

From Spiral to Spline: Optimal Techniques in Interactive Curve Design

by

Raphael Linus Levien

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Engineering—Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Carlo Séquin, Chair
Professor Jonathan Shewchuk
Professor Jasper Rine

Fall 2009

From Spiral to Spline: Optimal Techniques in Interactive Curve Design

Copyright © 2009

by

Raphael Linus Levien

Abstract

From Spiral to Spline: Optimal Techniques in Interactive Curve Design

by

Raphael Linus Levien

Doctor of Philosophy in Engineering—Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Carlo Séquin, Chair

A basic technique for designing curved shapes in the plane is *interpolating splines*. The designer inputs a sequence of control points, and the computer fits a smooth curve that goes through these points. The literature of interpolating splines is rich, much of it based on the mathematical idealization of a thin elastic strip constrained to pass through the points. Until now there is little consensus on which, if any, of these splines is ideal. This thesis explores the properties of an ideal interpolating spline. The most important property is *fairness*, a property often in tension with *locality*, meaning that perturbations to the input points do not affect sections of the curve at a distance. The idealized elastic strip has two serious problems. A sequence of co-circular input points results in a curve deviating from a circular arc. For some other inputs, no solution (with finite extent) exists at all.

The idealized elastic strip has two properties worth preserving. First, any ideal spline must be *extensional*, meaning that the insertion of a new point on the curve shouldn't change its shape. Second, curve segments between any two adjacent control points are drawn from a two-parameter family (and this property is closely related to good locality properties). A central result of this thesis is that any spline sharing these properties also has the property that all segments between two control points are cut from a single, fixed generating curve. Thus, the problem of choosing an ideal spline is reduced to that of choosing the ideal generating curve. The Euler spiral has excellent all-around properties, and, for some applications, a log-aesthetic curve may be even better.

Shapes in applications such as font outlines contain extra features such as corners and transitions between straight lines and smooth curves. Attaching additional constraints to control points expresses these features, and, carefully applied, give the designer a richer palette of curve types.

The splines presented in this thesis are entirely practical as well, especially for designing fonts. Sophisticated new numerical techniques compute the splines at interactive speeds, as well as convert to optimized cubic Bézier representation.

Professor Carlo Séquin
Dissertation Committee Chair

To my father.

Contents

Contents	ii
List of Figures	vii
List of Tables	xi
Acknowledgements	xii
1 Introduction	1
2 Properties of splines	6
2.1 Fairness	6
2.1.1 Empirical testing of fairness	7
2.2 Continuity	8
2.3 Roundness	9
2.4 Extensionality	10
2.5 Existence and uniqueness	11
2.6 Locality	13
2.7 Transformational invariance	13
2.8 Counting parameters	14
2.9 Monotone curvature	14
2.10 Summary	15
3 The Elastica	16
3.1 Statement of the elastica problem	17
3.2 Variational study	17
3.2.1 The Euler–Lagrange equation	17
3.2.2 Application of the Euler–Lagrange equation to the elastica	18

3.3	A family of solutions	19
3.4	Equilibrium of moments	21
3.5	Pendulum analogy	23
3.6	Equilibrium of forces: Finite element approach	25
3.7	Solutions with length and endpoint constraints	29
3.8	Elastica with general constraints	29
3.9	Angle constraints	32
3.10	Elastica as an interpolating spline	35
3.11	SI-MEC	35
3.12	Real splines	39
4	Two-parameter splines	42
4.1	Extensional two-parameter splines have a generating curve	43
4.2	Properties of the MEC spline	45
4.3	SI-MEC	46
4.4	Cubic Lagrange (with length parametrization)	47
4.5	Ikarus (biarc)	47
4.6	Euler’s spiral	48
4.6.1	Euler’s spiral as a spline primitive	51
4.6.2	Variational formulation of Euler’s spiral	51
4.6.3	Euler spiral spline in use	52
4.6.4	Existence and uniqueness of Euler spiral spline	53
4.7	Hobby’s spline	57
4.8	Splines from fixed generating curves	60
4.8.1	Log-aesthetic curve splines	61
4.9	Piecewise SI-MEC splines, or Kurgla 1	63
4.10	Circle splines	64
4.11	Summary	65
5	History of the elastica	67
5.1	Jordanus de Nemore – 13th century	68
5.2	Galileo sets the stage – 1638	68
5.3	Hooke’s law of the spring – 1678	69
5.4	James Bernoulli poses the elastica problem – 1691	70

5.5	James Bernoulli partially solves it – 1692	71
5.6	James Bernoulli publishes the first solution – 1694	73
5.7	Daniel Bernoulli proposes variational techniques – 1742	77
5.8	Euler’s analysis – 1744	78
5.8.1	Moments	83
5.9	Elliptic integrals	85
5.10	Laplace on the capillary – 1807	86
5.11	Kirchhoff’s kinetic analogy – 1859	87
5.12	Closed form solution: Jacobi elliptic functions – 1880	89
5.12.1	Inflectional solutions	90
5.12.2	Non-inflectional solutions	91
5.13	Max Born – 1906	91
5.14	Influence on modern spline theory – 1946 to 1965	92
5.15	Numerical techniques – 1958 through today	93
6	History of the Euler spiral	95
6.1	James Bernoulli poses a problem of elasticity – 1694	95
6.2	Euler characterizes the curve – 1744	98
6.3	Euler finds the limits – 1781	100
6.4	Relation to the elastica	101
6.5	Fresnel on diffraction problems – 1818	103
6.6	Talbot’s railway transition spiral – 1890	106
6.7	Mathematical properties	107
6.8	Use as an interpolating spline	108
6.9	Efficient computation	109
7	Four-parameter curves	110
7.1	Why four parameters?	110
7.2	Minimum Variation Curve	114
7.2.1	Euler–Poisson equation	114
7.2.2	Euler–Poisson solution of MVC	115
7.3	Polynomial spiral spline	117
7.4	Generalized constraints	119

8	Numerical toolbox	122
8.1	Numerical integration of polynomial spirals	123
8.1.1	Polynomial approximation of the integral	124
8.1.2	Higher-order polynomial approximations	125
8.1.3	Hybrid approach	126
8.2	Determining Euler spiral parameters	127
8.3	Global constraint solver	129
8.3.1	Two-parameter solver	130
8.3.2	Four-parameter solver	131
8.4	2-D Lookup table	132
9	Converting to Bézier curves	135
9.1	Parameters	136
9.2	Error metrics	136
9.3	Simple conversion	138
9.4	Equal arc length, equal arm length solution	139
9.5	Fully optimized solution	140
9.6	Global optimum	140
9.6.1	Hybrid error metric	144
9.6.2	Simple subdivision	145
9.6.3	Smarter subdivision	145
9.6.4	Alternative optimum algorithm	147
10	Applications in font design	149
11	Space curves	160
11.1	Curves in space	161
11.2	Minimum energy curve in space	162
11.2.1	MEC as spline in 3-space	163
11.3	Mehlum’s spiral	163
11.4	Log-aesthetic space curves	165
11.5	Surfaces	165
12	Conclusion	167

List of Figures

1.1	A mechanical spline.	2
1.2	Rectangular elastica.	3
1.3	Euler spiral.	3
1.4	Conversion to optimized Bézier curves.	4
2.1	Test instrument for fairness user study.	7
2.2	Survey results for aesthetic curve preference.	8
2.3	Continuity \neq fairness.	9
2.4	MEC roundness failure.	10
2.5	Circle spline extensionality failure.	10
2.6	Nonexistence of solution to MEC.	11
2.7	Uniqueness failure in Euler spiral spline.	12
2.8	G^2 spline has better locality than G^4	13
2.9	Attneave's curvature experiment [5].	15
3.1	The family of elastica solutions.	20
3.2	Curvature plots for the family of elastica solutions.	22
3.3	Equilibrium of moments.	23
3.4	An oscillating pendulum.	24
3.5	Examples of pendulum analogy.	26
3.6	Finite element approximation of elastica as chain of struts and pivots.	27
3.7	Forces on a single strut.	27
3.8	Forces on a single pivot.	27
3.9	Forces on a single pivot point in a chain.	28
3.10	Values of λ for varying chord lengths.	30
3.11	Multiplicity of solutions for same tangent constraints.	31

3.12	Forces acting on rectangular elastica.	32
3.13	Angle-constrained minimum energy curves.	33
3.14	Curvature plots, angle-constrained minimum energy curves.	34
3.15	MEC and SI-MEC energies for varying lengths ($\pi/2$).	36
3.16	MEC and SI-MEC energies for varying lengths ($\pi/3$).	37
3.17	SI-MEC and MEC curve energy for ($\pi/2, -\pi/2$) angle constraints.	38
3.18	A real spline.	39
3.19	High-friction spline constraints on a real spline.	40
3.20	Bricks as high friction constraints on a real spline.	41
4.1	Incremental construction of the generator curve of a two-parameter spline.	43
4.2	MEC roundness failure.	46
4.3	Typical letter-shape using Ikarus.	48
4.4	Euler’s spiral	49
4.5	Euler’s spiral spline primitive over its parameter space.	50
4.6	Closed Euler example.	53
4.7	Complex Euler example.	54
4.8	Difficulty with straight-to-curved joins.	55
4.9	Multiple windings for Euler spiral segment connecting two points (from [82]).	56
4.10	Hobby spline primitive over its parameter space.	58
4.11	Curvature plot of Hobby spline over its parameter space.	59
4.12	Test instrument for fairness user study.	61
4.13	Perturbation fall-off rate as a function of aesthetic curve exponent.	62
4.14	MEC, SI-MEC, and Euler spiral compared.	64
4.15	Circle spline extensionality failure.	65
5.1	Galileo’s 1638 problem.	68
5.2	Hooke’s figure on compound elasticity.	69
5.3	Bernoulli poses the elastica problem.	71
5.4	Drawing from James Bernoulli’s 1692 Med. CLXX, with modern reconstruction.	72
5.5	Bernoulli’s 1694 publication of the elastica.	73
5.6	Bernoulli’s justification for his formula for curvature.	74
5.7	A simplified diagram of moments.	75
5.8	Huygens’s 1694 objection to Bernoulli’s solution.	77

5.9	Euler’s elastica figures, Tabula III.	81
5.10	Euler’s elastica figures, Tabula IV.	82
5.11	The family of elastica solutions.	84
5.12	Bernoulli’s lemniscate.	85
5.13	Maxwell’s figures for capillary action.	87
5.14	An oscillating pendulum.	88
5.15	Rectangular elastica as roulette of hyperbola.	90
5.16	Born’s experimental apparatus for measuring the elastica.	92
6.1	Euler’s spiral as an elasticity problem.	96
6.2	Bernoulli’s construction.	97
6.3	Euler’s drawing of his spiral, from Tabula V of the Additamentum.	98
6.4	Reconstruction of Euler’s Fig. 17, with complete spiral superimposed.	99
6.5	The figure from Euler’s “De valoribus integralium”, with modern reconstruction. . .	101
6.6	Euler’s spiral, rectangular elastica, and cubic parabola.	102
6.7	Diffraction through a slit.	103
6.8	Cornu’s plot of the Fresnel integrals.	105
6.9	Talbot’s Railway Transition Spiral.	107
7.1	Eccentricity 0.56199 oval with G^2 and G^4 continuity.	111
7.2	The suitcase corners problem.	112
7.3	Bracketed serifs and straight-to-curve joins.	113
7.4	One-way constraints enforce locality.	113
7.5	G^2 spline has better locality than G^4	118
7.6	Matching nonzero curvatures.	119
7.7	A one-way constraint creates a cyclic dependency.	120
7.8	Examples of different kinds of constraint points.	120
8.1	Precision vs. timing for different integration strategies.	126
8.2	Determining Euler spiral parameters.	128
8.3	Python code for computing Euler spiral parameters.	130
8.4	Pseudocode for computing four-parameter splines.	132
8.5	Example of four-parameter solver.	133
8.6	Contour plot of curvature lookup table for Euler spiral.	134

9.1	A cubic Bézier.	136
9.2	Two circle approximations with equal distance error.	137
9.3	Euler spiral approximation drawn using .03 and .001 error threshold, respectively.	138
9.4	Three Bézier solutions: fast, equal arc length, and fully optimized. Curvature plot is the top graph, the Bézier is on bottom.	139
9.5	Approximation errors over parameter space for cubic Bézier fit.	141
9.6	Local minima of cubic Bézier fitting, $t=[0.5, 1.0]$. Curvature plots are the top two graphs, the Béziers are on the bottom.	142
9.7	Local minima of cubic Bézier fitting, $t=[0.5, 0.85]$. Curvature plots are the top graphs, the Béziers are on the bottom.	143
9.8	Conversion into optimized cubic Béziers.	144
9.9	Four different levels of optimization: subdivision in half with equal arm length, subdivision in half with fully optimized Bézier segments, greedy subdivision with fully optimized segments, and fully optimized.	146
10.1	Selected glyphs from Inconsolata.	150
10.2	Selected glyphs from Inconsolata.	151
10.3	Selected glyphs from Baskerville Italic.	153
10.4	Selected glyphs from Baskerville Italic.	154
10.5	Lowercase ‘w’ from Baskerville Italic, scan with outline.	155
10.6	Selected glyphs from Cecco.	156
10.7	Selected glyphs from Cecco.	157
10.8	Typical letter-shape using IKARUS, with Spiro comparison.	158
10.9	Interpolation masters.	158
10.10	Interpolation between extremes.	158
10.11	Interpolation between extremes, superimposed.	159
10.12	Interpolation between cubic Béziers fails to preserve G^1 -continuity.	159
11.1	Space curves rendered into ironwork.	160
11.2	Curvature as a function of arc length in the Mehlum spiral.	164
12.1	Contour plot of curvature lookup table for $\alpha = 2$ log-aesthetic curve.	168

List of Tables

4.1	Comparison of properties of two-parameter splines.	66
5.1	Euler's characterization of the space of all elastica.	80
7.1	Constraints for curve points, and their corresponding formulae.	121
10.1	Constraints used to draw font outlines.	152

Acknowledgements

Thanks to Derek Waters for the image of the spline and weights in Figure 1.1. The photograph of architectural ironwork shown in Chapter 11 is by Terry Ross of drawmetal.com, as is the original work.

This historical sketch in Chapter 5 draws heavily on Truesdell’s history in an introduction to Ser 2, Volume X of Euler’s *Opera Omnia* [86]. Indeed, a careful reading of that work reveals virtually all that needs to be known about the problem of the elastica and its historical development through the time of Euler.

I am deeply grateful to Alex Stepanov, Seth Schoen, Martin Meijering and Ben Fortson for help with the translations from the Latin in Chapters 5 and 6. Special mention is also due the excellent “Latin words” program by William Whitaker, which lets anyone with a rudimentary knowledge of Latin and a good deal of patience and determination puzzle out the meaning of a text. Of course, all errors in translation are my own responsibility.

Patricia Radelet has been most helpful in supplying high-quality images for James Bernoulli’s original figures, reproduced here as Figures 5.4 and 5.5, and for providing scans of the original text of his 1694 publication.

Thanks also to the Stanford Special Collections Library for giving me access to Prof. Forsythe’s notes and correspondence, which were invaluable in tracing the influence of the elastica through early work on non-linear splines.

Thanks to Karl Berry for supporting the development of the Inconsolata font, and for building the T_EX configuration files so it could easily be used in this thesis.

I gratefully appreciate Prof. Norimasa Yoshida for a careful reading and helpful comments on an earlier draft of this thesis.

And of course, many thanks are due to my thesis committee for providing feedback invaluable for refining earlier drafts, and especially my advisor, Prof. Carlo Séquin.

Chapter 1

Introduction

A basic design task is drawing smooth curves in the plane. Curves are needed in purely aesthetic applications such as making vector drawings, purely industrial applications such as the cross-sections of airplane wings or railway track layouts, and also applications both functional and artistic, such as font design.

There are several approaches to curve design. The most popular is drawing with Bézier curves, usually cubic. These curves are versatile and very easy to compute and manipulate, but they are somewhat difficult to learn, and it is especially difficult to get smooth curves. In expert hands, the results are good, but for less experienced designers, adding more control points to tweak a curve often leads to more lumpiness.

Another approach is to sample real-world data (for example, from a sculpture), then fit a curve to it. Since these points are noisy, it's most common to fit an *approximating spline* through them – the spline goes near the points but not necessarily through them.

The main focus of this thesis is *interpolating splines*, or curves that go through the points given, in sequence. With a good interpolating spline, the designer can enter a relatively sparse sequence of data points, and the result is a smooth curve.

Interpolating splines have their roots in the mechanical spline, a thin strip of flexible material constrained to go through the chosen points with weighted implements (called “ducks” or “whales” due to their shape), but otherwise free. An example is shown in Figure 1.1. Such a strip will naturally find its minimal energy configuration. Indeed, one approach to interpolating splines is simply to simulate the mechanical spline. As we shall see, such an approach is severely flawed.

A part of the problem with interpolating splines is that there are many plausible constructions, with quite a few in actual use, but no real consensus on which one is best, or even how to adequately measure or compare. Many of the interpolating splines in use seem to have been chosen based on expedience, simple representation, or cheap computational cost. However, these criteria are not really justified, not only because Moore’s law has made even intricate computations feasible at interactive speeds, but also because excellent practical splines exist with a simple representation and an efficient implementation.

This thesis will closely examine the question of what interpolating spline is *ideal* out of all

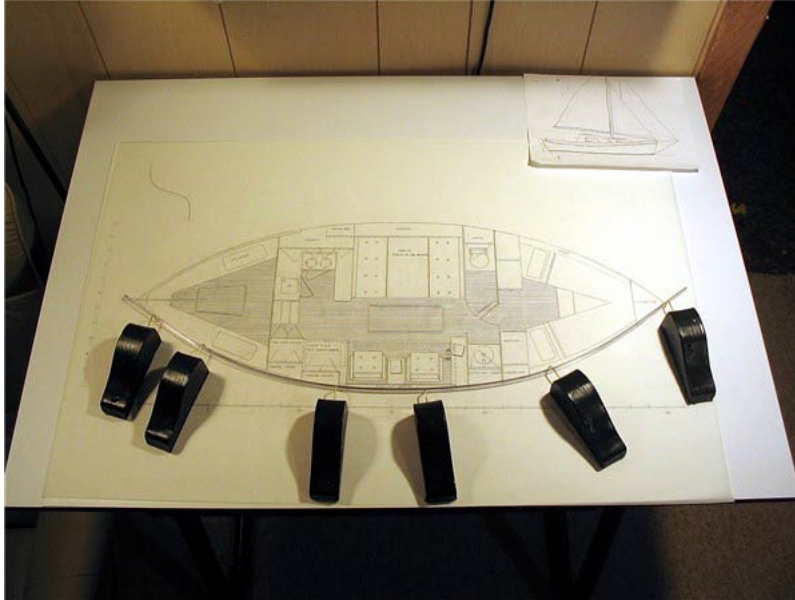


Figure 1.1. A mechanical spline.

possibilities. Such a focus reduces the space of alternatives to consider, excluding all the many rough and crude approximations.

By far, the most important property of an interpolating spline is *fairness*, (also known, less technically, as smoothness). Indeed, one way to define an interpolating spline is to choose a fairness metric (such as the total bending energy of a flexible strip) and optimize it, given that the curve passes through the control points. However, this approach may compromise other important properties, such as robust convergence, locality (intuitively, lack of ripples), stability, and others. Chapter 2 lists all these desirable properties of interpolating splines and describes them in detail.

The mechanical spline is the basis for essentially all following work in splines. Its mathematical idealization is the *Minimum Energy Curve* (MEC), and the curve describing segments between adjacent control points is the *elastica*. This curve has surprising and deep properties, and there are a number of ways to fully understand it – as the result of an energy minimization, as the limit of a finite element system consisting of rigid beams connected by flexible springs, and as the physical analogue of a pendulum (which explains why the curve is periodic). Chapter 3 explores all these characterizations and the connections between them.

Yet, while the MEC is the optimum of a particular functional (bending energy), it is clearly not the best possible spline. Experiment shows that bending energy does not accurately predict perceived smoothness. This finding is also supported by theoretical analysis – in particular, the MEC spline lacks *roundness*, the property that control points arranged on a circular arc yields that arc. Further, optimizing this particular metric yields a spline with poor robustness – for many inputs, no optimum solution exists at all. Even so, the MEC has many desirable properties, some of which are worth preserving. Two in particular stand out. An optimum spline must be *extensional*, meaning that adding an additional control point on the curve of an existing spline preserves the shape of the curve. Further, the MEC has the property that the family of curve segments between any two adjacent control points is a manifold with a two-dimensional parameter space. Chapter

4 argues that, for two-dimensional drawing applications, any “best” spline should also have this property.

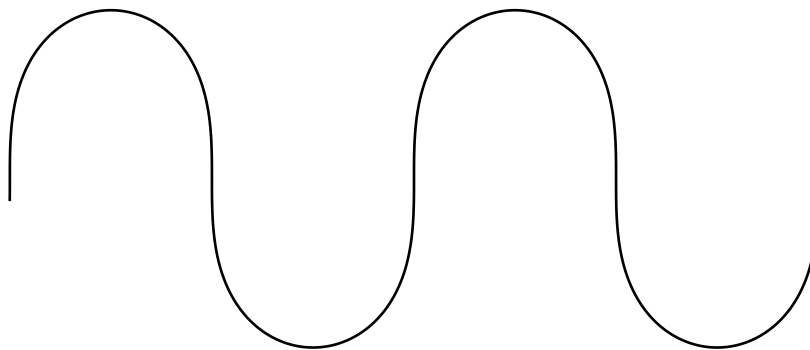


Figure 1.2. Rectangular elastica.

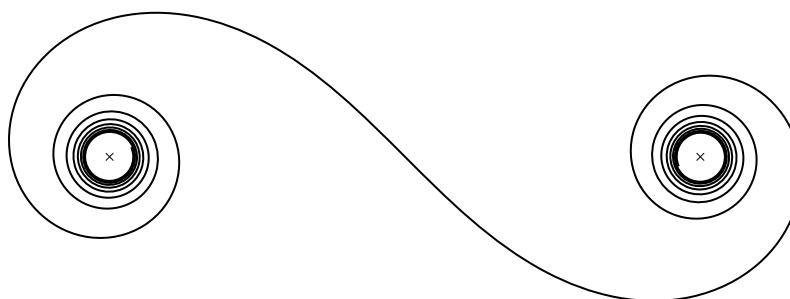


Figure 1.3. Euler spiral.

Chapter 4 also presents one of the central technical results of this thesis: any spline that is both extensional and two-parameter is derived from a single, rigid generating curve, with each segment cut from the curve, then rotated, scaled, and translated into place to align with the endpoints. In the MEC, this curve is the *rectangular elastica* (shown in Figure 1.2), but other choices yield better splines. In particular, the *Euler spiral* (shown in Figure 1.3) produces results similar to the MEC, but fixing the roundness and robustness problems. Chapter 4 then concludes by exploring the potential for the family of *log-aesthetic* curves to produce splines even slightly “better” than the Euler spiral spline, improving simultaneously fairness (as determined by empirical measurement) and locality.

The two basic underlying curves for two-parameter splines, the elastica and the Euler spiral, both have rich histories, dating back hundreds of years. But they have not been fully explained in one concise source until now. For both of these curves, Jacob Bernoulli wrote down the precise mathematical description and partially explored the nature of the curves, while Euler characterized them fully. Chapter 5 deeply explores the history of the elastica, including its subsequent role in founding the field of study of elliptical integrals, which are now also useful for giving a closed-form, analytical solution to the shape of the curve. Similarly, Chapter 6 explores the history of the Euler spiral, including its repeated rediscovery and applications in other diverse fields, such as solving diffraction problems and layout of railroad tracks.

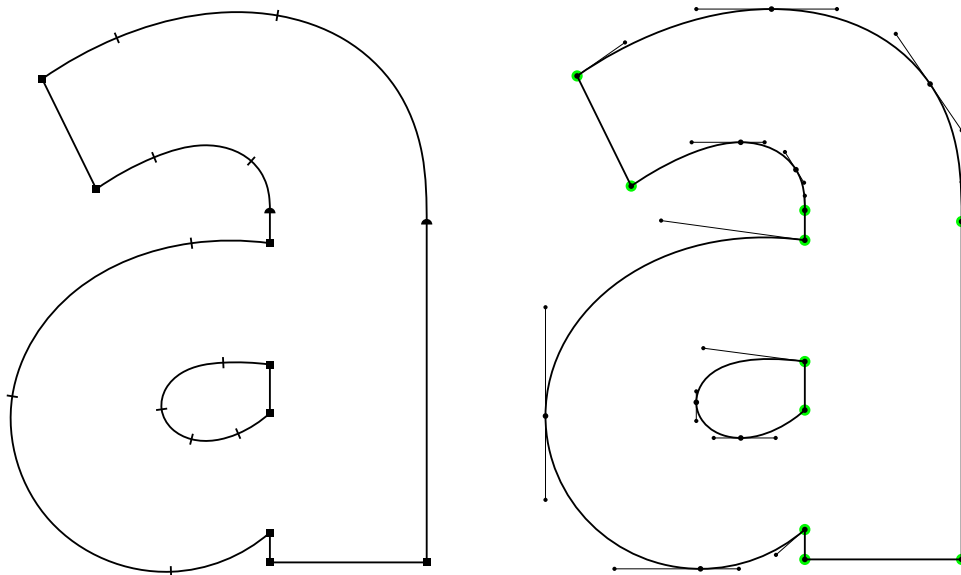


Figure 1.4. Conversion to optimized Bézier curves.

Two-parameter splines are ideal as basic interpolating splines used for planar curve design, but there are applications where a higher degree of geometric continuity is desired. Chapter 7 presents the logical extension to a four-parameter space, providing G^4 -continuity, but at the cost of reduced locality. The underlying curve family is also useful for satisfying a richer family of constraints than merely that the curve passes through the control points. In particular, such constraints can specify smooth transitions from straight sections to curves, a frequent pattern in font designs. The chapter also explores the connection of this spline to the *Minimum Variation Curve*, a known spline based on minimizing a metric defined in terms of the *variation* of curvature rather than curvature itself, as in the MEC.

None of these theoretical results about optimal splines would be much use without an efficient, practical implementation. Chapter 8 provides a toolbox of such techniques, including: efficiently computing the integral defining the Euler spiral and its generalization to a four-parameter spline; solving the parameters of this curve to meet given tangent constraints; and providing an efficient Newton-style solver for simultaneously solving all the constraints required for a given spline. Together, these numerical techniques add up to a very fast solver for splines, entirely suitable for interactive work even in limited computational environments.

Even after solving the global spline and determining an analytical curve that meets the constraints, the resulting curve is still not in a form that can be widely used. Euler spirals (and their cubic polynomial generalization, even more so) are considered “special functions” and are very unlikely to be present as primitives in graphics files formats, CAD applications, and so on. Thus, another important component for actually using these splines in practice is conversion to a more widely-supported format, of which by far the most popular is cubic Bézier curves. Chapter 9 describes how to convert these curves into Bézier form, with a tradeoff between the computational effort expended on optimization and the conciseness of the representation. For contexts when the size of the representation matters (for example, when downloading fonts over the network), the chapter presents a fully optimized technique, yielding the best possible Bézier representation with

respect to the given error threshold. Figure 1.4 shows an example of a letterform converted into such an optimized Bézier representation.

A major motivation for this work was to improve tools for designing fonts. I designed several complete fonts entirely using the splines described in this thesis, and one, *Inconsolata*, is released and is widely used. Chapter 10 presents some of these font designs, and also discusses particular aspects of splines relevant to the specific application of font design, including prospects for more easily creating fonts with *variation*, for example, a smooth blending between light and heavy weights.

While the primary focus of this thesis is planar curves, the principles are also useful for splines in space, as explored in Chapter 11. An intrinsic formulation of planar curves is a relationship between arc length and curvature. The corresponding intrinsic form for a space curve is a relationship between arc length and both curvature and torsion. Three parameters are sufficient to attain G^2 -continuity of space curves, as opposed to two for planar curves. The MEC generalizes easily from planar to space curves, but of course has the same limitations as the planar MEC. A beautiful curve primitive due to Mehlum [60] generalizes the Euler spiral to space curves. In addition to its excellent properties as a spline primitive, it has a beautiful geometric construction in terms of rolling up an Euler spiral on the surface of the sphere.

Chapter 12 concludes and summarizes the main results of this thesis. The spline curves presented in the following chapters are practical and well-suited to design tasks such as fonts. Theoretical analysis characterizes large regions of the space of all possible splines, suggesting that the ideal interpolating spline is, in fact, drawn from an easy-to-understand family. In addition, the discovery that all extensional, two-parameter curves are derived from a characteristic generating curve unifies previously disparate areas of the literature, in particular the Euler spiral spline and variational techniques. It is hoped that these beautiful curves will become more popular in the future, now that their properties are better known, and their implementation need not be much more complex or computationally expensive than various crude polynomial approximations have been in the past.

Chapter 2

Properties of splines

The space of all possible interpolating splines is infinite, and there is no consensus in the spline literature of which is “best.” However, a number of properties have emerged as good criteria for evaluating interpolating splines. Foremost among these is *fairness*, which is a technical term for smoothness. There is no agreed-upon formal definition for fairness, but it is associated with smoothly varying curvature.

Although fairness is important, other properties matter too. For example, some splines (including those modeling the mechanical spline) do not exist for arbitrary sequences of points. This chapter explores fairness deeply, and lists other desirable properties of splines. In subsequent chapters, the evaluation of splines will be based on the properties described below.

2.1 Fairness

In almost all applications, the most important property of a spline is *fairness*, also known as smoothness. Fairness is ultimately a property of the way humans perceive curves, not an abstract mathematical property. As in Justice Potter Stewart’s definition of pornography, “I know it when I see it.”¹

Even so, there are a number of metrics that correlate reasonably well with perception of fairness. Knowledge of the human visual system and empirical evidence show that fairness is intimately tied to *curvature*, which of course is an objective mathematical property. Curves with regions of extremely high curvature are clearly not fair. In addition, while more subtle, discontinuities in curvature and regions with large curvature variation are also perceived as lack of fairness.

Curvature is central in many other application domains as well. There is compelling evidence that the human visual system perceives curves in terms of curvature features, motivating a substantial body of literature on *shape completion*, or inferring missing segments of curves when shapes

¹Justice Potter Stewart, concurring opinion in *Jacobellis v. Ohio* 378 U.S. 184 (1964), regarding possible obscenity in *The Lovers*.

are occluded in the visual field. A recent example, advocating the Euler spiral spline, is Kimia et al. [44].

2.1.1 Empirical testing of fairness

Bending energy has been proposed as a fairness metric, and minimizing that metric is the definition of the MEC spline (which is discussed in considerable detail throughout this thesis). However, does it accurately measure human perception of smoothness?

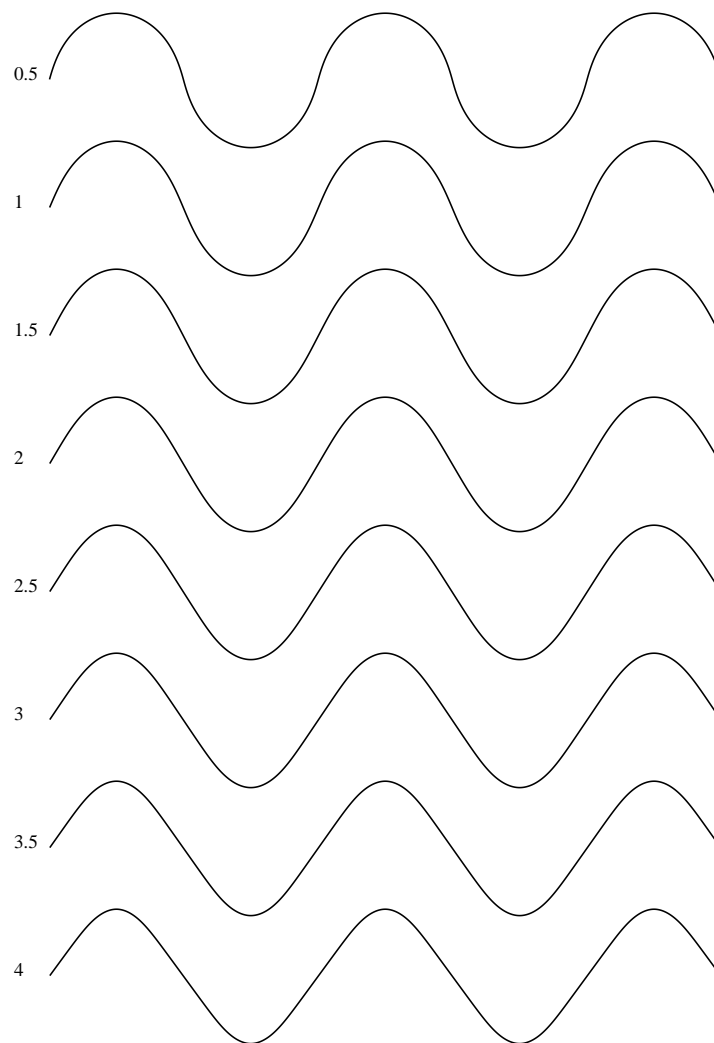


Figure 2.1. Test instrument for fairness user study.

To address this question, I did a very simple user study, asking respondents to choose the fairest curve from a range of possibilities, shown in Figure 2.1.² The individual curves are all generated with the formula $\kappa = cs^\alpha$, with α the varying parameter and c chosen to make the curvature

²The discussion thread is on the web at <http://typophile.com/node/52821>

continuous at the endpoints. These are part of a family of so-called “log-aesthetic curves” [63], discussed in more detail in Section 4.8.1.

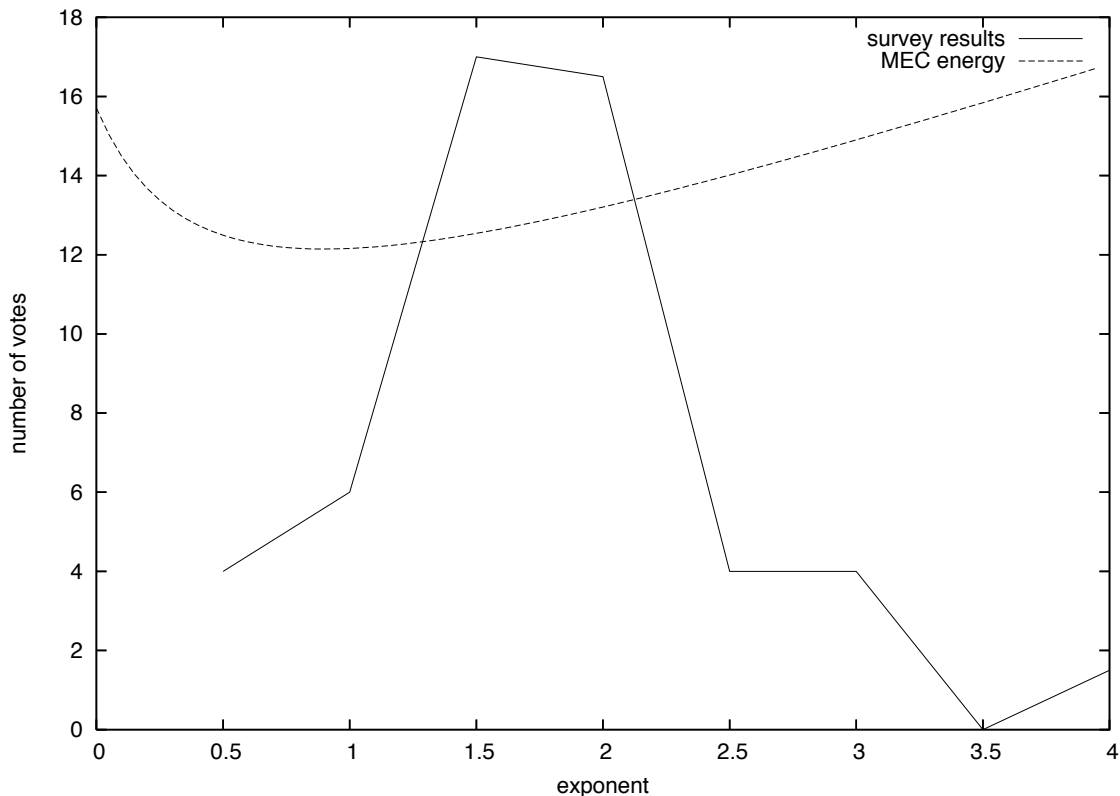


Figure 2.2. Survey results for aesthetic curve preference.

Among the curves in this family, minimal bending energy is obtained with an exponent α near 0.9; thus the Euler spiral, with the exponent $\alpha = 1$, also yields very low bending energy. The on-line study then asked respondents to select the “fairest” curve from the choices shown in Figure 2.1. The results are plotted in Figure 2.2, along with the MEC bending energy. The mean preferred exponent is 1.78; this is a curve for which the bending energy is 6% higher than the Euler spiral. These results confirm the belief that the concept of fairness is rather fuzzy, and that it varies significantly among different designers and graphic artists. Nonetheless, the preference for an exponent in the range 1.5 to 2 is marked. In particular, this study establishes that MEC energy does not in fact accurately predict the perception of fairness.

2.2 Continuity

A concrete property related to fairness is the *degree of continuity*, essentially the number of higher derivatives that exist. A parametrized polygon, for example, has a first derivative everywhere, but that first derivative is not continuous. Thus, a polygon has only G^0 -continuity (the property of being a connected curve). All splines worthy of the name have at least G^1 -continuity, meaning that

the second derivative exists everywhere, but relatively fewer have G^2 -continuity, corresponding to continuous curvature. It is worth remembering, though, that wild but continuous swings in curvature are no fairer in practice than tiny jumps at the control points.

Intuitively, curvature is the position of the steering wheel when driving a car along the curve. Therefore, G^2 continuity is equivalent to the lack of jerks of the steering wheel.

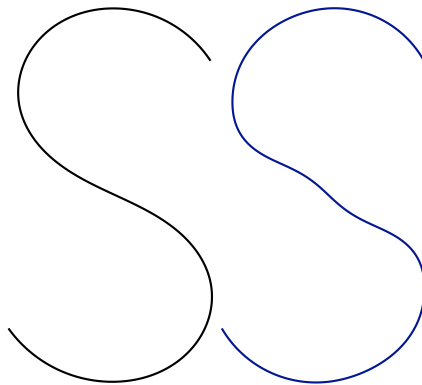


Figure 2.3. Continuity \neq fairness.

Figure 2.3 illustrates clearly that continuity is not the same as fairness. The spline on the left is an Euler spiral spline, with G^2 continuity, and the spline on the right is a circle spline, with G^3 . It seems clear that the one on the left, the G^2 spline, is fairer, because it exhibits less variation in curvature.

2.3 Roundness

One specific aspect of fairness is *roundness*. A round spline always yields a circular arc when the control points are co-circular. It seems intuitively obvious that a circular arc is the fairest (or smoothest) curve.

It makes sense for circular arcs to be the result of splines, as the circle is often found in nature and is also essential in mechanics. However, many splines merely produce an approximation, some better than others. In particular, any spline based on polynomials cannot be exactly round.

Roundness is not an all-or-nothing property. Even the splines that do not generate perfect circular arcs produce curves that are fairly close. Exactly *how* close can be quantified, as we will see in the discussions of individual splines.

Splines other than polynomials can also fail the roundness property. For example, the curve minimizing the bending energy of a thin strip (the Minimum Energy Curve) fails to describe a perfect circular arc. The deviation from roundness, interpolating three control points placed on an equilateral triangle, is shown in Figure 2.4, and will be discussed in more detail in Section 4.2. In this figure, the circle is on the inside, and the MEC describes a curve with somewhat larger total arc length.

