newton himself carried out the simplification; again, see Knoebel (2007, pp. 182–185).

Finally, consider another particularly important special case in which the particle travels at *unit speed*, so that

$$|\mathbf{v}| = \sqrt{\dot{x}^2 + \dot{y}^2} = 1,$$

and $\delta s = \delta t$. Thus the change from the general figure [8.6b] is that $\delta\mathbf{v}$ is now orthogonal to $\mathbf{v}$, tangent to the unit circle. As illustrated in [8.7], during the time δt the tip of $\mathbf{v}$ now traces out an arc $\delta\varphi$ of the unit circle, its height rising by $\delta\dot{y} \asymp \ddot{y}\,\delta t = \ddot{y}\,\delta s$.

From the ultimate similarity of the two shaded triangles, we deduce that

$$\frac{\delta\varphi}{\ddot{y}\,\delta s} \asymp \frac{1}{\dot{x}}.$$

Thus, recalling that $\kappa \asymp (\delta\varphi/\delta s)$, (8.6) reduces to an extremely simple formula which (rather strangely) is not readily found in standard texts: for a *unit-speed* orbit,

$$\boxed{\kappa = \ddot{y}/\dot{x}.} \tag{8.7}$$

Likewise, from the same triangles, $\kappa = -\ddot{x}/\dot{y}$. (See Ex. 1 for a less illuminating proof via calculation.)

8.5 Example: Newton's *Tractrix*

The constant negative Gaussian curvature of the pseudosphere can in fact be traced back to the curvature of Newton's tractrix, which generates it.

Given the parametric representation of the tractrix (see Ex. 16 on p. 88), its curvature can be found via a routine calculation [exercise] based on (8.6). However, a geometric analysis of the problem is more elegant *and* furnishes the answer in a form that will prove more useful in studying

the pseudosphere itself. Rather than employ any of the general formulas we have developed thus far, we shall instead present an argument[5] that is tailored to this particular curve.

Let

$$\begin{aligned} \rho_1 &= \textit{radius of curvature of the generating tractrix}, \\ \rho_2 &= \textit{the segment } pl \textit{ of the normal from the tractrix to its axis}, \end{aligned}$$

as illustrated in [8.8]. (Later we will explain that ρ_2 is also a radius of curvature, hence the use of the same Greek letter for both distances.)

By definition, the tractrix in this figure has tangents of constant length R. At the neighbouring points p and q, [8.8] illustrates two such tangents, pa and qb, containing angle $\bullet$. The corresponding normals po and qo therefore contain the same angle $\bullet$. Note that ac has been drawn perpendicularly to qb.

Now let's watch what happens as q coalesces with p, which itself remains fixed. In this limit, o is the centre of the circle of curvature, pq is an arc of this circle, and ac is an arc of a circle of radius R centred at p. Thus,

$$\rho_1 \asymp op \quad \text{and} \quad \frac{pq}{op} \asymp \bullet \asymp \frac{ac}{R},$$

and so

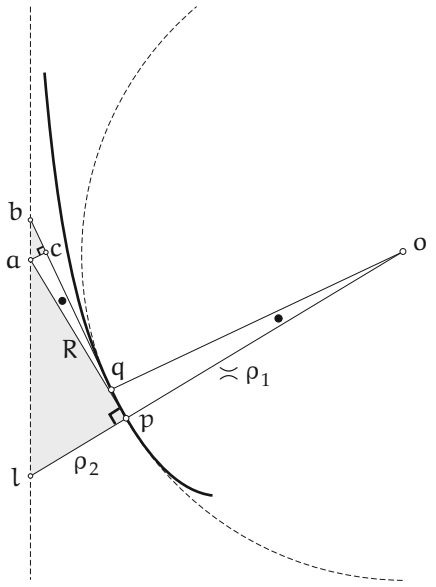
$$\frac{ac}{pq} \asymp \frac{R}{\rho_1}.$$

Next we appeal to the defining property $pa = R = qb$ of the tractrix to deduce that

$$bc \asymp pq.$$

Finally, using the fact that the triangle abc is ultimately similar to the triangle lap, we deduce that

$$\frac{R}{\rho_1} \asymp \frac{ac}{pq} \asymp \frac{ac}{bc} \asymp \frac{\rho_2}{R}.$$



[8.8] *The tractrix has* $\kappa = (\rho_2/R^2)$.

Thus,

$$\boxed{\kappa = \frac{1}{\rho_1} = \frac{\rho_2}{R^2}.} \tag{8.8}$$

[5] Previously published in VCA, p. 295.

Chapter 9

Curves in 3-Space

The foregoing analysis of plane curves can be applied, with only minor modification, to curves that escape from the plane and twist their way through 3-dimensional space.

The essential insight is that even for such a twisting 3-dimensional orbit, each infinitesimal segment nevertheless lies in a plane (to which our previous analysis therefore applies), but this instantaneous plane of motion (called the ***osculating plane***) no longer remains fixed in space but instead rotates about the particle's direction of motion as the particle travels along the curve. The rate of rotation of the osculating plane is called the ***torsion***, denoted τ.

[9.1] The Frenet Frame. *Pressing the cardboard against the bent wire curve reveals the osculating plane, with* **N** *pointing toward the centre of curvature within that that plane. As we move the frame along the curve, the* **torsion** τ *is the rate at which the plane's normal vector* **B** *(called the* **binormal***) rotates about* **T**.

There can be no substitute for direct physical experience of mathematical facts: do not merely try to imagine the following experiment—please *do* it!

Physically construct the contraption shown in [9.1], as follows. Cut a short section of a narrow drinking straw and tape or glue it down on an inflexible piece of stiff cardboard (or anything else flat and rigid). On the cardboard, draw a vector **T** extending the segment of straw, and draw an equally long vector **N** perpendicular to the straw, as shown. Finally, construct a third vector of equal length and attach it as the plane's normal vector, marked **B** in [9.1].

Take a sturdy piece of wire (perhaps a metal coat hanger, cut open) and bend it into any non-planar curve that suits your fancy. Now thread the wire through the section of straw, and press the cardboard against the curve so that it becomes the osculating plane. Once fitted to the curve in this way, the orthonormal set of vectors $(\mathbf{T}, \mathbf{N}, \mathbf{B})$ is called the ***Frenet frame*** of the curve.

Holding the wire fixed in space (ideally by enlisting the help of a friend), slide your Frenet frame along the curve, all the while keeping the cardboard in contact with the bent wire, as best you can, thereby ensuring that it continues to represent the osculating plane. In this manner you will experience τ as the rate at which you must rotate your piece of cardboard so as to keep it in contact with your curved wire.

Alternatively, if you are only interested in how the curvature κ varies along the curve, hold the cardboard fixed in one hand, with the thumb of that hand holding the wire down on the plane, so as to ensure it is the osculating plane. With your other hand, you may then gradually pull the wire through the straw, all the while pressing it against the cardboard with your thumb, and watch the changing curvature of the piece of curve on the fixed cardboard plane in your hand.

Returning to our previous discussion of plane curves for a moment, recall that the plane of the curve is spanned by **T** and by **N**, and further recall that the direction of **N** can be deduced from the rate of change of **T**, via (8.5): $\mathbf{T}' = \kappa\mathbf{N}$. (Recall that here the derivative is with respect to arc length, which is the same as the time derivative only if the particle travels at unit speed.)

In the present case of a twisting 3-dimensional curve, we can turn this around and *define* **N** to be the direction in which the tangent is turning:

$$\mathbf{N} \equiv \frac{\mathbf{T}'}{|\mathbf{T}'|}.$$

This normal **N** is called the ***principal normal***, to distinguish it from the infinitely many other "normals" that lie in the plane perpendicular to **T**; it is distinguished by the fact that it lies in the osculating plane, and that it points directly *at* the centre of curvature (instead of away from it). Thus, a little more explicitly than before, the osculating plane, within which the particle is momentarily moving, is the plane spanned by **T** and $\mathbf{T}'$.