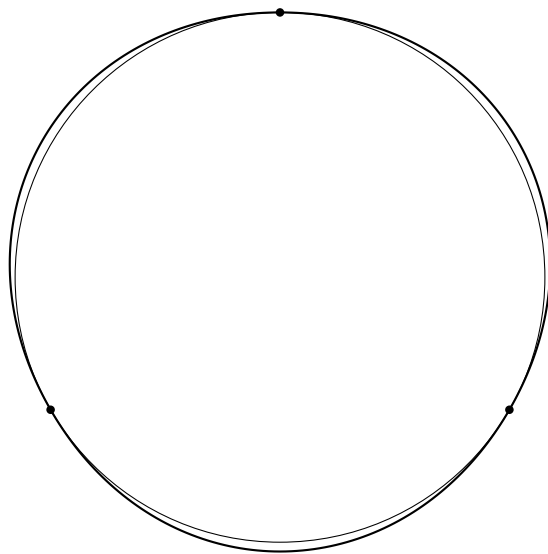


Figure 2.4. MEC roundness failure.

2.4 Extensionality

An *extensional* interpolating spline is one in which adding a new point to the control polygon, lying on the curve generated from the original control polygon, does not change the resulting curve. Unlike fairness, extensionality is not so much a property of the curve segments used to generate the spline as of the effect of changes to the control polygon (adding and deleting points) on the resulting curve.

Extensionality is closely related to the concept of choosing an optimal curve with respect to some fairness metric. If adding an on-curve point changes the curve, either the new or the old curve must not have been optimal. Therefore, splines defined in terms of minimizing a functional tend to be extensional. Extensionality is a powerful filter for weeding out crude approximations, but it does admit a variety of interesting splines in addition to those defined in terms of optimization.

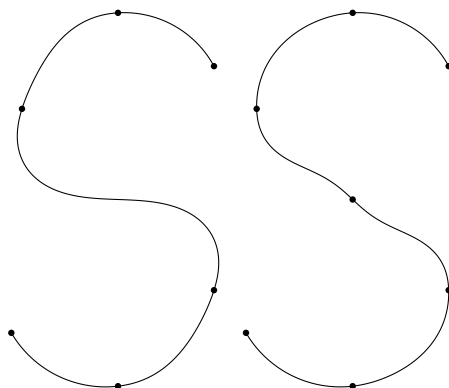


Figure 2.5. Circle spline extensionality failure.

Figure 2.5 shows a spline with particularly poor extensionality, the circle spline (described in more detail in Section 4.10). Adding the point in the center changes the tangent radically at that point, introducing new inflection points, and clearly degrading the overall quality. In this case, the problem arises because the tangents are derived from a local window; when three consecutive control points are co-linear, that line determines the tangent of the middle one even when it is not ideal given the global shape.

Roundness is a very weak extensionality property; for the special case of control points on a circle, adding an additional point on the circle does not change the resulting curve. However, even though these two properties are related, there are round splines that are not extensional (Ikarus [42]), and extensional splines that are not round (MEC).

2.5 Existence and uniqueness

One highly desirable property of a spline is that all control polygons yield interpolating curves. Especially in interactive applications, where control polygons may well pass through ill-defined or degenerate states on the way, it is inconvenient at best for there to be no solution.

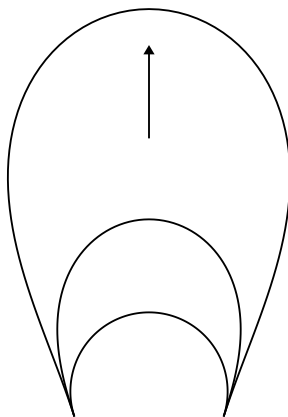


Figure 2.6. Nonexistence of solution to MEC.

Figure 2.6 shows a configuration in which the MEC spline has no viable solution. In this configuration, the tangent angles at the endpoints are constrained to be at greater than a right angle with respect to the chord, but the curve is otherwise unconstrained. In particular, the length of the curve is allowed to assume the value that minimizes overall bending energy. Intuition suggests that, for any given functional to optimize, *some* curve is better than none at all, so the lack of a solution is surprising.

Better intuition comes from considering the MEC as a model of a mechanical spline. Bending energy tends to decrease as the curve gets longer. In particular, the longer curves in Figure 2.6 have a lower bending energy than the shorter ones. Thus, the mechanical springiness applies a force to pull in more length. In a real spline, at some point this force becomes so small that it is easily

defeated by the friction at the control point, but in the mathematical idealization, it would continue indefinitely until the length became infinite and the bending energy zero.

Ironically, even though intuition would suggest that *some* curve is better than no curve at all, the splines with the worst existence properties are those defined in terms of minimizing a fairness functional. Apparently, fairness is not the *only* measure of quality. In general, splines defined in terms of meeting continuity constraints at the control points tend to have better existence properties. For example, an interpolating Euler spiral spline always exists, as discussed in Section 4.6.4.

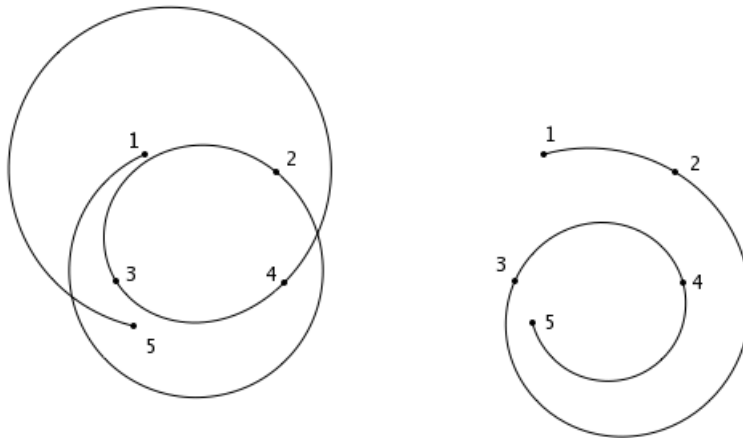


Figure 2.7. Uniqueness failure in Euler spiral spline.

Another desirable property is *uniqueness*, the principle that only one curve exists that satisfies the definition of the spline, for a fixed input. For splines that are defined by optimization, the concern is that there may be multiple local minima, where a local minimum is a solution that is locally optimum under smooth deformation of the curve while it still interpolates the control points. Correspondingly, for splines that are defined in terms of continuity constraints on parametrized curve segments, there may be multiple assignments of parameters that satisfy the constraints and still interpolate the points. Figure 2.7 shows an example, based on the Euler spiral spline. Both curves interpolate the same five input points, in the same order, both are composed of Euler spiral segments between each pair of adjacent control points, and both satisfy the G^2 -continuity constraints.

It is possible to fix the uniqueness problems of a spline by adding an additional (global) criterion, to prefer one solution to the other. Then, the revised spline can be defined as one that both satisfies the local constraints and also optimizes the global criterion. It may be difficult to properly implement a spline revised in this way, as search for a global optimum is often expensive, but at least such a spline is well defined.

As will be discussed in Section 4.7, Hobby used the criteria of existence and uniqueness to motivate the design of his spline. Using “mock curvature” guarantees existence of a unique solution, even though it has less desirable fairness properties.

2.6 Locality

When a control point moves, how much of the resulting curve changes? A small section in the neighborhood of the control point, perhaps, or does the entire spline change? This is the property of *locality*.

Some splines have *finite support*, a particularly strong form of locality. In these splines, moving a control point changes the curve only for a section bracketed by a finite number of control points on either side of the one moving. For example, in circle-splines this section consists of four segments bracketed by stepping out two control points in either direction.

However, finite support conflicts directly with extensionality, so in general we will measure locality in terms of how quickly the effect of moving a control point dies out as we move away from it. In splines with good locality, this effect falls off exponentially in the distance from the control point (measured by counting control points). Then, locality can be quantified by the ratio of effect from one control point to the next. Section 4.8.1 shows, this ratio is $2 + \sqrt{3} \approx 3.73$ for the Euler spiral spline (as well as the MEC).

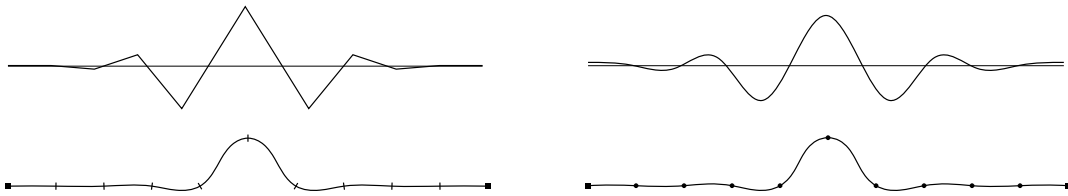


Figure 2.8. G^2 spline has better locality than G^4 .

Figure 2.8 shows two splines with different locality. The one on the left is an Euler spiral spline with G^2 -continuity, and the one on the right is its G^4 counterpart. In general, there is a tradeoff between the degree of continuity and the locality, so the latter spline has more wiggling farther from the perturbed point.

2.7 Transformational invariance

Another basic property is invariance under various transformations and symmetries. All functions worthy of being considered splines are invariant under rigid body transformations (rotation and translation) as well as uniform scaling. If the control polygon undergoes a rigid body transformation, then the resulting curve undergoes exactly the same transformation.

Further, all of the splines we consider in this thesis are invariant under mirror symmetry and front-to-back reversal of the order of the points in the control polygon. These are very common properties, but it is worth noting that quaternion splines computed on the surface of the sphere lack front-to-back symmetry.

A sometimes desirable property of splines is *affine invariance*, meaning that the shape of the spline is preserved through affine transformation of the control points. This property is observed

in many polynomial-based splines, but it conflicts with the roundness property. Consider that any round spline fits a circular arc to an input consisting of three control points. Yet, affine transformation of the resulting curve yields an ellipse, which in general would not match the spline resulting from the three transformed control points – which would also need to be a circular arc due to the roundness property. For the types of designs considered in this thesis, roundness is more important than affine invariance, so the latter will not be considered in detail.

2.8 Counting parameters

In an arbitrary spline, the family of curve segments between any two control points is potentially drawn from an infinite-dimensional space. In practice, most splines use curve segments chosen from a finite-dimensional manifold. Generally, there is a vector of real parameters that uniquely determines the shape of the curve between any two endpoints. In counting these parameters, we hold the endpoints fixed, and apply the rotation, scaling, and translation to make the curve coincide with these endpoints.

Thus, the degenerate spline consisting of straight lines is effectively a zero-parameter spline. A fair number of interesting splines fall into the category of two-parameter splines, and these are discussed further in Chapter 4. Well constructed splines in this class can achieve G^2 continuity.

2.9 Monotone curvature

The human visual system is particularly sensitive to minima and maxima of curvature. This property suggests that all curvature extrema should coincide with the given control points, and that the spline segments in between should exhibit monotone curvature. A fair amount of literature is concerned with constraining existing curves (such as cubic polynomials) to have monotone curvature. One such citation is “Measures of Fairness for Curves and Surfaces” [75, p. 93], which also discusses fairness metrics in considerable detail. The Euler spiral, in which curvature varies linearly with arc length, obviously has this property, as do other related curves. However, the MEC does not have this property; for symmetrical end conditions it has a tendency to increase arc length and to add a curvature maximum in the middle of the segment. The MVC also in general does not place curvature extrema to coincide with the control points.

Evidence that the visual system is sensitive to minima and maxima of curvature goes all the way back to Attneave’s seminal experiments in the mid 1950’s [5]. Attneave presented test subjects with somewhat random, blobby drawings and asked them to annotate which points along the outline were the most salient. An example is shown in Figure 2.9. The “porcupine quills” emanating from the shape represent the relative frequency that subjects identified the points. These strongly tend to coincide with curvature extrema. As such, in interactive curve design it is reasonable to expect designers to place these points first, then refine the curves with more control points only as needed. A good spline should accommodate such a designer by preserving the curvature extrema as marked.

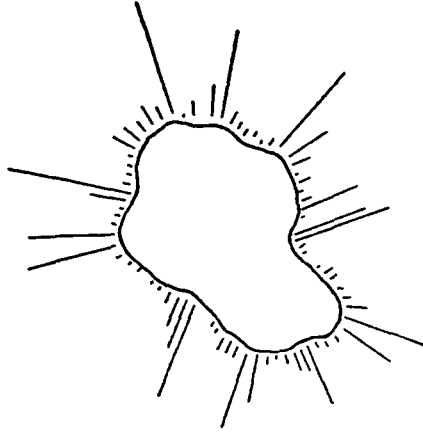


Figure 2.9. Attneave's curvature experiment [5].

2.10 Summary

The most important property for interpolating splines is fairness. Yet, other properties such as robustness and locality are also important.

Several previous authors have proposed a set of properties desirable in splines. Knuth's enumeration of such properties, for example, comprised invariance, symmetry, extensionality, locality, smoothness, and roundness [47, pp. 39–42].

Similarly, Henry Moreton in his Ph.D. thesis [64, ch. 2] gives a rather exhaustive list of properties, including most of the ones described above.

Chapter 3

The Elastica

The word *spline* originally refers to a thin strip of wood used to smoothly interpolate control points. A full understanding of splines requires an examination of the theory of such elastic strips. This theory, that of the *elastica*, has a long and rich history, much of which will be explored in the next chapter, but the literature is missing a concise, accessible treatment of the topic. This chapter fills that gap.

One difficulty of the literature is that there are many different ways to approach the elastica: as an equilibrium of moments in a static physical system; as an equilibrium of forces in the same physical system (analyzed in terms of a finite element approximation composed of rigid struts and pivot springs); as the curve that minimizes the total bending energy of that system; as the solution to an abstract differential equation; and through the analogue with another physical system, the pendulum, which happens to share the same differential equation.

Each approach sheds light on a different facet of the elastica. For instance, treating it as a statics problem is the easiest way to understand the effect of different constraints operating on the curve, and of course the physics interpretation is the only way to think about the differences between a real elastic lamina and the mathematical ideal – especially, the effect of friction at the constraint points. The variational formulation is most revealing when examining the minimization of some other metric than total bending energy. And the pendulum analogy is helpful for developing an intuitive understanding of the underlying differential equation, particularly the fact that it is periodic.

My approach is to pull together *all* of these approaches, letting each tell its own story. For a given problem involving the elastica, any one of these approaches may offer the simplest, most intuitive solution. Yet, the underlying mathematics is the same. In the equilibrium of moments formulation, the equation takes on a remarkably simple form: $\kappa = \lambda y$, where κ is the curvature, y is the Cartesian ordinate, with x the abscissa, and λ is a constant.

3.1 Statement of the elastica problem

The elastica models a thin strip of flexible material (traditionally wood, but spring metal will do). In the spirit of idealizing the problem, “thin” means that the thickness is negligible compared with the amount of bending. The mathematical idealization further assumes that the centerline of this strip lies entirely on a plane, and that there is no twisting, stretching, or compression.

Thus, the curve traced by the strip is characterized by its curvature at each point along its length. A convenient definition of curvature is the derivative of tangent angle θ with respect to arc length s . Unless otherwise stated, in this chapter the θ' notation will represent $d\theta/ds$, so curvature κ is written simply θ' .

Bending such a strip induces potential energy, much like that of a spring. According to the theory of elasticity, the bending energy is proportional to the square of the curvature. Thus, the total bending energy of a strip of length l is

$$E[\kappa(s)] = \int_0^l \kappa(s)^2 ds . \quad (3.1)$$

When completely unconstrained, the elastica will assume the shape of a straight line, in which the curvature is everywhere zero, and thus the total bending energy is also zero. When constrained, the bending energy will tend to the minimum possible under the constraints.

3.2 Variational study

The elastica equation 3.1 is particularly well suited to study by the calculus of variations [19], a technique for transforming problems specified in terms of infinite-dimensional functions into tractable ordinary differential equations. The calculus of variations was founded at the end of the 17th century, and was developed by Johann and Jacob Bernoulli, Leonhard Euler, and others.

In particular, Equation 3.1 yields readily to the Euler–Lagrange equation.

3.2.1 The Euler–Lagrange equation

If there exists a minimum value of the functional

$$v[y(x)] = \int_{x_0}^{x_1} F(x, y(x), y'(x)) dx , \quad (3.2)$$

subject to the n constraints, each of the form $1 \leq i \leq n$,

$$\int_{x_0}^{x_1} c_i(y(x)) dx = k_i , \quad (3.3)$$

then this minimum is characterized by a tuple of values λ_i such that

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} + \sum_i \lambda_i \frac{dc_i}{dy} = 0. \quad (3.4)$$

Therefore, the problem of finding a *function* minimizing the cost functional v in an infinite-dimensional space of functions is reduced to that of finding a tuple of values minimizing that functional. In general, such an approach yields easily to numerical techniques for both integrating the differential equation 3.4 and finding a set of values λ_i minimizing the cost functional.

3.2.2 Application of the Euler–Lagrange equation to the elastica

There are several ways to apply the Euler–Lagrange equation to find curves satisfying the elastica problem as defined by Equation 3.1. The main choice is that of independent and dependent variables (x and y in the formulation of Equation 3.2). At least three such pairs appear in the literature of the elastica: arc length and tangent angle, arc length and signed curvature, and Cartesian coordinates.

Of these, the choice of arc length s as the independent variable and tangent angle θ as the dependent yields the simplest variational solution. The most general endpoint constraints are specification of curve length, endpoint positions, and endpoint tangents. Of these, the endpoint position constraint requires Lagrange multipliers in the form of Equation 3.3. Assume without loss of generality that the starting point is $(0, 0)$ and the endpoint is (x, y) (simple translation will allow arbitrary starting point). Then, the endpoint constraint is expressed as the integral of the unit speed vector of direction $\theta(s)$:

$$\begin{aligned} \int_0^l \cos \theta(s) ds &= x, \\ \int_0^l \sin \theta(s) ds &= y. \end{aligned} \quad (3.5)$$

In terms of arc length and tangent, these constraints are easily expressible in the form of equation 3.3.

It's now easy to see why representing the curve in terms of tangent angle as a function of arc length yields most readily to the variational approach. In terms of curvature κ , the endpoint constraint is a double integral, which does not fit neatly into a formulation of constraints as a single integral of some function. Alternatively, Cartesian coordinates would make the endpoint constraints even simpler, but the variational function becomes more complex, requiring second derivatives as well as the first derivatives of the Euler–Lagrange equation. The first solution was indeed done using Cartesian coordinates, but this approach is definitely simpler.

Applying the variational technique to minimizing the bending energy of a strip, Equation 3.1, we arrive at the solution,

$$\theta'' + \lambda_1 \sin \theta + \lambda_2 \cos \theta = 0. \quad (3.6)$$

Equation 3.6 is the simplest form of the solution, but other forms can be readily derived, and often appear in the literature. Taking the derivative with respect to s ,

$$\theta''' + \lambda_1 \theta' \cos \theta - \lambda_2 \theta' \sin \theta = 0. \quad (3.7)$$

Conversely, multiplying both sides of 3.6 by $d\theta/ds$ and then taking the integral,

$$\frac{1}{2}(\theta')^2 - \lambda_1 \cos \theta + \lambda_2 \sin \theta + C = 0. \quad (3.8)$$

Combining 3.7 and 3.8, we can eliminate the Lagrange multipliers, retaining the single constant of integration C .

$$\theta''' + \frac{1}{2}(\theta')^3 + C\theta' = 0. \quad (3.9)$$

Since $\kappa = \theta'$, we can reformulate this differential equation entirely in terms of curvature:

$$\kappa'' + \frac{1}{2}\kappa^3 + C\kappa = 0. \quad (3.10)$$

Another formulation in the literature (for example, Eq. 12 of Lee and Forsythe [51]) multiplies this last equation by $1/\kappa$, yielding

$$\frac{1}{\kappa}\kappa'' + \frac{1}{2}\kappa^2 + C = 0. \quad (3.11)$$

3.3 A family of solutions

The equations derived in the previous section for the shape of the elastica don't have a single solution. Rather, they describe a family of solutions, characterized by a single scalar parameter. All such solutions are realizable by a physical elastica; none are purely an artifact of the mathematical approach.

For mathematical convenience, assign $\kappa = 1$ at the midpoint of the curve, and use λ_1 as the parameter in Equation 3.6, setting $\lambda_2 = 0$. If we write $\lambda_1 = \lambda \cos \phi$ and $\lambda_2 = \lambda \sin \phi$, then all other instances of the elastica can be transformed to meet these conditions by scaling and rotating the entire system so that $\phi = 0$. Since the assumption of $\phi = 0$ obviates the need to distinguish multiple λ parameters, it is convenient to drop the subscript ₁ and simply write λ , resulting in a simplified equation.

$$\theta'' + \lambda \sin \theta = 0. \quad (3.12)$$

Figure 3.1 shows a representative selection from this family, for various values of λ .

The remainder of this chapter is devoted to exploring these solutions and their physical meaning, but a few highlights are immediately apparent. All solutions have a mirror (even) symmetry at each curvature maximum. All solutions save $\lambda = 0.25$ are periodic. The solutions for values larger than 0.25 have inflections, and have odd symmetry around each inflection point. For values smaller than 0.25, there is no inflection; the curvature always remains the same sign. Note that the curve has a mirror symmetry for all values of λ .

As λ approaches 0, the elastica tends to a circle. As λ approaches ∞ , the elastica tends to a sine wave. The solution at $\lambda = 0.5$ is special, and is known as the *rectangular elastica*. The tangents

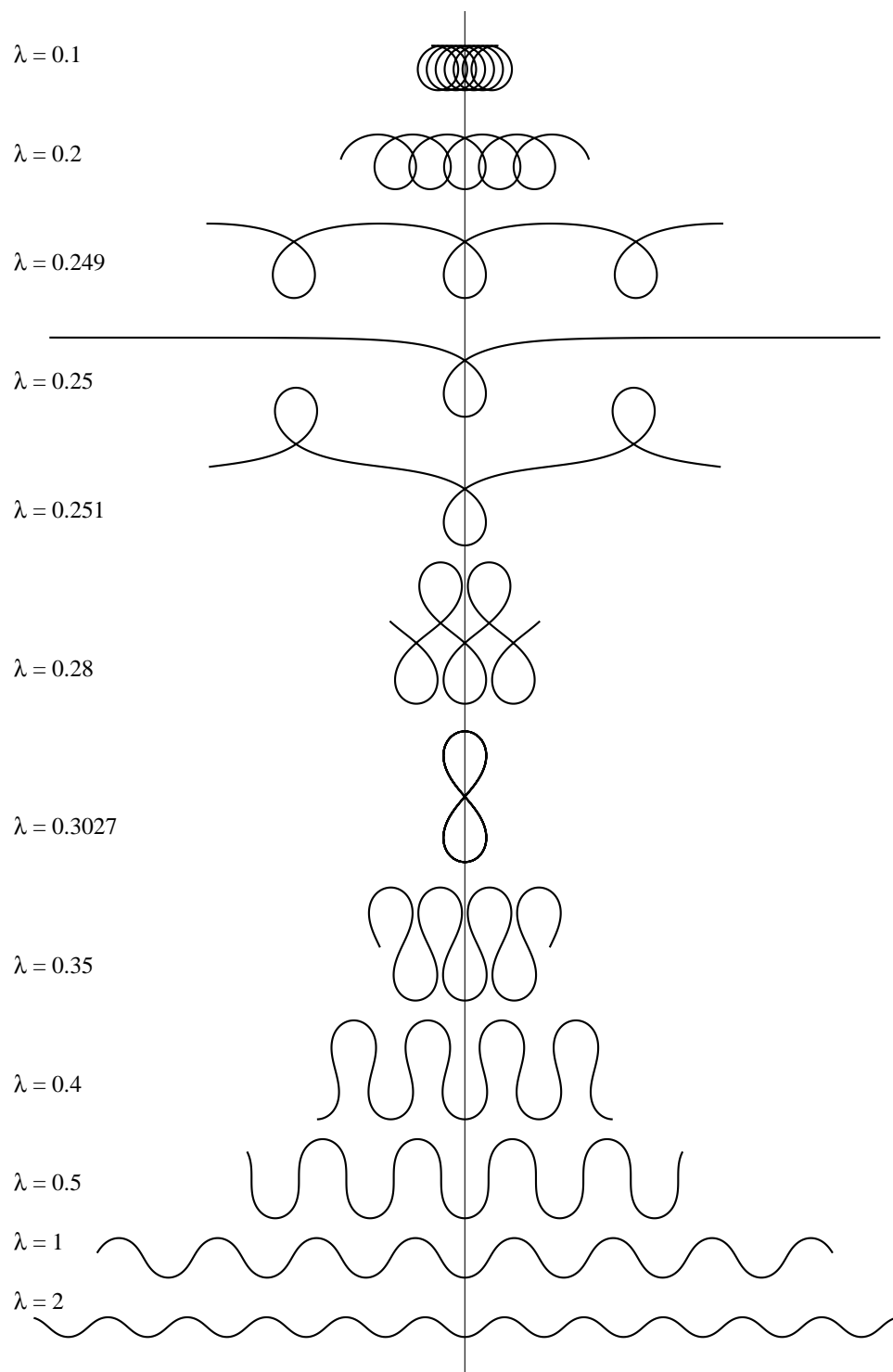


Figure 3.1. The family of elastica solutions.

at the inflection points of this curve are perpendicular to the horizontal axis; the curve is tangent to the rectangle formed by inflections and curvature extrema.

Figure 3.2 shows the curvature plots of the instances of the elastica family shown in Figure 3.1, plotted with curvature on the vertical axis and arc length on the horizontal. As is usual, inflection points on the curve correspond to the curvature crossing (or touching) the horizontal axis.

3.4 Equilibrium of moments

Analyzing the elastica using the theory of elasticity yields three physical quantities which must be in equilibrium along its length: tension (longitudinal force), shearing force, and bending moment. The tension and shearing forces are useful when working out the effect of constraints, but it is possible to derive the equation for the shape of the elastica from moments alone. Indeed, computing the moments is by far the easiest derivation of the elastica equation.

This section uses elementary principles of statics to carry out the derivation of the elastica from equilibrium of moments. A classic textbook introducing these concepts is *Statics*, by Goodman and Warner [28].

Consider the elastica in Figure 3.3. Two forces of magnitude \mathbf{F} and opposite directions push the ends of the elastica together. Because these forces are equal in magnitude, opposite in direction, and have parallel lines of action, they form a *couple* [28, p. 65] (the line of action is defined as a vector in the same direction as the force, emerging from the point where the force is applied).

At any point along the elastica, then, the *resultant moment* is simply the force of the couple times the distance from the line of action: $M = Fy$, where F is the magnitude of the force vector \mathbf{F} .

The simple theory of elasticity says that the relation between bending moment and curvature is linear. Thus, collecting all the constants together into a single λ yields this remarkably simple formulation of the shape of a general elastica,

$$\kappa = \lambda y . \tag{3.13}$$

This equation is readily shown to be equivalent to the variational solution above. Take the derivative with respect to arc length s , and note that $y' = \sin \theta$. Then,

$$\kappa' = \lambda \sin \theta , \tag{3.14}$$

and this equation is equivalent to Equation 3.12, assuming without loss of generality that the line of action is parallel to the horizontal axis.

All elastica solutions are equivalent to the system shown in Figure 3.3, but in non-inflectional cases the line of action of the couple is virtual, in that it doesn't touch the curve.

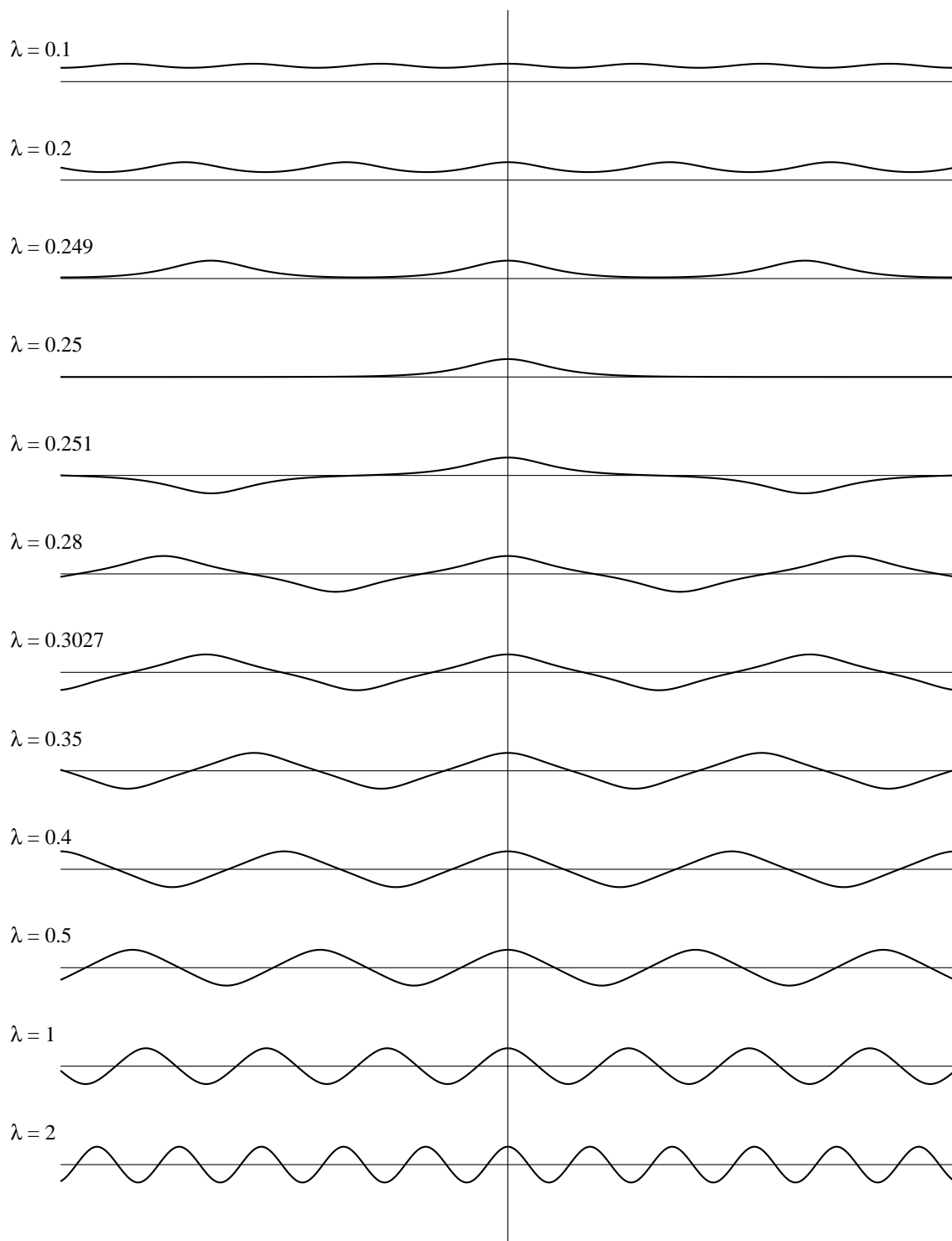


Figure 3.2. Curvature plots for the family of elastica solutions.

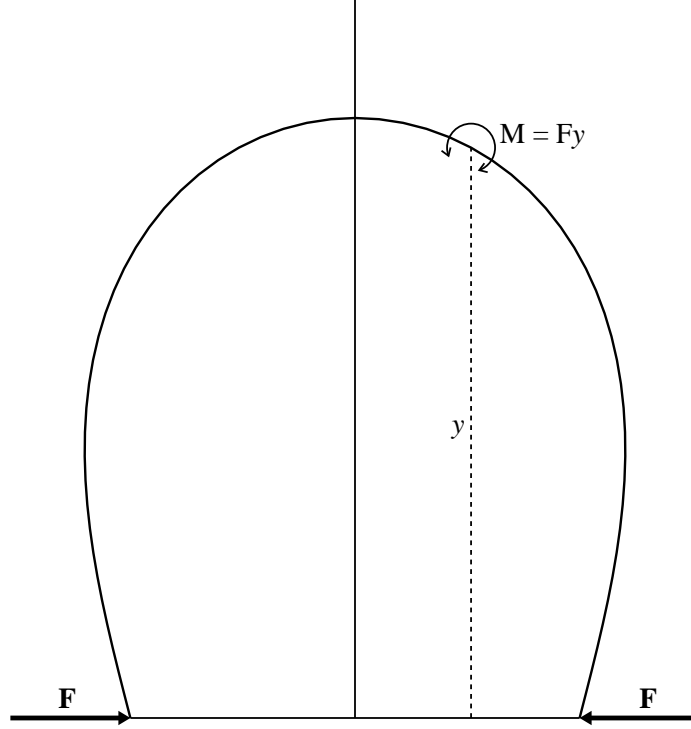


Figure 3.3. Equilibrium of moments.

3.5 Pendulum analogy

The differential equation 3.12 for the shape of the elastica is mathematically equivalent to that of the dynamics of a simple swinging pendulum. This kinetic analog is useful for developing intuition about the solutions of the elastica equation. In particular, it's easy to see that all solutions are periodic, and that the family of solutions is characterized by a single parameter (modulo scaling, translation, and rotation of the system).

Consider, as shown in Figure 3.4, a weight of mass m at the end of a pendulum of length r , with the angle from vertical specified as a function of time $\theta(t)$.

Then the velocity of the mass is $r\theta'$ (where, in this section, the $'$ notation represents derivative with respect to time), and its acceleration is $r\theta''$. The net force of gravity, acting with the constraint of the pendulum's angle is $mg \sin \theta$, and thus we have

$$F_{\text{net}} = ma = mr\theta'' = mg \sin \theta . \quad (3.15)$$

With the substitution $\lambda_1 = -g/r$, and the replacement of arc length s for time t , this equation becomes equivalent to Equation 3.12, the equation of the elastica. Note that angle (as a function of arc length) in the elastica corresponds to angle (as a function of time) in the pendulum, and the elastica's curvature corresponds to the pendulum's angular momentum.

The swinging of a pendulum is perhaps the best-known example of a periodic system. Further, it is intuitively easy to grasp the parameter space of the system. Transformations of scaling (varying

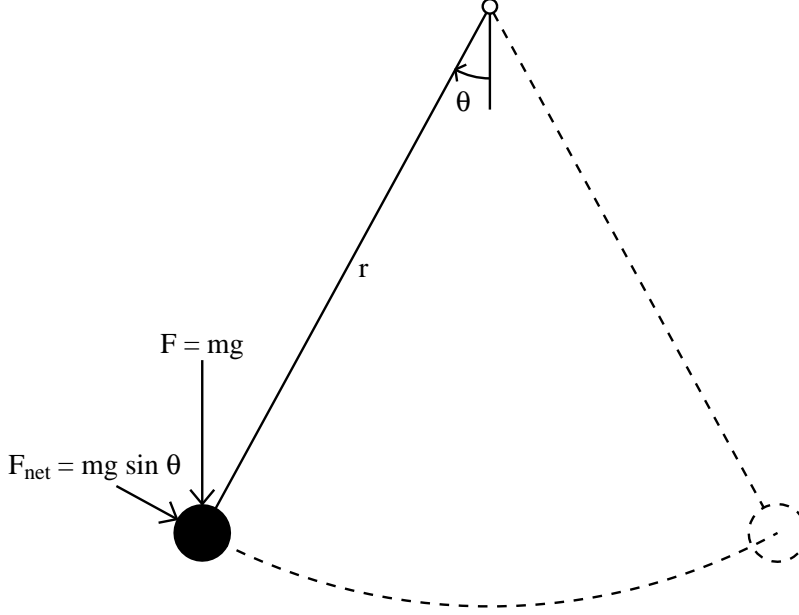


Figure 3.4. An oscillating pendulum.

the pendulum length) and translation (assigning different phases of the pendulum swing to $t = 0$) leave the solutions essentially unchanged. Only one parameter, how high the pendulum swings, changes the fundamental nature of the solution.

The solutions of the motion of the pendulum, as do the solutions of the elastica equation 3.12, form a family characterized by a single scalar parameter, once the trivial transforms of rotation and scaling are factored out. Without loss of generality, let $t = 0$ at the bottom of the swing (i.e. maximum velocity) and let the pendulum have unit length. The remaining parameter is, then, the ratio of the total energy of the system (which remains unchanged over time) to the potential energy of the pendulum at the top of the swing (the maximum possible), in both cases counting the potential energy at the bottom of the swing as zero. Thus, the total system energy is also equal to the kinetic energy at the bottom of the swing.

For mathematical convenience, define the parameter λ as one fourth the top-of-swing potential energy divided by the total energy. This convention is equivalent to varying the force of gravity while keeping the maximum kinetic energy fixed, which may be justified by noting that the zero-gravity case (which would require infinite kinetic energy if we were assuming unit gravity) is far more relevant in the context of splines (it corresponds, after all, to a circular arc, so any elastica-based spline meeting the criterion of roundness must certainly have this solution) than the opposite of zero kinetic energy.

The potential energy at the top of the pendulum is $2mgr$. The kinetic energy at the bottom of the swing is $\frac{1}{2}mv^2$, where, as above, $v = r\theta'$. Thus, according to the definition above,

$$\lambda = \frac{1}{4} \frac{2mgr}{\frac{1}{2m(r\theta')^2}} = \frac{g}{r(\theta')^2}.$$

In other words, assuming (without loss of generality) unit pendulum length and unit velocity at the bottom of the swing, λ simply represents the force of gravity.

Figure 3.2, the curvature plots for solutions of the elastica equation, thus also shows the velocity of the pendulum as a function of time for various solutions of the motion of the pendulum. For each value of λ , velocity (θ') is plotted against time. The bottom of the swing (maximum velocity) is the center of the plot, and 25 units of time are plotted to either side.

Note that the period has a singularity at $\lambda = 0.25$. For values of λ greater than 0.25, the pendulum swings back and forth periodically, velocity reaching zero at the highest point in the swing. For values less than 0.25, the pendulum winds round and round, slowing at the top of the cycle, but never actually reaching zero. At exactly 0.25, the kinetic energy at the bottom is equal to the potential energy at the top, so the velocity is zero at the top.

In the kinetic analogy with the elastica, velocity of the pendulum corresponds to curvature. Thus, values of λ above $\frac{1}{4}$ have inflection points. Figure 3.5 shows the relationship graphically, including the right angle made by the pendulum at the top of its swing for the $\lambda = 0.5$ rectangular elastica case.

Note that while $\lambda = 0.3027$ is a special value for the elastica (it loops over itself and occupies finite space), there is no special physical meaning in the pendulum analogy.

3.6 Equilibrium of forces: Finite element approach

When an elastic strip is in a minimal energy configuration, it is in mechanical equilibrium, meaning that at all points along its length, the sum of forces on that point is zero.

The standard derivation of the equations of mechanical equilibrium from the theory of elasticity [52] expresses relations between shear forces, tension (longitudinal) forces, and flexural couples. The equations are straightforward, but the relevant physics requires appeal to results in the theory of elasticity. This section, therefore, presents a simplified finite-element model with only bending and tension forces.

The finite element model is a chain consisting of alternating rigid struts and pivots with torsion springs. Such a chain, with spring torque indicated schematically at each pivot, is shown in Figure 3.6.

Each strut generates a tension force at both ends, in the same direction as the strut's length. In Figure 3.7, this tension force is indicated by the quantity T . By convention, compression force is denoted with a negative sign.

At each pivot is a spring generating forces tending to straighten out the chain, i.e. towards a turning angle of zero. In Figure 3.8, the turning angle is denoted by θ and the force generated by the moment (or spring torque) by M . At the endpoints, the force generated by the moment is normal to the strut joining the center to the endpoint. According to Newton's third law, the sum of all forces arising from the element must be zero, so there is a balancing force at the center point equal to the vector sum of the forces at the endpoints, pointing in the opposite direction.