With these conventions in place, the acceleration of a unit-speed particle is always directed toward the centre of curvature and its magnitude is the curvature. It therefore makes sense to rechristen this acceleration as the ***curvature vector*** $\boldsymbol{\kappa}$ of this unit-speed particle:

$$\boldsymbol{\kappa} \equiv \kappa \mathbf{N}. \tag{9.1}$$

As shown in [9.1], the orientation in space of the osculating plane is conveniently encoded in its unit normal vector, denoted **B**, and called the ***binormal*** of the curve:

$$\mathbf{B} \equiv \frac{\mathbf{T}\times\mathbf{T}'}{|\mathbf{T}'|} = \mathbf{T}\times\mathbf{N}.$$

As we have said, the torsion is the rate of rotation of the osculating plane about the direction of motion, **T**. Equivalently, it is the rate of turning of **B** about **T**.

We note the following simple but fundamental fact:

> *When a unit vector begins to rotate, its tip moves on the unit sphere, within the tangent plane to the unit sphere at that point, and therefore in a direction* perpendicular *to the vector itself.* (9.2)

As **B** rotates, its tip must therefore begin to move in a plane parallel to the osculating, cardboard $(\mathbf{T}, \mathbf{N})$-plane. Note, however, that here **B** cannot tip in the **T** direction, and its rate of change is therefore purely in the **N** direction; make sure you can see this geometrically, perhaps with the assistance of your toy Frenet frame. This can of course also be shown by calculation: see Exercise 2. Thus

$$\mathbf{B}' = -\tau \mathbf{N}.$$

NOTATIONAL NOTE: The minus sign that is included in this definition of τ does not appear to us to have anything to recommend it, but since the great majority of authors appear to include it, we shall let discretion be the better part of valour!

As for the curvature, the foregoing analysis applies with only minor change. As before, **T** spins within the osculating plane at a rate given by κ, and **N** spins with it, as in [8.5]. But now **N** does not merely rotate within the osculating plane, for it must also remain orthogonal to **B**, so it

rotates out of the osculating plane at the same rate as **B** rotates about **T**. In other words, applying the geometric idea [8.5] twice, and in accord with (9.2), the total rate of change of **N** is given by

$$\mathbf{N}' = -\kappa\,\mathbf{T} + \tau\,\mathbf{B}.$$

The rotation of the entire Frenet frame as we move along the curve can therefore be summarized as the following matrix equation, known as the ***Frenet–Serret Equations***:[1]

$$\begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}' = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix} = [\Omega] \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}. \tag{9.3}$$

Let us retain a firm grasp on the geometry underlying the structure of $[\Omega]$: the vanishing leading diagonal is merely the algebraic manifestation of (9.2); likewise, the skew-symmetry ($[\Omega]^{\mathsf{T}} = -[\Omega]$) is merely the algebraic manifestation of [8.5].

The matrix $[\Omega]$ itself tells us how the $(\mathbf{T}, \mathbf{N}, \mathbf{B})$ frame rotates from one moment to the next. If we watch the frame move along the curve for a short time δt, then

$$\text{new frame after } \delta t \asymp [I + [\Omega]\,\delta t]\,[\text{original frame}]\,.$$

For more on the rotation of the frame, see Exercise 3.

[1] Independently discovered by Frenet in 1847 and by Serret in 1851.

Chapter 10

The Principal Curvatures of a Surface

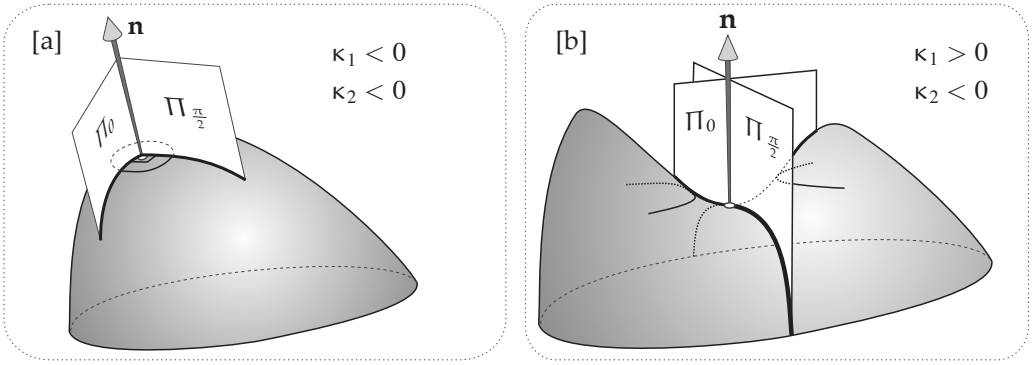
10.1 Euler's Curvature Formula

In Act I we immediately and anachronistically launched into a discussion of Gauss's revolutionary 1827 conception of the intrinsic geometry of surfaces, and with it his associated concept of the *intrinsic* curvature $\mathcal{K}$. But, historically, the natural progression from Newton's investigation of the extrinsic curvature of plane curves was the study of the *extrinsic* curvature of surfaces: how do they bend within the surrounding space?

It was Euler in 1760 who made the first fundamental breakthrough. We will first describe his discovery, and then prove it afterwards. At a point p on the surface $\mathcal{S}$ he considered the plane curve $\mathcal{C}_\theta$ through p obtained by intersecting $\mathcal{S}$ with a plane Π_θ that rotates about the surface normal $\mathbf{n}_p$ at p. See [10.1], which illustrates two orthogonal positions of this plane Π_θ, on two different kinds of surface. Here θ denotes the angle of rotation of Π_θ, starting from an arbitrary (at least for now) initial direction. As Π_θ rotates, the shape of the intersection curve $\mathcal{C}_\theta$ changes, and therefore its curvature $\kappa(\theta)$ at p will (in general) vary too.

Before continuing, we should explain that $\kappa(\theta)$ has a *sign* attached to it, according to this convention: the vector from p to the centre of curvature c of $\mathcal{C}_\theta$ is defined to be $\frac{1}{\kappa(\theta)}\mathbf{n}$. Thus if c lies in the direction of $+\mathbf{n}$ then $\kappa(\theta)$ is positive, while if it lies in the direction of $-\mathbf{n}$ then $\kappa(\theta)$ is negative. Of course there are actually two opposite (equally valid) choices for $\mathbf{n}$. Reversing the choice of $\mathbf{n}$ reverses the sign of κ. Note that the principal normal (as defined in the previous section) of $\mathcal{C}_\theta$ is therefore $\mathbf{N} = [\text{sign of } \kappa(\theta)]\,\mathbf{n}$.

As θ varies, let κ_1 and κ_2 denote the maximum and minimum values of $\kappa(\theta)$. Euler's elegant and important discovery was that these extreme values of the curvature (the so-called ***principal curvatures***) will always occur in *perpendicular* directions, which are called the ***principal directions***. Furthermore, if we choose $\theta = 0$ to coincide with the direction that has curvature κ_1, he found



[10.1] [a] *If κ_1 and κ_2 have the same sign then the surface locally resembles a hill, and a slice parallel to (and close to) the tangent plane yields an ellipse (or misses the surface entirely).* **[b]** *If κ_1 and κ_2 have opposite sign then the surface locally resembles a saddle. A parallel slice above the tangent plane yields both branches of a hyperbola (as illustrated); slicing below the tangent plane yields both branches of an orthogonal hyperbola (not shown).*

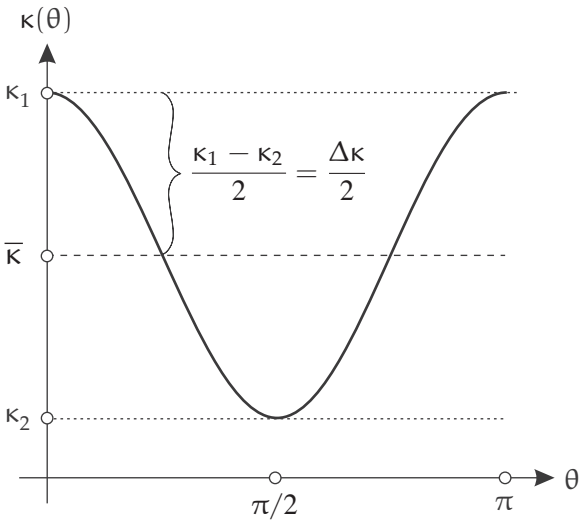
$$\textbf{\textit{Euler's Curvature Formula}}: \quad \kappa(\theta) = \kappa_1 \cos^2\theta + \kappa_2 \sin^2\theta. \tag{10.1}$$

Although this is the standard form of the formula found in most modern texts, there is a superior way of writing it, which was in fact the form in which Euler[1] himself first published it.

Substituting $\cos^2\theta = (1+\cos 2\theta)/2$ and $\sin^2\theta = (1-\cos 2\theta)/2$, we obtain

$$\kappa(\theta) = \overline{\kappa} + \tfrac{\Delta\kappa}{2}\cos 2\theta, \tag{10.2}$$

where $\overline{\kappa} \equiv \left[\frac{\kappa_1+\kappa_2}{2}\right]$ is the ***mean curvature***, about which the value oscillates, and $(\Delta\kappa/2) \equiv (\kappa_1 - \kappa_2)/2$ is the amplitude of the oscillation. See the graph [10.2]. Note that the extremal nature of κ_1 and κ_2, together with the orthogonality of the principal directions, can be deduced directly from this formula. Furthermore, this form makes manifest the geometric fact that rotating the normal plane by π returns it to the same position, so the variation in curvature has period π.



[10.2] *Euler's Curvature Formula tells us that as we vary the angle θ of the normal section $\mathcal{C}_\theta$, its curvature $\kappa(\theta)$ oscillates sinusoidally, achieving its maximum and minimum values in perpendicular directions.*

As illustrated in [10.1a], if κ_1 and κ_2 have the same sign then $\kappa(\theta)$ always shares this same sign; i.e., the surface locally resembles a hill. But if κ_1 and κ_2 have opposite signs then $\kappa(\theta)$ changes sign; i.e., $\mathcal{C}_\theta$ flips from one side of the tangent plane to the other, and, as shown in [10.1b], the surface locally resembles a saddle.

Of course it is possible that $\kappa_1 = \kappa_2$, in which case $\kappa(\theta) = \text{const.}$ at p; in this case p is called ***umbilic***, and the surface locally resembles a sphere. In general it can be shown that such umbilics can only occur in isolated places on a surface. Of course the sphere of radius R is a strong exception to this: *every* point is an umbilic, with $\kappa(\theta) = 1/R$. But it can be proved that the sphere is the *only* surface with this property.

In the exceptional case that the umbilic has $\kappa_1 = 0 = \kappa_2$, the shape of the surface surrounding the point can be much more complex. This will be discussed and illustrated in Section 12.4.

10.2 Proof of Euler's Curvature Formula

We will now provide a geometric proof[2] of Euler's Curvature Formula. Choose p to be the origin of the Cartesian (x, y, z) coordinates, and let the x and y axes be chosen to lie in the tangent planne

[1] The more familiar form (10.1) was derived from Euler's result about 50 years later, by Charles Dupin, in 1813. See Knoebel (2007, p. 188).

[2] The proof in the lovely article by Aleksandrov (1969) is similar, but it requires two calculations, which are here replaced with geometry. The now-standard idea of using the quadratic approximation relative to the tangent plane is simpler than Euler's original approach; it was discovered by Jean-Baptiste Meusnier in 1776. See Knoebel (2007, p. 194).

T_p at p. Then locally the surface can be represented by an equation of the form $z = f(x, y)$, such that $f(0,0) = 0$ and $\partial_x f = 0 = \partial_y f$ at the origin. Expanding $f(x, y)$ into a Taylor series, we deduce that as x and y tend to zero,

$$z \asymp ax^2 + by^2 + cxy. \tag{10.3}$$

Slicing through the surface with planes $z = \text{const.} = k$ parallel to T_p, and very close to it, therefore yields intersection curves whose equations (as k goes to zero) are quadratics, $ax^2 + by^2 + cxy \asymp k$, and which are therefore (ultimately) conic sections.

Figure [10.1] illustrates the fact that these conics are ellipses if κ_1 and κ_2 have the same sign, and that they are hyperbolas if κ_1 and κ_2 have opposite signs. In both cases, *the conics have two perpendicular axes of symmetry that are independent of the height* k *of the slicing plane*. This follows from the homogeneous quadratic nature of the equation. For example, quadrupling the height of the slice just doubles the size of the conic, without changing its shape: $k \to 4k$ and $(x, y) \to (2x, 2y)$ yield the same equation as before.

The use of this conic to quantify the amount and type of bending of S at p goes back to Charles Dupin in 1813, and it is called the ***Dupin indicatrix*** in his honour. The point p itself is called ***elliptic***, ***hyperbolic***, or ***parabolic*** according to the type of the Dupin indicatrix at p.

Crucially, the symmetry of the conic sections implies that the surface itself has local mirror symmetry in two perpendicular directions.[3] We can now derive Euler's Curvature Formula and deduce that these two perpendicular planes of symmetry are in fact the *same* planes that yield the maximum and minimum curvatures; i.e., these local mirror symmetry directions are the same as the principal directions. To summarize what we shall prove,

> *An infinitesimal neighbourhood of a generic point of a surface has mirror symmetry in two perpendicular planes (both containing the surface normal), and the perpendicular directions in which these plane intersect the tangent plane are the principal directions of maximum and minimum curvature.*

(10.4)

Refining our coordinate system, we now align the x and y axes with these symmetry directions. Since the equation (10.3) is now invariant under the reflections $x \mapsto -x$ and $y \mapsto -y$ it follows that $c = 0$, and the local equation of the surface therefore becomes

$$z \asymp ax^2 + by^2. \tag{10.5}$$

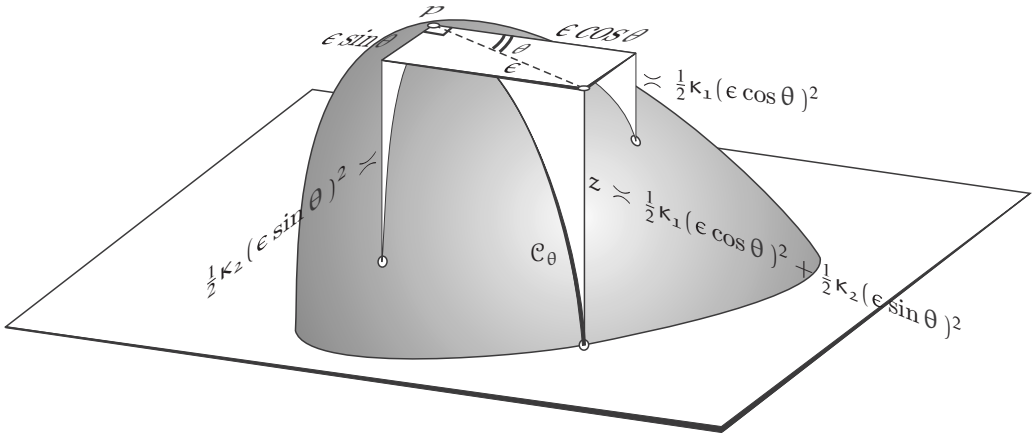
To find the geometric meaning of the coefficients a and b we now refer back to [8.2] and view it as depicting the intersection of Π_θ with S: the curve $\mathcal{C}$ is now $\mathcal{C}_\theta$, and the tangent $\mathcal{T}$ is now the intersection of the tangent plane T_p with Π_θ, and the deviation σ of the curve from its tangent is now simply the height z of the curve above the tangent plane.