The spring at the pivot point is assumed to be purely elastic, which is to say that the moment (torque) is proportional to the turning angle. Using c for this constant of proportionality (which,

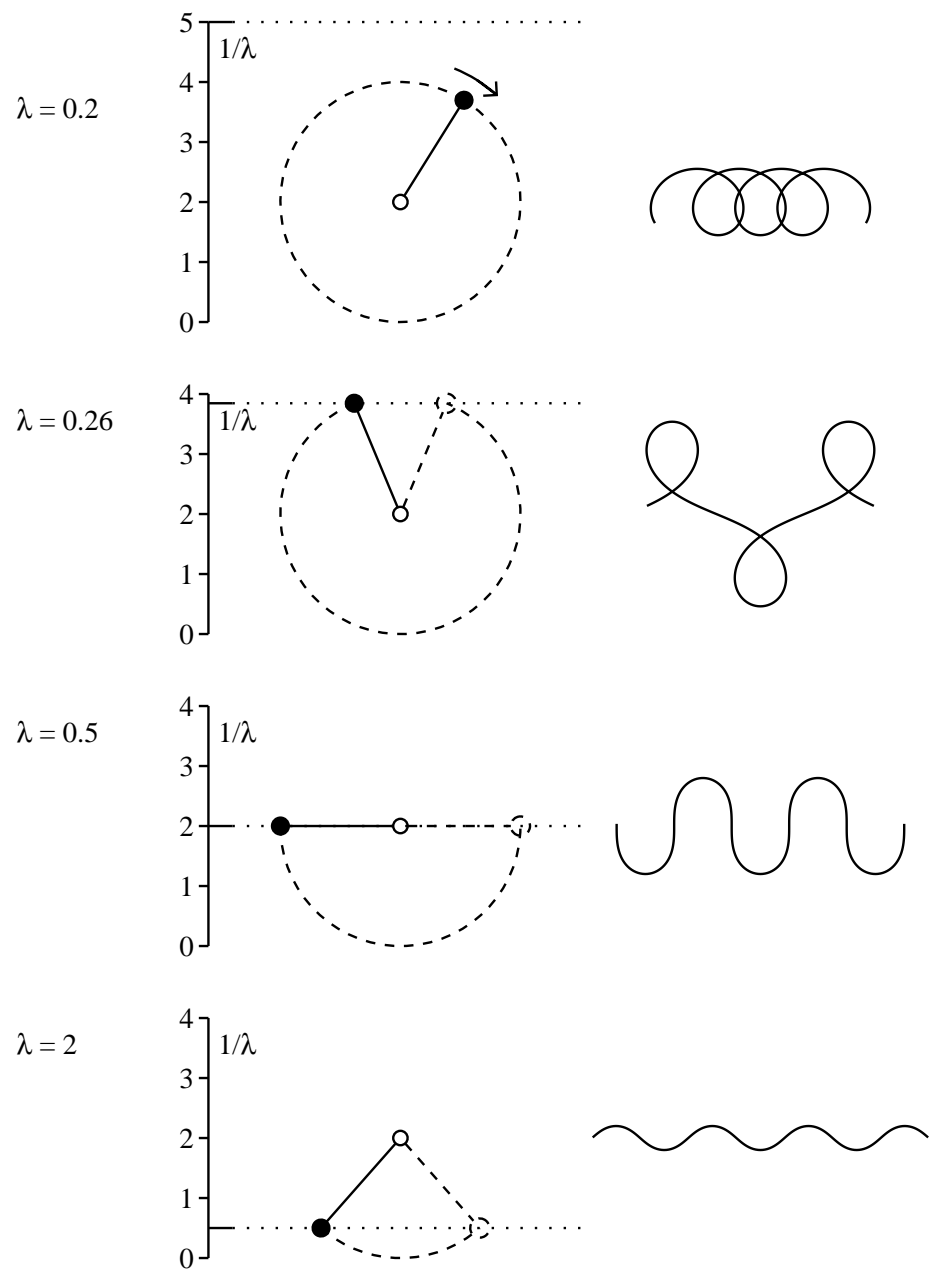


Figure 3.5. Examples of pendulum analogy.

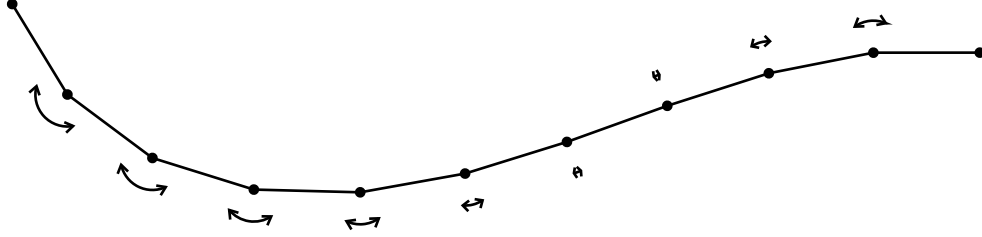


Figure 3.6. Finite element approximation of elastica as chain of struts and pivots.



Figure 3.7. Forces on a single strut.

in a real physical model, would be Young's modulus of elasticity times the second moment of the section of the beam about the axis of bending),

$$M_i = c\Delta\theta_i . \quad (3.16)$$

As shown in Figure 3.9, when these primitive elements are linked together in a chain, there are five force vectors acting on a single pivot point: the tension force from each adjoining strut (T_0 and $-T_1$), the moment forces from the previous and next pivots in the chain (M_0 and M_2), and the balancing force $-2M_1 \cos(\Delta\theta_1/2)$.

In flatland, the equilibrium condition that these vector forces sum to zero induces two scalar equations. When the axes are chosen to be parallel and perpendicular to one of the struts, these equations have a particularly simple form. Arbitrarily choosing the strut labeled 1 in Figure 3.9, the parallel component is

$$-T_1 + \cos \Delta\theta_1 T_0 - \sin \Delta\theta_1 (M_0 - M_1) = 0 , \quad (3.17)$$

and the perpendicular component is

$$\sin \Delta\theta_1 T_0 + \cos \Delta\theta_1 (M_0 - M_1) + M_2 - M_1 = 0 . \quad (3.18)$$

In the continuous limit of this finite element model, the individual turning angles at the pivots

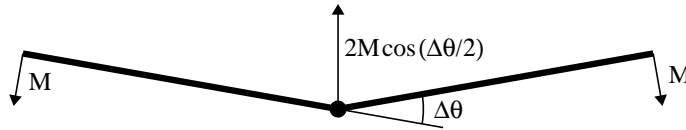


Figure 3.8. Forces on a single pivot.

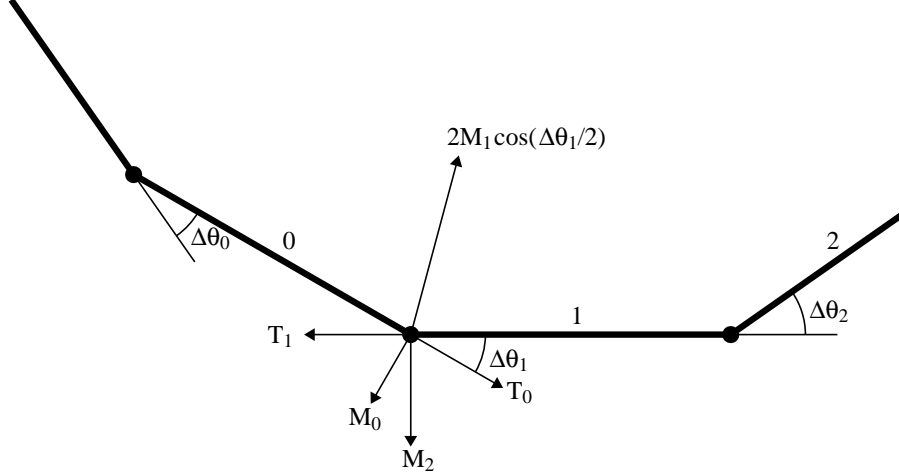


Figure 3.9. Forces on a single pivot point in a chain.

$\Delta\theta_i$ tend to zero, justifying the usual linear approximations $\sin \Delta\theta \approx \Delta\theta$ and $\cos \Delta\theta \approx 1$. The result of using these approximations, and substituting the elastic pivot assumption of Equation 3.16 is

$$-T_1 + T_0 - c\Delta\theta_1(\Delta\theta_0 - \Delta\theta_1) = 0 , \quad (3.19)$$

$$\Delta\theta_1 T_0 + c(\Delta\theta_0 + \Delta\theta_2 - 2\Delta\theta_1) = 0 . \quad (3.20)$$

In the limit as $\Delta\theta$ approaches zero (as the number of links in the chain increases), these difference equations become the differential equations

$$-\frac{dT}{ds} + c\kappa \frac{d\kappa}{ds} = 0 , \quad (3.21)$$

$$\kappa T + c \frac{d^2\kappa}{ds^2} = 0 . \quad (3.22)$$

Equation 3.21 is readily integrated, yielding

$$T = \frac{c}{2}\kappa^2 + C . \quad (3.23)$$

Substituting this into Equation 3.22, and dividing by the common constant c , yields a restatement of Equation 3.10.

$$\kappa'' + \frac{1}{2}\kappa^3 + C\kappa = 0 . \quad (3.24)$$

The fact that the constant c , related to the coefficient of elasticity, factors out has the physical interpretation that the elastica takes the same shape regardless of the value of this coefficient, as long as it is consistent across the length of the elastica.

It is satisfying but not surprising that the variational and the mechanical equilibrium treatments of the elastica problem result in the same equations. The theory of conservative systems states that at a stable equilibrium, the energy is at a local minimum. In addition to confirming the variational derivation, the mechanical equilibrium equations prove a useful physical interpretation for the constant C in terms of curvature and tension (longitudinal force). In particular, if C is zero, then the tension vanishes at inflection points.

3.7 Solutions with length and endpoint constraints

Now that we have the mathematical solution to the elastica equation, including plots for a variety of values of the parameter λ , we turn our attention to particular solutions corresponding to real elastic strips under given constraints. Can real elastic strips take on all values of λ , or is some of the solution space a mathematical artifact?

Without loss of generality, assume that the length of the strip is 1, and that the endpoints lie on the horizontal axis (all other cases can be recovered through scaling, rotation, and translation). Three parameters remain: the tangent angles at the endpoints with respect to the horizontal axis, and the distance between the endpoints (chord length).

Figure 3.10 shows one linear section of this three dimensional parameter space: the endpoint tangents are fixed at $\pi/2$ from horizontal, and the chord length varies from 0 to 1.

Several properties are apparent: λ varies smoothly from about 0.34 to a maximum of 0.5, then falls quickly to 0, where the solution is a precise half-circle, then increases asymptotically to 0.25 as the strip is pulled tighter and tighter towards its full length. For these endpoint angles, the λ maximum at 0.5 also corresponds to zero curvature (inflection points) at the endpoints.

Note that the above discussion only reflects the *principal solution* of the elastica with respect to the constraints, i.e. the one with the minimum total arc length from the original curve. The principal solution always has at most one inflection point. The periodic nature of the elastica implies that additional solutions may also exist, with more than one inflection point. In general, these additional solutions are not interesting for the purpose of generating spline curves, but they are valid solutions. Figure 3.11 shows a number of these solutions for the end constraints $\theta_0 = \pi/4$, $\theta_1 = \pi/4$, $\lambda = 1/2$. It is interesting that some of these solutions lack symmetry even though the constraints themselves are symmetrical.

3.8 Elastica with general constraints

Working towards the application of an elastica to splines, this section considers the effect of general constraints on an elastica. A goal is to develop physical intuition, so the constraints are expressed in terms of forces.

There are four basic types of constraints. Each constraint, can either rotate freely or fix the tangent angle of the elastica. Further, it can either allow the elastica to slide freely, or can fix the relative arc length of the elastica. A segment between two constraints that fix relative arc length has

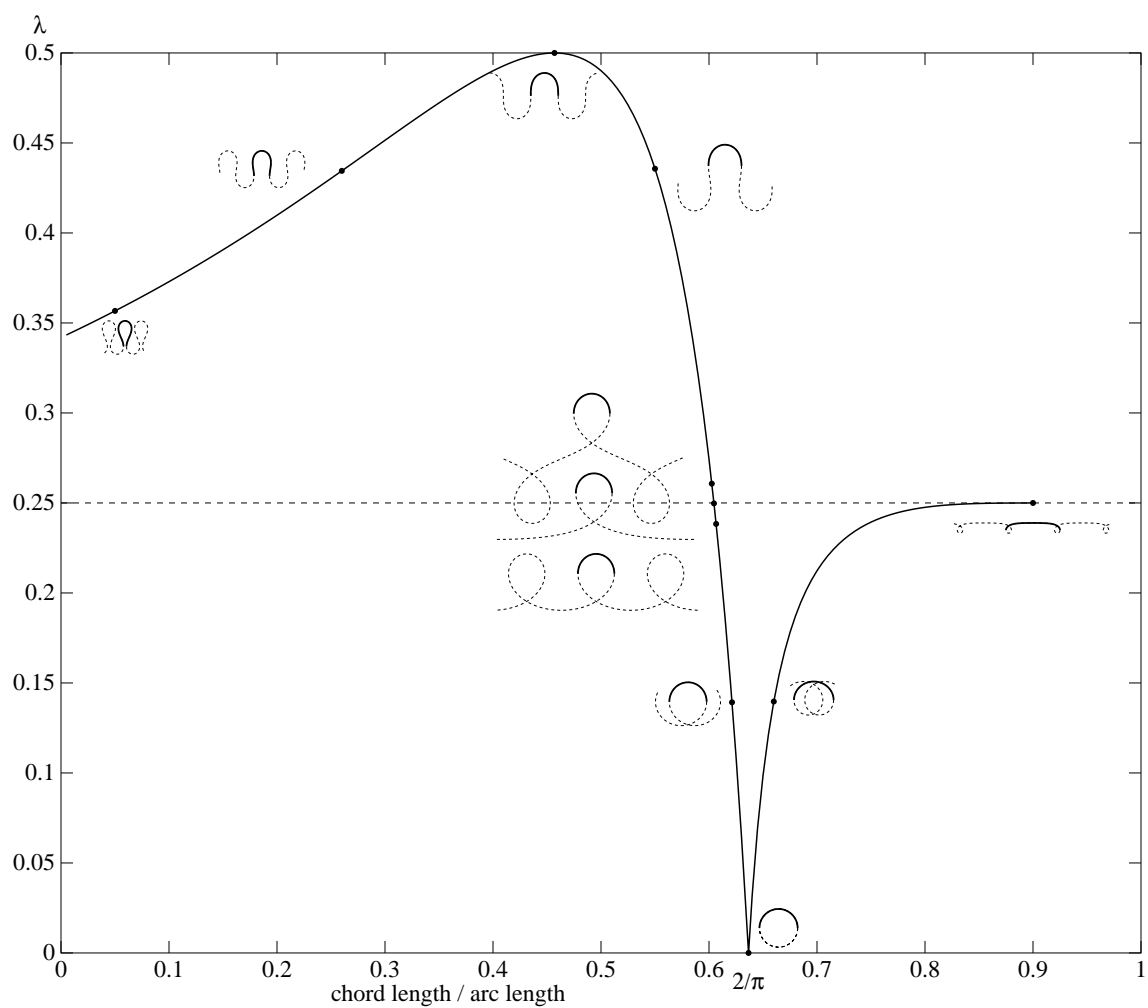


Figure 3.10. Values of λ for varying chord lengths.

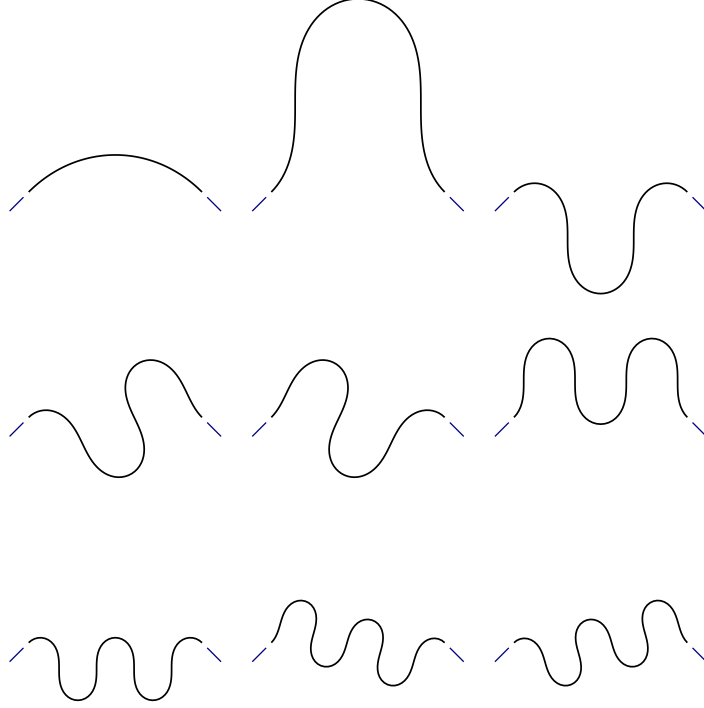


Figure 3.11. Multiplicity of solutions for same tangent constraints.

fixed arc length, even if there are sliding constraints in between. In all these types of constraints, the position of the constraint is fixed, and the elastica is constrained to go through that point.

A position constraint, freely rotating, and with the elastica free to slide through the point, exerts only a normal force on the elastica. If the tangent angle is fixed, the constraint exerts a normal force on the elastica. Similarly, if the relative arc length is fixed, the constraint exerts a longitudinal force.

Result 1 *A normal force acting on an elastica preserves G^2 -continuity across the point where the force is applied, but creates a discontinuity in the first derivative of curvature.*

Result 2 *A shear force (torque) acting on an elastica preserves G^1 -continuity of the elastica but creates a discontinuity in curvature.*

Result 3 *A segment of an elastica with zero longitudinal force at the endpoints takes the form of a rectangular elastica. Nonzero longitudinal force induces values of λ other than 0.5.*

Proofs of these results are not included here, as they are known in the literature of the elastica. An accessible source for them, in the context of interpolating splines, is Lee and Forsythe [51].

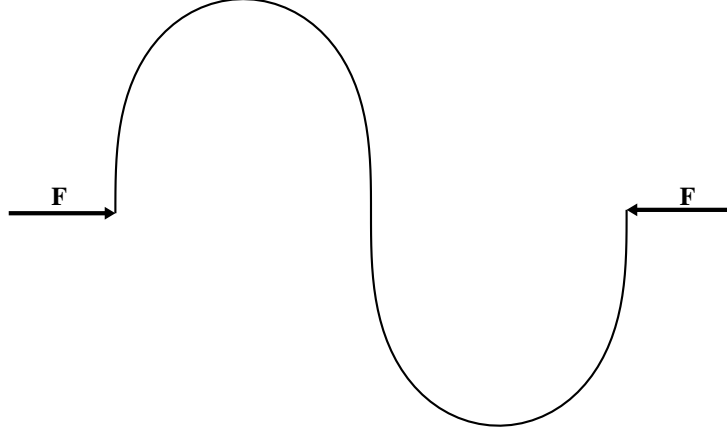


Figure 3.12. Forces acting on rectangular elastica.

Considering a full cycle of the rectangular elastica, as shown in Figure 3.12, it is clear that the longitudinal forces are zero at the endpoints, because the curve is perpendicular to the force \mathbf{F} acting on it.

3.9 Angle constraints

The general constraint approach of the previous section yields a specific result relevant to using the elastica as a spline.

Result 4 *An elastica segment with an angle constraint at each endpoint, but allowed to slide freely through those constraints, is a section cut from the rectangular elastica.*

Equivalently, the minimum energy curve (of arbitrary length) with given endpoints and given angles at the endpoints is a section cut from the rectangular elastica.

Figure 3.13 shows the full parameter space of angle-constrained minimum energy curves. For sufficiently large angles, no such solution exists; the envelope for which the MEC exists is shown as well. This envelope was determined by implementing a very robust numerical solver and measuring the boundary at which the solver did not converge. As far as is known, there is no analytical formula for the envelope's shape. Only half of the complete parameter space is shown, to avoid clutter. The envelope is symmetrical around the $\theta_{left} = \theta_{right}$ diagonal.

Similarly, Figure 3.14 shows the curvature plots of all angle-constrained MECs within their parameter space. It is worth noting that, for small angles, the curvature is very nearly linear with arc length.

Note that the curvatures at the endpoints are fully determined by the angles at the endpoints.

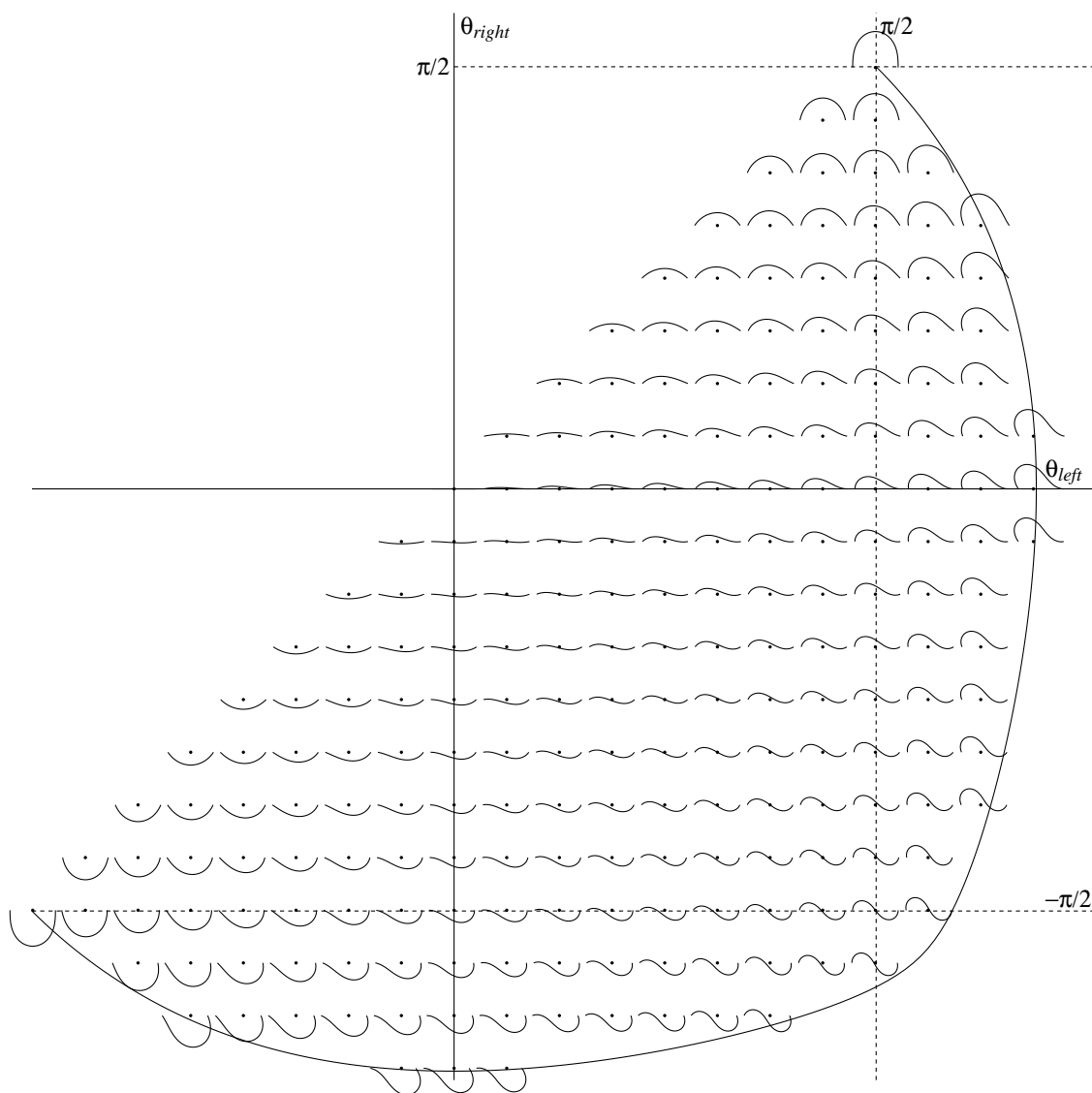


Figure 3.13. Angle-constrained minimum energy curves.

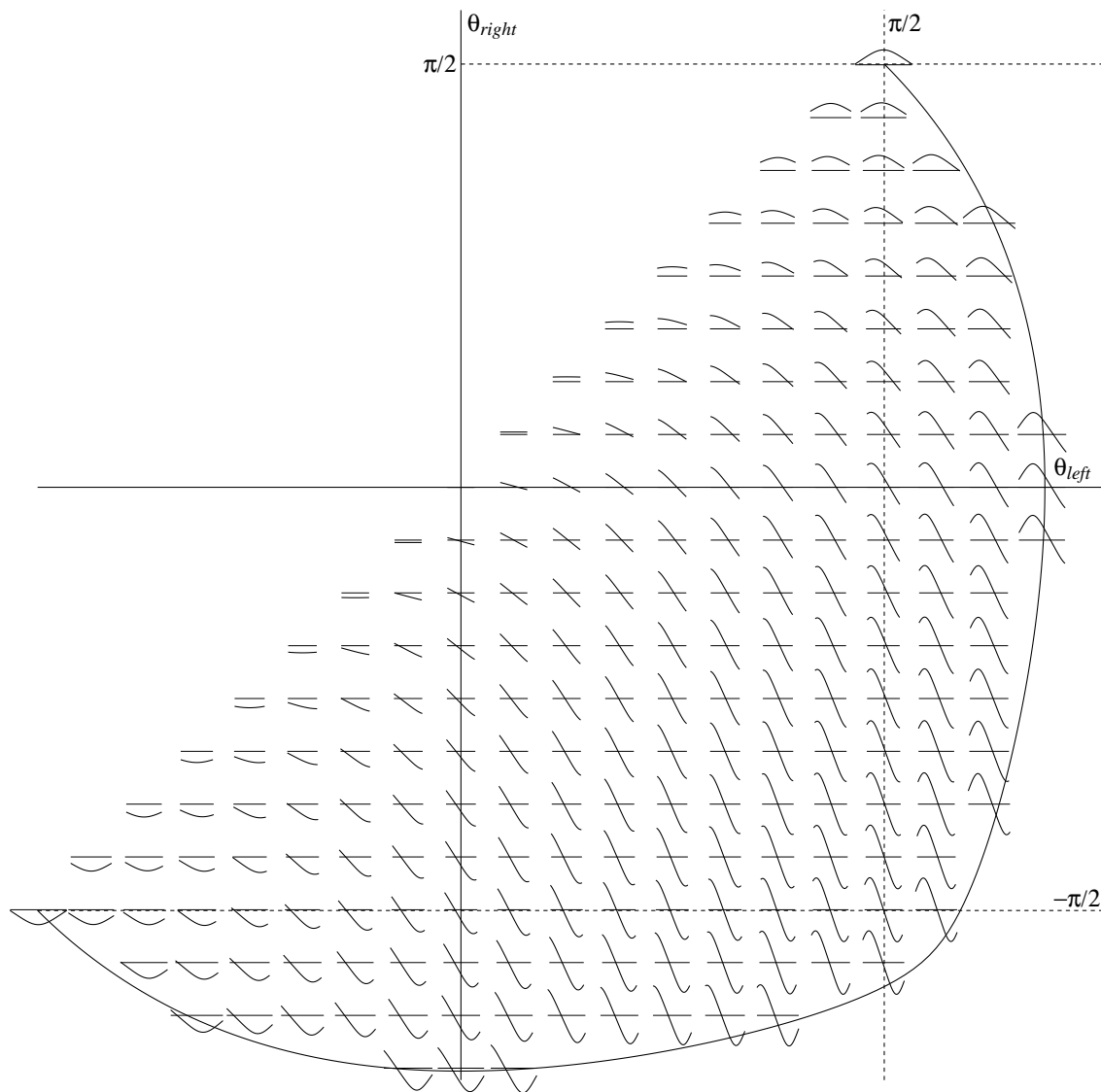


Figure 3.14. Curvature plots, angle-constrained minimum energy curves.

3.10 Elastica as an interpolating spline

The classical formulation of an interpolating spline is the minimum energy curve passing through a sequence of control points. Equivalently, it is an elastica passing through a sequence of constraints, each of which rotates freely and allows the elastica to slide freely through it.

Because each of these constraints exerts only a normal force, the curve is G^2 continuous. Since the elastica is unbent past either endpoint, the curvature at the endpoints is zero (this is also known as the “natural” endpoint condition).

Each segment of the spline, bounded by two control points, is an instance of an angle-constrained MEC as described in the previous section.

To find an interpolating spline, it is necessary and sufficient to find a global solution assigning a tangent angle to each point so that the resulting curvatures are continuous across each point (and zero at the endpoints).

3.11 SI-MEC

To address some of the shortcomings of the MEC as a spline, Moreton and Séquin proposed the Scale Invariant Minimum Energy Curve (SI-MEC) [65]. It is defined as the curve minimizing the functional

$$\text{SIMEC}[\kappa(s)] = l \int_0^l \kappa(s)^2 ds . \quad (3.25)$$

The SI-MEC derives its name from the fact that the functional is invariant with respect to scaling of the curve. The MEC, by contrast, is inversely proportional to the scale factor. However, this scaling property has little, if any, physical meaning, as the resulting shape is invariant to scaling of the entire physical system in either case.

The SI-MEC functional of Equation 3.25 is equivalent to the bending energy of a lamina with length normalized to unity. Thus, a SI-MEC with angle constraints at the endpoints is equivalent to this physical system: the endpoint constraints slide on a rail, but have their angles fixed. The lamina is of fixed length and is not allowed to slide through the endpoint constraints.

The SI-MEC will, for any given constraints, be shorter than the corresponding MEC. Otherwise, the SI-MEC would have even lower MEC energy than the MEC curve, which is impossible by definition.

Any elastica meeting the given angle constraints can be evaluated in its MEC energy and its SI-MEC energy. In short, the MEC energy is the bending energy when the chord length is held fixed and the arc length varies, and the SI-MEC energy is the bending energy when the arc length is held fixed.

Consider first the case of symmetrical angle constraints. Figure 3.15 shows the MEC and SI-MEC energy for all elastica meeting the angle constraints $(\pi/2, \pi/2)$, of chord length 0 to 1 (assuming unit arc length). The MEC energy is minimized, as expected, for $\lambda = 0.5$, the rectangular elastica. The SI-MEC energy is minimized for the circle, $\lambda = 0$. Figure 3.16 shows a similar plot

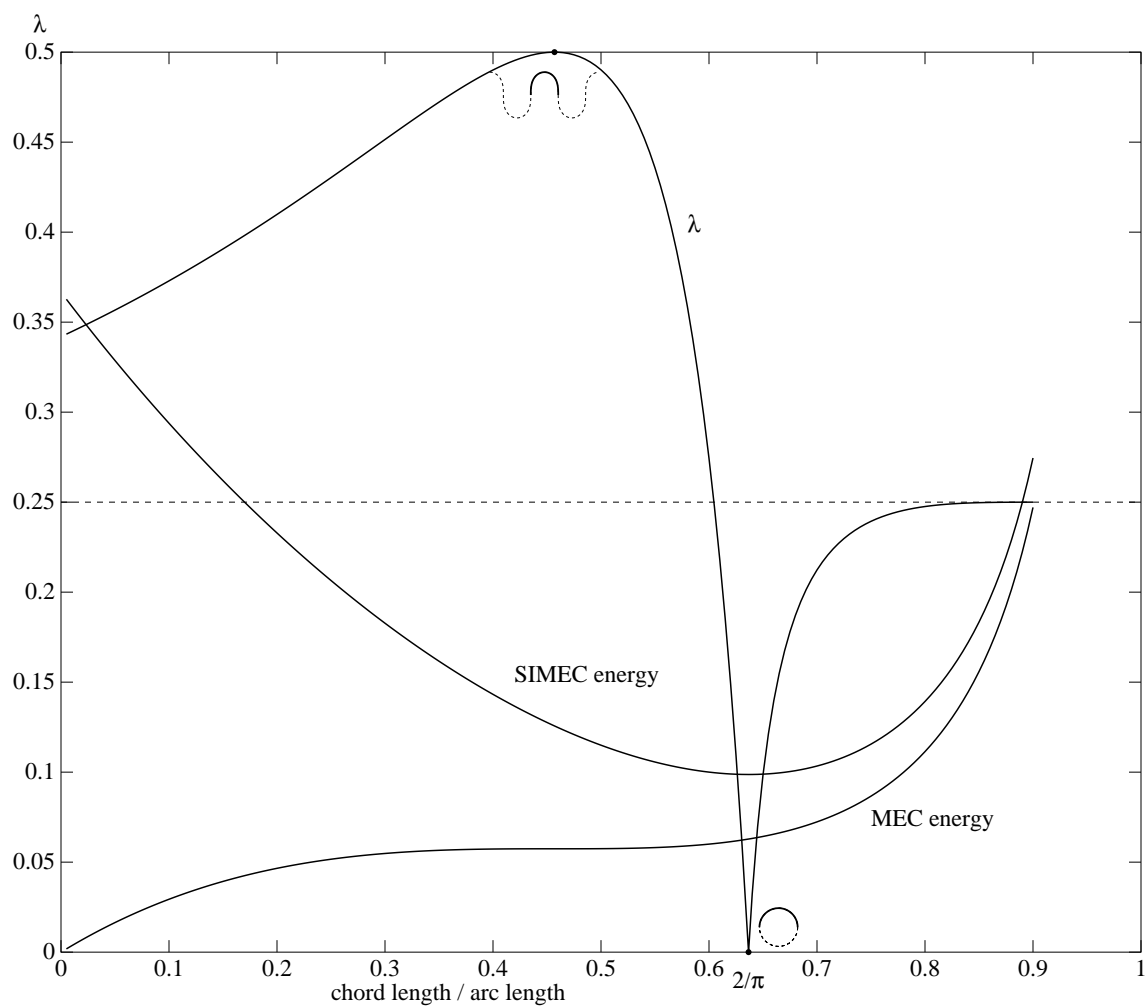


Figure 3.15. MEC and SI-MEC energies for varying lengths ($\pi/2$).

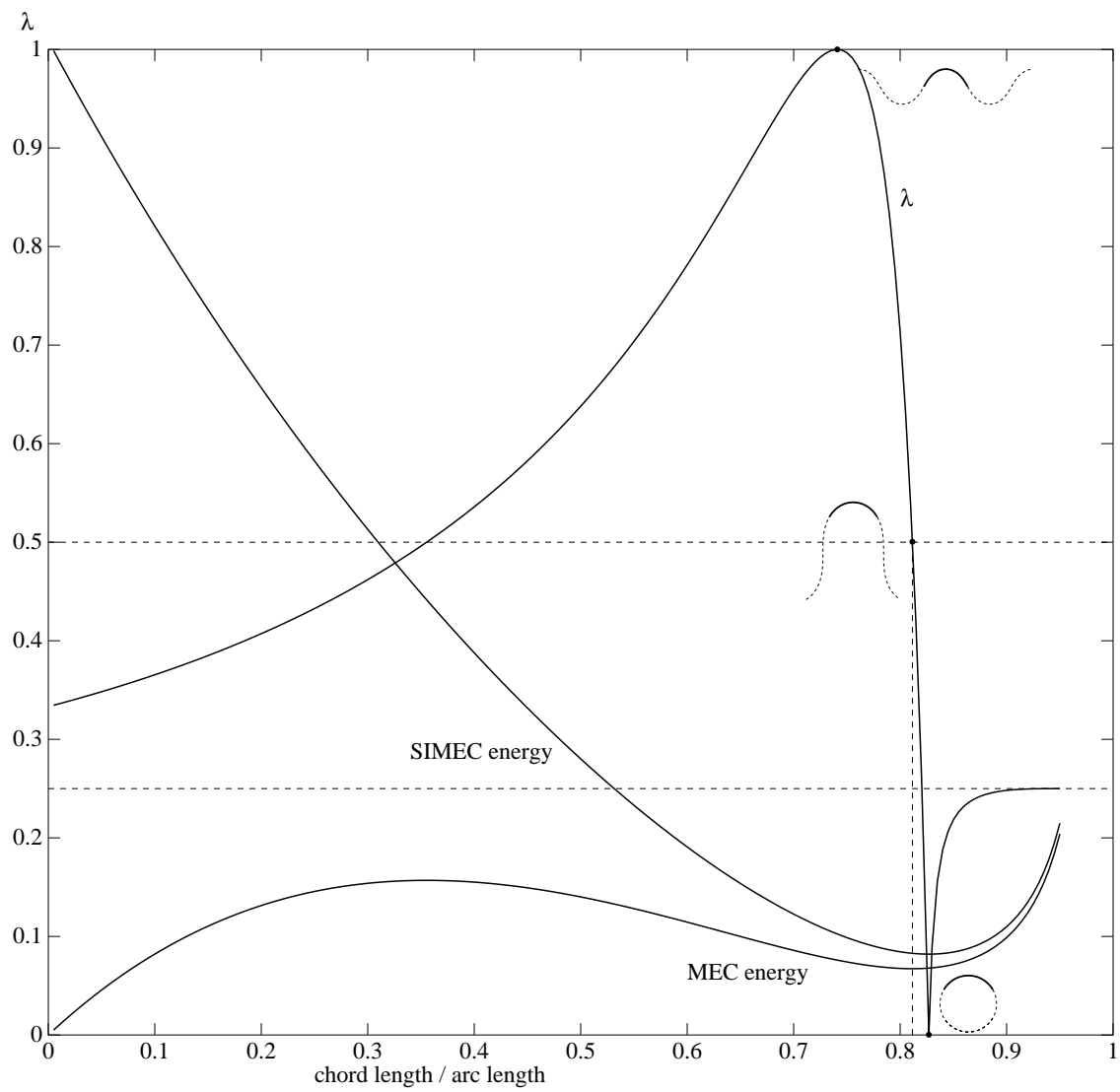


Figure 3.16. MEC and SI-MEC energies for varying lengths ($\pi/3$).

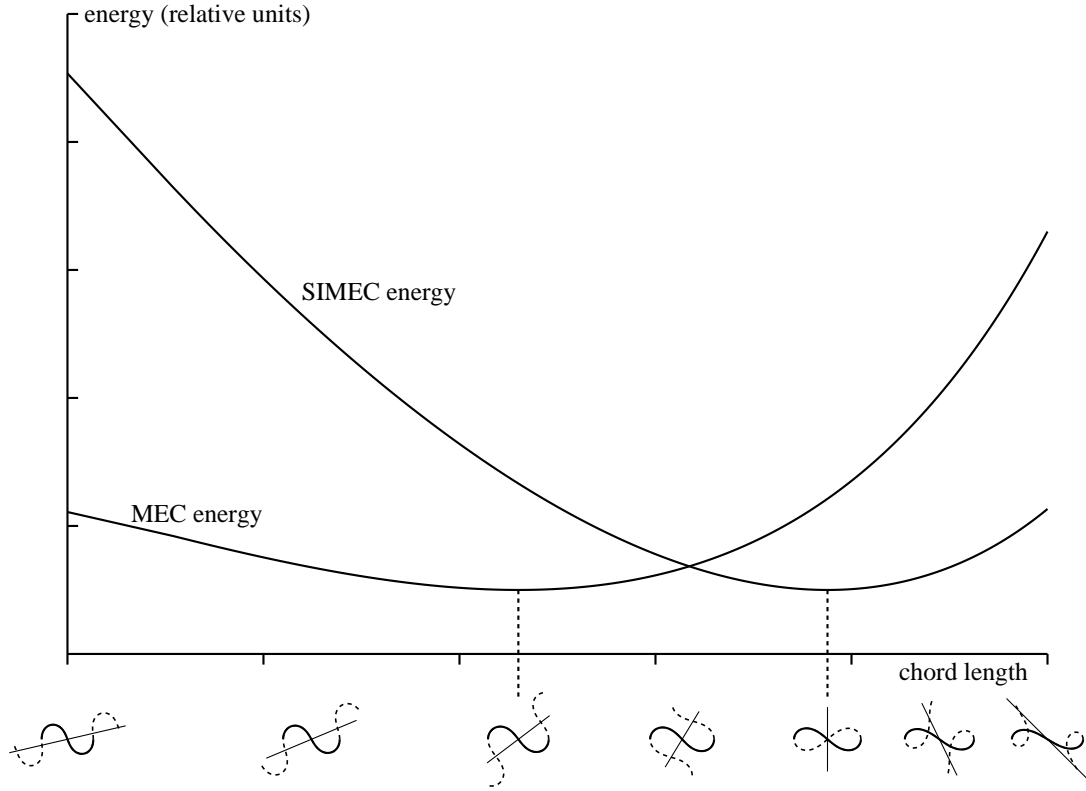


Figure 3.17. SI-MEC and MEC curve energy for $(\pi/2, -\pi/2)$ angle constraints.

for angle constraints $(\pi/3, \pi/3)$, showing that as the angles become less extreme, the difference in chord length between MEC and SI-MEC becomes less pronounced.

Figure 3.17 plots the MEC and SI-MEC energy as a function of the chord length / arc length ratio, given angle constraints of $(\pi/2, -\pi/2)$.

Along with each representative elastica curve is shown the continuation of the elastica curve (with dashed lines), and the line of action (of which the length is proportional to the force of action applied by the constraints).

Note that the line of action for the curve minimizing the SI-MEC energy is vertical. In fact, the line of action is always perpendicular to the chord for a SI-MEC with angle constraints at the endpoints, as can be deduced from the equilibrium of forces in the physical model. If the action had nonzero component along the chord, the constraints would slide closer together or farther apart along the rail.

Several interesting properties follow immediately from the vertical line of action. Based on equilibrium of moments, the curvature varies linearly along the chord length. For symmetrical angle constraints, the curvature is also symmetrical, and thus must be equal at the two endpoints. Thus, the curvature is constant, and the curve is a circle.

For the angle constraints $(\pi/2, -\pi/2)$, the SI-MEC is a segment of the lemnoid elastica. This is also true for $\pi/2$ and the lemnoid elastica angle ($\approx .71$ radians).

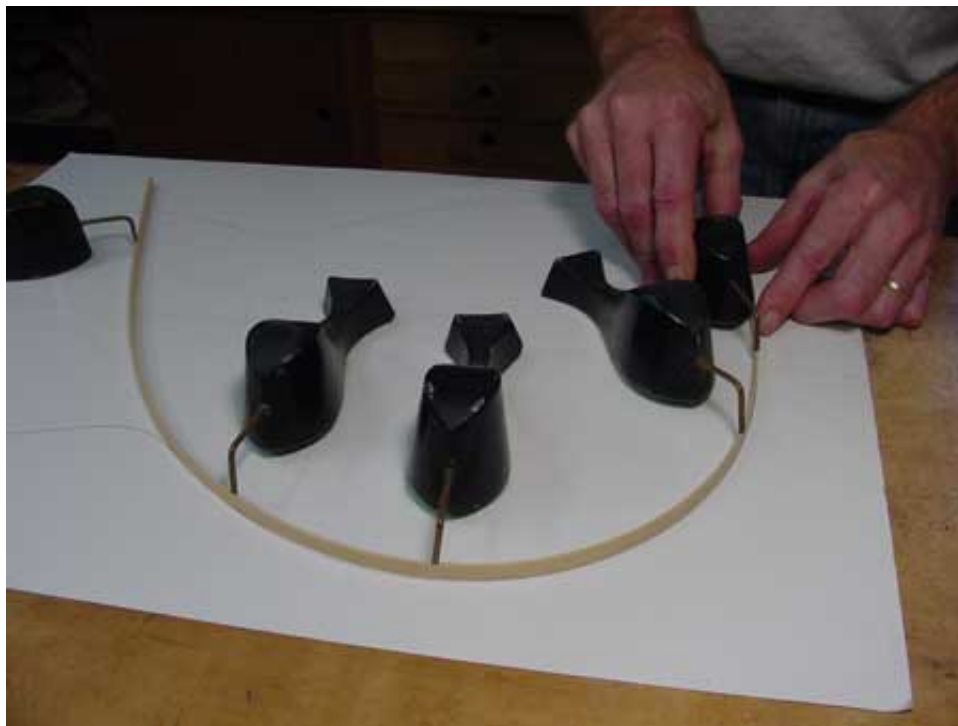


Figure 3.18. A real spline.

Further, all such angle-constrained SI-MEC curves have $\lambda \leq 0.5$, for the $\lambda > 0.5$ cases are single-valued in x for any given y (where the line of action is drawn as vertical), and of course both endpoints must lie on the x axis.

Because curvature is linear along the chord length, for angle constraints with absolute value $\leq \pi/2$, curvature is monotonic with arc length. Contrast with the MEC, which has a curvature peak for all symmetric and nearly-symmetric angle constraints, even as the angles approach zero. As described in Section 2.9, monotonic curvature is a desirable property for curve segments used in interpolating splines.

The domain where the angle-constrained SI-MEC has a solution is about double the size of the MEC's. In the symmetrical case, it reaches a singularity for the full circle (π, π) (exactly double that of the MEC), and in the anti-symmetrical case the singularity is a full figure-eight around the lemnoid elastica, approximately $(5.57, -5.57)$, as opposed to the MEC's $(2.22, -2.22)$.

The SI-MEC has some problems when generalizing from angle constraints at the endpoints to multiple position constraints. In particular, it does not form an extensional spline.

3.12 Real splines

As mentioned, the MEC is a mathematical idealization of a real wooden spline, the kind shown in Figure 3.18 that has been used in shipbuilding for hundreds of years.



Figure 3.19. High-friction spline constraints on a real spline.

For small angles, the two agree closely. But for large angles, how accurately does the MEC model a real spline? The fact that the MEC does not exist for certain angle constraints is particularly suspicious.

The key to this question is, I believe, tension forces and friction. In a real spline, small tension forces are countered by friction at the control points.

Indeed, shipbuilders use traditional spline weights in two different ways. In one such way, the spline weight pushes against the side of the spline, and the spline is free to slide with fairly low friction.

Another common way to use spline weights is to place the point of the pin on the spline curve, as shown in Figure 3.19. In this way, the weights effectively constrain the arc length as well as the position, leaving only rotation entirely free.

It is understood that applying such constraints indiscriminately yields curves of poor fairness (applying tension at the constraints yields curvature discontinuity). Thus, designers use an iterative technique to improve fairing. Once the curve is laid out roughly, each weight is lifted, the spline makes small, fairly localized adjustments when it is free from the weight, and then the weight is replaced, in turn. Of course, if the resulting curve diverges too much from the desired shape, the spline is pushed when the weight is replaced, or a new weight is added.

Iterative local adjustment is reminiscent of successive over-relaxation, a common numerical technique for solving global systems of equations. Note that, as a weight is lifted, the total arc length bounded by the two neighboring weights remains constant. Thus, this process does not inevitably converge to the MEC solution, which has longer arc length than a spline under some tension. Solutions similar to the SI-MEC, or to circular arcs, remain stable.

Friction, then, seems an appealing technique for improving the robustness of the pure MEC, and for making it conform to circles more closely. Shipbuilders have built up a tradition for using



Figure 3.20. Bricks as high friction constraints on a real spline.

splines effectively. For the spline, they often prefer poplar or white pine for the material, about a quarter inch thick (but thinner for more detailed work involving smaller radii, or thicker for large boat lines). For the weights, they usually prefer a pound or two so it holds the spline firmly, a rounded shape resembling a duck or whale for easier handling, and a configuration of the pin so that the weight itself is separated from the spline. Even so, the traditional weight has no different effect on the spline than would a brick (and simple bricks are pressed into service by do-it-yourself boatbuilders, as well, as in Figure 3.20). Aside from shape of the weight, though, all these factors affect the degree to which the real spline models a MEC, especially the friction at the constraints.

A number of researchers have been tempted by the possibility of using tension to make elastica-based splines better than the pure MEC. However, there is no clear way to accomplish this. Real splines use spline weight and wood thickness to set the free parameters, but on a computer these parameters would either need to be set explicitly by the user, or in some other way – at worst, a computer representation of physical quantity such as “one pound weight” or “quarter inch thick white pine” are “voodoo constants,” which are well known for limiting the scope of usefulness of a system.

All these problems are solved, though, when the elastica is left behind, and the Euler spiral is used in its stead. As described in Chapters 5 and 6, dedicated to the history of these curves, their fates are closely intertwined. In fact, the complete mathematical characterization of both curves can be traced to the same 1744 treatise by Leonhard Euler. But the details will wait for Chapter 5. The next chapter explores the properties of G^2 -continuous interpolating splines in general, and quantifies exactly why the Euler spiral is optimal as a curve primitive for interpolation.

Chapter 4

Two-parameter splines

The broad goal of this work is to find the *best* spline, or at least provide a menu of best choices, each suited to a specific requirement. In attempting to characterize the space of all splines, including making a taxonomy of existing splines in the literature, a remarkable number share this property:

A two-parameter spline is defined as one for which each curve segment between two knots is uniquely determined by the two angles between tangent and chord at the endpoints of the segment.

Examples of two-parameter splines include the Minimum Energy Curve, the Euler spiral spline, Mehlum’s KURGLA 1, Hobby’s spline, Ikarus, and Séquin’s circle spline. Note that the count of parameters excludes the scaling, rotation, and translation required to get the curve segment to coincide with the two endpoints – those additional parameters are uniquely determined by the constraints that the curve segment fits the endpoints, providing no additional flexibility for the shape of the curve.

Any given two-parameter spline can be characterized by two properties: the family of curves parametrized by the endpoint tangent angles, and the algorithm used to determine the tangent angles. For example, the Euler spiral spline uses a curve family with linearly varying curvature, and uses a global solver to ensure G^2 -continuity to determine the tangents.

Of the two-parameter splines listed above, the MEC and the Euler spiral are extensional. In the case of the MEC, this property follows directly from the variational statement – if adding a new on-curve point would produce a new curve with a lower cost functional, that curve would have satisfied the old constraints as well. In addition, both splines are based on cutting segments from a fixed *generating curve*, the rectangular elastica or the Euler spiral.

These properties are not simply coincidental. In fact, all two-parameter, extensional splines correspond to a single generating curve, and any generating curve can be used as the basis of a two-parameter extensional spline. The two families, then, are essentially equivalent.

4.1 Extensional two-parameter splines have a generating curve

A central result of this thesis is that for any extensional two-parameter spline there exists a generator curve so that all curve segments between adjacent control points can be cut from that generator curve, subject to scaling, rotation, translation, and mirror image transformations to make the endpoints of the cut segments align with the endpoints in the spline. Thus, any spline in this family can be defined simply in terms of the generator curve. In some ways, this definition circles back to one of the earliest formulations of nonlinear splines, “After looking carefully into the relevant equations, one realizes that the schemes of [53] and [23] do not approximate mechanical splines more closely than other curves – nor does it seem particularly desirable to have them do so. For example, they approximate equally well to Hermite interpolation by segments of *Euler’s spirals*, joined together with continuous curvature.” [10, p. 171]. The Euler spiral is but one of a family of desirable curves for constructing extensional interpolating splines.

Theorem 1 *In an extensional, G^2 -continuous, two-parameter spline, there exists a curve such that for any spline segment (parametrized by angles θ_0 and θ_1) there exist two points on the curve (s_0 and s_1) such that the segment of the curve, when transformed by rotation, scaling and translation so that the endpoints coincide, also coincides along the length of the segment.*

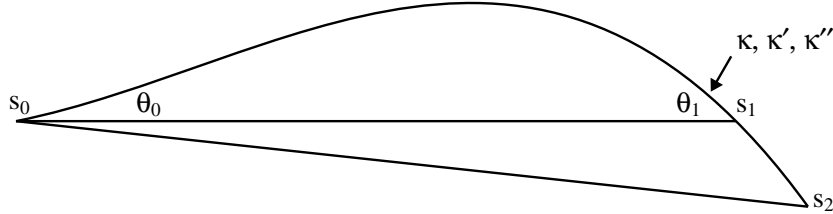


Figure 4.1. Incremental construction of the generator curve of a two-parameter spline.

A typical section of curve is shown in Figure 4.1. For some given endpoint angles, the generated curve segment is the section from s_0 to s_2 . Because of the extensionality, for any s_1 along this curve, the subdivided curve from s_0 to s_1 (which is the result of applying θ_0 and θ_1 endpoint constraints) must be consistent with the original segment. Modulo scaling, the curvature κ and its derivatives κ' and κ'' must of course also match between these two segments.

The proof of this theorem revolves around the quantity κ'/κ^2 , which is also the limit of $6(\theta_1 - \theta_0)/(\theta_0 + \theta_1)^2$ as θ_0 and θ_1 both approach zero, i.e. as the length of the curve segment becomes infinitesimal (this limit follows from Equation 8.31, which is derived in detail in Section 8.2). This quantity represents the amount of curvature variation for a segment of a set amount of curvature – it is invariant to similarity transformations including uniform scaling. As such, for a given generator curve, it uniquely identifies a point on the curve (modulo periodicity, if the curve is periodic as opposed to monotone in curvature).

We propose that for any two-parameter spline that is also extensional, the *second* derivative of curvature, suitably normalized for scale invariance, κ''/κ^3 , is uniquely determined given a particular

value of κ'/κ^2 . If there are two curve segments with different values for this second derivative, then it is possible to cut subsegments from each with identical θ_0 and θ_1 values. However, the value of κ''/κ^3 remains invariant even after such a cutting, so there are two different values for this. Yet, the two parameter property states that there exists only a single curve for a given θ_0, θ_1 pair – a contradiction!

Thus, there is a process for constructing the generator curve. Start with arbitrary κ and κ' values sampled from some point on the curve, then continually derive κ'' based on the lookup process suggested by the above relation: find a curve segment that has the correct value for κ'/κ^2 , then sample κ''/κ^3 at that point. Continually integrate (in the sense of an ordinary differential equation), and the result is a curve.

The final component of the argument is that for any given θ_0, θ_1 pair, there exists a segment cut from this curve such that it is equal to the spline segment. Determine κ'/κ^2 for one side (s_0) of the spline segment, and find that point on the generator curve. Continue extending the curve until the other side is reached. Since at each point κ''/κ^3 must be equal between the two curves, the curves themselves must also be equal.

This argument depends on the assumption that the curve is a continuous function of the parameters θ_0 and θ_1 . We conjecture that this assumption follows from the assumption of G^2 -continuity and extensionality, but do not have a rigorous proof. It is also possible the continuity assumptions could be weakened and the theorem would still be valid.

Both the two-parameter and extensionality conditions are necessary for the generator curve property to hold, and there are counterexamples when either is not present. The Ikarus and Metafont splines are two-parameter curves, but not extensional, and they do not have a single generator curve. The Minimum Variation Curve (discussed in much more detail in Section 7.2) is extensional but does not fit into a two-parameter space, and also does not have a single generator curve.

The Minimum Energy Curve and the Euler spiral spline have a number of features in common, among them that all segments between two adjacent control points are fully determined by two parameters. This count of parameters does not include the scaling, rotation, and translation needed to make the two endpoints coincide – just the shape of the curve once the positions of the endpoints are fixed.

In fact, the parameter space for each such segment is described by the tangent angles at endpoints. Since our definition of “interpolating spline” requires G^1 -continuity, determining a single tangent angle at each control point suffices. Thus, combining a two-parameter curve family with a technique for determining tangent angles at the control points yields an interpolating spline.

Note that there are also several splines in the literature designed to approximate the MEC (quite accurately in the case of small angles), but do not fit into the two-parameter framework. In general, cubic polynomial curves have four parameters. In a standard cubic spline, the curve segments can be any cubic polynomial, so the parameter space is four dimensional. Parametrizing by chord length approximates the MEC more closely but does not reduce the dimensionality of the parameter space.

Similarly, the scale invariant MEC (dubbed SI-MEC in [65]) has segments that are piecewise elastica, and there is an additional parameter global to the spline (and associated with tension forces at the endpoints, in a physical model of the spline) that increases the dimension space of individual segments from the two of the pure MEC to the three admitted by the elastica.

While the previous chapter examined the elastica with constrained endpoints, a more general problem is the minimum energy curve that interpolates a sequence of control points. More formally, we seek, over the space of all curves that pass through the control points (x_i, y_i) in order, the one that minimizes the MEC energy functional, restated here:

$$E_{\text{MEC}}[\kappa(s)] = \int_0^l \kappa^2 ds \quad (4.1)$$

This problem is the mathematical idealization of a mechanical spline through control points fixed in position but free to rotate, and with an infinitely thin beam free to slide through them with zero friction.

The earliest discussion of this more general problem is probably Birkhoff and de Boor in 1964. Over the next few years, several authors published algorithms for computing these “nonlinear splines.” Probably the most illuminating and accessible discussion of the early nonlinear spline theory is the 1971 report of Forsythe and Lee [51]. A brief summary of their results follows:

- Each segment between knots is a MEC with $\lambda = 0.5$.
- Curvature is continuous across knots (G^2 -continuity).
- Curvature is zero at endpoints (in the open case).
- The spline is extensional.

For each segment between knots, the elastica with λ fixed has two remaining parameters, corresponding to the tangent angles at the endpoints. Recall that Figure 3.13 shows instances over the entire two-dimensional space in which such solutions exist, as well as the boundary of that solution space. In particular, in the case of symmetrical endpoint tangents, the curve is *not* a circular arc, so the MEC spline does not have the roundness property. Forsythe and Lee found this unappealing, and propose adding an arc length penalty term to the functional,

$$E[\kappa(s)] = \int_0^l \kappa(s)^2 + c \, ds . \quad (4.2)$$

As we now know, this change admits the solutions to the elastica equation other than $\lambda = 0.5$. Forsythe and Lee derive the value for c so that the solution is a circular arc (corresponding to $\lambda = 0$), and speculate “whether such an approach could be generalized is an open question.” Answering their question is a primary focus of this chapter.

4.2 Properties of the MEC spline

The most important property the MEC spline is lacking is roundness. For three control points placed in an equilateral triangle (shown in Figure 4.2, with the spline drawn in a thick line, and the circle going through the control points thin), the containing radius for the spline is approximately 1.035 times the radius of the circle going through the control points. With four points, this radius

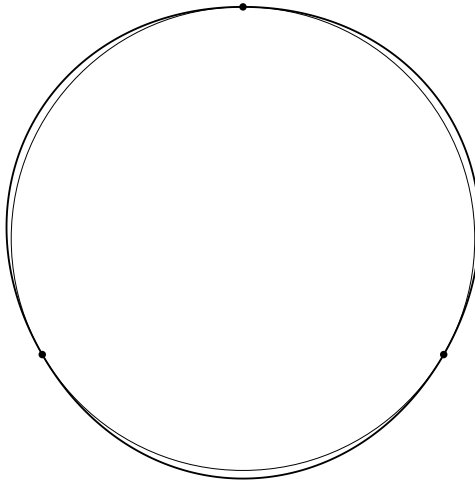


Figure 4.2. MEC roundness failure.

deviation decreases to about 1.009, and asymptotically converges as $O(n^4)$ with the number of points.

Another unpleasant property of the MEC spline is the failure to converge to a stable solution. This property is intimately related to the relatively small envelope enclosing the solution space – for symmetric solutions the maximal endpoint tangent for $\lambda = 0.5$ is $\pi/2$, and it is not much better for any of the non-symmetric solutions.

4.3 SI-MEC

Several researchers have tried to address these shortcomings in the MEC spline. One is the scale invariant minimum energy curve (or SI-MEC) proposed by Moreton and Séquin [65]. This spline is defined as the curve minimizing the SI-MEC functional

$$E[\kappa(s)] = l \int_0^l \kappa(s)^2 ds . \quad (4.3)$$

The functional is termed “scale invariant” because its value is invariant as the entire curve is scaled up. By contrast, the MEC functional is proportional to the reciprocal of the scaling constant. The MEC functional can be seen as rewarding increases in the length of the curve, which accounts for the roundness failure – the actual curve obtained is slightly longer than the circle, which accounts for the lower MEC functional. In the SI-MEC, the bias toward longer curves is removed, and, in fact, the spline is round when the control points are co-circular.

The SI-MEC minimizes the MEC functional for a given length (a curve with a lower MEC energy at the same length would also obviously have a lower SI-MEC cost functional). Thus, an appealing physical interpretation is that of a mechanical spline with a bit of tension pulling at the endpoints, making the overall curve slightly shorter. Such a curve is generally *not* two-parameter

in the formulation above, and, in fact, all values of the elastica are possible (thus, it can be seen as a three-parameter spline). The SI-MEC has an unpleasant global coupling (equivalent to finding the tension to apply to the mechanical spline) that makes a global solver considerably slower than the corresponding pure MEC [65].

The SI-MEC is extensional, which follows directly from its variational formulation as the curve minimizing an energy functional, such that it passes through all the control points. It shares G^2 -continuity and similar locality properties with the MEC.

4.4 Cubic Lagrange (with length parametrization)

The cubic Lagrange spline is defined as follows (see [49, p. 23] for more detail), given a sequence of input data (t_i, z_i) , where z_i can be interpreted as the vector (x_i, y_i) :

$$L_{i,3}(t) = \sum_{j=i-1}^{i+2} z_j \frac{\omega_{i-1,3}(t)}{(t-t_j)\omega'_{i-1,3}(t_j)}, \quad \omega_{i-1,3}(t) = \prod_{j=i-1}^{i+2} (t-t_j) \quad (4.4)$$

To obtain the Lagrange polynomial normalized by chord lengths, set successive t_{i+1} so that $t_{i+1} - t_i = |z_{i+1} - z_i|$. In other words, the parametrization is the same as the chord length. (Set $t_0 = 0$ for convenience, but the shape of the spline doesn't depend on it.)

This spline is a reasonable approximation to MEC when the angles of the control polygon are small. It is also not round, as it is based on a polynomial approximation. It shares locality properties with the MEC as well.

Further, the spline is not extensional, as adding additional points perturbs the chord length parametrization.

Without chord length parametrization, the cubic Lagrange spline is extensional, but prone to wild excursions when the control points are not spaced evenly.

The cubic Lagrange spline is sometimes used on its own, but its main interest here is as the basis of the Ikarus spline, discussed in the next section.

4.5 Ikarus (biarc)

The spline used by Peter Karow's Ikarus system [42] also falls into the category of two-parameter splines. It was intended for font design, and was particularly well suited for interpolating between "masters" of different weights.

The Ikarus spline uses two completely different mechanisms to achieve the two basic elements of the two parameter spline: the determination of tangents and the shapes of the curve segments between control points. To determine the tangents, first a cubic Lagrange polynomial (with normalization based on the chord lengths, as described in Section 4.4) is fit through the control points. Then the polynomial curve is discarded (but the tangents preserved), and, for each segment, a curve defined by biarcs is substituted. An arbitrary biarc has one additional degree of freedom, cor-

responding to the split point where the two arcs join, and Ikarus defines this somewhat arbitrarily to provide reasonable results.

Even though the Lagrange polynomial fit can easily incorporate inflection points, the biarc primitive cannot, so inflection points must be explicitly marked in the input.

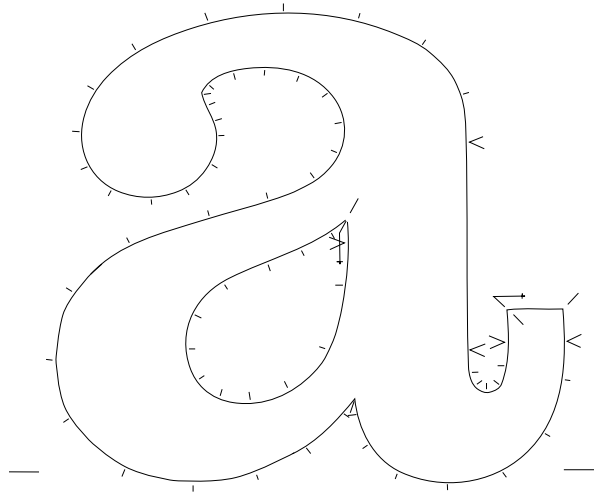


Figure 4.3. Typical letter-shape using Ikarus.

The spline is not extensional, as adding points perturbs the chord length parametrization. Further, roundness is fairly decent, as the biarc becomes a circular arc when the endpoint tangents are symmetrical. Continuity is only G^1 , but the biarc segments keep the spline from exhibiting extreme variations in curvature typical of purely polynomial-based splines. The spline is reported to work well when the control points are dense. Karow [43] recommends one control point for every 30 degrees of arc. A typical example is shown in Figure 4.3 (reproduced from [43]). See also Figure 10.8 for a comparison with the splines of this thesis.

4.6 Euler's spiral

One particularly well-studied two-parameter spline is Euler's spiral, also known as the spiral of Cornu or the clothoid. More precisely, it is the spline formed by joining piecewise Euler's spiral segments with curvature (G^2) continuity at the knots.

Euler's spiral, plotted in Figure 4.4 over the range $[-5, 5]$, is most easily characterized by its intrinsic definition: curvature is linear in arc length.

$$\kappa(s) = \frac{s}{2}. \quad (4.5)$$

Integrating the intrinsic equation yields this equation, known as the Fresnel integrals (historically, one each for the real and complex part):

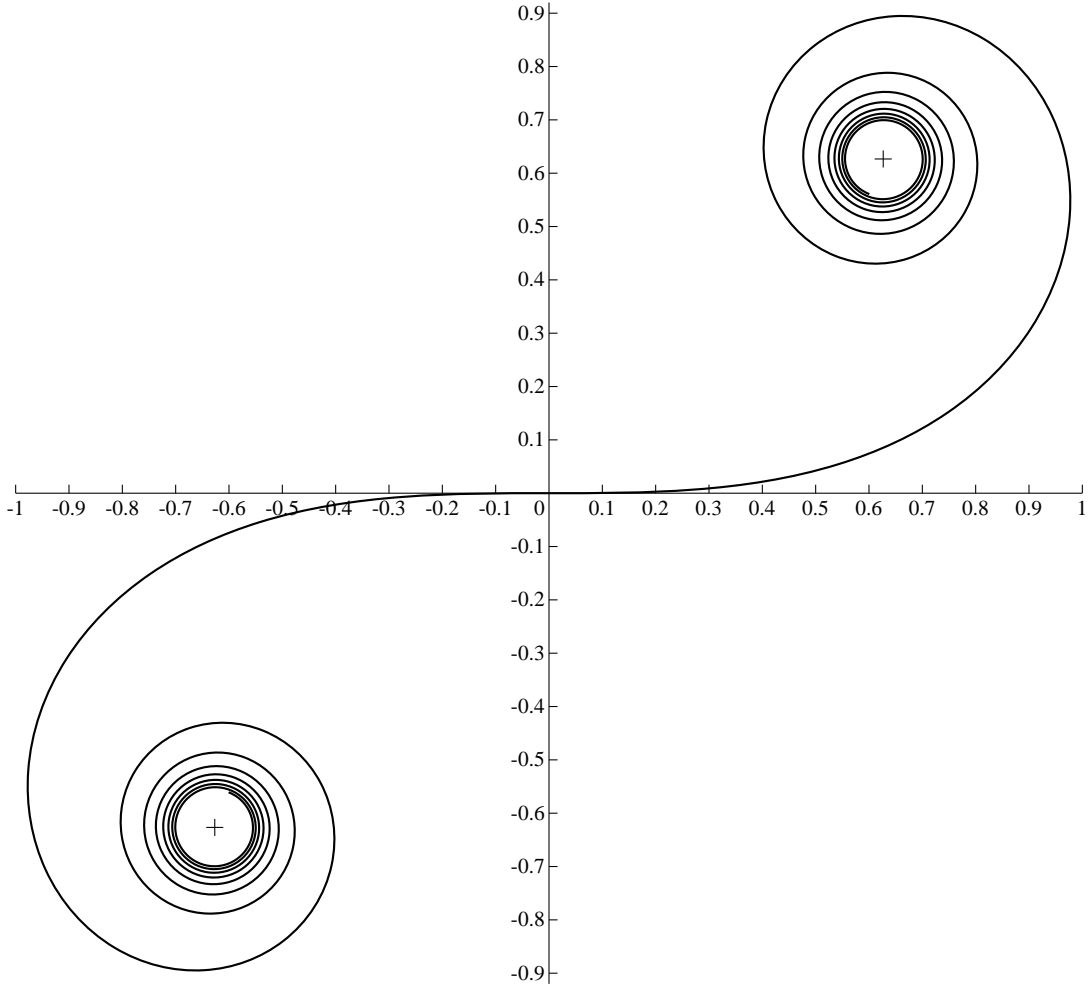


Figure 4.4. Euler's spiral

$$x(s) + iy(s) = \int_0^s e^{it^2} dt . \quad (4.6)$$

This integral is also commonly expressed with a scale factor applied to the parameter. The resulting curve is similar; its size is $\sqrt{\frac{2}{\pi}}$ times that of the first formulation, with the velocity similarly scaled so that it also has unit velocity. The traditional notation for these integrals is $C(s)$ and $S(s)$, which evoke cosine and sine, respectively.

$$C(s) + iS(s) = \int_0^s e^{i\pi t^2/2} dt . \quad (4.7)$$

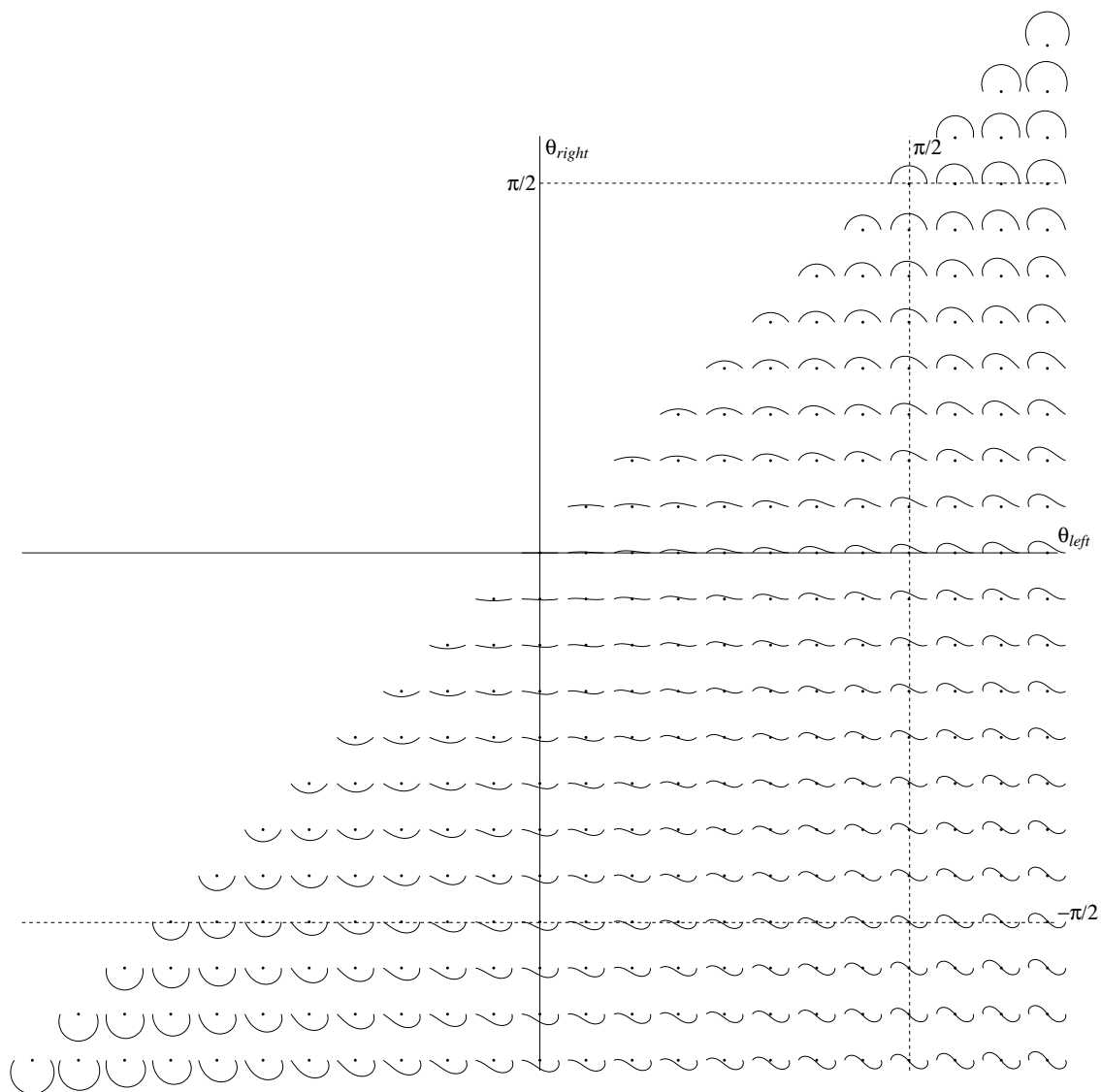


Figure 4.5. Euler's spiral spline primitive over its parameter space.

4.6.1 Euler’s spiral as a spline primitive

Euler’s spiral is readily adapted for use as a two-parameter primitive for a spline. Given tangent angles at the endpoints, find parameters t_0 and t_1 from which to “cut” a segment from the spiral, so that the resulting segment meets the endpoint constraints (for now, assume that such parameters can be found; an efficient algorithm will be given in Section 8.2).

The behavior of this primitive is shown in Figure 4.5. Compare to Figure 3.13, showing the corresponding behavior for the MEC spline. Note, among other things, that the Euler’s spiral spline primitive is defined and well-behaved over a much wider range of endpoint tangent angles.

4.6.2 Variational formulation of Euler’s spiral

The spline literature has generally been split into two camps: variational techniques, and the use of explicitly stated curves such as the Euler spiral. Ironically, Euler’s spiral was born of the study of elastic springs just as was the elastica, but for the most part the literature does not examine Euler’s spiral from a variational point of view. This section ties these two threads together, showing that the Euler spiral does indeed satisfy some fundamental variational properties.

First, as noted by Kimia et al [44], the Euler spiral is among the solutions of the equation for the MVC functional, i.e. the curve that minimizes the MVC energy functional subject to various endpoint constraints:

$$E_{MVC}[\kappa(s)] = \int_0^l \left(\frac{d\kappa}{ds} \right)^2 ds \quad (4.8)$$

This result can be readily derived by treating the problem as an Euler variational problem expressed in terms of κ , i.e. $F(x, y, y') = y'^2$, with s for x and κ for y in the terms of Equation 3.2, as stated in Chapter 3. The corresponding Euler–Lagrange equation is particularly simple:

$$\kappa'' = 0 \quad (4.9)$$

The solution family is thus $\kappa = k_0 + k_1 s$, which are simple similarity transformations of the basic Euler spiral.

While the simplicity of this result is pleasing, it is unsatisfying for a variety of reasons. First, because the variational equation is written in terms of κ rather than θ , the boundary conditions are written in terms of curvature rather than tangent angle. Similarly, the additional terms corresponding to Equation 3.5 representing the *positional* endpoint constraints are missing. Recall that without these constraints, the only solution to the MEC equation is a straight line.

So this result only tells us that the Euler spiral is the MVC-optimum curve satisfying *curvature* constraints on the endpoints, but with unconstrained endpoint positions and tangent angles. As we shall explore more deeply in Section 7.2.2, the solution to the MVC under other constraint configurations will in general not be an Euler spiral, a point not clearly made in the presentation by Kimia et al.

A more fundamental problem is that the degree of continuity doesn’t match. A spline made of

piecewise Euler spiral segments has G^2 -continuity, the same as a MEC spline, while the solution to the MVC functional with positional constraints for internal knots has G^4 -continuity.

So the Euler spiral is a reasonably good first-order approximation to the MEC functional, and is a special case in the solution space of the MVC functional, but might there be a functional for which it is actually the optimum? In a sense, this is the reverse of the usual variational problem, in which a function is sought that minimizes a given function. Here, we seek a functional which happens to be minimized by the given function, in this case the Euler spiral.

We conjecture that there is indeed a functional: the maximum rate of curvature change over the entire arc length,

$$E_{\text{MAXVC}}[\kappa(s)] = \max \left| \frac{d\kappa}{ds} \right|. \quad (4.10)$$

This functional is related to the MVC; while the MVC is the 2-norm of curvature variation, E_{MAXVC} corresponds to the ∞ -norm.

4.6.3 Euler spiral spline in use

How does the Euler spiral perform in the intended application domain of drawing glyph outlines? Experimentation yields encouraging results. One such example, a lowercase ‘c’ traced from 14 point metal Linotype Estienne, is shown in Figure 4.6. Estienne is a distinguished book face designed by the eminent English printer George W. Jones, and released by Linotype in 1930 [58, p. 133]. It is not currently available in digital form, making it an attractive candidate for digitization.

The ‘c’ is a simple closed curve, reasonably smooth everywhere, but with regions of high curvature at both terminals. Below the glyph itself is a curvature plot, beginning at the knot marked by an arrow and continuing clockwise. The long flat section on the left is the inner contour, the sharp spike is the lower terminal, and the other long flat section (which is both longer and lower curvature) is the outer contour. Note that the curve has two inflection points, corresponding to zero crossings on the curvature plot and marked on the outline with short lines.

A more complex example is shown in Figure 4.7, in this case a lowercase ‘g’ from the same font. In this case, there are three closed curves; the outer shape and two “holes.” While the top hole is a simple smooth curve, the other curves contain both G^2 -continuous smooth knots and “corner points”, with only G^0 -continuity. Such curves can be factored into sections consisting of the corner points at the endpoints and the smooth knots in the interior, then joined together. Note that sections with no smooth knots are straight lines (there is one at the top of the “ear”), and sections with a single smooth knot are circular arcs (two of these make up the rest of the ear). More generally, each spline segment bordered by a corner point is a circular arc, with zero variation of curvature.

Experience with drawing a significant number of glyphs is encouraging. The Euler spiral spline can represent a wide range of curves expressively and with an economy of points. Further, in an interactive editor, the correspondence between the input (the placement of knots and control points) seems considerably more intuitive than the standard Bézier curve editing techniques.

The trickiest curves to draw using pure Euler’s spiral splines are transitions between straight lines and curves, as typical in the shape of a “U.” The difficulty is illustrated in Figure 4.8. Es-

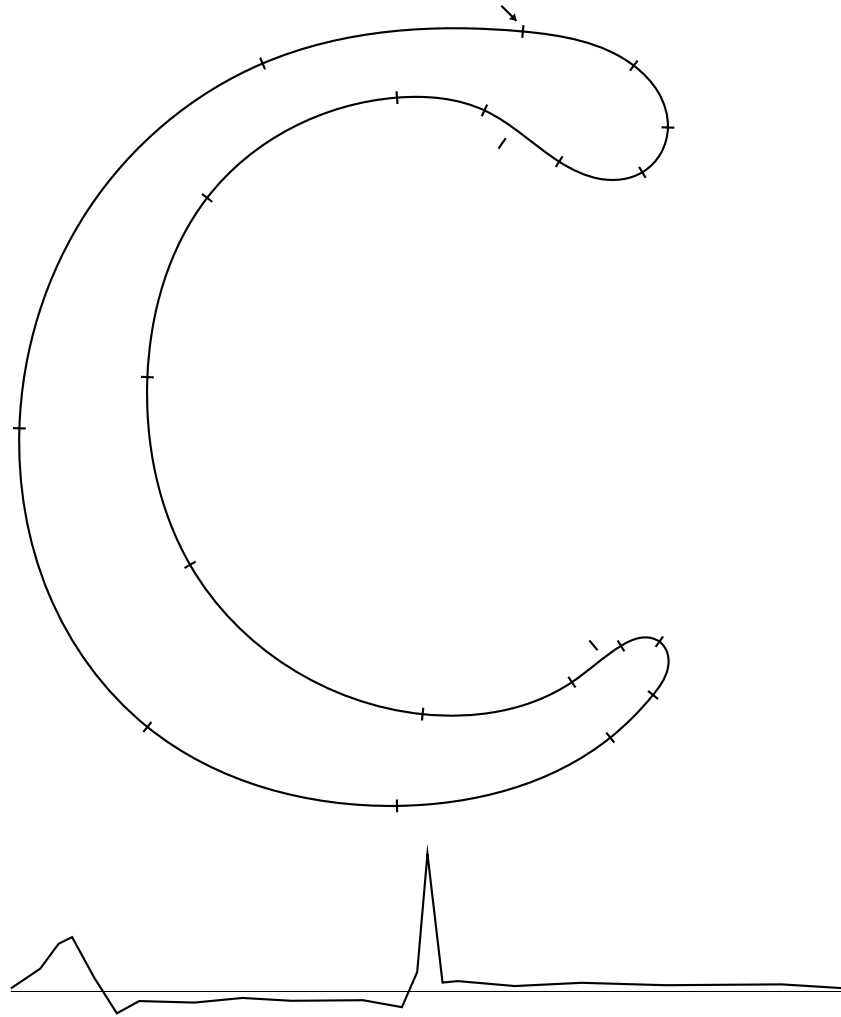


Figure 4.6. Closed Euler example.

essentially, the problem is that the segment adjoining the corner point tends to be a circular arc segment, rather than being constrained to be a straight line. It's possible, by careful interactive placement, to decrease this curvature so that it's almost a straight line, but then any changes to the curved segment will undo the effect of this careful placement. Similar problems occur if the points are determined by interpolation between different masters, or some other parametric variation. We shall return to the problem of straight-curve joins in Section 7.4.