Let $\theta = 0$ correspond to the x-axis, and let $\kappa_1 = \kappa(0)$ be the curvature of $\mathcal{C}_0 =$ (the intersection of S with the xz-plane), having equation $z = ax^2$. Then the result (8.2) shows that $a = \frac{1}{2}\kappa_1$. In exactly the same way, defining $\kappa_2 = \kappa(\frac{\pi}{2})$ to be the curvature of the intersection curve with the yz-plane, we find that $b = \frac{1}{2}\kappa_2$. Thus (10.5) can be expressed more geometrically as

$$z \asymp \tfrac{1}{2}\kappa_1 x^2 + \tfrac{1}{2}\kappa_2 y^2. \tag{10.6}$$

Now consider [10.3], which depicts the curve $\mathcal{C}_\theta$ for a general angle θ. (This diagram assumes (and others to follow do as well) that the Gaussian curvature is positive, but the accompanying

[3] At least in the general case, where $\kappa_1 \neq \kappa_2$; the symmetry can be much more complex in the case of an umbilic, as we shall see in Section 12.4.



[10.3] *The principal curvatures tell us how quickly the surface falls away—downwards in this figure—from the tangent plane as we begin to travel in each of the two perpendicular principal directions. When we instead move a distance ϵ within the tangent plane in a general direction θ, the distance z that the surface falls away is simply the sum (according to (10.6)) of the falls due to each of these two components separately.*

reasoning applies equally well to negatively curved surfaces.) If we move a distance ϵ within T_p in the direction θ then we arrive at the illustrated point $x = \epsilon\cos\theta$, $y = \epsilon\sin\theta$. Thus inserting (10.6) into (8.2) yields

$$\kappa(\theta) \asymp 2\left[\frac{z}{\epsilon^2}\right] \asymp 2\left[\frac{\frac{1}{2}\kappa_1(\epsilon\cos\theta)^2 + \frac{1}{2}\kappa_2(\epsilon\sin\theta)^2}{\epsilon^2}\right] = \kappa_1\cos^2\theta + \kappa_2\sin^2\theta,$$

proving Euler's Curvature Formula, and thereby establishing the extremal nature of the curvatures κ_1 and κ_2 associated with the orthogonal directions of local mirror symmetry.

10.3 Surfaces of Revolution

If we rotate a plane curve $\mathcal{C}$ about a line L within its plane, then we obtain a surface of revolution $\mathcal{S}$ for which the principal directions are easily identified. Figure [10.4] illustrates this in the particular case in which $\mathcal{C}$ is the tractrix, L is its axis, and therefore $\mathcal{S}$ is the pseudosphere.

Clearly $\mathcal{S}$ has mirror symmetry in each plane through its axis L, so the intersection of such a plane with $\mathcal{S}$ (which is simply a copy of $\mathcal{C}$), yields one of the principal directions within $\mathcal{S}$.

At a point p on $\mathcal{S}$ let us choose this direction along the copy of $\mathcal{C}$ to correspond to $\theta = 0$, so that using the previous notation $\mathcal{C} = \mathcal{C}_0$. Thus the first principal curvature $\kappa_1 = (1/\rho_1)$ is simply (up to sign) the curvature of the original plane curve $\mathcal{C}$.

The second principal direction at p must therefore be the direction within $\mathcal{S}$ perpendicular to this plane through L. The radius of curvature ρ_2 associated with this second principal direction is in fact simply the distance pl, in the direction of the normal $\mathbf{n}$ to the surface, from p to the point l on the axis L:

$$\rho_1 = \textit{radius of curvature } op \textit{ of the generating curve } \mathcal{C}, \tag{10.7}$$

$$\rho_2 = \textit{the distance } pl \textit{ along the normal } \mathbf{n} \textit{ to the axis } L. \tag{10.8}$$

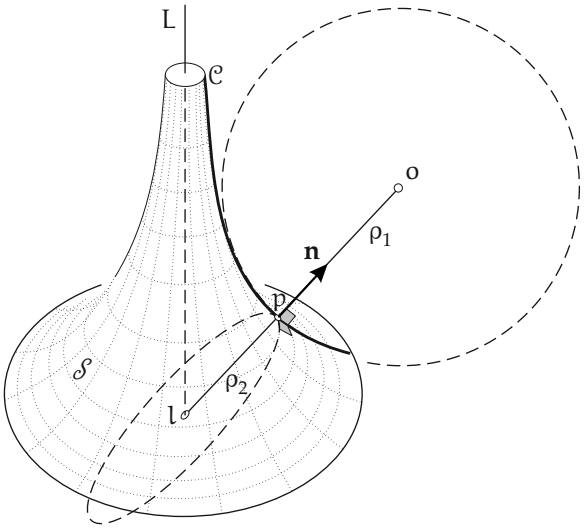
To confirm this, let us use the same notation as before and denote the intersection of $\mathcal{S}$ with this perpendicular plane as $\mathcal{C}_{(\pi/2)}$. If we rotate lp about L then it sweeps out a cone and p moves within $\mathcal{S}$ along the circular edge of this cone, which initially coincides with $\mathcal{C}_{(\pi/2)}$.

Now recall Newton's original method of locating the centre of curvature as the intersection of neighbouring normals of a curve. In the present case of $\mathcal{C}_{(\pi/2)}$, the neighbouring normals are generators of this cone, meeting at l, thereby confirming that $\kappa_2 = \pm(1/\rho_2)$, the sign depending on the choice of $\mathbf{n}$.

Returning to [10.4], with the illustrated choice of the normal vector $\mathbf{n}$, we find for the

$$\textit{pseudosphere}: \quad \kappa_1 = +\frac{1}{\rho_1} \quad \text{and} \quad \kappa_2 = -\frac{1}{\rho_2} = -\frac{\rho_1}{R^2}, \tag{10.9}$$

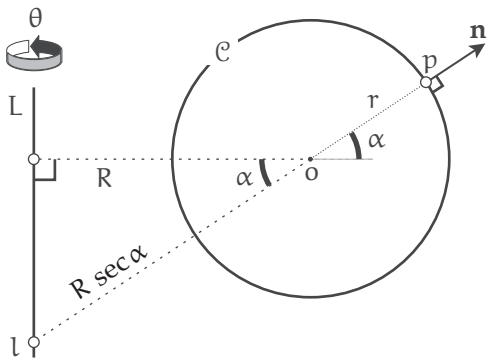
by virtue of (8.8). (Of course, quite generally, reversing the direction of $\mathbf{n}$ reverses the sign of both principal curvatures.) Note that we have now kept our promise and have explained the notation ρ_2 that we saw in [8.8].