4.6.4 Existence and uniqueness of Euler spiral spline

One of the serious shortcomings of the MEC spline is lack of existence for all configurations of control points (see Section 2.5 for a general discussion). An appealing feature of the Euler spiral spline is that it solutions exist in many cases when no such solution exists for the MEC spline.

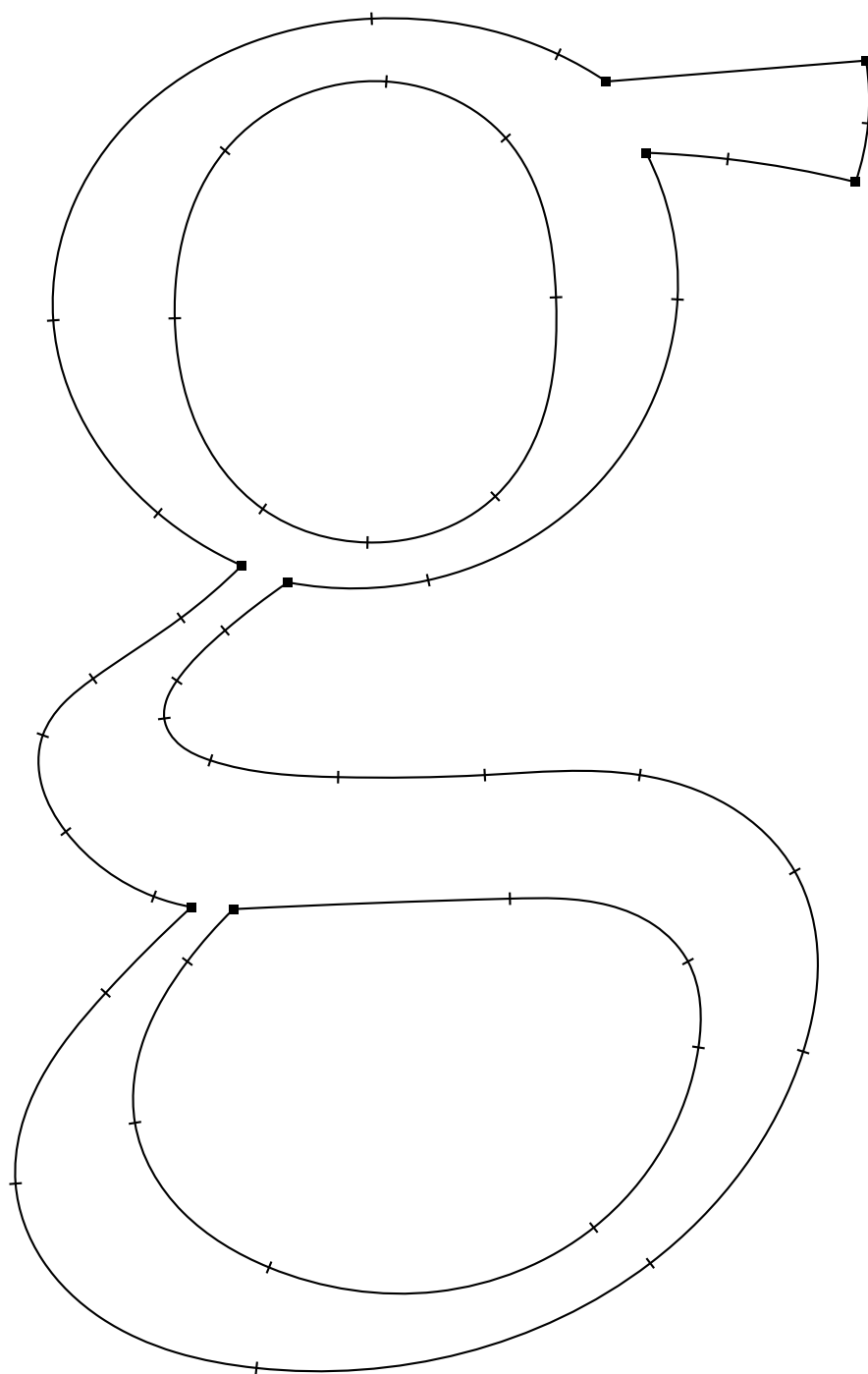


Figure 4.7. Complex Euler example.

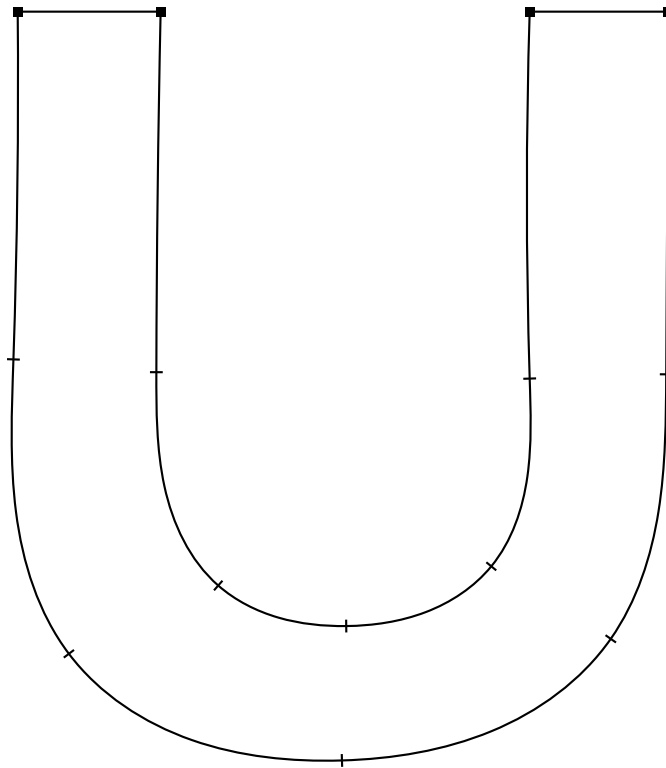


Figure 4.8. Difficulty with straight-to-curved joins.

The obvious question is then, does a solution to the Euler spiral spline always exist, no matter what sequence of points are given? Also, is the solution unique?

A partial answer is given in a 1982 paper by Stoer [82].¹ Stoer considered a slightly weaker formulation of the Euler spiral spline: each segment is piecewise an Euler spiral segment, and there is G^2 -continuity, but there are no additional constraints on the endpoints. Given these relaxed constraints, there is always a solution, in fact infinitely many. In other words, for any given sequence of control points, there exist infinitely many assignments of Euler spiral parameters such that the resulting segments have G^2 -continuity.

The proof is dependent on the following theorem:

Stoer's theorem 2.25 (restated): For all points P_0 and P_1 , tangent angles θ_0 , and curvatures κ_0 , there exist countably many different Euler spiral segments connecting P_0 to P_1 such that the tangent at P_0 is θ_0 and the curvature at P_0 is κ_0 .

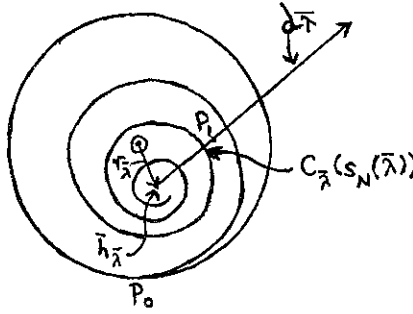


Figure 4.9. Multiple windings for Euler spiral segment connecting two points (from [82]).

Stoer's proof is based on the fact that an Euler spiral segment with gently varying curvature may wind around arbitrarily many times before reaching P_1 . In the limit, the total winding number of the spiral segment increases by approximately 2π for each successive solution. Such multiple windings are shown in Figure 4.9, reproduced from [82, p. 329].

The result that there are infinitely many piecewise-Euler spiral, G^2 -continuous interpolating splines follows readily from Stoer's theorem 2.25. Start at the first point, choosing arbitrary tangent and curvature, then choose one of the solutions guaranteed by the theorem, and repeat until the final point is reached.

With arbitrary choices, such a procedure will tend to produce splines with excessive winding. Stoer proposed using the MEC energy to select a single "best" spline from the myriad alternatives. Stoer also gave an algorithm based on the standard band-diagonal Jacobian matrix, and claimed that it gave good results in general, but did not include a proof that the algorithm achieves a global optimum in all cases.

In addition, using the global minimum of MEC energy will also have an effect on the curvature at the endpoints. Since minimizing MEC over all curves assigns zero curvature to endpoints (see

¹Although, the reader is warned, this paper contains numerous errors, including in the statement of the crucial Theorem 2.25.

Section 3.10), and the Euler spline primitive is a fairly good approximation of the MEC primitive, at least in the small-angle assumption the curvature at the endpoints would be approximately zero.

It is not currently known whether a solution always exists in the presence of additional endpoint constraints, such as zero curvature variation for the terminal segments. Obviously, with the methods of Stoer it is possible to apply such a constraint to one side, but it's not clear whether there, for all input sequences of points, an assignment of tangent for the left side such that the curvature variation is zero on the right side.

The numerical techniques proposed later in this thesis (Section 8.3 converge on a solution (and a solution with no excess windings, as well) for reasonable input sequences of control points, but can fail to converge when turning angles become excessive.

4.7 Hobby's spline

A significant interpolating spline, especially in font design, is Hobby's spline as implemented in Metafont [37]. It is perhaps easiest to understand this spline as a very computationally efficient approximation to the Euler spiral spline of the previous section.

The definition of the primitive segment, with endpoint tangent angles θ_0 and θ_1 , is

$$\begin{aligned}
a &= \sqrt{2}, \\
b &= \frac{1}{16}, \\
c &= (3 - \sqrt{5})/2, \\
\alpha &= a(\sin \theta_0 - b \sin \theta_1)(\sin \theta_1 - b \sin \theta_0)(\cos \theta_0 - \cos \theta_1), \\
\rho &= \frac{2+\alpha}{1+(1-c)\cos \theta_0 + c \cos \theta_1}, \\
\sigma &= \frac{2-\alpha}{1+(1-c)\cos \theta_1 + c \cos \theta_0}, \\
x_0, y_0 &= 0, 0, \\
x_3, y_3 &= 1, 0, \\
x_1, y_1 &= \frac{\rho}{3\tau} \cos \theta_0, \frac{\rho}{3\tau} \sin \theta_0, \\
x_2, y_2 &= 1 - \frac{\sigma}{3\tau} \cos \theta_1, \frac{\sigma}{3\tau} \sin \theta_1.
\end{aligned} \tag{4.11}$$

The (x_i, y_i) are the coordinates for a standard cubic Bézier curve. The adjustable parameter τ corresponds closely to the “tension” parameter of B-splines, and its default value is 1. In the Metafont sources for Computer Modern, the most common other value found is 0.9, and that is only for a few glyphs, including lowercase ‘c’ [46, p. 311].

The behavior of the Hobby spline primitive over its parameter space can be seen in Figure 4.10 (see Figure 3.13 for comparison to the MEC). It is defined for all values of θ_0 and θ_1 , but the figure only shows values up to 2.156 radians (123.5 degrees).

Similarly, the curvature plot is shown in Figure 4.11. For relatively small angles, its approximation to linear curvature is good, but sharp curvature peaks are apparent at large angles.

While the Euler spiral spline is defined as the one which is a piecewise Euler spiral and G^2 -continuous at the control points, the Hobby spline uses a slightly different criterion. Hobby introduces the concept of *mock curvature*, which is the linear approximation (taking only the first-order term of the exact equation) of endpoint curvature as a function of the endpoint angles θ_0 and θ_1 .

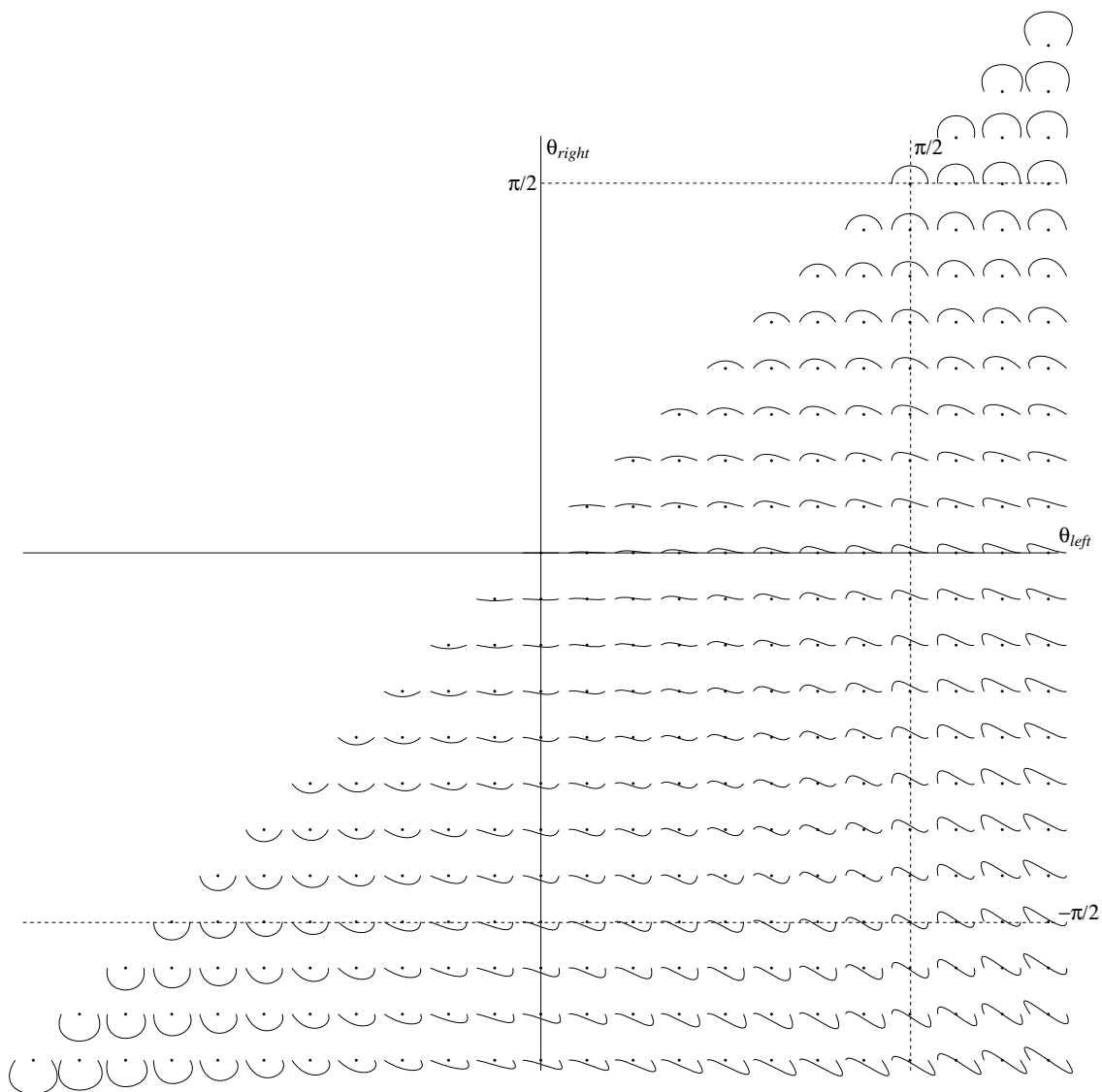


Figure 4.10. Hobby spline primitive over its parameter space.

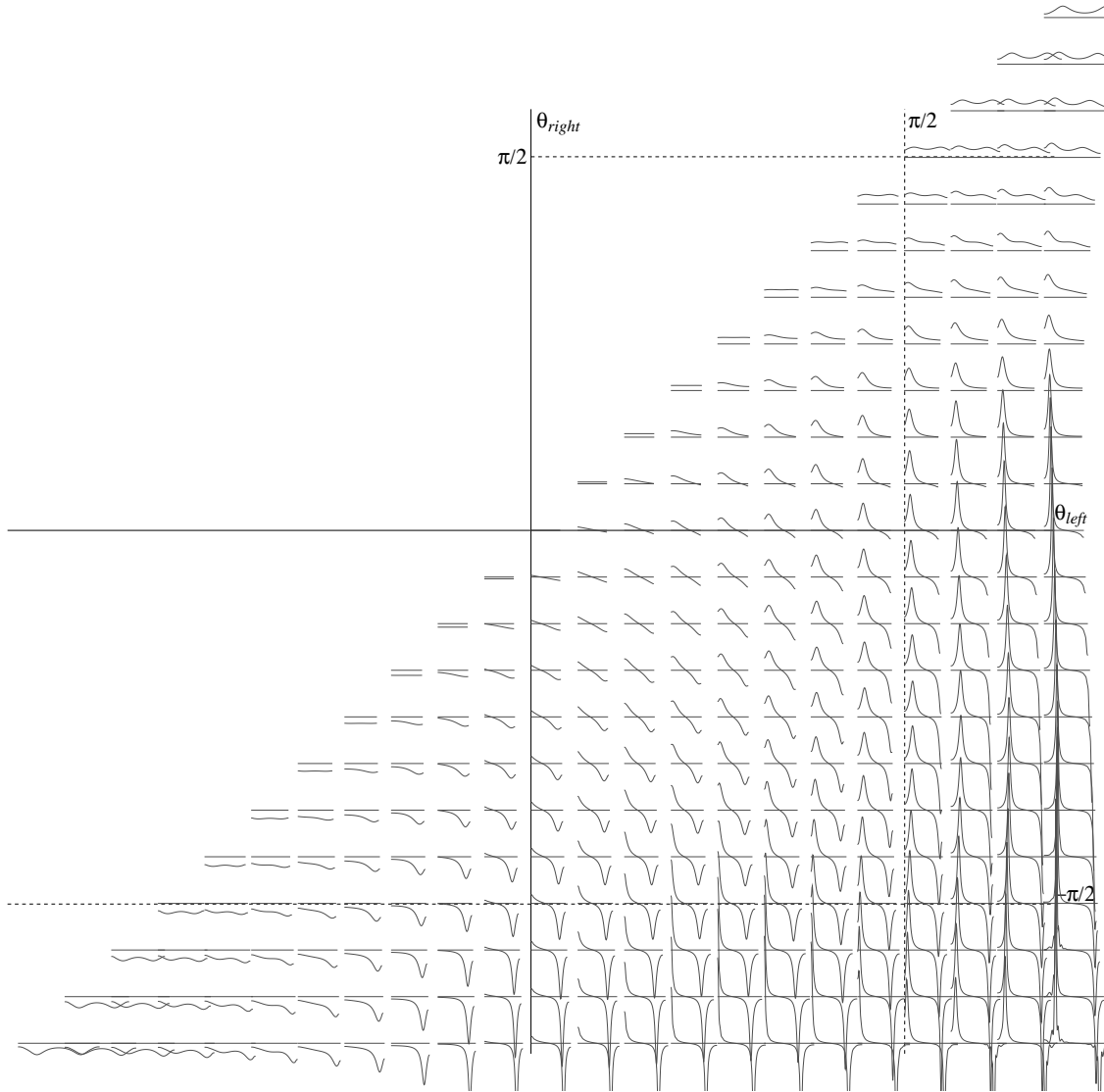


Figure 4.11. Curvature plot of Hobby spline over its parameter space.

The global solution to the spline is the assignment to the vector of tangent angles so that mock curvature is equal on either side of each knot.

A very important consequence is that solving the global spline is a *linear* problem. In fact, since each tangent angle affects two neighboring segments, and each segment in turn contributes to the mock curvature constraint at its two endpoints, the entire system of constraints can be represented as a tridiagonal linear equation, which is readily solvable in $O(n)$ using textbook techniques. Hobby also proves that the matrix is nonsingular over a reasonable range of values for τ , including $\tau = 1$. Thus, Hobby’s is one of the few “advanced” splines for which there is a solution for *all* configurations of input points.

Evaluating the spline properties, we find that the Hobby spline does fairly well. For small angles, it closely approximates the Euler spiral spline. It is not round, but converges to roundness as $O(n^6)$ as opposed to the MEC, which is only $O(n^4)$. Similarly, the extensionality and continuity of curvature are only approximate, and only good at small angles. However, as noted above, it is an exceedingly robust spline. It is thus a good choice if computational resources are limited, robustness is key, and the input control points are closely enough spaced that angles are small.

Another two-parameter curve family closely related to the Hobby spline is the primitive used in Potrace [79]. While Potrace is not an interpolating spline (rather, it is a tool for automatically converting bitmaps into vector curve outlines), its use of a primitive curve family is conceptually similar. Like the Hobby spline, the potrace primitive curve family is based on cubic Bézier curves (inherently four parameters), with mathematical equations to determine the additional two parameters.

4.8 Splines from fixed generating curves

The MEC has the property that all segments between control points can be cut from a single generating curve, the rectangular elastica (here, “cut” means selecting points on the generating curve, then performing the unique rotation, scaling and translation to make these endpoints match the corresponding control points). This property can be derived from the equilibrium-of-forces formulation of the elastica, and the fact that the constraints at the control points exert zero tension, but these facts are not intuitively obvious.

Similarly, in the Euler spiral spline, segments between control points are cut from the Euler spiral (“cut” may also include a mirror flip, because this curve lacks a mirror symmetry).

What do splines cut from a single generating curve have in common? Why do so few other splines have this property? Might there be some additional flexibility that can be gained by choosing segments from a curve family rather than a fixed generating curve?

The key to answering these questions is extensionality. In fact, as demonstrated in Section 4.1, the class of two-parameter extensional splines is exactly equal to the class of two-parameter splines cut from a generating curve.

Other properties, such as roundness, can then be derived from properties of the generating curve. For example, for the spline to be round, the curve must tend to a circular shape in the limit. More precisely, $\lim_{s \rightarrow \infty} \kappa'(s)/\kappa(s)^2 = 0$. The Euler spiral meets this condition because κ tends to infinity, while κ' is fixed. Another possibility would be a curve tending to fixed curvature in the

asymptote, in which case κ' would tend to 0. The MEC, by contrast, clearly does not have this property.

The envelope of endpoint chord angles (as plotted in Figure 3.13 for the MEC) also has an effect. A small envelope yields failure to exist. A generating curve without an inflection (such as a parabola) excludes all antisymmetric endpoint angle configurations from this envelope, even for tiny angles.

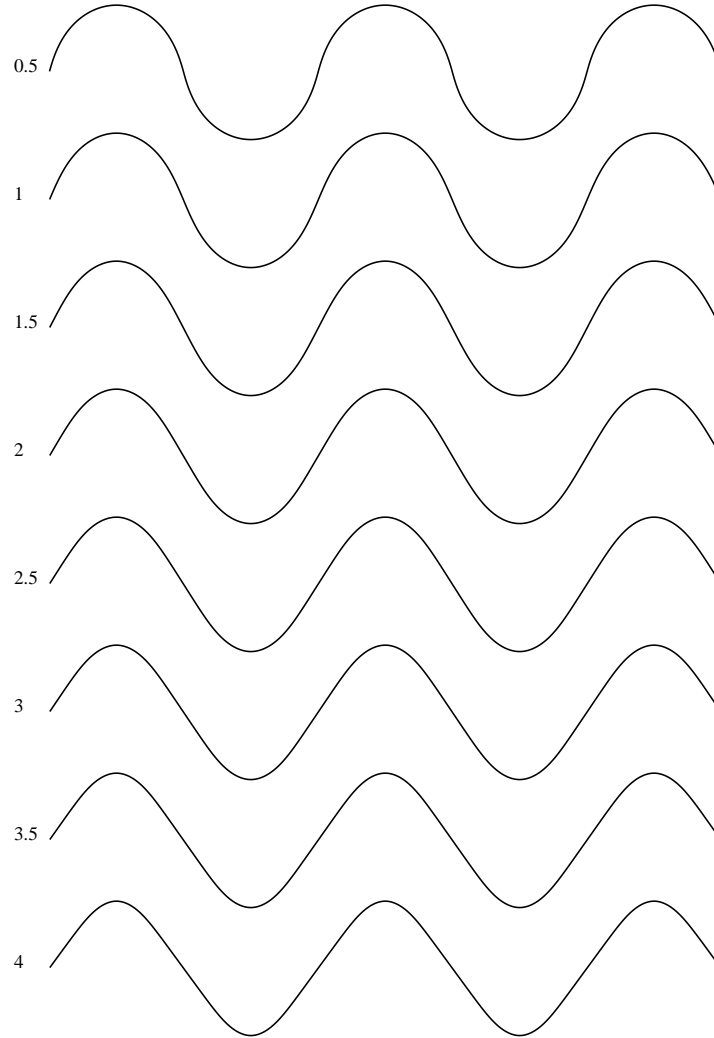


Figure 4.12. Test instrument for fairness user study.

4.8.1 Log-aesthetic curve splines

A very promising family of candidates for generator curves is the *aesthetic curve*, also known as the *log-aesthetic curve* [63]. These curves are defined as having a straight line when κ'/κ is plotted against κ on a log-log scale. In addition to a number of special cases (including the equiangular spiral and Nielsen's spiral), the general formula for aesthetic curves is $\kappa = s^\alpha$, where the exponent

α is the parameter that distinguishes the members of this family. An intuitive motivation for the aesthetic curve is that the eye is not as sensitive to absolute variation of curvature as it is to relative variation. In areas of high curvature, high variation of curvature is more easily tolerated; but in low-curvature areas, even small variations in curvature are visible.

The literature on aesthetic curves doesn't tend to include curves with inflection points – on a log-log plot, these are off the scale. However, since it is well known that the aesthetic curve with $\alpha = 1$ is the Euler spiral, curves with inflection points are known to be part of the family, but it's possible to generalize the formula to $\kappa = \text{sign}(s)|s|^\alpha$ to admit a continuous range of exponents. A family of such curves (used to interpolate a zigzag sequence of control points) is shown in Figure 4.12.

As described in Section 2.1.1, an empirical user study establishes that MEC energy is not an accurate predictor of perceived fairness. MEC energy is minimized at $\alpha = 0.9$, so if MEC energy were an accurate predictor of fairness, that would have been the preferred value. Since the mean preferred value was $\alpha = 1.78$, that hypothesis is refuted.

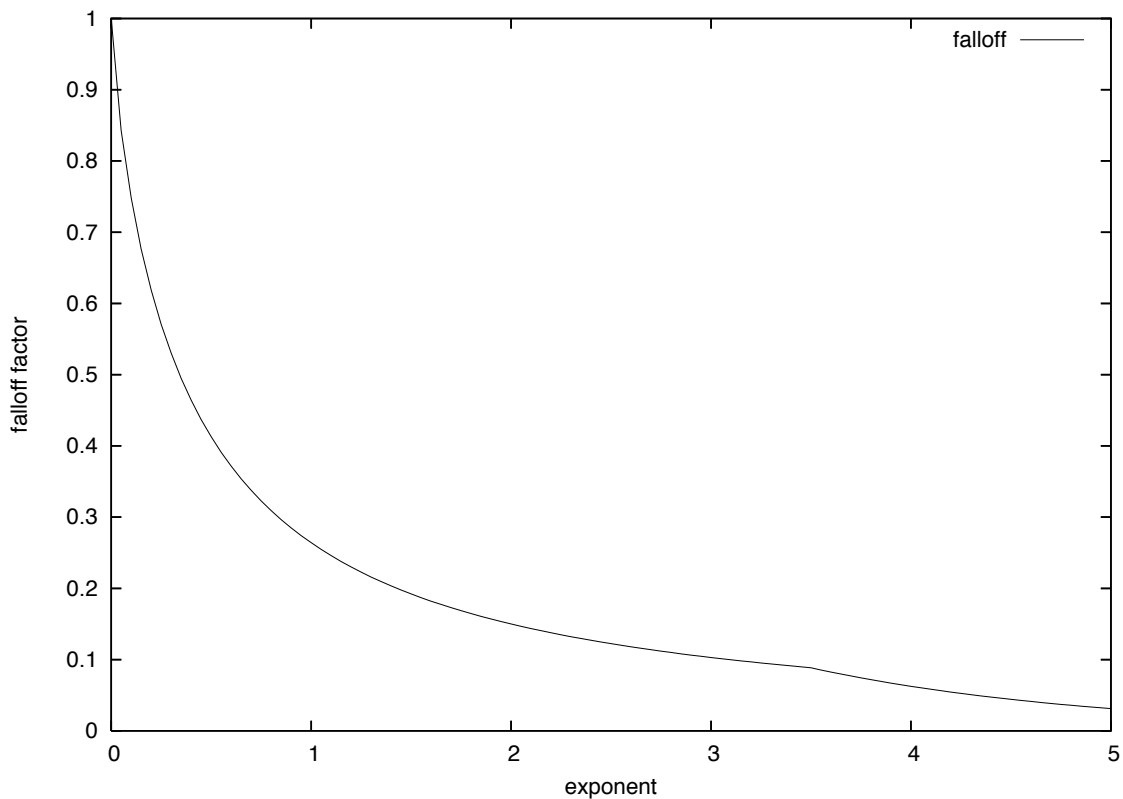


Figure 4.13. Perturbation fall-off rate as a function of aesthetic curve exponent.

The exponent of the aesthetic curve used as the generator curve in a spline also has a profound effect on the locality of the resulting spline. In general, the weaker the curvature variation near the inflection point of the generator curve, the better the locality. The aesthetic curve family varies from very steep to very shallow curvature variation across its parameter space. The falloff factor for a single perturbed point, a quantitative measurement of locality, is plotted against the exponent of the aesthetic curve in Figure 4.13. For $\alpha = 1$, which represents the Euler spiral, the amplitude

of the ripple diminishes by 0.268 from one crest to the next one. This ratio can be determined analytically by using the small-angle approximation and solving for $\kappa_0\theta_1 = \kappa_1\theta_0$ for the endpoints of a segment, which yields a value of $1/(2 + \sqrt{3})$ for the Euler spiral. Note also that the MEC spline has the same locality properties as the Euler spiral spline. For the user-preferred exponent of 1.78, the falloff factor is 0.173, noticeably smaller and thus better. It is possible to achieve even better locality at the expense of higher curvature variation, if that is desired.

4.9 Piecewise SI-MEC splines, or Kurgla 1

The Autokon system of Mehlum [59] was inspired by the goal of simulating a real, mechanical spline, but its KURGLA 1 algorithm actually implemented something a bit different, with both advantages and disadvantages with respect to the MEC it was intended to mimic.

Mehlum based his design on the fact that the curvature of an elastica is linear in distance from the line of action. His solver approximated the integration of this intrinsic equation as a sequence of circular arcs, the curvature of each determined by the distance from the line of action.

As we’ve seen, the general elastica has three degrees of freedom; two for the position and angle of the line of action with respect to the control point chord, and one for the amplitude of the curvature variation. Note that inflections occur at the line of action. In this formulation, choosing all these parameters to minimize the total bending energy is a tricky global optimization problem, and KURGLA 1 did not attempt it.

In particular, it nailed down the *angle* of the line of action, choosing it to be perpendicular to the chord. In this way, the parameter space was reduced from three to two, and the problem became much more tractable. In particular, the curve segments were determined by the tangents at the endpoints, and the remaining global problem of choosing tangents for G^2 -continuity yields to simple Newton-style solution.

As shown in Section 3.11, an elastica segment with the line of action perpendicular to the chord is the one minimizing the SI-MEC energy for the angle constraints at those endpoints. Thus, the KURGLA 1 algorithm effectively computes a piecewise SI-MEC.

One consequence is that KURGLA 1 is round, unlike the pure MEC, since the SI-MEC describes a circular arc in the case of symmetrical endpoint angle constraints. Another consequence is considerably better robustness than the pure MEC, because the envelope of parameter space is approximately twice as large.

However, the price paid for these improvements is loss of extensionality; when a new on-curve point is added, it generally changes the chords, and thus the lines of action for the resulting elastica sections.

Thus, though it is based on the elastica, and has some desirable properties, KURGLA 1 is neither a true energy minimizing spline like the MEC, nor does its extensionality and robustness compare well to the Euler spiral spline. Indeed, Mehlum was aware of these limitations, and his subsequent KURGLA 2 algorithm is the earliest known implementation of the Euler spiral spline.²

The MEC, SI-MEC, and Euler spiral are closely related. One way to visualize this relation is

²Mehlum wrote to George Forsythe on July 9, 1971, mentioning “Fresnel integral splines” as recent work. The work was published in 1974 [59].

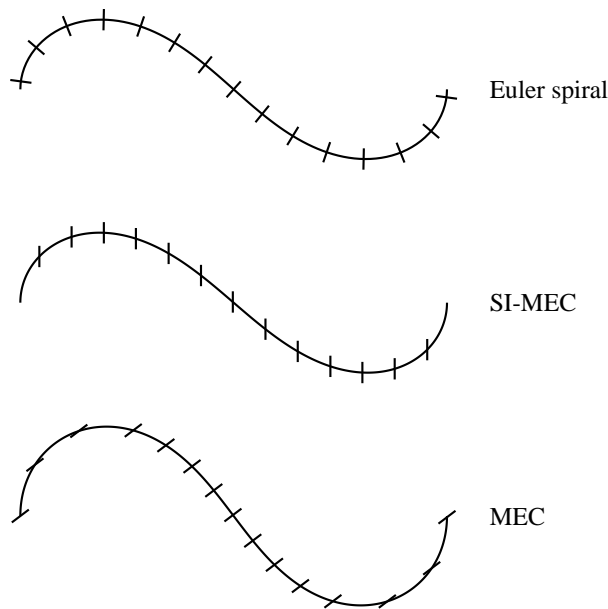


Figure 4.14. MEC, SI-MEC, and Euler spiral compared.

to consider the gradient of curvature. In an Euler spiral, curvature increases linearly along the arc length of the curves. In a SI-MEC defined by two endpoint tangents, curvature increases linearly along the chord between the endpoints. And in a pure MEC, curvature increases along the line tangent to the inflection point. Figure 4.14 shows these gradients of curvature visually – along each of the curves, tick marks meter the rate of change of curvature, and the tick mark is angled so that the rate of curvature change is constant between parallel lines. This figure also shows that the SI-MEC and Euler spiral are visually very similar.

4.10 Circle splines

Interpolating splines have a tradeoff between extensionality and locality. The two-parameter splines described above favor extensionality. It is possible to make different choices, and favor locality instead. An excellent example is the “circle spline” of Séquin, Lee, and Yin [80].

In the circle spline, the tangent angle at each control point is chosen as the tangent of the circle going through the control point and its two neighbors. Each segment is then constructed using a blending of these circles from the tangents at the endpoints. Thus, moving a control point only affects four segments – two on either side.

The construction of the primitive curves is robust and smooth. In fact, the construction guarantees zero κ' at the endpoints, and that the curvature is identical to that of the circle used to determine tangents, so the resulting spline has G^3 -continuity.

Unfortunately, even though the spline has good locality and G^3 -continuity, it is not particularly smooth for all inputs, and adding additional on-curve control points can cause serious deviations

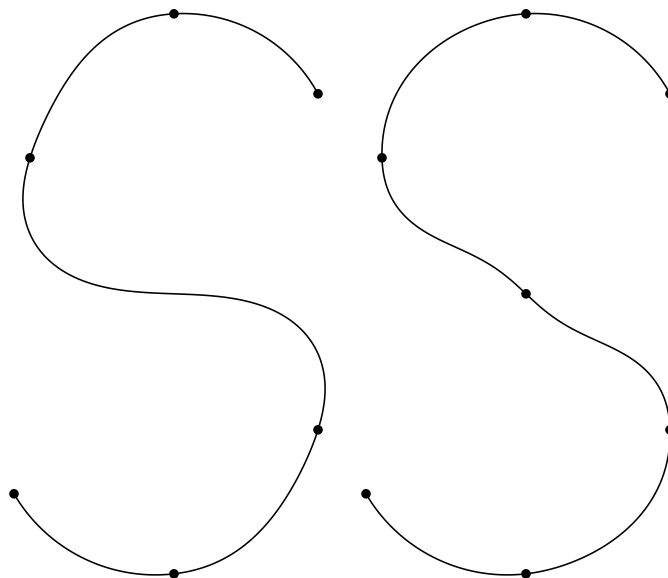


Figure 4.15. Circle spline extensionality failure.

from extensionality, in many cases adding inflections. Figure 4.15 clearly shows the lack of extensionality; the two paths differ only by the addition of the central point in the right-hand figure.

4.11 Summary

This chapter described an important class of interpolating splines, those with a primitive curve described by a two-dimensional parameter space. A great many splines in the literature fall into this category, notably the Minimum Energy Curve.

A central result of this work is that any two-parameter, extensional spline is defined in terms of a fixed generating curve. These properties hold across a large family of desirable splines, and makes the choice of an optimum spline a much simpler problem. Choosing a most preferred generating curve can be done empirically. By contrast, the variational approach would require developing an accurate model of the perception of curve fairness. Empirical testing establishes that the most common such metric, bending energy, does not coincide accurately with user preference.

Table 4.11 summarizes the properties of several splines in this two-parameter family. All of these are known and appear in the literature.

- MEC is the interpolating spline minimizing the total bending energy. Each segment between control points is a piece of a rectangular elastica, and the tangents are chosen so that G^2 continuity is preserved across each control point. (Section 3.10)
- Euler spiral is the interpolating spline for which each segment between control points is a section of an Euler spiral, i.e. curvature is linear in arc length. Like the MEC, the tangents are chosen for G^2 -continuity. (Section 6.8)

Curve type	Tangents	Continuity	Extensionality	Roundness	Robustness	Locality
MEC	G^2	G^2	+++	--	--	+
Euler spiral	G^2	G^2	+++	+++	++	+
log-aesthetic	G^2	G^2	+++	+++	-	++
Hobby	$\approx G^2$	$\approx G^2$	-	-	+++	+
Biarc (Ikarus)	Lagrange	G^1	--	+++	++	+
KURGLA 1	G^2	G^2	-	+++	+	+
Circle	local	G^3	---	+++	+++	+++

Table 4.1. Comparison of properties of two-parameter splines.

- log-aesthetic is the two-parameter extensible spline based on log-aesthetic curves with an exponent near $\alpha = 2$. (Section 4.8.1)
- Ikarus (Section 4.5).
- Hobby is a piecewise cubic spline developed for the Metafont system. It is best understood as a polynomial approximation to the Euler spiral spline, with better robustness but lower performance on other metrics such as smoothness and extensionality [37] (Section 4.7).
- KURGLA 1 (Section 4.9).
- Circle spline (Section 4.10).

For the basic interpolating spline problem, two-parameter extensible splines seem likely to be the best possible choice. From this family, while the MEC (based on the rectangular elastica) is the original inspiration for virtually the entire field of splines, the Euler spiral spline is superior and has excellent all-around properties, and is especially well-suited for interactive use due to its stability. The elastica and Euler spiral are particularly important and beautiful curves. In addition, they have a fascinating and intertwined history, explored in detail in the next two chapters.

Chapter 5

History of the elastica

This chapter tells the story of a remarkable family of curves, known as the elastica, Latin for a thin strip of elastic material. The elastica caught the attention of many of the brightest minds in the history of mathematics, including Galileo, the Bernoullis, Euler, and others. It was present at the birth of many important fields, most notably the theory of elasticity, the calculus of variations, and the theory of elliptic integrals. The path traced by this curve illuminates a wide range of mathematical style, from the mechanics-based intuition of the early work, through a period of technical virtuosity in mathematical technique, to the present day where computational techniques dominate.

There are many approaches to the elastica. The earliest (and most mathematically tractable), is as an equilibrium of moments, drawing on a fundamental principle of statics. Another approach, ultimately yielding the same equation for the curve, is as a minimum of bending energy in the elastic curve. A force-based approach finds that normal, compression, and shear forces are also in equilibrium; this approach is useful when considering specific constraints on the endpoints, which are often expressed in terms of these forces.

Later, the fundamental differential equation for the elastica was found to be equivalent to that for the motion of the simple pendulum. This formulation is most useful for appreciating the curve's periodicity, and also helps understand special values in the parameter space.

In more recent times, the focus has been on efficient numerical computation of the elastica (especially for the application of fitting smooth spline curves through a sequence of points), and also determining the range of endpoint conditions for which a stable solution exists. Many practical applications continue to use brute-force numerical techniques, but in many cases the insights of mathematicians (many working hundreds of years earlier) still have power to inspire a more refined computational approach.

5.1 Jordanus de Nemore – 13th century

In the recorded literature, the problem of the elastica was first posed in *De Ratione Ponderis* by Jordanus de Nemore (Jordan of the Forest), a thirteenth century mathematician. Proposition 13 of book 4 states that “when the middle is held fast, the end parts are more easily curved.” He then poses an incorrect solution: “And so it comes about that since the ends yield most easily, while the other parts follow more easily to the extent that they are nearer the ends, the whole body becomes curved in a circle.” [17]. In fact, the circle is one possible solution to the elastica, but not to the specific problem posed. Even so, this is a clear statement of the problem, and the solution (though not correct) is given in the form of a specific mathematical curve. It would be several centuries until the mathematical concepts needed to answer the question came into existence.

5.2 Galileo sets the stage – 1638

Three basic concepts are required for the formulation of the elastica as an equilibrium of moments: moment (a fundamental principle of statics), the curvature of a curve, and the relationship between these two concepts. In the case of an idealized elastic strip, these quantities are linearly related.

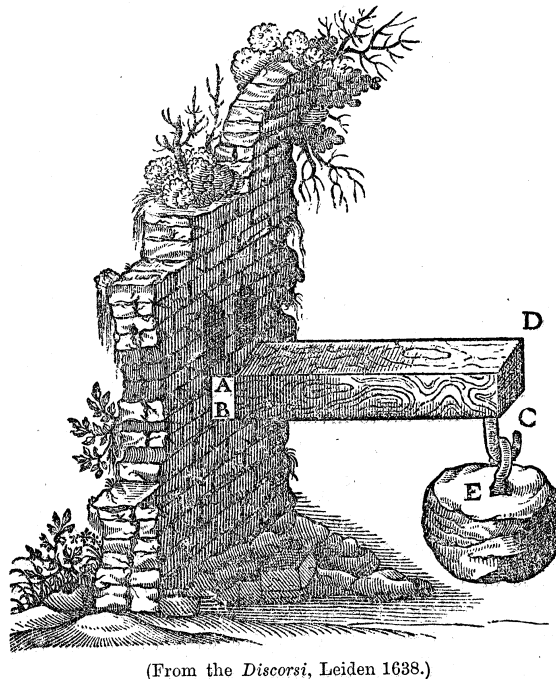


Figure 5.1. Galileo’s 1638 problem.

Galileo, in 1638, posed a fundamental problem, founding the mathematical study of elasticity. Given a prismatic beam set into a wall at one end, and loaded by a weight at the other, how much weight is required to break the beam? The delightful Figure 5.1 illustrates the setup. Galileo

considers the beam to be a compound lever with a fulcrum at the bottom of the beam meeting the wall at B. The weight E acts on one arm BC, and the thickness of the beam at the wall, AB, is the other arm, “in which resides the resistance.” His first proposition is, “The moment of force at C to the moment of the resistance... has the same proportion as the length CB to the half of BA, and therefore the absolute resistance to breaking... is to the resistance in the same proportion as the length of BC to the half of AB...”

From these basic principles, Galileo derives a number of results, primarily a scaling relationship. He does not consider displacements of the beam; for these types of structural beams, the displacement is negligible. Even so, this represents the first mathematical treatment of a problem in elasticity, and firmly establishes the concept of moment to determine the force on an elastic material. Many researchers elaborated on Galileo’s results in coming decades, as described in detail in Todhunter’s history [85].

One such researcher is Ignace-Gaston Pardies, who in 1673 posed one form of the elastica problem and also attempted a solution: for a beam held fixed at one end and loaded by a weight at the other, “it is easy to prove” that it is a parabola. However, this solution isn’t even approximately correct, and would later be dismissed by James Bernoulli as one of several “pure fallacies.” Truesdell gives Pardies credit for introducing the elasticity of a beam into the calculation of its resistance, and traces out his influence on subsequent researchers, particularly Leibniz and James Bernoulli [86, p. 50–53].

5.3 Hooke’s law of the spring – 1678

Hooke published a treatise on elasticity in 1678, containing his famous law. In a short Latin phrase posed in a cryptic anagram three years earlier to establish priority, it reads, “*ut tensio sic vis*”; that is, The Power of any Spring is in the same proportion with the Tension thereof... Now as the Theory is very short, so the way of trying it is very easie.” (Spelling and capitalization as in the original). In modern formulation, it is understood as $F \propto \Delta l$; the applied force is proportional to the change in length.

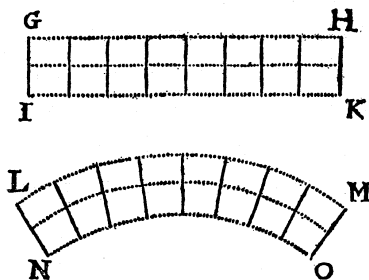


Figure 5.2. Hooke’s figure on compound elasticity.

Hooke touches on the problem of elastic strips, providing an evocative illustration (Figure 5.2), but according to Truesdell, “This ‘compound way of springing’ is the main problem of elasticity

for the century following, but Hooke gives no idea how to relate the curvature of one fibre to the bending moment, not to mention the reaction of the two fibres on one another.” [86, p. 55]

Indeed, mathematical understanding of curvature was still in development at the time, and mastery over it would have to wait for the calculus. Newton published results on curvature in his “Methods of Series and Fluxions,” written 1670 to 1671 [35, p. 232], but not published for several more years. Leibniz similarly used his competing version of the calculus to derive similar results. Even so, some results were possible with pre-calculus methods, and in 1673, Christiaan Huygens published the “Horologium oscillatorium sive de motu pendulorum ad horologia aptato demonstrationes geometrica,” which used purely geometric constructions, particularly the involutes and evolutes of curves, to establish results involving curvature. The flavor of geometric construction pervades much of the early work on elasticity, particularly James Bernoulli’s, as we shall see.

Newton also used his version of the calculus to more deeply understand curvature, and provided the formulation for radius of curvature in terms of Cartesian coordinates most familiar to us today [35],

$$\rho = (1 + y'^2)^{\frac{3}{2}}/y'' . \quad (5.1)$$

An accessible introduction to the history of curvature is the aptly named “History of Curvature” by Dan Margalit [56].

Given a solid mathematical understanding of curvature, and assuming a linear relation between force and change of length, working out the relationship between moment and curvature is indeed “easie,” but even Bernoulli had to struggle a bit with both concepts. For one, Bernoulli didn’t simply accept the linear law of the spring, but felt the need to test it for himself. And, since he had the misfortune to use catgut, rather than a more ideally elastic material such as metal, he found significant nonlinearities. Truesdell [88] recounts a letter from James Bernoulli to Leibniz on 15 December 1687, and the reply of Leibniz almost three years later. According to Truesdell, this exchange is the birth of the theory of the “curva elastica” and its ramifications. Leibniz had proposed: “From the hypothesis elsewhere substantiated, that the extensions are proportional to the stretching forces...” which is today attributed as Hooke’s law of the spring. Bernoulli questioned this relationship, and, as we shall see, his solution to the elastica generalized even to nonlinear displacement.

5.4 James Bernoulli poses the elastica problem – 1691

James Bernoulli posed the precise problem of the elastica in 1691:

“Si lamina elastica gravitatis espers AB, uniformis ubique crassitie & latitudinis, inferiore extremitate A alicubi firmetur, & superiori B pondus appendatur, quantum sufficit ad laminam eousque incurvandam, ut linea directionis ponderis BC curvatae laminae in B sit perpendicularis, erit curvatura laminae sequentis naturae:”

And then in cipher form:

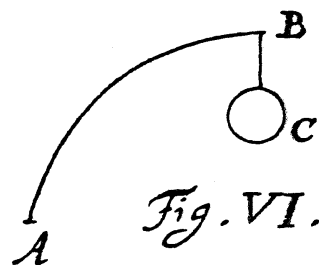


Figure 5.3. Bernoulli poses the elastica problem.

“Portio axis applicatam inter et tangentem est ad ipsam tangentem sicut quadratum applicatae ad constans quoddam spatium.”¹

Assuming a lamina AB of uniform thickness and width and negligible weight of its own, supported on its lower perimeter at A, and with a weight hung from its top at B, the force from the weight along the line BC sufficient to bend the lamina perpendicular, the curve of the lamina follows this nature:

The rectangle formed by the tangent between the axis and its own tangent is a constant area.

This poses one specific instance of the general elastica problem, now generally known as the *rectangular elastica*, because the force applied to one end of the curve bends it to a right angle with the other end held fixed.

The deciphered form of the anagram is hardly less cryptic than the original, but digging through his 1694 explanation, it is possible to extract the fundamental idea: at every point along the curve, the product of the radius of curvature and the distance from the line BC is a constant, i.e. the two quantities are inversely proportional. And, indeed, that is the key to unlocking the elastica; the equation for the shape of the curve follows readily, given sufficient mathematical skill.

5.5 James Bernoulli partially solves it – 1692

By 1692, James Bernoulli had completely solved the rectangular case of the elastica posed earlier. In his *Meditatione CLXX* dated that year, titled “*Quadratura Curvae, e cujus evolutione describitur inflexae laminae curvatura*” [7], or, “Quadrature of a curve, by the the evolution of which is traced out the curve of a bent lamina” he draws a figure of that curve (reproduced here as Figure 5.4), and gives an equation for its quadrature².

¹The cipher reads “Qrzzumu bapt dxqopddbbp...” and the key was published in the 1694 *Curvatura Laminae* with the detailed solution to the problem. Such techniques for establishing priority may seem alien to academics today, but are refreshingly straightforward by comparison to the workings of the modern patent system.

²Today, we would say simply “integral” rather than quadrature.

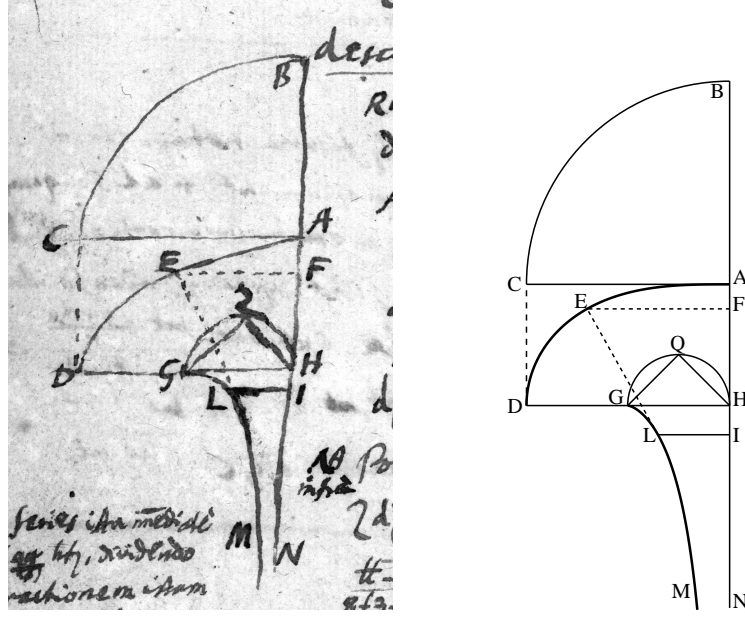


Figure 5.4. Drawing from James Bernoulli's 1692 Med. CLXX, with modern reconstruction.

Radius circuli $AB = a$, AED lamina elastica ab appenso pondere in A curvata, GLM illa curva, ex cujus evolutione AED describitur: $AF = y$, $FE = x$, $AI = p$, $IL = z$. Aequatio differentialis naturam curvae AED exprimens,

$$dy = \frac{xx \, dx}{\sqrt{a^4 - x^4}},$$

ut suo tempore ostendam:

A rough translation:

Let the radius of the circle $AB = a$, AED be an elastic lamina curved by a suspended weight at A , and GLM that curve, the involute of which describes AED . $AF = y$, $FE = x$, $AI = p$, $IL = z$. The differential equation for the curve AED is expressed,

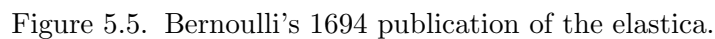
$$dy = \frac{x^2 \, dx}{\sqrt{a^4 - x^4}}, \quad (5.2)$$

as I will show in due time.

Equation 5.2 is a clear, simple statement of the differential equation for the rectangular instance of the elastica family, presented in readily computable form; y can be obtained as the integral of a straightforward function of x .

Based on the figure, AED is the elastica, and GLM is its evolute. Thus, the phrase “evolutione describitur” appears to denote taking the *involute* of the curve GLM to achieve the elastica AED ;

5.6 James Bernoulli publishes the first solution – 1694



73

1694 *Curvatura Laminae Elasticae*³. See Truesdell [86, pp. 88–96] for a detailed description of this work, which we will only outline here.

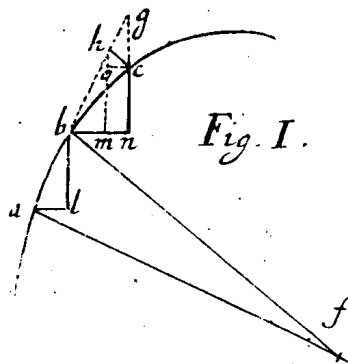


Figure 5.6. Bernoulli's justification for his formula for curvature.

Bernoulli begins the paper by giving a general equation for curvature (illustrated in Figure 5.6), which he introduces thus, “in simplest and purely differential terms the relation of the evolvent of radius of the osculating circle of the curve... Meanwhile, since the immense usefulness of this discovery in solving the velaria, the problem of the curvature of springs we here consider, and other recondite matters makes itself daily more and more manifest to me, the matter stands that I cannot longer deny to the public the golden theorem.” Bernoulli’s formulation is not entirely familiar to the modern reader, as it mixes infinitesimals somewhat promiscuously. He set out this formula for the radius of curvature z :

$$z = \frac{dx \, ds}{ddy} = \frac{dy \, ds}{ddx} \quad (5.3)$$

From Bernoulli’s tone, it is clear he thought this was an original result, but Huygens and Leibniz were both aware of similar formulas; Huygens had already published a statement and proof in terms of pure Cartesian coordinates (which would be the form most familiar today). However, in spite of this knowledge, both considered the problem of the elastica impossibly difficult. Huygens, in a letter to Leibniz dated 16 November 1691, wrote, “I cannot wait to see what Mr. Bernoulli the elder will produce regarding the curvature of the spring. I have not dared to hope that one would come out with anything clear or elegant here, and therefore I have never tried.” [86, p. 88, footnote 4]

Bernoulli’s treatment of the elastica is fairly difficult going (as evidenced by the skeptical reaction and mistaken conclusions from Leibniz and others), but again, it is possible to tease out the central ideas. First, the idea that the moment at any point along the curve is proportional to the

³*Curvatura Laminae Elasticae. Ejus Identitas cum Curvatura Lintei a pondere inclusi fluidi expansi. Radii Circulorum Osculantium in terminis simplicissimis exhibiti, una cum novis quibusdam Theorematis huc pertinentibus, &c.*, or “The curvature of an elastic band. Its identity with the curvature of a cloth filled out by the weight of the included fluid. The radii of osculating circles exhibited in the most simple terms; along with certain new theorems thereto pertaining, etc.”. Originally published in the June 1694 *Acta Eruditorum* (pp. 262–276), it is collected in the 1744 edition of his *Opera* [8, p. 576–600], now readily accessible online.

distance from the line of force. Bernoulli writes in a 1695 paper⁴ explaining the 1694 publication (and referring to Figure 5.5 for the legend): “I consider a lever with fulcrum Q , in which the thickness Qy of the band forms the shorter arm, the part of the curve AQ the longer. Since Qy and the attached weight Z remain the same, it is clear that the force stretching the filament y is proportional to the segment QP .” Next, Bernoulli carefully separated out the force and the elongation, and, in fact, allows for a completely arbitrary functional relationship, not necessarily linear (this function is represented by the curve AFC in Figure 5.5, labeled “Linea Tensionum”). “And since the elongation is reciprocally proportional to Qn , which is plainly the radius of curvature, it follows that Qn is also reciprocally proportional to x .” A central argument here is that the moment (and hence curvature, assuming a linear relationship) is proportional solely to the amount of force and the distance from the line of that force; the shape of the curve doesn’t matter.

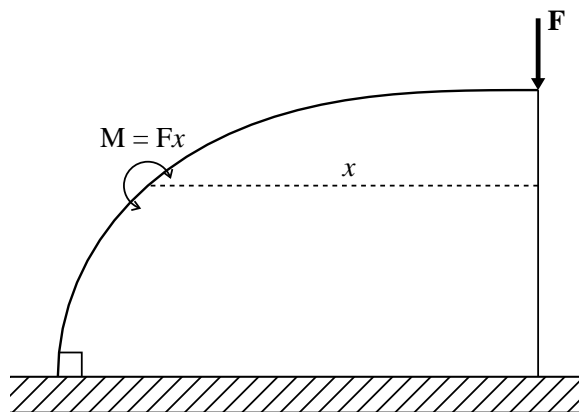


Figure 5.7. A simplified diagram of moments.

The moments can be seen in the simplified diagram in Figure 5.7, which shows the curve of the elastica itself (corresponding to AQR in Figure 5.5), the force F on the elastica (corresponding to AZ in Figure 5.5), and the distance x from the line of force (corresponding to PQ in Figure 5.5). According to the simple lever principle of statics, the moment is equal to the force F applied on the elastica times the distance x from the line of force.

Here we present a slightly simplified version of Bernoulli’s argument (see Truesdell [86, p. 92] for a more complete version). Write the curvature as a function of x , understanding that in the idealized case it is linear, i.e., $f(x) = cx$. Then, using Equation 5.3, the “golden theorem” for curvature:

$$\frac{d^2y}{dx ds} = f(x) . \quad (5.4)$$

Integrating with respect to dx and assuming $dy/ds = 0$ at $x = s = 0$ gives

⁴ *Explicationes, Annotationes et Additiones ad eas quae in actis superiorum annorum de curva elastica, isochrona paracentrui, et velaria, hinc inde memorata, et partim controversa leguntur; ubi de linea mediarum directionum, aliisque novis.* Originally published in *Acta. Eruditorum*, Dec. 1695, pp. 537–553, and reprinted in the 1744 *Complete Works* [8, pp. 639–663].

$$\frac{dy}{ds} = \int_0^x f(\xi) d\xi = S(x). \quad (5.5)$$

Substituting the standard identity (which follows algebraically from $ds^2 = dx^2 + dy^2$)

$$\frac{dy}{dx} = \frac{dy/ds}{\sqrt{1 - (dy/ds)^2}} \quad (5.6)$$

we get

$$\frac{dy}{dx} = \frac{S(x)}{\sqrt{1 - S(x)^2}}. \quad (5.7)$$

And in the ideal case where $f(x) = 2cx$, we have $S(x) = cx^2$, and thus

$$\frac{dy}{dx} = \frac{cx^2}{\sqrt{1 - c^2x^4}}. \quad (5.8)$$

Which is the same as the Equation 5.2, with straightforward change of constants. As mentioned before, the use of both ds and dx as the infinitesimal is strange by modern standards, even though in this case it leads to a solution quite directly. A parallel derivation using similar principles but only Cartesian coordinates can be found in Whewell’s 1833 *Analytical Statics* [89, p. 128]; there, Bernoulli’s use of ds rather than dx can be seen as the substitution of variable that makes the integral tractable.

It is worth noting that, while James Bernoulli’s investigation of the mathematical curve describing the elastica is complete and sound, as a more general work on the theory of elasticity there are some problems. In particular, Bernoulli assumes (incorrectly) that the inner curve of the elastica (AQR in Figure 5.5) is the “neutral fiber,” preserving its length as the elastica bends. In fact, determining the neutral fiber is a rather tricky problem whose general solution can still be determined only with numerical techniques (see Truesdell [86, pp. 96–109] for more detail). Fortunately, in the ideal case where the thickness approaches zero, the exact location of the neutral fiber is immaterial, and all that matters is the relation between the moment and curvature.

The limitations of Bernoulli’s approach were noted at the time [24]. Huygens published a short note in the *Acta eruditorum* in 1694, very shortly after Bernoulli’s publication in the same forum, illustrating several of the possible shapes the elastica might take on, and pointing out that Bernoulli’s quadrature only expressed the rectangular elastica. His accompanying figure is reproduced here as Figure 5.8. The shapes are shown from left to right in order of increasing force at the endpoints, and shape A is clearly the rectangular elastica.

Bernoulli acknowledged this criticism (while pointing out that he described these other cases explicitly in his paper, something that Huygens apparently overlooked) and indicated that his technique could be extended to handle these other cases (by using a non-zero constant for the integration of Equation 5.5), and went on to give an equation with the general solution, reproduced by Truesdell [86, p. 101] as:

$$\pm dy = \frac{(x^2 \pm ab)dx}{\sqrt{a^4 - (x^2 \pm ab)^2}} \quad (5.9)$$

TAB. X. ad A. 1694. pag. 339.

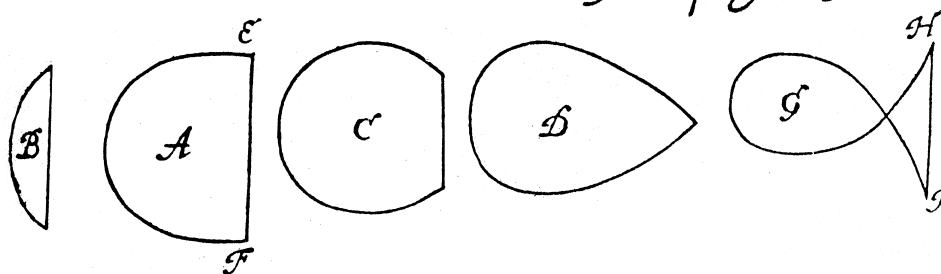


Figure 5.8. Huygens's 1694 objection to Bernoulli's solution.

It is not clear that the publication of this more general solution had much impact. Bernoulli did not graph the other cases, nor did he seem to be aware of other important properties of the solution, notably its periodicity, nor the fact that it includes solutions with and without inflections. (Another indication that these results were not generally known is that Daniel Bernoulli, in a November 8, 1738 letter to Leonhard Euler during a period of intense correspondence regarding the elastica, wrote, "Apart from this I have since noticed that the idea of my uncle Mr. James Bernoulli includes all elasticas." [86, p. 109])

Without question, the Bernoulli family set the stage for Euler's definitive analysis of the elastica. The next breakthrough came in 1742, when Daniel Bernoulli (the nephew of James) proposed to Euler solving the general elastica problem with the technique that would finally crack it: variational analysis. This general problem concerns the family of curves arising from an elastic band of arbitrary given length, and arbitrary given tangent constraints at the endpoints.

5.7 Daniel Bernoulli proposes variational techniques – 1742

Daniel Bernoulli, in an October 1742 letter to Euler [6], discussed the general problem of the elastica, but had not yet managed to solve it himself:

Ich möchte wissen ob Ew. die curvaturam laminae elasticae nicht könnten sub hac facie solviren, dass eine lamina datae longitudinis in duobus punctis positione datis fixirt sey, also dass die tangentes in istis punctis such positione datae seyen. ... Dieses ist die idea generalissima elasticarum; hab aber sub hac facie noch keine Solution gefunden, als per methodum isoperimetricorum, da ich annehme, dass die vis viva potentialis laminae elasticae insita müsse minima seyn, wie ich Ew. schon einmal gemeldet. Auf diese Weise bekomme ich eine aequationem differentialem 4ti ordinis, welche ich nicht hab genugsam reduciren können, um zu zeigen, dass die aequatio ordinaria elastica general sey.

A rough translation into English reads:

I'd like to know whether you might not solve the curvature of the elastic lamina under this condition, that on the length of the lamina on two points the position is fixed, and

that the tangents at these points are given. ... This is the idea of the general elastica; I have however not yet found a solution under this condition by the isoperimetric method, given my assumption that the potential energy of the elastic lamina must be minimal, as I've mentioned to you before. In this way I get a 4th order differential equation, which I have not been able to reduce enough to show a regular equation for the general elastica.

This previous mention is likely his 7 March 1739 letter to Euler, where he gave a somewhat less elegant formulation of the potential energy of an elastic lamina, and suggested the “isoperimetric method,” an early name for the calculus of variations. Many founding problems in the calculus of variations concerned finding curves of fixed length (hence isoperimetric), minimizing or maximizing some quantity such as area enclosed. Usually additional constraints are imposed to make the problem more challenging, but, even in the unconstrained case, though the answer (a circle) was known as early as Pappus of Alexandria around 300 A.D, rigorous proof was a long time coming.

In any case, Bernoulli ends the letter with what is likely the first clear mathematical statement of the elastica as a variational problem in terms of the stored energy⁵:

Ew. reflectiren ein wenig darauf, ob man nicht könne, sine interventu vectis, die curvaturam ABC immediate ex principiis mechanicis deduciren. Sonsten exprimire ich die vim vivam potentialem laminae elasticae naturaliter rectae et incurvatae durch $\int \frac{ds}{RR}$, sumendo elementum ds pro constante et indicando radium osculi per R . Da Niemand die methodum isoperimetricorum so weit perfectionniret, als Sie, werden Sie dieses problema, quo requiritur ut $\int \frac{ds}{RR}$ faciat minimum, gar leicht solviren.

You reflect a bit on whether one cannot, without the intervention of some lever, immediately deduce the curvature of ABC from the principles of mechanics. Otherwise, I'd express the potential energy of a curved elastic lamina (which is straight when in its natural position) through $\int \frac{ds}{RR}$, assuming the element ds is constant and indicating the radius of curvature by R . There is nobody as perfect as you for easily solving the problem of minimizing $\int \frac{ds}{RR}$ using the isoperimetric method.

Daniel was right. Armed with this insight, Euler was indeed able to definitively solve the general problem within the year (in a letter dated 4 September 1743, Daniel Bernoulli thanks Euler for mentioning his energy-minimizing principle in the “supplemento”), and this solution was published in book form shortly thereafter.⁶

5.8 Euler's analysis – 1744

Euler, building on (and crediting) the work of the Bernoullis, was the first to completely characterize the family of curves known as the elastica, and published this work as an Additamentum

⁵That said, James Bernoulli in 1694 expressed the equivalence of the elastica with the lintearia, which has a variational formulation in terms of minimizing the center of gravity of a volume of water contained in a cloth sheet, as pointed out by Truesdell [86, p. 201].

⁶Truesdell also points out that this formulation wasn't entirely novel; Daniel Bernoulli and Euler had corresponded in 1738 about the more general problem of minimizing $\int r^m ds$, and they seemed to be aware that the special case $m = -2$ corresponded to the elastica [86, p. 202]. The progress of knowledge, seen as a grand sweep from far away, often moves in starts and fits when seen up close.

(appendix) [20] in his landmark book on variational techniques. His treatment was quite definitive, and holds up well even by modern standards. This *Additamentum* is a truly remarkable work, deserving of deep study. Fortunately for the reader, it is accessible both in facsimile due to the Euler archive, and also in a 1933 English translation by Oldfather [68].

Closely following Daniel Bernoulli's suggestion, Euler expressed the problem of the elastica very clearly in variational form. He wrote [20, p. 247],

ut, inter omnes curvas ejusdem longitudinis, quæ non solum per puncta A & B transeant, sed etiam in his punctis a rectis positione datis tangantur, definiatur ea in qua sit valor hujus expressionis $\int \frac{ds}{RR}$ minimus.

In English:

That among all curves of the same length that not only pass through points A and B but are also tangent to given straight lines at these points, it is defined as the one minimizing the value of the expression $\int \frac{ds}{RR}$.

Here, s refers to arc length, exactly as is common usage today, and R is the radius of curvature (“radius osculi curvæ” in Euler’s words), or κ^{-1} in modern notation. Thus, today we are more likely to write that the elastica is the curve minimizing the energy $E[\kappa]$ over the length of the curve $0 \leq s \leq l$, where

$$E[\kappa(s)] = \int_0^l \kappa(s)^2 ds. \quad (5.10)$$

While today we would find it more convenient to work in terms of curvature (intrinsic equations), Euler quickly moved to Cartesian coordinates, using the standard definitions $ds = \sqrt{1 + y'^2} dx$ and $\frac{1}{R} = \frac{y''}{(1 + y'^2)^{3/2}}$, where y' and y'' represent dy/dx and d^2y/dx^2 , respectively. Thus, the variational problem becomes finding a minimum for

$$\int \frac{y''^2}{(1 + y'^2)^{5/2}} dx. \quad (5.11)$$

This equation is written in terms of the first and second derivatives of $y(x)$, and so the simple Euler–Lagrange equation does not suffice. Daniel Bernoulli had run into this difficulty, as he expressed in his 1742 letter. In hindsight, we now know that expressing the problem in terms of tangent angle as a function of arc length yields to a first-derivative variational approach, but this apparently was not clear to either Bernoulli or Euler at the time.

In any case, by 1744, Euler had discovered what is now known as the Euler–Poisson equation, capable of solving variational problems in terms of second derivatives, and he could apply it straightforwardly to Equation 5.11. (See Section 7.2.1 for more discussion of the Euler–Poisson equation and its application to a related variational problem.) He thus derived the following general equation, where a and c are parameters. (See Truesdell for a more detailed explanation of Euler’s derivation [86, pp. 203–204].)

$$\frac{dy}{dx} = \frac{a^2 - c^2 + x^2}{\sqrt{(c^2 - x^2)(2a^2 - c^2 + x^2)}}. \quad (5.12)$$

This equation is essentially the same as James Bernoulli's general solution, Equation 5.9.

Euler then goes on to classify the solutions to this equation based on the parameters a and c . To simplify constants, we propose the use of a single λ to replace both a and c ; this formulation has a particularly simple interpretation as the Lagrange multiplier for the straightforward variational solution of Equation 5.10.

$$\lambda = \frac{a^2}{2c^2} \quad (5.13)$$

Note also that this parameter has a straightforward geometrical interpretation. The tangent angle at the inflection point (relative to the line of action) is $\cos^{-1}(1 - .5/\lambda)$. Euler also worked with the slope of the curve at the inflection point, using the mathematically equivalent formulation of Equation 5.13.

Euler observed that there is an infinite variety of elastic curves, but that “it will be worth while to enumerate all the different kinds included in this class of curves. For this way not only will the character of these curves be more profoundly perceived, but also, in any case whatsoever offered, it will be possible to decide from the mere figure into what class the curve formed ought to be put. We shall also list here the different kinds of curves in the same way in which the kinds of algebraic curves included in a given order are commonly enumerated.” [20, §14]. According to the note in Oldfather's translation, Euler is referring to Newton's famous classification of cubic curves [68, p. 152, Note 5]. In any case, Euler finds nine such classes, enumerated in the table below:

Euler's species #	Euler's Figure	Euler's parameters	λ	
1		$c = 0$	$\lambda = \infty$	straight line
2	6	$0 < c < a$	$0.5 < \lambda$	
3		$c = a$	$\lambda = 0.5$	rectangular elastica
4	7	$a < c < a\sqrt{1.65187}$	$.30269 < \lambda < .5$	
5	8	$c = a\sqrt{1.65187}$	$\lambda = .30269$	lemnoid
6	9	$a\sqrt{1.65187} < c < a\sqrt{2}$	$.25 < \lambda < .30269$	
7	10	$c = a\sqrt{2}$	$\lambda = .25$	syntractrix
8	11	$a\sqrt{2} < c$	$0 < \lambda < .25$	
9		$a = 0$	$\lambda = 0$	circle

Table 5.1. Euler's characterization of the space of all elastica.

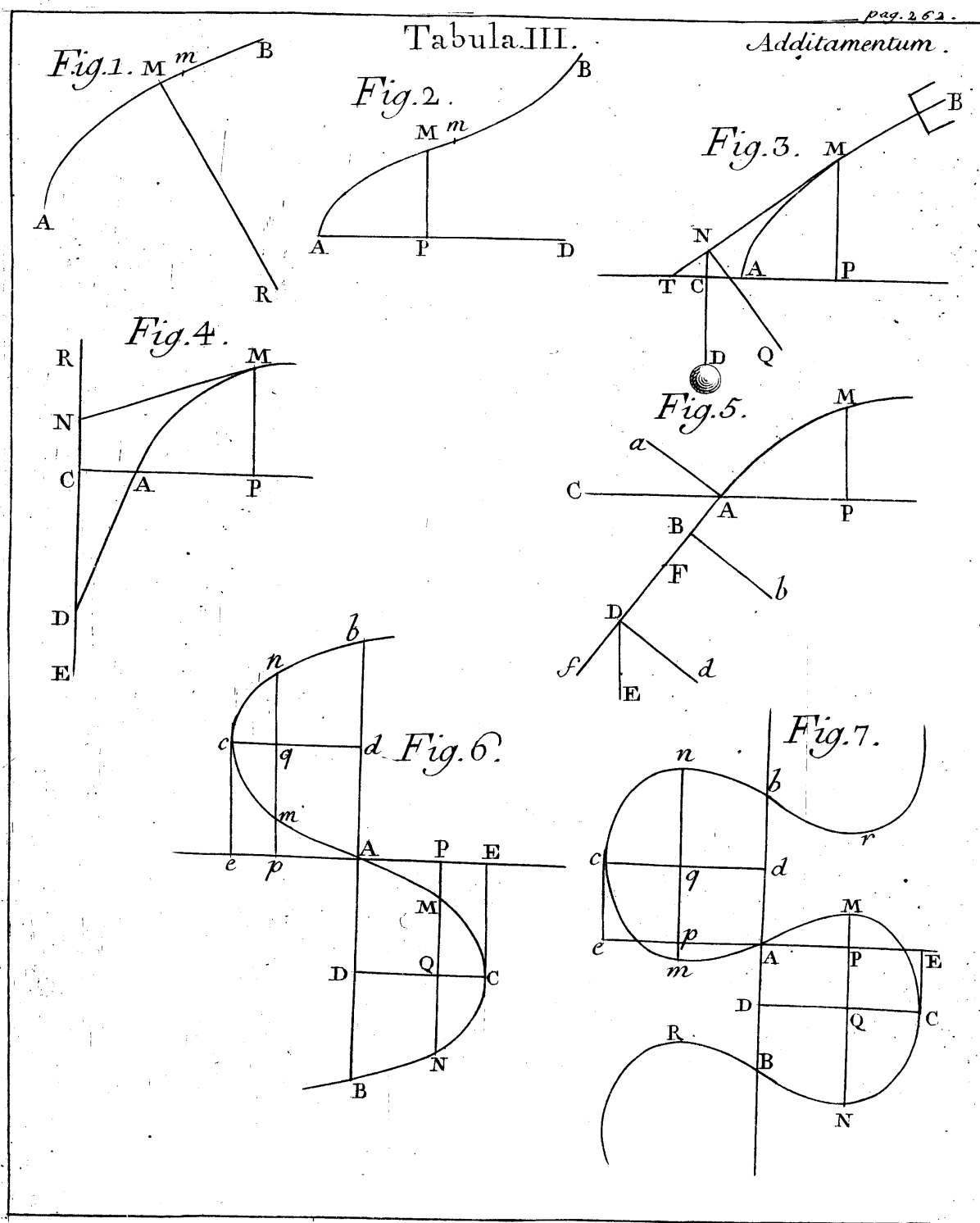


Figure 5.9. Euler's elastica figures, Tabula III.

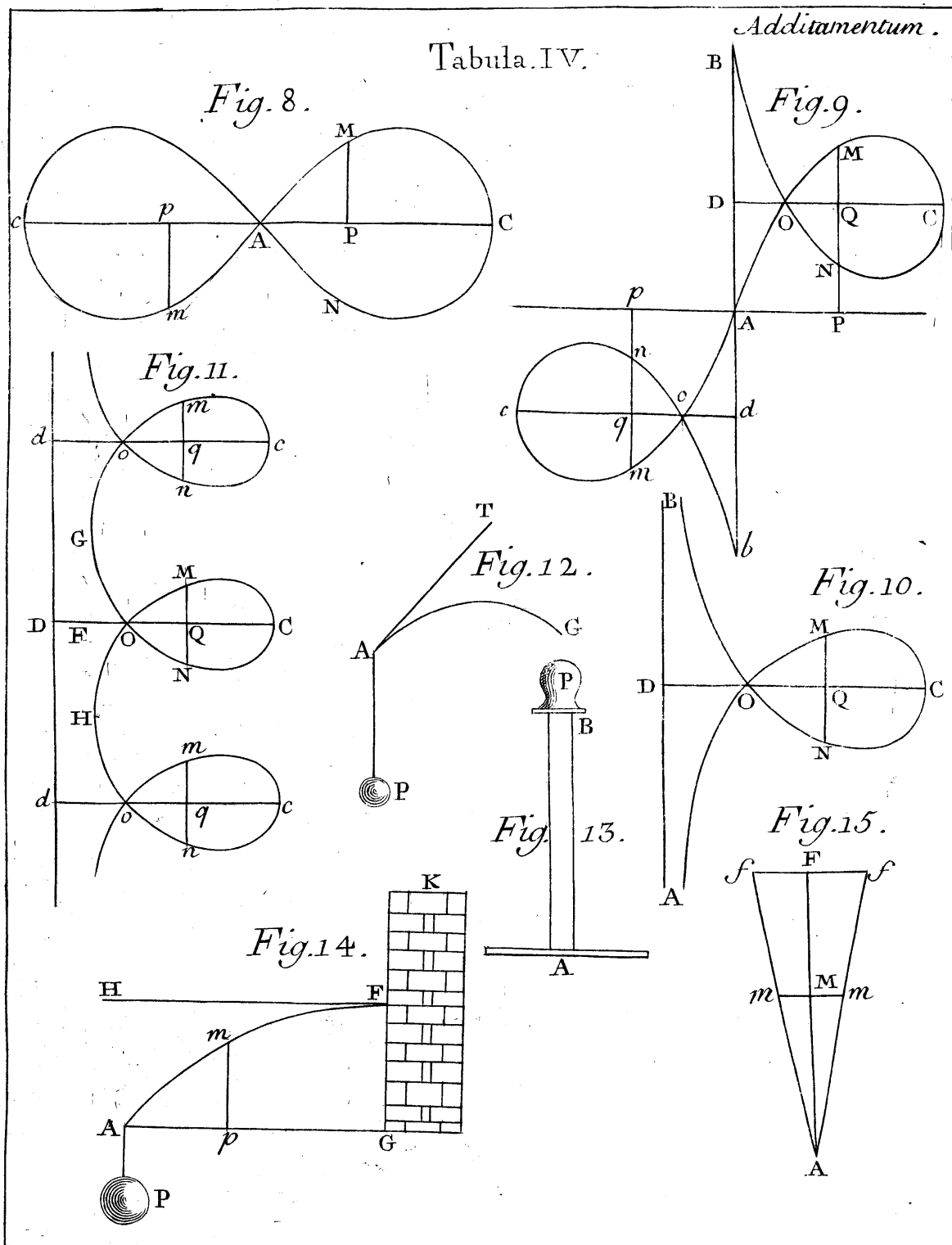


Figure 5.10. Euler's elastica figures, Tabula IV.

Euler includes figures for six of the nine cases, reproduced here in Figures 5.9 and 5.10. Of the three remaining, species #1 is a degenerate straight line, and species #9 is a circle. Species #3, the *rectangular elastica*, is of special interest, so it is rather disappointing that Euler did not include a figure for it. Possibly, it was considered already known due to Bernoulli's publication. Note also that in this special case, $c = a$, the equation becomes equivalent to Bernoulli's Equation 5.2.

Compare these figures with the computer-drawn Figure 5.11. It is remarkable how clearly Euler was able to visualize these curves, even 250 years ago. It is likely that the distortions and inaccuracies are due primarily to the draftsman engraving the figures for the book rather than to Euler himself; they display lack of symmetry that Euler clearly would have known. In fact, Euler computed a number of values to seven or more decimal places, including the values of a and c for the lemnoid shape (species #5).