[10.4] *The principal radii of curvature ρ_1 and ρ_2 of a surface of revolution, here illustrated with the pseudosphere.*

For our second example, consider the torus (doughnut) obtained by rotating a circle $\mathcal{C}$ of radius r and centre o about a line L at distance R from o. See [10.5]. We imagine that $\mathcal{C}$ is traced by a particle p that rotates about o, the radius op making angle α with the horizontal. From the figure, with the illustrated choice of $\mathbf{n}$, we see that for this

$$\textit{torus}: \quad \kappa_1 = -\frac{1}{r} \quad \text{and} \quad \kappa_2 = -\frac{1}{r + R\sec\alpha}. \tag{10.10}$$



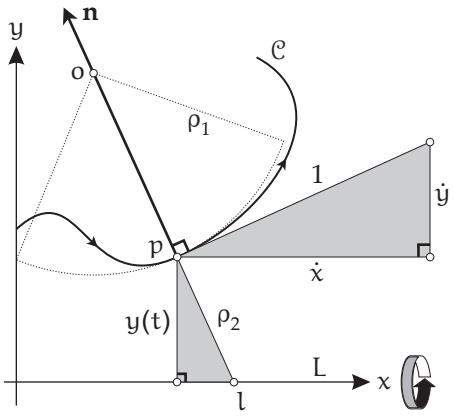
[10.5] *For the torus obtained by rotating the circle $\mathcal{C}$ about L, the principal radii of curvature are $\rho_1 = -r$ and $\rho_2 = -(r + R\sec\alpha)$.*

Leaving behind specific examples, suppose now that $\mathcal{C}$ is a general curve, but *traced at unit speed* by a particle moving in the (x, y)-plane, whose position at time time is $x = x(t)$ and $y = y(t)$. See [10.6].

Let us take L to be the horizontal x-axis, and rotate $\mathcal{C}$ about this axis to generate a surface $\mathcal{S}$. Let us choose $\mathbf{n}$ pointing to the *left* of the direction of motion, as illustrated. Using (10.7) and (8.7), we deduce the first principal curvature for a curve traced at

$$\boxed{\textit{unit speed}: \quad \kappa_1 = \dddot{y}/\dot{x}.} \tag{10.11}$$

As we previously noted regarding (8.7), the result (10.11) may equally well be expressed as $\kappa_1 = -\dddot{x}/\dot{y}$. Check for yourself [exercise] that these formulas yield the correct sign for κ_1, no matter where we are on the curve.



[10.6] *The principal radii of curvature* ρ_1 *and* ρ_2 *of a surface of revolution generated by rotating* $\mathcal{C}$ *about* L *(the* x*-axis).*

Finally, according to (10.8), $\rho_2 = pl$. By appealing to the similarity of the two shaded triangles in [10.6], we find that $(y/\rho_2) = (\dot{\dot{x}}/1)$, and so we deduce the second principal curvature for a curve traced at

$$\textit{unit speed}: \qquad \kappa_2 = -\dot{\dot{x}}/y. \tag{10.12}$$

Check for yourself [exercise] that this formula yields the correct sign for κ_2, no matter where we are on the curve.

Finally, we note an important lesson of the above analysis:

Let the curve $\mathcal{C}$ *be rotated about the line* L *to generate the surface of revolution* $\mathcal{S}$*. Then the parts of* $\mathcal{C}$ *that are concave* towards L *generate parts of* $\mathcal{S}$ *that have* positive *curvature, and the parts of* $\mathcal{C}$ *that are concave* away *from* L *generate parts of* $\mathcal{S}$ *that have* negative *curvature. Inflection points of* $\mathcal{C}$ *generate circles on* $\mathcal{S}$ *where the curvature vanishes; these circles separate the regions of opposite curvature.* (10.13)

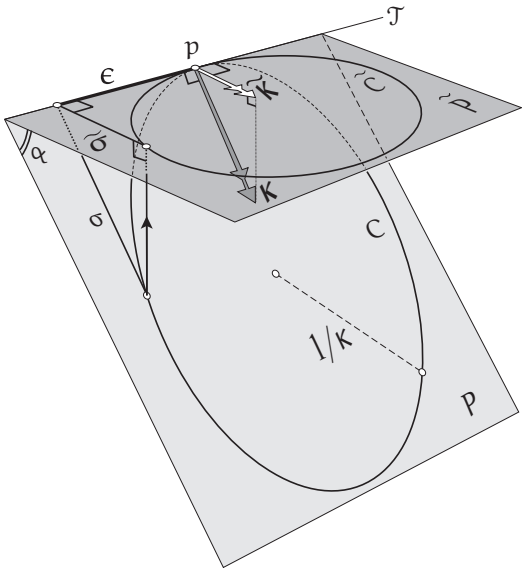
Chapter 11

Geodesics and Geodesic Curvature

11.1 Geodesic Curvature and Normal Curvature

To us, as inhabitants of the Earth's surface, a great circle is not only analogous to a line in that it provides the shortest route between two points, but it also appears to be *straight*: it has no apparent curvature. If you walk in a "straight line" across a seemingly flat desert, you are actually walking along such a great circle, whose curvature in 3-space is $1/(\text{radius of the Earth})$. How shall we reconcile these two conflicting views of one and the same curve?

The answer, in short, is that the full curvature in 3-space of a *general* curve within a *general* surface $\mathcal{S}$ can be decomposed into two components: one within the surface (visible to its inhabitants), and another perpendicular to the surface (invisible to the inhabitants). The visible component within the surface is called the ***geodesic curvature***, denoted κ_g; the invisible component perpendicular to the surface is called the ***normal curvature***, denoted κ_n.



[11.1] *When the circle* C *in the plane* P *is projected orthogonally onto the plane* $\widetilde{\mathrm{P}}$*, at angle* α *to* P*, the distance from the common tangent* $\mathcal{T}$ *contracts by* $\cos\alpha$*, and therefore the size of the curvature vector at* p *does too:* $\widetilde{\kappa} = \kappa\cos\alpha$.

If the osculating plane is perpendicular to the surface, i.e., contains the surface normal $\mathbf{n}$, then *all* of the curvature is "normal curvature" ($\kappa_n = \kappa$) and *none* of it is visible "geodesic curvature" within the surface ($\kappa_g = 0$). This was the case for our great circle on the surface of the Earth.

Now suppose instead that you are standing in middle of a seemingly flat desert, and you trace a very small circle of radius r in the sand at your feet. You thus obtain a seemingly planar curve with very large curvature $\kappa = 1/r$. But of course the desert is part of the curved surface of the Earth, and from this perspective all that is special about your curve is that its osculating plane almost coincides with Earth's tangent plane at your location. In this case we have almost the reverse of the former case: almost all of the curvature is now geodesic curvature, but in fact there is *still the same amount* of invisible normal curvature as before, as we shall explain shortly.

In fact the general case may be thought of as simply an appropriate mixture of the two extreme cases described just previously.

The essential point really has little to do with surfaces, *per se*, rather it has to do with how the curvature of a plane curve changes when it is projected (casts a shadow) on another plane. See [11.1], which shows a circle C of radius $(1/\kappa)$ within a plane P, the tangent at p being $\mathcal{T}$.

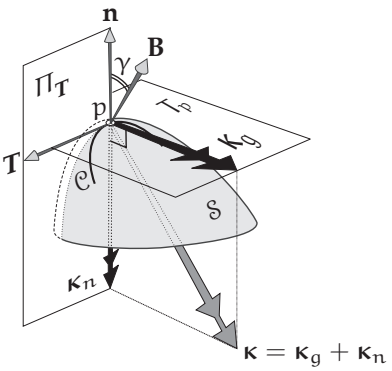
The figure also shows the orthogonal projection $\widetilde{C}$ of C onto a second plane $\widetilde{P}$ through $\mathcal{T}$, the angle between the planes being α. As you probably know [or exercise] $\widetilde{C}$ is in fact an ellipse, the original circle C having been compressed in a direction perpendicular to $\mathcal{T}$. Thus it is clear that the curvature $\widetilde{\kappa}$ of $\widetilde{C}$ at p is *less* than the original curvature κ.

More precisely, the figure shows that under this projection the distance from the common tangent $\mathcal{T}$ undergoes a compression by $\cos\alpha$, so that $\widetilde{\sigma} = \sigma\cos\alpha$. Thus, by virtue of (8.2),

$$\widetilde{\kappa} \asymp \frac{2\widetilde{\sigma}}{\epsilon^2} = \frac{2\sigma\cos\alpha}{\epsilon^2} \asymp \kappa\cos\alpha.$$

By thinking of C as the circle of curvature of a general curve $\mathcal{C}$, we see that this formula applies to $\mathcal{C}$ as well:

If a plane curve $\mathcal{C}$ (with tangent $\mathcal{T}$ at p) is projected orthogonally onto a second plane that passes through $\mathcal{T}$ and makes angle α with the first plane, then at p the projected curve $\widetilde{\mathcal{C}}$ has $\widetilde{\kappa} = \kappa\cos\alpha$. (11.1)



[11.2] *The net acceleration $\boldsymbol{\kappa}$ can be decomposed into the geodesic curvature vector $\boldsymbol{\kappa}_g$ tangent to the surface, and the normal curvature vector $\boldsymbol{\kappa}_n$ perpendicular to the surface: $\boldsymbol{\kappa} = \boldsymbol{\kappa}_g + \boldsymbol{\kappa}_n$.*

This result can also be understood vectorially, and perhaps even more intuitively, in terms of motion. As we have discussed, if a particle traverses $\mathcal{C}$ at unit speed then it is appropriate to rename the particle's acceleration as the ***curvature vector*** $\boldsymbol{\kappa} = \kappa\mathbf{N}$, for this acceleration vector points towards the centre of curvature and its magnitude equals the curvature.

Next, observe [exercise] that *the acceleration of the projection equals the projection of the acceleration*. Note, however, that the projected orbit $\widetilde{\mathcal{C}}$ will in general *not* be traced at constant speed, and its acceleration vector will therefore have a component tangent to $\widetilde{\mathcal{C}}$ and will therefore *not* be orthogonal to $\widetilde{\mathcal{C}}$, in contrast to $\widetilde{\boldsymbol{\kappa}}$. (Remember, $\widetilde{\boldsymbol{\kappa}}$ is the acceleration the particle *would* have had if it traversed $\widetilde{\mathcal{C}}$ at *constant unit speed*.) This breakdown of orthogonality also makes sense geometrically: in [11.1] a radius of C projects (in general) to a line in $\widetilde{P}$ that is no longer orthogonal the image ellipse $\widetilde{C}$. That said, p is an exception to this, and as the particle and its shadow momentarily move together at unit speed through p along $\mathcal{T}$,

$$\widetilde{\boldsymbol{\kappa}} = (\text{projection of } \boldsymbol{\kappa} \text{ onto } \widetilde{P}),$$

as illustrated in [11.1]. By equating the lengths of these vectors, (11.1) follows immediately.

Now let us return to the original problem, in which $\mathcal{C}$ is a general curve on a surface $\mathcal{S}$. See [11.2]. As before, let $\mathbf{T}$ denote the unit tangent to $\mathcal{C}$ at p, let T_p be the tangent planethe tangent plane to $\mathcal{S}$ at p, and let us also introduce $\Pi_{\mathbf{T}}$ as the normal plane spanned by $\mathbf{n}$ and $\mathbf{T}$. Let the osculating plane be inclined at angle γ to the tangent plane T_p; equivalently, γ is the angle between the binormal $\mathbf{B}$ of $\mathcal{C}$ and surface normal $\mathbf{n}$ of $\mathcal{S}$. Thus κ_g and κ_n are the curvatures of the projections of $\mathcal{C}$ onto T_p and $\Pi_{\mathbf{T}}$, respectively. Then (11.1) implies that

$$\kappa_g = \kappa\cos\gamma \quad \text{and} \quad \kappa_n = \kappa\sin\gamma. \tag{11.2}$$

Again, this can also be understood in terms of acceleration. Imagine particles traversing the projections of $\mathcal{C}$ onto T_p and $\Pi_{\mathbf{T}}$ at unit speed. Their respective accelerations will then be the

geodesic curvature vector $\boldsymbol{\kappa}_g$ and the ***normal curvature vector*** $\boldsymbol{\kappa}_n$, pointing toward the respective centres of curvature in the tangent and normal planes, and with magnitudes equal to κ_g and κ_n, respectively. As illustrated in [11.2], the acceleration can then be decomposed into these two orthogonal components:

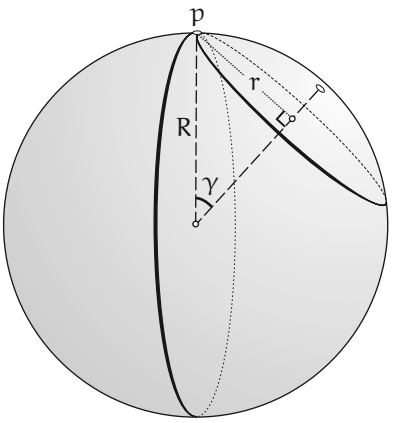
$$\boldsymbol{\kappa} = \boldsymbol{\kappa}_g + \boldsymbol{\kappa}_n. \tag{11.3}$$

From this (11.2) follows immediately.

11.2 Meusnier's Theorem

Consider the family of all curves (such as $\mathcal{C}$) on $\mathcal{S}$ that pass through p in the particular direction $\mathbf{T}$. The inhabitants of $\mathcal{S}$ may freely draw curves of this type that are tightly curved, slightly curved, or not curved at all: κ_g can be freely and arbitrarily specified. However, the same is *not* true of the normal curvature κ_n: the bending of the surface itself *forces* all curves within it to bend in the $\mathbf{n}$ direction.

In fact, as Meusnier realized in 1779, the surface forces all such curves to bend by the *same* amount: κ_n is independent of the specific curve $\mathcal{C}$, so, in particular, κ_n must equal the curvature of the normal section in this direction. More intuitively, this says that all of these curves locally have the *same projection* onto $\Pi_{\mathbf{T}}$: near p, these projections all look like the normal section, which in turn looks like a circle within $\Pi_{\mathbf{T}}$ of radius $(1/\kappa_n)$ centred at $(1/\kappa_n)\mathbf{n}$.



[11.3] *Slicing a sphere of radius* R *through the north pole* p *with a plane at angle* γ *to the horizontal yields a circle of radius* $r = R\sin\gamma$.

Even granted this, however, that this curve-independent normal curvature must in general depend on the *direction* $\mathbf{T}$ of the family of curves as they pass through p. Thus we may write the common normal curvature of the family as a function $\kappa_n(\mathbf{T})$ of this direction. (For example, if $\mathbf{e}_1$ and $\mathbf{e}_2$ are the principal directions, then $\kappa_n(\mathbf{e}_{1,2}) = \kappa_{1,2}$ are the principal curvatures.) Combining this (as yet) unproven claim with (11.2), we can state

> **Meusnier's Theorem.** *All curves that pass through a point* p *of a curved surface in the same direction* $\mathbf{T}$ *have the same normal curvature* $\kappa_n(\mathbf{T})$ *as the normal section in that direction. If the osculating plane at* p *of one such curve makes angle* γ *with the tangent plane at* p*, and its curvature there is* κ_γ*, then* $\kappa_\gamma \sin\gamma = \kappa_n(\mathbf{T})$ *is independent of* γ. (11.4)

Before giving a general argument, consider the sphere of radius R, for in this special case the truth of the theorem is easy to visualize. See [11.3]. Taking p to be the north pole, the normal section is a great circle (a meridian) of radius R and hence the curvature $\kappa_n = (1/R)$. Slicing through the north pole with a plane at angle γ to the horizontal tangent plane at p, the intersection with the sphere will be a circle of radius $r = R\sin\gamma$, and hence $\kappa_\gamma = 1/r = 1/(R\sin\gamma)$.