Note that for $c^2 < 2a^2$, or $\lambda > .25$, the inflectional cases, the equation is well-defined for $-c < x < c$. In these cases, the parameter c is half the width of the figure. In the non-inflectional cases ($\lambda < .25$), the equation has no solutions at $x = 0$, and in Euler's figure the curve is drawn to the right of the y axis.

The curve described by species #7, the only non-periodic solution, was known to Poleni in 1729, and is also known as the "syntractrix of Poleni," or, in French, "la courbe des forçats" (the curve of convicts, or galley slaves). Euler gave its equation (in section 31) in closed form (without citing Poleni):

$$y = \sqrt{c^2 - x^2} - \frac{c}{2} \log \frac{c + \sqrt{c^2 - x^2}}{x} . \quad (5.14)$$

5.8.1 Moments

Euler was also clearly aware of the simple moment approach to the elastica, and in section 43 of the *Additamentum*, demonstrated its equivalence to the variational approach. To do so, he manipulated the quadrature formulation of the elastica, Equation 5.12, grouping all the terms involving first and second derivatives of the curve into a single term $Sq(1 - p^2)^{-3/2}$. Recall that $p = \frac{dy}{dx}$ and $q = \frac{dp}{dx} = \frac{d^2y}{dx^2}$. Thus, this term is S times the curvature, yielding an equation relating curvature to Cartesian coordinates. Euler explains:

"But $\frac{-(1+pp)^{3:2}}{q}$ is the radius of curvature R ; whence, by doubling the constants β and γ , the following equation will arise:

$$\frac{S}{R} = \alpha + \beta x - \gamma y . \quad (5.15)$$

This equation agrees admirably with that which the second or direct method supplies. For let $\alpha + \beta x - \gamma y$ express the moment of the bending power, taking any line you please as an axis, to which moment the absolute elasticity S , divided by the radius of curvature R must be absolutely equal. Thus, therefore, not only has the character of the elastic curve observed by the celebrated BERNOULLI been most abundantly demonstrated, but also the very great utility of my somewhat difficult formulas has been established in this example."

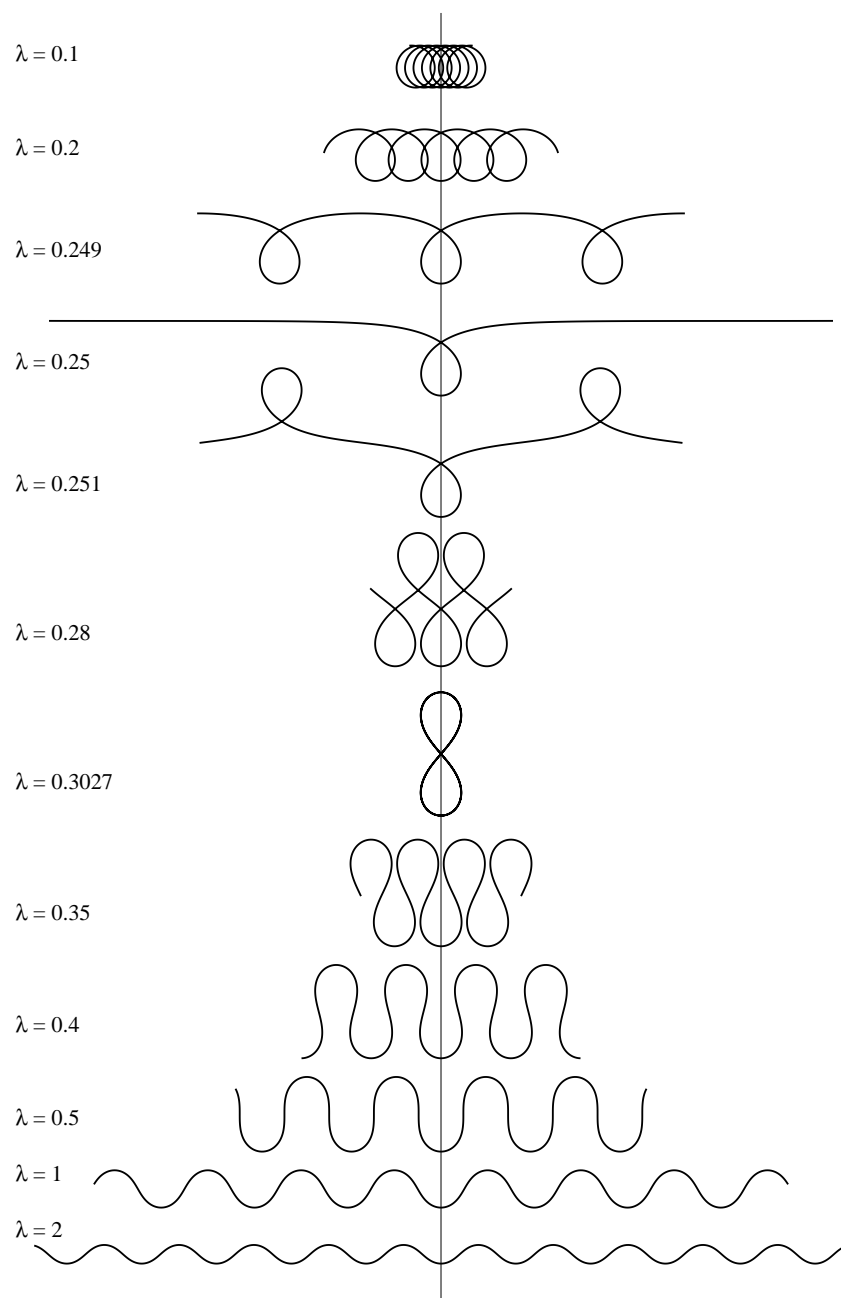


Figure 5.11. The family of elastica solutions.

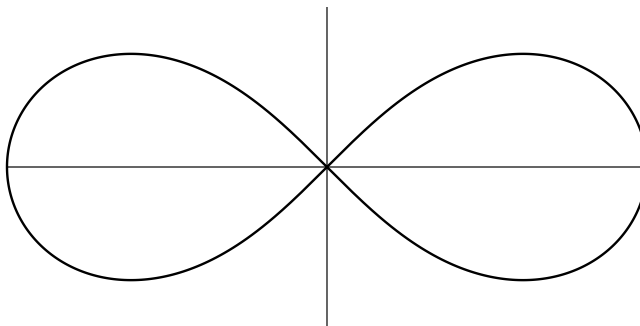


Figure 5.12. Bernoulli's lemniscate.

This result certainly is confirmation of the variational technique, but in fairness it must be pointed out that James Bernoulli was able to derive essentially the identical equation (again, note the similarity between Equations 5.9 and 5.12) by using a combination of mechanical insight and clever integration.

5.9 Elliptic integrals

The elastica, having been present at the birth of the variational calculus, also played a major role in the development of another branch of mathematics: the theory of elliptic functions.

Even as the quadratures of these simple curves came to be revealed, analytic formulae for their lengths remained elusive. The functions known by the first half of the 18th century were insufficient to determine the length even of a curve as well-understood as an ellipse.

James Bernoulli set himself to this problem and was able to pose it succinctly (and even compute approximate numerical values), if not fully solve it himself [81]. If the quadrature of a unit-excursion rectangular elastica is given by the equation

$$y = \int_0^x \frac{x^2 dx}{\sqrt{1-x^4}}, \quad (5.16)$$

then the arc length is given by:

$$s = \int_0^x \frac{dx}{\sqrt{1-x^4}}. \quad (5.17)$$

This integral is now called the “Lemniscate integral”, because of its connection with the lemniscate, another beautiful curve studied by James Bernoulli (Figure 5.12). The length of the lemniscate is equal to that of the rectangular elastica; while Equation 5.17 gives the arc length of the elastica as a function of the x -coordinate, this nearly identical equation relates arc length to the radial coordinate r in the lemniscate:

$$s = \int_0^r \frac{dr}{\sqrt{1-r^4}}. \quad (5.18)$$

Among the lemniscate's other representations, its implicit equation in Cartesian coordinates is a simple polynomial (no such corresponding formulation exists for the elastica),

$$(x^2 + y^2)^2 = x^2 - y^2. \quad (5.19)$$

Bernoulli approximated the integral for the arc length of the lemniscate using a series expansion and determined upper and lower numerical bounds, but felt the calculation of them would not fall to standard analytical techniques. He wrote, “I have heavy grounds to believe that the construction of our curve depends neither on the quadrature nor on the rectification of any conic section.” [87]. Here, “quadrature” means the area under the curve, or the indefinite integral, and “rectification” means the computation of the length of the curve. Bernoulli's prediction would not turn out to be entirely accurate; as we shall see, the curve would later be expressed in terms of Jacobi elliptic functions, which in turn are deeply related to the question of determining the arc length (rectification) of the ellipse.

Fagnano took up the problem of finding the length of the lemniscate, and achieved some impressive results. Indeed, Jacobi fixes the date for the birth of elliptic functions as 23 December 1751, when Euler was asked to review Fagnano's collected works. However, Euler had begun study of elliptic integrals as early as 1738, when he wrote to the Bernoullis that he had “noticed a singular property of the rectangular elastica” having unit excursion [87]:

$$\text{length} \cdot \text{height} = \int_0^1 \frac{dx}{\sqrt{1-x^4}} \cdot \int_0^1 \frac{x^2 dx}{\sqrt{1-x^4}} = \frac{1}{4}\pi. \quad (5.20)$$

After receiving Fagnano's work, Euler caught fire and started his remarkable research, first on lemniscate integrals, then on the more general problem of elliptic integrals, especially the discovery of the addition theorems for elliptic functions in the 1770s. The reader interested in more details, including mathematical derivations, is directed to Sridharan's delightful historical sketch [81].

These integrals would ultimately be the basis for closed-form solutions of the elastica equation, with both curvature and Cartesian coordinates given as “special functions” of the arc length parameter, as will be described in Section 5.12.

5.10 Laplace on the capillary – 1807

Remarkably, the elastica appears as yet another shape of the solution of a fundamental physics problem – the capillary. Pierre Simon Laplace investigated the equation for the shape of the capillary in his 1807 *Supplément au dixième livre du Traité de mécanique céleste. Sur l'action capillaire*⁷. See I. Grattan Guinness [31, p. 442] for a thorough description of these results. In particular, Laplace considered the surface of a fluid trapped between two vertical plates, and obtained the equation (722.16 in [31])

⁷reprinted in Volume 4 of Laplace's Complete Works [50, pp. 349–401]

$$z'' = 2(\alpha z + b^{-1})(1 + z'^2)^{3/2}. \quad (5.21)$$

Here, z represents height, and z' and z'' represent first and second derivatives with respect to the horizontal coordinate. With suitable renaming of coordinates and substitution of Equation 5.1 for curvature in Cartesian coordinates, this equation can readily be seen to be equivalent to Euler's Equation 5.15.

Laplace also recognized that at least one instance of his equation was equivalent to the elastica. He derives Equation 5.7 (save that Laplace writes Z for S) and noted that it is identical to the elastic curve [50, p. 379].

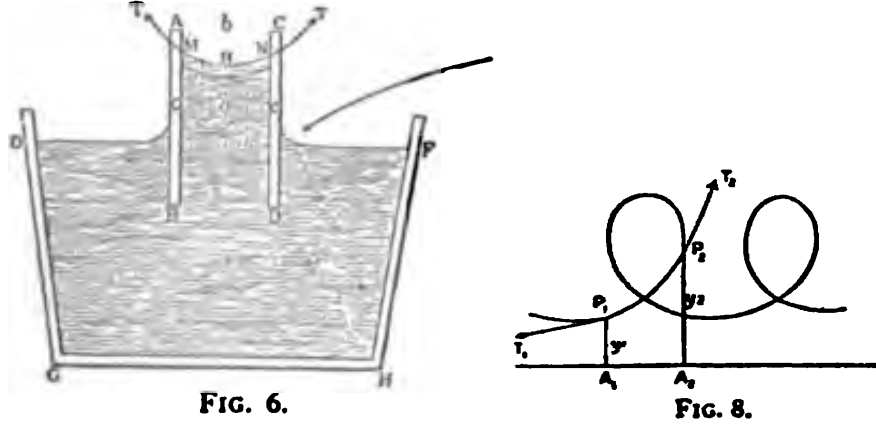


Figure 5.13. Maxwell's figures for capillary action.

It is not clear when the general equivalence between the capillary surface and the elastica was first appreciated. The special case of a single plate is published in *The Elements of Hydrostatics and Hydrodynamics* [62, p. 32] in 1833. In any case, by 1890 James Clerk Maxwell was fully aware of it, including it in his article "Capillary Action" in the 9th edition of the *Encyclopædia Britannica* [57]. His figures (reproduced here as Figure 5.13) illustrate the problem and clearly show a non-inflectional, periodic instance of the elastica.

5.11 Kirchhoff's kinetic analogy – 1859

Surprisingly, the differential equation for the elastica, expressing curvature as a function of arc length, are equivalent to those of the motion of the pendulum, as worked out by Kirchhoff in 1859 (see D'Antonio [17] and also Goss [30] for historical detail). Using simple variational techniques, taking arc length as the independent variable and the angle θ from the horizontal coordinate as the dependent, Equation 5.10 yields

$$\theta'' + \lambda_1 \sin \theta + \lambda_2 \cos \theta = 0. \quad (5.22)$$

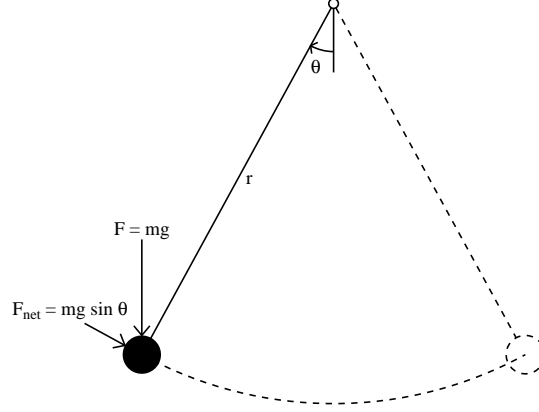


Figure 5.14. An oscillating pendulum.

Here, the sine and cosine arise from the specification of endpoint location constraints, which are described as an integral of the sine and cosine of θ over the length of the curve.

This differential equation for the shape of the elastica is mathematically equivalent to that of the dynamics of a simple swinging pendulum. This kinetic analog is useful for developing intuition about the solutions of the elastica equation. In particular, it's easy to see that all solutions are periodic, and that the family of solutions is characterized by a single parameter (modulo scaling, translation, and rotation of the system).

Consider, as shown in Figure 5.14, a weight of mass m at the end of a pendulum of length r , with angle from the vertical specified as a function of time $\theta(t)$.

Then the velocity of the mass is $r\theta'$ (where, in this section, the $'$ notation represents derivative with respect to time), and its acceleration is $r\theta''$. The net force of gravity, acting with the constraint of the pendulum's angle is $mg \sin \theta$, and thus we have

$$F_{\text{net}} = ma = mr\theta'' = mg \sin \theta . \quad (5.23)$$

With the substitution $\lambda_1 = -g/r$, and the replacement of arc length s for time t , this equation becomes equivalent to Equation 5.22, the equation of the elastica. Note that angle (as a function of arc length) in the elastica corresponds to angle (as a function of time) in the pendulum, and the elastica's curvature corresponds to the pendulum's angular momentum.

The swinging of a pendulum is perhaps the best-known example of a periodic system. Further, it is intuitively easy to grasp the parameter space of the system. Transformations of scaling (varying the pendulum length) and translation (assigning different phases of the pendulum swing to $t = 0$) leave the solutions essentially unchanged. Only one parameter, how high the pendulum swings, changes the fundamental nature of the solution.

The solutions of the motion of the pendulum, as do the solutions of the elastica Equation 5.22, form a family characterized by a single scalar parameter, once the trivial transforms of rotation and scaling are factored out. Without loss of generality, let $t = 0$ at the bottom of the swing (i.e. maximum velocity) and let the pendulum have unit length. The remaining parameter is then the ratio of the total energy of the system (which remains unchanged over time) to the potential

energy of the pendulum at the top of the swing (the maximum possible), in both cases counting the potential energy at the bottom of the swing as zero. Thus, the total system energy is also equal to the kinetic energy at the bottom of the swing.

For mathematical convenience, define the parameter λ as one fourth the top-of-swing potential energy divided by the total energy. (The constant is chosen so that λ matches the Lagrange multiplier λ_1 in Equation 5.22.)

The potential energy at the top of the pendulum is $2mgr$. The kinetic energy at the bottom of the swing is $\frac{1}{2}mv^2$, where, as above, $v = r\theta'$. Thus, according to the definition above,

$$\lambda = \frac{1}{4} \frac{2mgr}{\frac{1}{2m(r\theta')^2}} = \frac{g}{r(\theta')^2} .$$

In other words, assuming (without loss of generality) unit pendulum length and unit velocity at the bottom of the swing, λ simply represents the force of gravity. And this λ is the same as defined in Equation 5.13.

5.12 Closed form solution: Jacobi elliptic functions – 1880

The closed-form solutions of the elastica, worked out by Saalschütz in 1880 [76], rely heavily on Jacobi elliptic functions. (See also [17] for more historical development, and Greenhill [33] for a contemporary presentation of these results in English, suggesting the kinetic analog was a major motivator to deriving the closed form solutions.)

The *Jacobi amplitude* $\text{am}(u, m)$ is defined as the inverse of the Jacobi elliptic integral of the first kind,

$$\begin{aligned} \text{am}(u, m) &= \text{the value of } \phi \text{ such that} \\ u &= \int_0^\phi \frac{1}{\sqrt{1 - m \sin^2 t}} dt . \end{aligned} \tag{5.24}$$

Here, m is known as the *parameter*. Jacobi elliptic functions are also commonly written in terms of the *elliptic modulus* $k = \sqrt{m}$.

From this amplitude, the doubly periodic generalizations of the trigonometric functions follow:

$$\text{sn}(u, m) = \sin(\text{am}(u, m)) , \tag{5.25}$$

$$\text{cn}(u, m) = \cos(\text{am}(u, m)) , \tag{5.26}$$

$$\text{dn}(u, m) = \sqrt{1 - m \sin^2(\text{am}(u, m))} . \tag{5.27}$$

Note that when m is zero, sn and cn are equivalent to \sin and \cos , respectively, and when m is unity, sn and cn are equivalent to \tanh and sech ($1/\cosh$) [1, §16.6, p. 571].

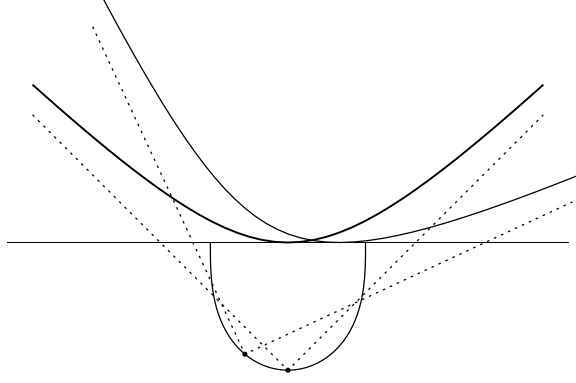


Figure 5.15. Rectangular elastica as roulette of hyperbola.

5.12.1 Inflectional solutions

The closed form solutions are given separately for inflectional and non-inflectional cases. The intrinsic form, curvature as a function of arc length, is fairly simple:

$$\frac{d\theta}{ds} = \kappa = 2\sqrt{m} \operatorname{cn}(s, m) . \quad (5.28)$$

Expressing the curve in the form of angle as a function of arc length is also straightforward:

$$\sin \frac{1}{2}\theta = \sqrt{m} \operatorname{sn}(s, m) . \quad (5.29)$$

Through an impressive feat of analysis, the curve can be expressed in Cartesian coordinates as a function of arc length. First, we'll need the elliptic integral of the second kind,

$$E(\phi, k) = \int_0^\phi \sqrt{1 - m \sin^2 \theta} \, d\theta , \quad (5.30)$$

$$\begin{aligned} x(s) &= s - 2E(\operatorname{am}(s, m), m) , \\ y(s) &= -2\sqrt{m} \operatorname{cn}(s, m) . \end{aligned} \quad (5.31)$$

In Love's presentation of these results, $E(\operatorname{am}(s, m), m)$ is defined in terms of an integral of $\operatorname{dn}(u, m)$. The equivalence with Equation 5.30 follows through a standard identity of the Jacobi elliptical functions ([1, p. 576], 16.26.3),

$$E(\operatorname{am}(s, m), m) = \int_0^s (\operatorname{dn}(u, m))^2 \, du . \quad (5.32)$$

The solution in terms of elliptic integrals opened the way to a few more curious results. For one, the *roulette* of the center of a rectangular hyperbola is a rectangular elastica [33]. More precisely, let the hyperbola roll along a line without slipping. Then, the curve traced by its center is a rectangular

elastica, as shown in Figure 5.15. The general term for a roulette formed from a conic section is a Sturm’s roulette.⁸

5.12.2 Non-inflectional solutions

The equations in the non-inflectional case are similar.

$$\frac{d\theta}{ds} = \kappa = \frac{2}{\sqrt{m}} \operatorname{dn}\left(\frac{s}{\sqrt{m}}, m\right), \quad (5.33)$$

$$\sin \frac{1}{2}\theta = \operatorname{sn}\left(\frac{s}{\sqrt{m}}, m\right). \quad (5.34)$$

For the Cartesian version, and for a more detailed derivation, refer to Love [52].

Truesdell considers the popularity of the elliptic function approach to be a mixed blessing for mechanics and for mathematics. Where Euler’s method used direct nonlinear thinking based on physical principles, much of the elliptic literature explored properties of special functions, which “came to be ends of research rather than means to solve a natural problem.” [87] Indeed, elliptic functions are rarely used today for computation of elastica, in favor of numerical methods. For example, I’ve used a simple fourth-order Runge–Kutte differential equation solver (computing Equation 5.22) to draw the figures for this work, due to its good convergence and efficiency and simple expression in code. One particularly unappealing aspect of the elliptic approach is the sharp split between inflectional and non-inflectional cases, while one differential equation smoothly covers both cases.

Even so, Jacobi elliptic functions are now part of the mainstream of special functions, and fast algorithms for computing them are well known [71]. Elliptic functions are still the method of choice for the fastest computation of the shape of an elastica.

5.13 Max Born – 1906

In spite of the equation for the general elastica being published as early as 1695, the curves had not been accurately plotted until Max Born’s 1906 Ph.D. thesis, “Investigations of the stability of the elastic line in the plane and in space” [11].

Born also constructed an experimental apparatus using weights and dials to place the elastic band in different endpoint conditions, and used photographs to compare the equations to actual shapes. An example setup is shown in Figure 5.16, which shows an inflectional elastica under fairly high tension, $\lambda \approx 0.26$.

Years later, Born wrote, “...I felt for the first time the delight of finding a theory in agreement with measurement – one of the most enjoyable experiences I know.” [12, p. 21]. Born used the best modern mathematical techniques to address the problem of the elastica, and, among other things, was able to generalize it to the three-dimensional case of a wire in space, not confined to the plane.

⁸Visit <http://www.mathcurve.com/courbes2d/sturm/sturm.shtml> for an animated demonstration of this property.

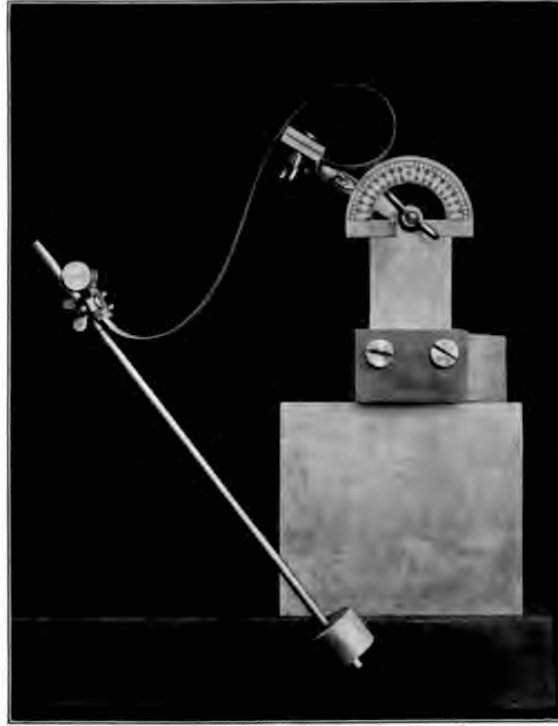


Figure 5.16. Born's experimental apparatus for measuring the elastica.

5.14 Influence on modern spline theory – 1946 to 1965

Mechanical splines made of wood or metal have long been an inspiration for the mathematical concept of spline (and for its name). Schoenberg's justification for cubic splines in 1946 was a direct appeal to the notion of an elastic strip. Schoenberg's main contribution was to define his spline in terms of a variational problem closely approximating Equation 5.10, but making the small-deflection approximation.⁹

3.1 Polynomial spline curves of order k . A spline is a simple mechanical device for drawing smooth curves. It is a slender flexible bar made of wood or some other elastic material. The spline is placed on the sheet of graph paper and held in place at various points by means of certain heavy objects (called “dogs” or “rats”) such as to take the shape of the curve we wish to draw. Let us assume that the spline is so placed and supported as to take the shape of a curve which is nearly parallel to the x -axis. If we denote by $y = y(x)$ the equation of this curve then we may neglect its small slope y' , whereby its curvature becomes

$$1/R = y''/(1 + y'^2)^{3/2} \approx y''$$

The elementary theory of the beam will then show that the curve $y = y(x)$ is a polygonal line composed of cubic arcs which join continuously, with a continuous first and second

⁹Note: what follows is the corrected version as appears in his selected papers [78].

derivative¹⁰. These junction points are precisely the points where the heavy supporting objects are placed.

While Schoenberg’s splines are excellent for fitting the values of functions (the problem of approximation), it was clear that they were not ideal for representing shapes. Birkhoff and de Boor wrote in 1965 [10] that “linearized interpolation schemes have a basic shortcoming: they are *not intrinsic* geometrically because they are not invariant under rigid rotation. Physically it seems more natural to replace linearized spline curves by *non-linear* splines (or “elastica”), well known among elasticians,” citing Love [52] as an authority. They also stated the result that in a mechanical spline constrained to pass through the control points only by pure *normal forces*, only the rectangular elastica is needed.

Birkhoff and de Boor also noted some shortcomings in the elastica as a replacement for the mechanical spline, particularly the lack of an existence and uniqueness theory for non-linear spline curves with given endpoints, end-slopes, and sequences of internal points. Indeed, they state that “nor does it seem particularly desirable” to have techniques for intrinsic splines approximate the elastica, and they proposed other techniques, such as Hermite interpolation by segments of Euler spirals joined together with continuous curvature.

According to notes taken by Stanford professor George Forsythe [22], in the conference presentation of this work, they referenced Max Born’s Ph.D. thesis [11], and also described a goal of the work as providing an “automatic French curve.” The presentation must have made an impression on Forsythe, judging from the laconic sentence, “Deep.” Forsythe, working with his colleague Erastus Lee from the Mechanical Engineering college at Stanford, would go on to analyze the nonlinear spline considerably more deeply, deriving it from both the variational formulation of Equation 5.10 and an exploration of moments, normal forces, and longitudinal forces [51]. That publication also included a result for energy minimization in the case where the elastica is a closed loop.

5.15 Numerical techniques – 1958 through today

The arrival of the high-speed digital computer created a strong demand for efficient algorithms to *compute* the elastica, particularly to compute the shape of an idealized spline constrained to pass through a sequence of *control points*.

Birkhoff and de Boor [10] cited several approximations to non-linear splines, including those of Fowler and Wilson [23] and MacLaren [53] at Boeing in 1959, and described their own work along similar lines. None of these was a particularly precise or efficient approximation.

Around the same time, and very possibly working independently, Mehlum and others designed the Autokon system [59] using an approximation to the elastica (called the KURGLA 1 algorithm, and discussed in more detail in Section 4.9). However, this algorithm was not an accurate simulation of the mechanical spline, as it did not minimize the total bending energy of the curve, and, in fact, could generate (approximations to) the entire family of elastica curves in service of its splines. An objection to this technique is that the spline was not *extensional*; adding a new control point coinciding with the generated spline would slightly perturb the curve. Thus, this approximate elastica was soon replaced by the KURGLA 2 algorithm implementing precisely the Euler spiral Hermite interpolation suggested by Birkhoff and de Boor.

¹⁰Schoenberg is indebted for this suggestion to Professor L. H. Thomas of Ohio State University.

A sequence of papers refining the numerical techniques for computing the elastica-based non-linear spline followed: Glass in 1966 [26], Woodford in 1969 [90], Malcolm in 1977 [54]. Most of these involve discretized formulations of the problem. Interestingly, Malcolm reports that Larkin developed a technique based on direct evaluation of the Jacobi elliptic integrals which is nonetheless “probably quite slow” compared with the discretized approaches, “due to the large number of transcendental functions which must be evaluated.” Malcolm’s paper contains a good survey of known numerical techniques, and is recommended to the reader interested in following this development in more detail.

In spite of the fairly rich literature available on the elastica, at least one researcher, B. K. P. Horn in 1981, seems to have independently derived the rectangular elastica from the principle of minimizing the strain energy [38], going through an impressive series of derivations, and using the full power of elliptic integral theory, to arrive at exactly the same integral as Bernoulli had derived almost three hundred years previously.

Edwards [18] proposed a spline solution in which the elastica in the individual segments are computed using Jacobi elliptic functions. He also gave a global solver based on a Newton technique, each iteration of which solves a band-diagonal Jacobian matrix. A particular concern was exploring the cases when no stable solution exists, a weakness of the elastica-based spline compared to other techniques. Edwards claimed that his numerical techniques would always converge on a solution if one exists.

Chapter 6

History of the Euler spiral

This chapter traces the history of the Euler spiral, a beautiful and useful curve known by several other names, including “clothoid,” and “Cornu spiral.” The underlying mathematical equation is also most commonly known as the Fresnel integral. The profusion of names reflects the fact that the curve has been discovered several different times, each for a completely different application: first, as a particular problem in the theory of elastic springs; second, as a graphical computation technique for light diffraction patterns; and third, as a railway transition spiral. Through all the names, the curve retains its aesthetic and mathematical beauty as Euler had clearly visualized. Its equation is related to the Gamma function, the Gauss error function (erf), and is a special case of the confluent hypergeometric function.

The Euler spiral is defined as the curve in which the curvature increases linearly with arc length. Changing the constant of proportionality merely scales the entire curve. Considering curvature as a signed quantity, it forms a double spiral with odd symmetry, a single inflection point at the center, as shown in Figure 4.4. According to Alfred Gray, it is “one of the most elegant of all plane curves.” [32]

6.1 James Bernoulli poses a problem of elasticity – 1694

The first appearance of the Euler spiral is as a problem of elasticity, posed by James Bernoulli in the same 1694 publication as his solution to a related problem, that of the elastica.

The elastica is the shape defined by an initially straight band of thin elastic material (such as spring metal) when placed under load at its endpoints. The Euler spiral can be defined as something of the inverse problem; the shape of a pre-curved spring, so that when placed under load at one endpoint, it assumes a straight line.

The problem is shown graphically in Figure 6.1. When the curve is straightened out, the moment at any point is equal to the force F times the distance s from the force. The curvature at the point in the original curve is proportional to the moment (according to elementary elasticity theory).

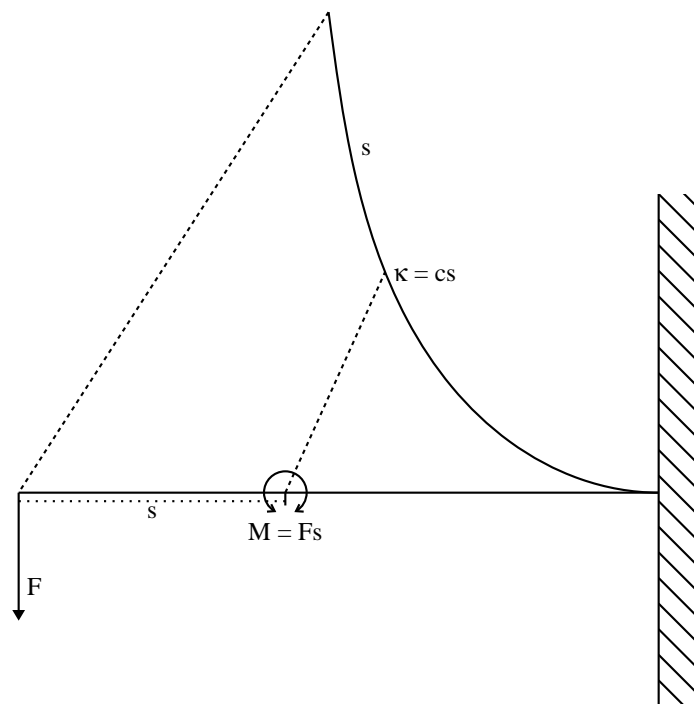


Figure 6.1. Euler's spiral as an elasticity problem.

Because the elastic band is assumed not to stretch, the distance from the force is equal to the arc length. Thus, curvature is proportional to arc length, the definition of the Euler spiral.

James Bernoulli, at the end of his monumental 1694 *Curvatura Laminae Elasticae* (see Section 5.6), presenting the solution to the problem of the elastica, sets out a number of other problems he feels may be addressed by the techniques set forth in that paper, for example, cases where the elastica isn't of uniform density or thickness.

A single sentence among a list of many, poses the mechanics problem whose solution is the Euler spiral: To find the curvature a lamina should have in order to be straightened out horizontally by a weight at one end¹.

The same year, Bernoulli wrote a note containing the integral² entitled "To find the curve which an attached weight bends into a straight line; that is, to construct the curve $a^2 = sR$ ".

Quia nominatis abscissa = x , applicata = y , arcu curvæ s , & posita ds constante, radius circuli oscularis, curvedini reciproce proportionalis, est $dxds : -ddy$; habebitur, ex hypothesi, hæc æquatio $-aaddy = sdsdx$

¹"Nec non qualem Lamina curvaturam habere debeat, ut ab appenso onere, vel proprio pondere, vel ab utroque simul in rectam extedatur." The translation here is due to Raymond Clare Archibald [4].

²The 1694 original (Latin title "Invenire curvam, quae ab appenso pondere flectitur in rectam; h.e. construere curvam $aa = sz$ " is No. CCXVI of Jacob Bernoulli's *Thoughts, notes, and remarks*. An expanded version was published in slightly expanded form as No. XX of his "Varia Posthuma," which were collected in volume 2 of his *Opera*, published in 1744 (and available online through Google Books) [9, pp. 1084–1086]. Both works are also scheduled to be published in Volume 6 of *Die Werke von Jacob Bernoulli*.

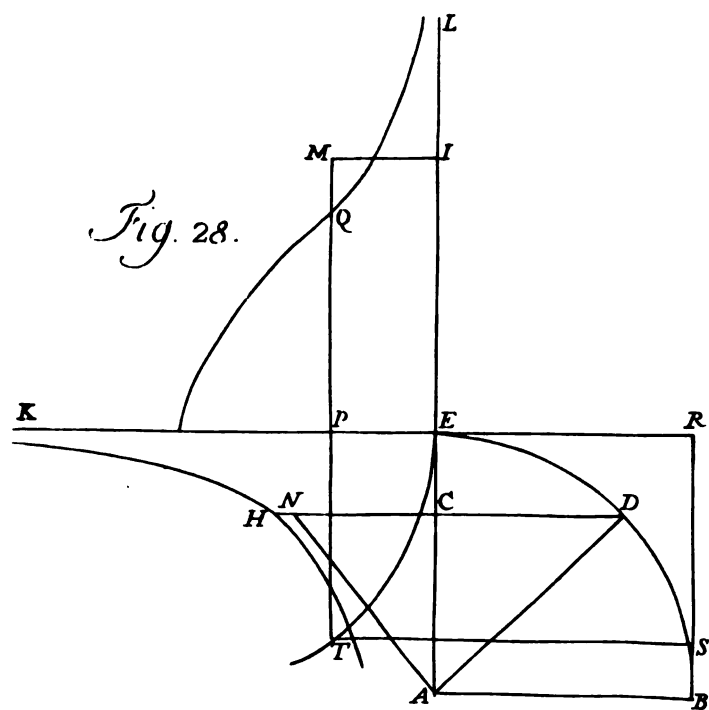


Figure 6.2. Bernoulli's construction.

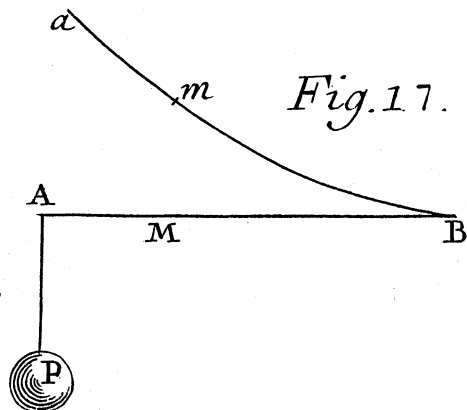


Figure 6.3. Euler's drawing of his spiral, from Tabula V of the Additamentum.

Translated into English (and with slightly modernized notation):

Let us call the abscissa x , the ordinate y , the arc length s , and hold ds constant. Then, the radius of the osculating circle, which is proportional to the reciprocal of the moment, is $dx ds/d^2y$. Thus, we have, by hypothesis, the equation $-a^2 d^2y = s ds dx$.

The remainder of the note is a geometric construction of the curve. According to Truesdell [86, p. 109], “it is not enlightening, as it does not reveal that the curve is a spiral, nor is this indicated by his figure.” Nonetheless, it is most definitely the equation for the Euler spiral. This construction is illustrated in Figure 6.2. The curve of interest is ET , and the others are simply scaffolding from the construction. However, it is not clear from the figure that Bernoulli truly grasped the shape of the curve. Perhaps it is simply the fault the draftsman, but the curve ET is barely distinguishable from a circular arc.

In summary, Bernoulli had written the equation for the curve, but did not draw its true shape, did not compute any values numerically, and did not publish his reasoning for *why* the equation was correct. His central insight was that curvature is additive; more specifically, the curvature of an elastic band under a moment force is its curvature in an unstressed state plus the product of the moment and a coefficient of elasticity. But he never properly published this insight. In editing his work for publication in 1744, his nephew Nicholas I Bernoulli wrote about the equation $s = -a^2/R$, “I have not found this identity established” [86, p. 108].

6.2 Euler characterizes the curve – 1744

The passage introducing the Euler spiral appears in section 51 of the Additamentum³, referring to his Fig. 17, which is reproduced here as Figure 6.3:

³Additamentum 1 to *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici lattissimo sensu accepti* [20]. The quotes here are based on Oldfather's 1933 translation [68], with additional monkeying by the author.

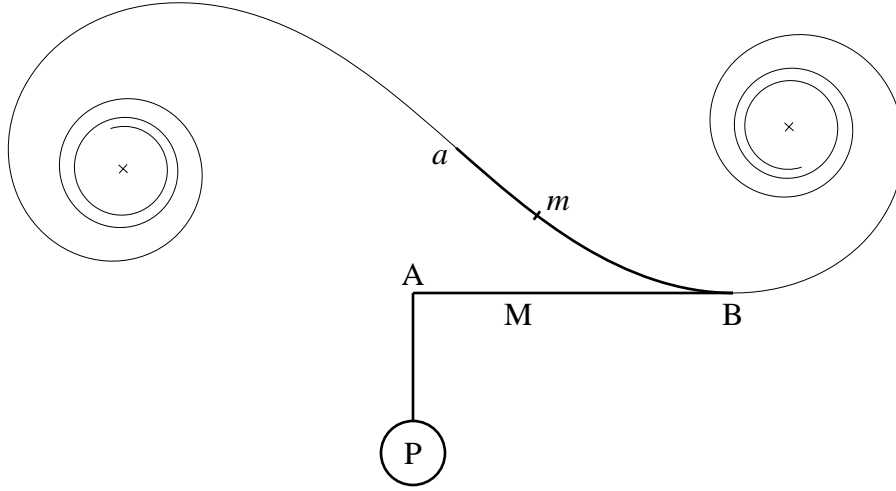


Figure 6.4. Reconstruction of Euler's Fig. 17, with complete spiral superimposed.