Returning to the introductory example of a small circle drawn in the sand, imagine that γ tends to zero, in which case the radius of the circle shrinks to zero, and the length of its curvature vector tends to infinity. But, simultaneously, this curvature vector is tending towards orthogonality with the surface normal, so that less and less of it projects onto that direction. These two effects cancel each other out *exactly*, so that the projection of the curvature vector onto the surface normal has *constant* magnitude:

$$\kappa_\gamma \sin\gamma = \frac{1}{R} = \kappa_n,$$

confirming this instance of the theorem.

For the general case, imagine that $\mathcal{C}$ is traced by a particle at unit speed, and let $\mathbf{T}$ denote its velocity vector, not just at p but along the entire orbit. Then, by (11.3),

$$\dot{\mathbf{T}} = \boldsymbol{\kappa} = \boldsymbol{\kappa}_g + \boldsymbol{\kappa}_n \quad \Longrightarrow \quad \kappa_n(\mathbf{T}) = \dot{\mathbf{T}} \bullet \mathbf{n}.$$

But for the same essential reason as in [8.5], $\dot{\mathbf{T}} \bullet \mathbf{n} = -\mathbf{T} \bullet \dot{\mathbf{n}}$, so

$$\kappa_n(\mathbf{T}) = -\mathbf{T} \bullet \dot{\mathbf{n}}. \tag{11.5}$$

But $\dot{\mathbf{n}}$ is the rate of change of the surface normal in the $\mathbf{T}$ direction, which is independent of $\mathcal{C}$, thereby proving the theorem.

11.3 Geodesics are "Straight"

We have defined geodesics via their length-minimizing property within the surface. But the geodesics of the Euclidean plane (lines) may also be recognized by their *straightness*. Likewise, on the sphere we have just seen that the geodesics (great circles) not only provide the shortest routes, but they too are "straight," in the sense that none of their curvature is visible to inhabitants of the surface: $\kappa_g = 0$. In fact this connection between length minimization and intrinsic straightness is universal:

> *Geodesics appear to be straight lines to the inhabitants of the surface: they are intrinsically "straight" in the sense that their geodesic curvature vanishes:* $\kappa_g = 0$ *at every point of the geodesic.*

To begin to understand this, imagine a guitar string in its straight-line equilibrium position L being plucked with a pick whose position in space is p. As the pick pulls the string away from L, the string forms a triangle in the plane Π containing p and L. Once released, the net force, resulting from the string's compulsion to shorten its length, acts within Π and the resulting motion back towards L all takes place within that plane. Next, suppose that instead of pulling the string away from L into a sharp triangle in Π, we pull it away into the form of a gentle convex curve lying in Π, then the resulting forces and motion will again clearly reside within Π.

Now let us return to our curved surface and picture our geodesic as a guitar string stretched taut over the (frictionless) surface to connect two fixed points. The string is at rest: it is in equilibrium on the surface, already having slid over the surface, contracting to become as short as possible. Now focus attention on a very short segment apb of the geodesic. This segment will *almost* lie in a single plane Γ, namely, the one through a, p, and b. As a and b approach p, the

limiting position of Γ is the osculating plane at p, which we shall denote Π_p. By the forgoing reasoning, the net length-shrinking force $\mathbf{F}_p$ acting on apb will ultimately lie within Π_p.

If $\mathbf{F}_p$ were to have any component tangent to the surface, the string would be free to move in this direction, thereby reducing its length. But since the string is already as short as possible, this cannot happen! Thus, having no component within the surface, the force $\mathbf{F}_p$ (lying within Π_p) must be directed perpendicularly to the surface, along the normal $\mathbf{n}_p$. Thus we have reached an important characterization[1] of the geodesic in terms of the *extrinsic* geometry of the surface:

> *At every point p of a geodesic, the osculating plane Π_p contains the surface normal $\mathbf{n}_p$, and hence the geodesic curvature vanishes: $\kappa_g = 0$.* (11.6)

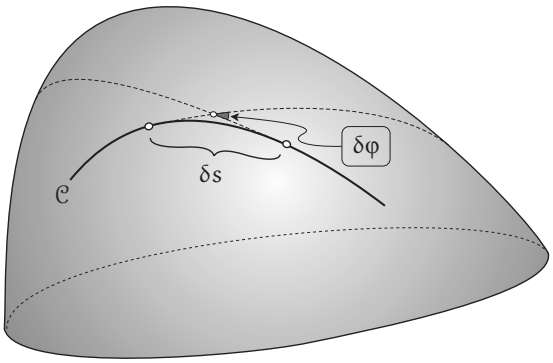
Intrinsic and extrinsic geometry appear to be belong to entirely different worlds, and yet we see here that the two are strangely entangled. Later we shall witness even deeper and more mysterious connections between these two worlds.

11.4 Intrinsic Measurement of Geodesic Curvature

We have just seen that vanishing geodesic curvature characterizes geodesics, but we may also use geodesics as an intrinsic tool with which to measure the geodesic curvature of a *non*geodesic curve, for which $\kappa_g \neq 0$.

In (8.4) we saw that the curvature of a curve in the Euclidean plane may be viewed as the rate of rotation of its tangent line with respect to distance along the curve. Well, this original construction (illustrated in [8.4]) also makes perfectly good sense to the inhabitants of the surface.

From *their* perspective there is no change at all to the construction. See [11.4]. They draw tangent "lines" (dashed geodesics) to the curve at neighbouring points δs apart, find the angle $\delta\varphi$ at which these tangents intersect, then calculate the curvature (in the limit that the points merge) as $\kappa_g \asymp (\delta\varphi/\delta s)$.



[11.4] *The surface's inhabitants can measure the geodesic curvature κ_g of the curve $\mathcal{C}$ by constructing tangent geodesics to the curve at neighbouring points δs apart, then finding the angle $\delta\varphi$ at which these tangents intersect. As the points merge, $\kappa_g \asymp (\delta\varphi/\delta s)$.*

But what they call *the* curvature of the curve, we recognize as being only one part of the curvature, its geodesic curvature κ_g; the normal curvature κ_n is invisible and unknowable to them. The only other difference between their intrinsic perspective and our extrinsic perspective (looking down on the surface) is that what they call "straight lines" we recognize as geodesics within their surface.

Of course if $\mathcal{C}$ is itself a geodesic then both tangents coincide with $\mathcal{C}$, so $\delta\varphi = 0$, and therefore $\kappa_g = 0$, as it should.

---

[1] This characterization was first discovered by Johann Bernoulli in 1697, who then taught it to his student, Euler.

11.5 A Simple Extrinsic Way to Measure Geodesic Curvature

In Act I we discussed the fact that if we peel away from a curved surface a narrow strip centred on a geodesic G, then when that strip is laid down on a flat plane it becomes a straight line $\widetilde{G}$. See [1.11] on page 12. Thus the intrinsic straightness ($\kappa_g = 0$) of G on the surface manifests itself as ordinary straightness in the plane: the curvature $\widetilde{\kappa}$ of the flattened strip $\widetilde{G}$ vanishes.

If we likewise remove from the surface a narrow strip centred on a *non*geodesic curve $\mathcal{C}$, for which $\kappa_g \neq 0$, then flatten it onto the plane, we obtain a plane curve $\widetilde{\mathcal{C}}$ for which its ordinary curvature $\widetilde{\kappa} \neq 0$. See [1.12] on page 13.

So how is the geodesic curvature $\kappa_g(p)$ at a particular point p on the surface related to the curvature $\widetilde{\kappa}(\widetilde{p})$ of the flattened strip at the corresponding point $\widetilde{p}$ of $\widetilde{\mathcal{C}}$ in the plane?

> *Let $\kappa_g(p)$ denote the geodesic curvature at p of a curve $\mathcal{C}$ on a curved surface, and let $\widetilde{\kappa}(\widetilde{p})$ denote the curvature at the corresponding point of the plane curve $\widetilde{\mathcal{C}}$ into which $\mathcal{C}$ is carried when a narrow strip centred on $\mathcal{C}$ is flattened onto the plane. Then*
>
> $$\kappa_g(p) = \widetilde{\kappa}(\widetilde{p}).$$

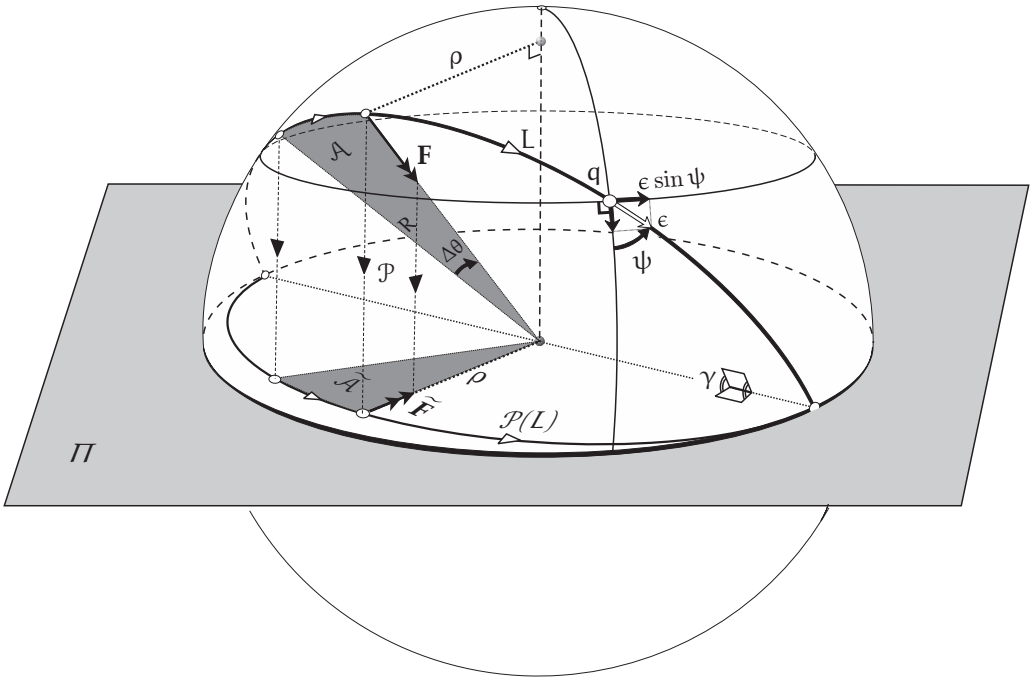
To see this, take the intrinsic construction [11.4] and imagine peeling narrow strips off the surface around the entire construction: around $\mathcal{C}$, and also around both of the small dashed segments of tangent geodesics meeting at angle $\delta\varphi$. When the small triangle of connected peeled strips is laid flat in the plane, we have returned to the original construction [8.4] for measuring curvature. The flattened geodesics have become straight lines in the plane, and since these lines are still tangent to $\widetilde{\mathcal{C}}$, they are tangent lines. But neither δs nor $\delta\varphi$ are altered in the flattening process, so $\kappa_g(p) \asymp (\delta s/\delta\varphi) \asymp \widetilde{\kappa}(\widetilde{p})$, as claimed. To see this, take the intrinsic construction [11.4] and imagine peeling narrow strips off the surface around the entire construction: around $\mathcal{C}$, and also around both of the small dashed segments of tangent geodesics meeting at angle $\delta\varphi$. When the small triangle of connected peeled strips is laid flat in the plane, we have returned to the original construction [8.4] for measuring curvature. The flattened geodesics have become straight lines in the plane, and since these lines are still tangent to $\widetilde{\mathcal{C}}$, they are tangent lines. But neither δs nor $\delta\varphi$ are altered in the flattening process, so $\kappa_g(p) \asymp (\delta s/\delta\varphi) \asymp \widetilde{\kappa}(\widetilde{p})$, as claimed.

11.6 A New Explanation of the Sticky-Tape Construction of Geodesics

We originally used the length-minimizing property of geodesics to explain the construction (1.7), whereby a geodesic is constructed by rolling a narrow strip of sticky tape down onto the surface, starting at an arbitrary point, and heading off in an arbitrary direction of our choosing. But, as we have been discussing, geodesics can also be characterized by their *straightness*: vanishing geodesic curvature. We now use this property to provide a new, second explanation of our geodesic construction.

Consider a narrow, straight strip of tape, $\mathcal{L}$, lying flat in the plane, with centre line L. The line L is extrinsically straight in $\mathbb{R}^3$; i.e., $\boldsymbol{\kappa} = \mathbf{0}$. It is also intrinsically straight within $\mathcal{L}$: $\boldsymbol{\kappa}_g = \mathbf{0}$. Now let us pick up $\mathcal{L}$ and isometrically bend and twist it in space into any shape we please. The centre line L remains intrinsically straight within whatever new form $\mathcal{L}$ now takes, so $\boldsymbol{\kappa}_g = \mathbf{0}$, still. Thus the curvature vector is normal to the strip all along L:

$$\boldsymbol{\kappa} = \boldsymbol{\kappa}_g + \boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n.$$



[11.5] *As the particle* q *travels along the geodesic* L *at unit speed, the radius sweeps out area at a constant rate, and therefore its projection onto* Π *does too. It follows that* $\rho \sin\psi$ *is constant along* L*, which is a special case of Clairaut's Theorem.*

Now roll the strip down onto a smooth surface $\mathcal{S}$. As we explained on page 14, only the centre line L can actually make contact with $\mathcal{S}$, but at each point of contact, p, the tangent plane of the strip coincides with the tangent plane of the surface: $\mathcal{T}_p(\mathcal{L}) = \mathcal{T}_p(\mathcal{S})$.

But we just established that $\boldsymbol{\kappa}$ is normal to the strip $\mathcal{L}$ all along L, but this means that it is also normal to the surface, $\mathcal{S}$. Thus, viewed as a curve within $\mathcal{S}$, the geodesic curvature of L vanishes, and it is indeed a geodesic of $\mathcal{S}$.

11.7 Geodesics on Surfaces of Revolution

11.7.1 Clairaut's Theorem on the Sphere

Geodesics are extremely hard to find explicitly on all but the simplest surfaces. However, there exists one general class of surfaces for which we can give an explicit geometrical recipe (called *Clairaut's Theorem*) for the paths of the geodesics: surfaces of revolution.

Of such surfaces, the sphere is one of the simplest; the Ancients already knew its geodesics—the great circles. We now look afresh at these great circles, and expose a hidden property that they possess, one that we will then be able to generalize to *all* surfaces of revolution.

Consider the sphere [11.5], thought of as the surface of revolution generated by rotating a semicircle about the vertical z-axis. As this semicircle (or indeed a general generating curve) rotates about the axis, its creates the so-called ***meridians*** of the surface. Equivalently, these are the curves of intersection of the surface with planes through its axis of symmetry, and the same is true on a general surface of revolution. On the sphere these meridians are the great circles through the poles, perhaps better known as circles of longitude. It is no accident that these meridians on the sphere are geodesics; as we already noted in the footnote on page 53, and as we shall discuss shortly, the meridians on a general surface of revolution are necessarily geodesics, too.