51. Hence the figure amB , which the lamina must have in its natural state, can be determined, so that by the force P , acting in the direction AP , it can be unfolded into the straight line AMB . For letting $AM = s$, the moment of the force acting at the point M will equal Ps , and the radius of curvature at M will be infinite by hypothesis, or $1/R = 0$. Now the arc am in its natural state being equal to s , and the radius of curvature at m being taken as r , because this curve is convex to the axis AB , the quantity r must be made negative. Hence $Ps = Ekk/r$, or $rs = aa$, which is the equation of the curve amB . [20, §51]

In modern terms (and illustrated by the modern reconstruction of Euler's drawing, Figure 6.4), the lamina in this case is not straight in its natural (unstressed) state, as is the case in his main investigation of the elastica, but begins with the shape amB . At point B , the curve is held so the tangent is horizontal (i.e. point B is fixed into the wall), and a weight P is suspended from the other end of the lamina, pulling that endpoint down from point A to point a , and overall flattening out the curve of the lamina.

The problem posed is this: what shape must the lamina amB take so that it is flattened into an exactly straight line when the free end is pulled down by weight P ? The answer derived by Euler appeals to the simple theory of moments: the moment at any point M along the (straightened) lamina is the force P times the distance s from A to M .

The curvature of the curve resulting from the original shape stressed by force P is equal to the original curvature plus the moment Ps divided by the lamina's stiffness Ek^2 . Since this resulting curve must be a straight line with curvature zero, the solution for the curvature of the original curve is $\kappa = -Ps(Ek^2)^{-1}$. Euler flips the sign for the curvature and groups all the force and elasticity constants into one constant a for convenience, yielding $1/r = \kappa = s/aa$.

From this intrinsic equation, Euler derives the curve's quadrature (x and y as a function of the arc length parameter s), giving equations for the modern expression of the curve (formatting in this case preserved from the original):

$$x = \int ds \sin. \frac{ss}{2aa}, \text{ \& } y = \int ds \cos. \frac{ss}{2aa} \quad (6.1)$$

Euler then goes on to describe several properties of the curve, particularly, “Now from the fact that the radius of curvature continuously decreases the greater the arc $am = s$ is taken, it is manifest that the curve cannot become infinite, even if the arc s is taken infinite. Therefore the curve will belong to the class of spirals, in such a way that after an infinite number of windings it will roll up a certain definite point as a center, which point seems very difficult to find from this construction.” [20, §52]

Euler did give a series expansion for the above integral, but in the 1744 publication is not able to analytically determine the coordinates of this limit point, saying “Therefore analysis should gain no small advance, if someone were to discover a method to assign a value, even if only approximate, to the integrals [of Equation 6.1], in the case where s is infinite; this problem does not seem unworthy for geometers to exercise their strength.”

Euler also derived a series expansion for the integrals [20, §53], which still remains a viable method for computing them for reasonably small s (here $b = \sqrt{2}a$, chosen for notational convenience).

$$\begin{aligned} x &= \frac{s^3}{1 \cdot 3 \, b^2} - \frac{s^7}{1 \cdot 2 \cdot 3 \cdot 7 \, b^6} + \frac{s^{11}}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 11 \, b^{10}} - \frac{s^{15}}{1 \cdot 2 \cdot \dots \cdot 7 \cdot 15 \, b^{14}} + \&c. \\ y &= s - \frac{s^5}{1 \cdot 2 \cdot 5 \, b^4} + \frac{s^9}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 9 \, b^8} - \frac{s^{13}}{1 \cdot 2 \cdot \dots \cdot 6 \cdot 13 \, b^{12}} + \&c. \end{aligned} \quad (6.2)$$

6.3 Euler finds the limits – 1781

It took him about thirty-eight years to solve the problem of the integral's limits. In his 1781 “On the values of integrals extended from the variable term $x = 0$ up to $x = \infty$ ”⁴, he finally gave the solution, which he had “recently found by a happy chance and in an exceedingly peculiar manner”, of $x = y = \frac{a}{\sqrt{2}}\sqrt{\frac{\pi}{2}}$. This limit is marked with a cross in Figure 4.4, as is its mirror-symmetric twin.

The paper references a “Fig. 2”. Unfortunately, the original figure is not easy to track down. The version appearing in Figure 6.5 is from the 1933 edition of Euler's collected works. For reference, an accurately plotted reconstruction is shown alongside.

Euler's technique in finding the limits is to substitute $s^2/2a^2 = v$, resulting in these equivalences (this part of the derivation had already been done in his 1744 *Additamentum* [20, §54]):

⁴“De valoribus integralium a terminus variabilis $x = 0$ usque ad $x = \infty$ extensorum”, presented to the Academy at Petrograd on April 30, 1781, E675 in the Enneström index. Jordan Bell has recently translated this paper into English [21].

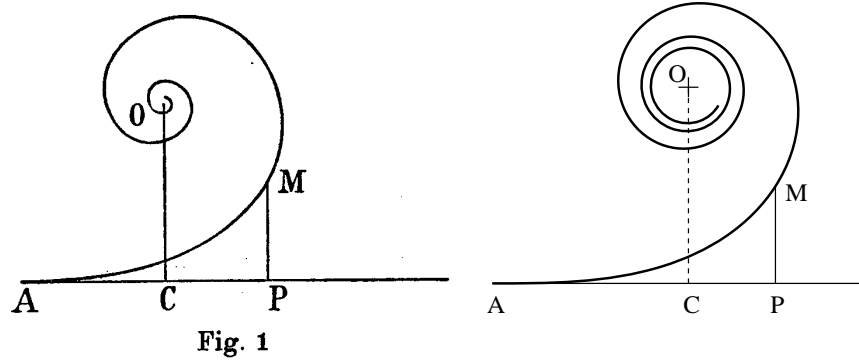


Figure 6.5. The figure from Euler’s “De valoribus integralium”, with modern reconstruction.

$$\begin{aligned}\int_0^\infty \sin \frac{x^2}{2a^2} dx &= \frac{a}{\sqrt{2}} \int_0^\infty \frac{\sin v}{\sqrt{v}} dv, \\ \int_0^\infty \cos \frac{x^2}{2a^2} dx &= \frac{a}{\sqrt{2}} \int_0^\infty \frac{\cos v}{\sqrt{v}} dv.\end{aligned}\tag{6.3}$$

To solve these integrals, Euler considers the Gamma function⁵ (which he calls Δ), defined as:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt\tag{6.4}$$

Through some manipulation, Euler solves a pair of fairly general integrals, of which the limit point of the Euler spiral will be a special case. Assuming $p = r \cos \alpha$, $q = r \sin \alpha$, he derives

$$\begin{aligned}\int_0^\infty t^{z-1} e^{-px} \cdot \cos qt \, dt &= \frac{\Gamma(z) \cos z\alpha}{r^z}, \\ \int_0^\infty t^{z-1} e^{-px} \cdot \sin qt \, dt &= \frac{\Gamma(z) \sin z\alpha}{r^z}.\end{aligned}\tag{6.5}$$

Euler sets $q = 1$, $p = 0$, $z = \frac{1}{2}$ and derives the limit of $\frac{a}{\sqrt{2}} \sqrt{\frac{\pi}{2}}$, which follows straightforwardly from the well-known value $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

6.4 Relation to the elastica

The Euler spiral can be considered something of a cousin to the elastica. Both curves were initially described in terms of elasticity problems. In fact, James Bernoulli was responsible for

⁵Euler published his discovery of the Gamma function in 1729, as *De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt*, Eneström index E19.

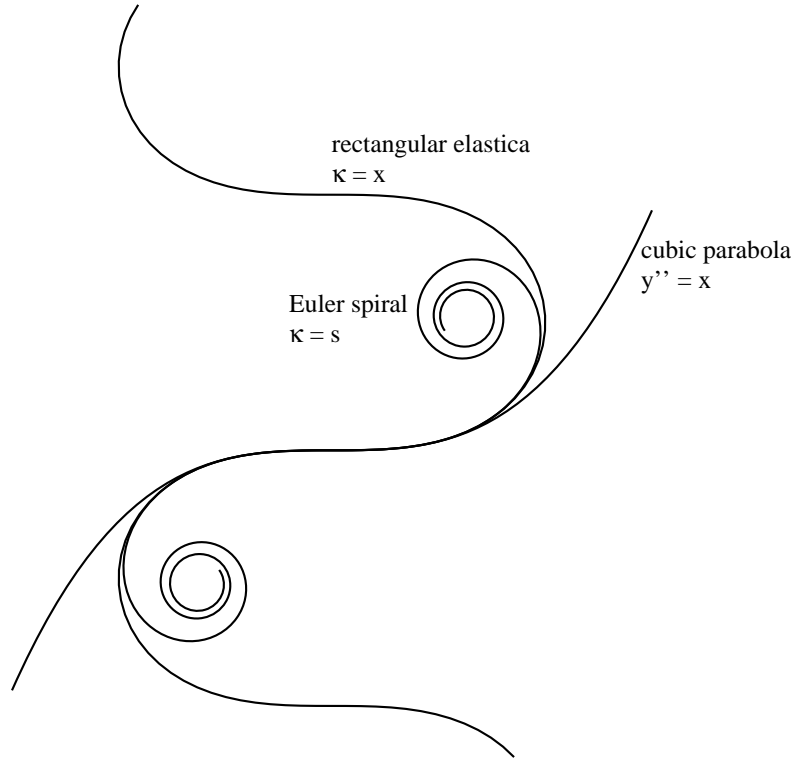


Figure 6.6. Euler’s spiral, rectangular elastica, and cubic parabola.

posing both problems, and Leonhard Euler described both curves in detail about fifty years later, in his 1744 *Additamentum* [20, p. 276].

Both curves also appear together frequently in the literature on spline curves, starting from Birkhoff and de Boor’s 1965 survey of nonlinear splines [10], through Mehlum’s work on the Autokon system [59], and Horn’s independent derivation of the rectangular elastica (the “curve of least energy”) [38].

The close relationship between the curves is also apparent in their mathematical formulations. The simplest equation of the elastica is $\kappa = x$, while that of the Euler spiral is $\kappa = s$ (here, κ represents curvature, x is a Cartesian coordinate, and s is the arc length of the curve). This similarity of equation is reflected in the similarity of shape, especially in the region where the arc is roughly parallel to the x axis, as can be seen in Figure 6.6, which shows both the rectangular elastica and Euler spiral, as well as the “cubic parabola,” the curve that results under the very small angle approximation $\kappa \approx y''$.

Similarly, both curves can be expressed in terms of minimizing a functional. The elastica is the curve that minimizes

$$E[\kappa(s)] = \int \kappa^2 ds. \quad (6.6)$$

The Euler spiral is one of many solutions that minimizes the L^2 -norm of the *variation* of curvature (known as the MVC, or minimum variation curve) [65]. It is, in fact, the optimal solution when the curvatures (but not the endpoint angles) are constrained.

$$E[\kappa(s)] = \int \left(\frac{d\kappa}{ds} \right)^2 ds . \quad (6.7)$$

It is also the curve that minimizes the L^∞ norm of the curvature variation, subject to endpoint constraints.

6.5 Fresnel on diffraction problems – 1818

Around 1818, Augustin Fresnel considered a problem of light diffracting through a slit, and independently derived integrals equivalent to those defining the Euler spiral⁶. At the time, he seemed to be unaware of the fact that Euler (and Bernoulli) had already considered these integrals, or that they were related to a problem of elastic springs. Later, this correspondence was recognized, as well as the fact that the curves could be used as a graphical computation method for diffraction patterns.

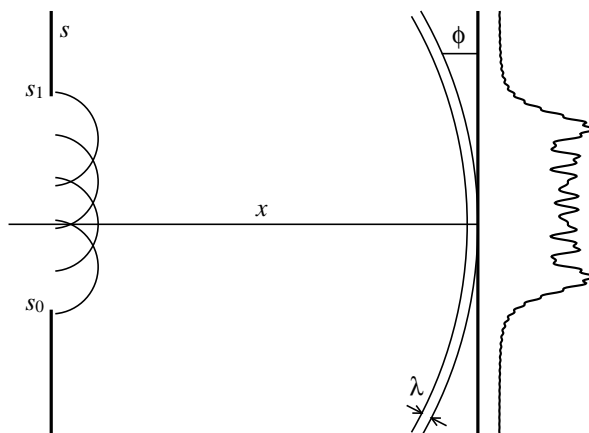


Figure 6.7. Diffraction through a slit.

The following presentation of Fresnel’s results loosely follows Preston’s 1901 *The Theory of Light* [72], which is among the earliest English-language accounts of the theory. Another readable account is Houstoun’s 1915 *Treatise on Light* [40].

Consider a monochromatic light source diffracted through a slit. Based on fundamental principles of wave optics, the wavefront emerging from the slit is the integral of point sources at each point along the slit, shown in Figure 6.7 as s_0 through s_1 . Assuming the wavelength is λ , the phase ϕ of the light emanating from point s reaching the target on the right is

⁶A. J. Fresnel, *Mémoire sur la diffraction de la lumière*, Annales de Chimie et de Physique, t.x., 288 1819. Henry Crew published a translation into English in 1900 [25]

$$\phi = \frac{2\pi}{\lambda} \sqrt{x^2 + s^2} . \quad (6.8)$$

Assuming that $s \ll x$, apply the simplifying approximation

$$\phi \approx \frac{2\pi}{\lambda} \left(x + \frac{1}{2} s^2 \right) . \quad (6.9)$$

Again assuming $s \ll x$, the intensity of the wave can be considered constant for all $s_0 < s < s_1$. Dropping the term including x (it represents the phase of the light incident on the target, but doesn't affect total intensity), and choosing units arbitrarily to simplify constants, assume $\lambda = \frac{1}{2}$, and then the intensity incident on the target is

$$I = \left[\int_{s_0}^{s_1} \cos \phi \, ds \right]^2 + \left[\int_{s_0}^{s_1} \sin \phi(s) \, ds \right]^2 = \left[\int_{s_0}^{s_1} \cos \frac{\pi}{2} s^2 \, ds \right]^2 + \left[\int_{s_0}^{s_1} \sin \frac{\pi}{2} s^2 \, ds \right]^2 . \quad (6.10)$$

The indefinite integrals needed to compute this intensity are best known as the *Fresnel integrals*:

$$\begin{aligned} S(z) &= \int_0^z \sin \left(\frac{\pi t^2}{2} \right) dt , \\ C(z) &= \int_0^z \cos \left(\frac{\pi t^2}{2} \right) dt . \end{aligned} \quad (6.11)$$

Choosing $a = 1/\sqrt{\pi}$, these integrals are obviously equivalent to the formula for the Cartesian coordinates of the Euler spiral, Equation 6.1. The choice of scale factor gives a simpler limit: $S(z) = C(z) = 0.5$ as $z \rightarrow \infty$. Fresnel gives these limits, but does not justify the result [25, p. 124].

Given these integrals, the formula for intensity, Equation 6.10, can be rewritten simply as

$$I = (S(s_1) - S(s_0))^2 + (C(s_1) - C(s_0))^2 . \quad (6.12)$$

Fresnel included in his 1818 publication a table of fifty values (with equally spaced s) to four decimal places.

Alfred Marie Cornu plotted the spiral accurately in 1874 [15] and proposed its use as a graphical computation technique for diffraction problems. His main insight is that the intensity I is simply the square of the Euclidean distance between the two points on the Euler spiral at arc length s_0 and s_1 . Cornu observes the same principle as Bernoulli's proposal of the formula for the integrals: *Le rayon de courbure est en raison inverse de l'arc* (the radius of curvature is inversely proportional to arc length), but he, like Fresnel, also seems unaware of Euler's prior investigation of the integral, or of the curve.

Today, it is common to use complex numbers to obtain a more concise formulation of the Fresnel integrals, reflecting the intuitive understanding of the propagation of light as a complex-valued wave

$$C(z) + iS(z) = \int_0^z e^{i\frac{\pi}{2}t^2} dt . \quad (6.13)$$

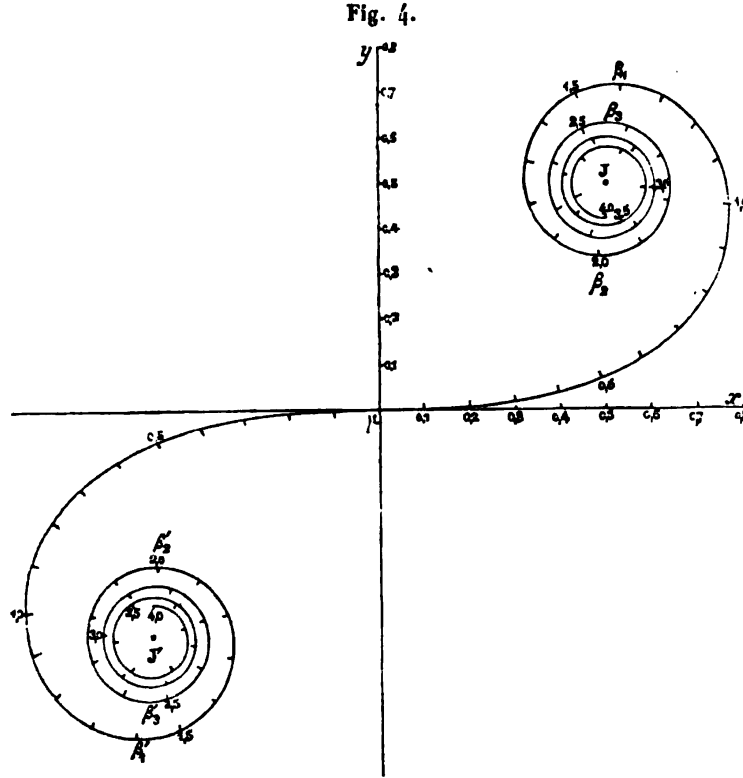


Figure 6.8. Cornu's plot of the Fresnel integrals.

Even though Euler anticipated the important mathematical results, the phrase “spiral of Cornu” became popular. At the funeral of Alfred Cornu on April 16, 1902, Henri Poincaré had these glowing words: “Also, when addressing the study of diffraction, he had quickly replaced an unpleasant multitude of hairy integral formulas with a single harmonious figure, that the eye follows with pleasure and where the spirit moves without effort.” Elaborating further in his sketch of Cornu in his 1910 *Savants et écrivains*, “Today, everyone, to predict the effect of an arbitrary screen on a beam of light, makes use of the spiral of Cornu.” (Translations from the original French are mine.)

Since apparently two names were not adequate, Ernesto Cesàro around 1886 dubbed the curve “clothoid”, after Clotho ($K\lambda\omega\theta\acute{\omega}$), the youngest of the three Fates of Greek mythology, who spun the threads of life, winding them around her distaff – since the curve spins or twists about its asymptotic points. Today, judging from the number of documents retrieved by keyword from an Internet search engine, the term “clothoid” is by far the most popular⁷. However, as Archibald wrote in 1917 [4], by modern standards of attribution, it is clear that the proper name for this beautiful curve is the Euler spiral, and that is the name used here throughout.

⁷As of 25 Aug 2008, Google search reports 17,700 results for “clothoid”, 5,660 hits for “Cornu spiral”, 935 hits for “Euler spiral”, and 17,800 for “Fresnel integral”.

6.6 Talbot’s railway transition spiral – 1890

The third completely independent discovery of the Euler spiral is in the context of designing railway tracks to provide a smooth riding experience. Over the course of the 19th century, the need for a track shape with gradually varying curvature became clear. William Rankine, in his *Manual of Civil Engineering* [73, p. 651], gives Mr. William Gravatt credit for the first such curve, about 1828 or 1829, based on a sine curve. The elastica makes another appearance, in a proposal about 1842 by William Froude to use circles for most of the curve, but ”a curve approximating the elastic curve, for the purpose of making the change of curvature by degrees.”

Charles Crandall [16, p. 1] gives priority for the “true transition curve” to Ellis Holbrook, in the *Railroad Gazette*, Dec. 3, 1880. Arthur Talbot was also among the first to approach the problem mathematically, and derived exactly the same integrals as Bernoulli and Fresnel before him. His introduction to “The Railway Transition Spiral” [84] describes the problem and his solution articulately:

A transition curve, or easement curve, as it is sometimes called, is a curve of varying radius used to connect circular curves with tangents for the purpose of avoiding the shock and disagreeable lurch of trains, due to the instant change of direction and also to the sudden change from level to inclined track. The primary object of the transition curve, then, is to effect smooth riding when the train is entering or leaving a curve.

The generally accepted requirement for a proper transition curve is that the degree-of-curve shall increase gradually and uniformly from the point of tangent until the degree of the main curve is reached, and that the super-elevation⁸ shall increase uniformly from zero at the tangent to the full amount at the connection with the main curve and yet have at any point the appropriate super-elevation for the curvature. In addition to this, an acceptable transition curve must be so simple that the field work may be easily and rapidly done, and should be so flexible that it may be adjusted to meet the varied requirements of problems in location and construction.

Without attempting to show the necessity or the utility of transition curves, this paper will consider the principles and some of the applications of one of the best of these curves, the railway transition spiral.

The Transition Spiral is a curve whose degree-of-curve increases directly as the distance along the curve from the point of spiral.

Thus, we have yet another concise statement of what Bernoulli and Euler wrote as $rs = aa$.

In his introductory figure (here reproduced as Figure 6.9), Talbot shows the spiral connecting a straight section tangent to point A to a circular arc LH (of which DLH continues the arc, and C is the center of the circle). The remainder of the paper consists of many tables of values for this spiral, as well as examples of its application to specific problems.

Talbot derives the basic equations in terms of the angle of direction (he uses Δ , apparently standing for “degrees”, having already used up θ to represent the angle BAL ; also, L is arc length, not to be confused with point L in the figure): $\Delta = \frac{1}{2}aL^2$ [84, p. 9]. He then expresses the x and y

⁸Super-elevation is the difference in elevation between the outer and inner rails, banking a moving train to reduce the lateral acceleration felt by passengers.

$$C(z) + iS(z) = z {}_1F_1\left(\frac{1}{2}; \frac{3}{2}; i\frac{\pi}{2}z^2\right). \quad (6.16)$$

The confluent hypergeometric function is

$${}_1F_1(a; b; z) = 1 + \frac{a}{b}z + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \dots = \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k} \frac{z^k}{k!}, \quad (6.17)$$

where $(a)_k$ and $(b)_k$ are *Pochhammer symbols*, also known as the “rising factorial,”

$$(a)_k = \frac{(a+k-1)!}{(a-1)!} = \frac{\Gamma(x+k)}{\Gamma(x)}. \quad (6.18)$$

Due to these equivalences, the Fresnel integrals are often considered part of a larger family of related “special functions” containing the Gamma function, erf, Bessel functions, and others.

6.8 Use as an interpolating spline

Several systems and publications have proposed the use of the Euler spiral as an interpolating spline primitive, but for some reason it has not become popular. Perhaps one reason is that many authors seem to consider the Euler spiral nothing but an approximation to a “true” spline based on the elastica.

An early reference to the Euler spiral for use as an interpolating spline is Birkhoff and de Boor’s 1965 survey of both linear and nonlinear splines [10]. There, they present the rectangular elastica as the curve simulating the mechanical spline, but also point out that an exact simulation of this curve is not “particularly desirable.” As an alternative, they suggest that “they approximate equally well to Hermite interpolation by segments of Euler’s spirals” [10, p. 172], and cite Archibald [4] as their source.

Probably the earliest actual application of Euler’s spiral for splines was in the Autokon system, developed in the beginning of the 1960’s. According to Even Mehlum [60], early versions of Autokon used the KURGLA 1 algorithm, which was a numerical approximation to the elastica. Later versions used Euler’s spiral, having derived it as a small curvature approximation to the elastica equation. (Note that KURGLA 1 is based on the elastica, but not quite the same as a true minimal energy curve (MEC), as the curve segments can take elastica forms other than the rectangular elastica). Mehlum writes: “There is hardly any visible difference between the curves the two algorithms produce, but KURGLA 2 is preferred because of its better computational economy. There is also a question of stability in the KURGLA 1 version if the total arc length is not kept under control. This question disappears in the case of KURGLA 2.”

Mehlum published both KURGLA algorithms in 1974 [59]. It is fairly clear that he considered the Euler spiral an approximation to true splines based on the elastica. He writes, “In Section 5 we make a ‘mathematical approximation’ in addition to the numerical, which makes the resulting curves of Section 5 slightly different from those of Section 4 [which is based on the elastica as a primitive]. The difference is, however, not visible in practical applications.”

The 1974 publication notes that the spline solution with linear variation of curvature and G^2 -continuity is a piecewise Cornu spiral, and that the Fresnel integrals represent the Cartesian coordinates for this curve, but does not cite sources for these facts.

Mehlum’s numerical techniques are based on approximating the Euler spiral using a sequence of stepped circular arcs of linearly increasing curvature.

Stoer [82] writes that the Euler spiral spline can be considered an approximation to the “true” problem of finding a minimal energy spline. His results are broadly similar to Mehlum’s, but he presents his algorithms in considerably more detail, including a detailed construction of a band-diagonal Jacobian matrix. He also presents an application of Euler spirals as a smoothing (approximating), rather than interpolating spline.

Stoer also brings up the point that there may be many discrete solutions to the interpolating spline problem, each G^2 continuous and piecewise Euler spiral, based on higher winding numbers. From these, he chooses the solution minimizing the total bending energy as the “best” one.

Later, Coope [14] again presents the “spiral spline” as a good approximation to the minimum energy curve, gives a good Newton approximation method with band-diagonal matrices for globally solving the splines, and notes that “for sensible end conditions and appropriately calculated chord angles convergence always occurs.” A stronger convergence result is posed as an open problem.

6.9 Efficient computation

Throughout the history of the Euler spiral, a major focus of mathematical investigation is to compute values of the integral efficiently. Fresnel, in particular, devotes many pages to approximate formulas for the definite integral when the limits of integration are close together [25], and used these techniques to produce his table. Many subsequent researchers throughout the 19th century refined these techniques and the resulting tables, including Knochenhauer, Cauchy, Gilbert, Peters, Ignatowsky, Lommel and Peters (see [4] for more detail and citations on these results).

Today, the problem can be considered solved. At least two published algorithms provide accurate results in time comparable to that needed to evaluate ordinary trigonometric functions.

One technique is that of the Cephys library [66], which uses a highly numerical technique, splitting the function into two ranges. For values of $s < 1.6$, a simple rational Chebyshev polynomial gives the answer with high precision. For the section turning around the limit point, a similar polynomial perturbs this approximation (valid for $z \gg 1$) into an accurate value,

$$C(z) + iS(z) \approx \frac{1+i}{2} - \frac{i}{\pi z} e^{i\frac{\pi}{2}z^2}. \quad (6.19)$$

Numerical Recipes [71] uses a similar technique to split the range. For the section $s < 1.5$ near the inflection point, the series given by Euler in 1744 (Equation 6.2) accurately computes the function. For the section spiraling around the limit point, the recipe calls for a continued fraction based on the erf function. Thanks to equivalences involving e^{iz^2} , this solution also converges to Equation 6.19 as $z \rightarrow \infty$ [71, p. 255].

Thus, the Fresnel integrals can be considered an ordinary “special function”, and the coordinates of the Euler spiral can be efficiently computed without fear of requiring significant resources.

Chapter 7

Four-parameter curves

For the basic problem of interpolating splines, two-parameter splines are ideal. However, when greater than G^2 -continuity is needed, or for designing particular shapes with straight-to-curve transitions, a four-parameter family of primitive curves is appropriate. Where a curve from a two-parameter family is determined by its tangent angles at the endpoints, a four-parameter curve is determined by both tangent angles and curvatures.

Section 7.1 motivates the need for four parameters in more detail. Section 7.2 discusses a well-known four-parameter spline, the Minimum Variation Curve, and presents a general equation for the shape of the MVC primitive curve. Section 7.3 presents an approximation to the MVC, where curvature is defined as a cubic polynomial of arc length. This spline has generally similar properties, but is computationally much more tractable and easier to analyze. Finally, Section 7.4 presents a technique for applying labels to control points representing different constraints, generalizing the idea of blended corner curves with G^2 -continuity to a richer set of configurations.

7.1 Why four parameters?

Chapter 4 contained a strong argument in favor of a two-parameter family of curves for interpolating splines. Why, then, consider a four-parameter family? When are two parameters not adequate?

First, while the human visual system appears to be sensitive only to G^2 -continuity, there are applications for which higher orders of continuity are important. For example, “reflection lines” resulting from the specular reflection of lines from a 3D surface show generally exhibit only C^{k-1} continuity when the underlying surface is G^k (see [39, p. 570] for a detailed discussion). Thus, to guarantee smooth reflection lines, the surface must be designed with a high order of continuity. Another application demanding higher order continuity is the design of rail layouts for high-speed trains. At the slower speeds characteristic of trains in the United States, G^2 -continuity is adequate, and clothoidal segments are the usual design norm, but in the rest of the world there has been recent attention on smoother curves [45].

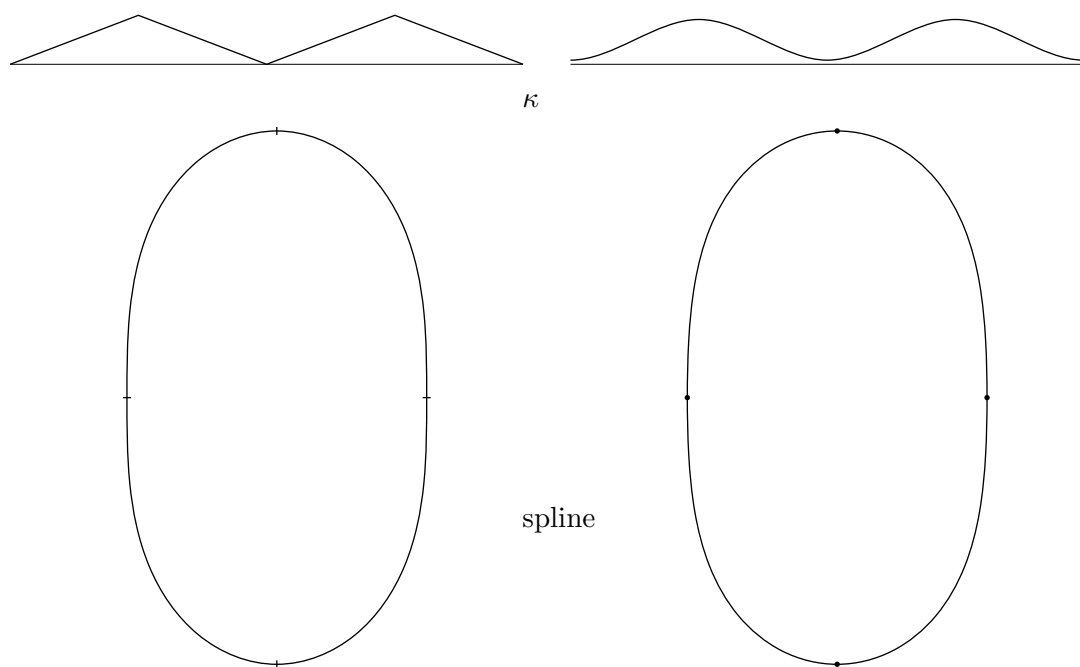


Figure 7.1. Eccentricity 0.56199 oval with G^2 and G^4 continuity.

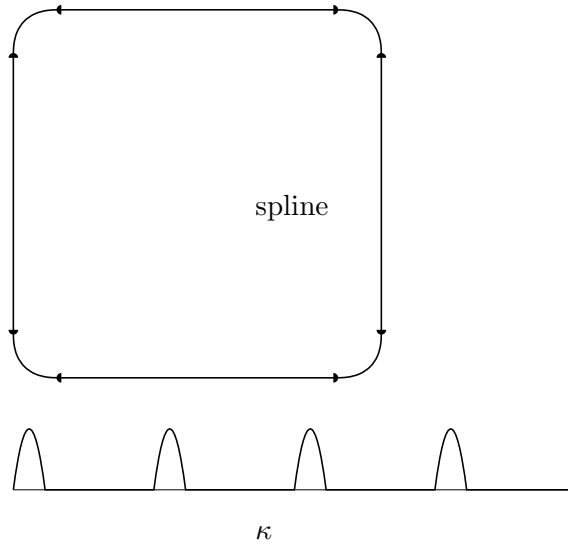


Figure 7.2. The suitcase corners problem.

Figure 7.1 shows an example of a simple closed curve (an ovoid shape with eccentricity 0.56199) drawn using both a G^2 -continuous spline and a G^4 -continuous spline. It is clear from the curvature plots that the curvature variation is smoother in the G^4 curve, but the points of discontinuity of curvature variation are not directly visible to the eye.

Second, many shapes are not best represented as a simple interpolating spline. The outline of a letterform is not, in general, a smooth undulating curve (although such examples do exist, such as the lowercase ‘c’ in Figure 4.6). In addition to sections of smooth curve, a letter form may also have sharp corners, sections of straight line, and, most importantly for this discussion, smooth transitions from straight lines to curved sections. A simple example is shown in Figure 7.2 (based on Fig. 4.10 in Moreton’s Ph.D. thesis [64]), illustrating the problem of blending smooth corners between sections of straight line.

In font design, similar situations arise for bracketed serifs, and for shapes such as the letter U, where straight sections blend smoothly with curves. A typical example showing both situations is shown in Figure 7.3, a capital U from the author’s font Cecco. The entire curve (with the exception of the sharp corners) is G^2 continuous, providing pleasing and smooth joins between the straight and curved sections.

A four-parameter curve family provides enough flexibility to meet such needs. It’s clear that a single section of Euler spiral is not adequate for one of the corners in Figure 7.2, as the curvature must start out zero, increase to a value large enough for the segment to turn by a right angle, then decrease again to zero.

Third, by annotating curves with additional constraints, it’s possible to improve locality. In particular, constraints that propagate curvature derivatives in one direction, but leave the parameters unconstrained on one side (which we call “one-way constraints”) can isolate sections of extreme changes of curvature, so that those sections do not influence nearby regions of gentle curvature change.

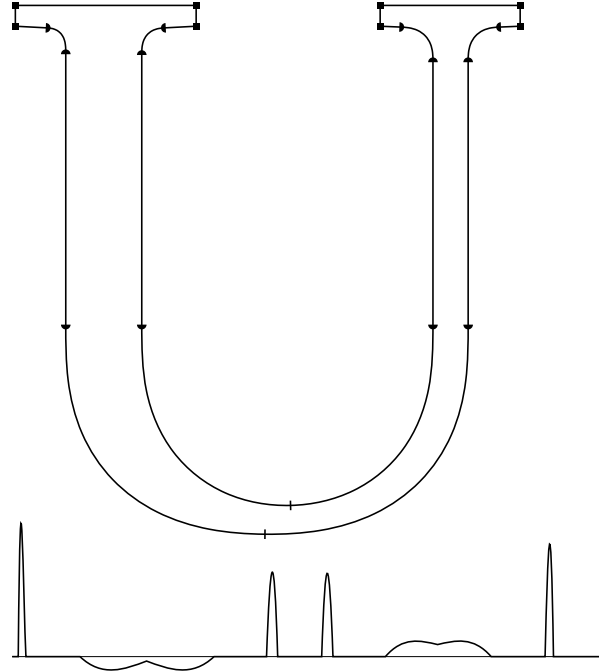


Figure 7.3. Bracketed serifs and straight-to-curve joins.

For example, Figure 7.4 shows how adding one-way constraints keeps the variation of curvature localized to just two segments of the overall spline, as opposed to the infinite extent of a pure interpolating spline. The central point can be perturbed arbitrarily without affecting the segments beyond the one-way constraints. In this figure, the top plots show curvature as a function of arc length, and the bottom plots show the control points (with symbols representing the constraints) with the interpolating curve. Note also that the curve with better locality also requires considerably more curvature (and variation of curvature) to fit the points.

There are two examples of a four-parameter primitive curve, both straightforward generalizations of two-parameter counterparts. One is the Minimal Variation Curve, which minimizes the L_2 norm of curvature variation. Another is the generalization of the Euler spiral so that curvature is an arbitrary cubic polynomial in arc length (the definition of the Euler spiral is a linear relationship). Both of these splines are G^4 -continuous and use a primitive curve which has four parameters. In

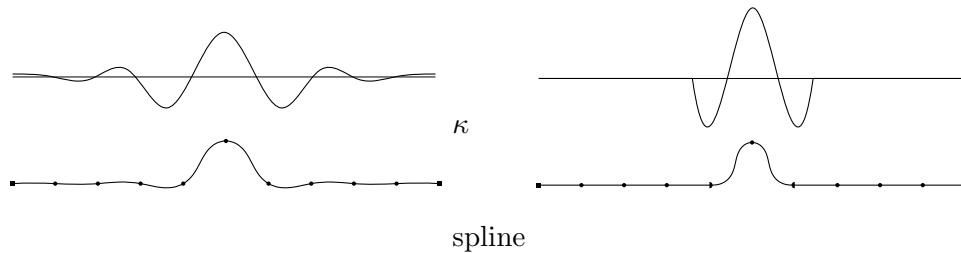


Figure 7.4. One-way constraints enforce locality.

practice, these curves are quite similar, and one can easily be considered an approximation of the other.

The parameter space of a two-parameter spline can be understood as the tangent angles of the endpoints, relative to the chord. Given arbitrary endpoint tangents, a two-parameter curve family finds a smooth curve meeting those constraints. Similarly, a four-parameter curve family finds a smooth curve meeting given tangent and curvature values at the endpoints.

7.2 Minimum Variation Curve

The Minimum Variation Curve (MVC) is defined as the curve minimizing the L_2 norm of the *variation* in curvature. Moreton's Ph.D. thesis [64] argued in favor of this spline to address known limitations of the MEC. In particular, the MEC lacks roundness, but since a circular arc has zero curvature variation, it is trivially the curve that minimizes the MVC cost functional when the input points are co-circular.

$$E_{\text{MVC}}[\kappa(s)] = \int_0^l (\kappa')^2 ds . \quad (7.1)$$

Most presentations of the MVC, including Moreton's, use a numerical approach, approximating the curve with some other approximation (in Moreton's case, quintic polynomials) and using a technique such as conjugate gradient descent to minimize the cost functional. Gumhold's more recent approach [34] is similar, also using conjugate gradient descent. However, a variational approach analogous to that of the MEC produces exact differential equations for the curve. The simple Euler–Lagrange equation is inadequate for this task, as it only handles functionals written in terms of a first derivative. Since the MVC functional is in terms of the second derivative of θ , the more general Euler–Poisson equation is needed. Section 7.2.1 recaps the general technique from the calculus of variations, then Section 7.2.2 applies it to the MVC and gives the solution. One formulation is the impressively simple $\kappa'' = \lambda y$, which is pleasingly analogous to $\kappa = \lambda y$ as the general solution to the elastica (Equation 3.13).

7.2.1 Euler–Poisson equation

Here we repeat the formulation of the general Euler–Poisson equation as given by Elsgolts [19]:

Let us investigate the extreme value of the functional

$$v[y(x)] = \int_{x_0}^{x_1} F(x, y(x), y'(x), \dots, y^{(n)}(x)) dx , \quad (7.2)$$

where we consider the function F differentiable $n + 2$ times with respect to all arguments and we assume that boundary conditions are of the form

$$y(x_0) = y_0, \ y'(x_0) = y'_0, \ \dots, \ y^{(n-1)}(x_0) = y_0^{(n-1)} ; \quad (7.3)$$

$$y(x_1) = y_1, \ y'(x_1) = y'_1, \ \dots, \ y^{(n-1)}(x_1) = y_1^{(n-1)} , \quad (7.4)$$

i.e. at the boundary points are given the values not only of the function but also of its derivatives up to the order $n - 1$ inclusive.

The function $y = y(x)$ that extremizes the functional given in Equation 7.2 must be a solution of the equation

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} + \frac{d^2}{dx^2} \frac{\partial F}{\partial y''} + \dots + (-1)^n \frac{d^n}{dx^n} \frac{\partial F}{\partial y^{(n)}} = 0 \quad (7.5)$$

7.2.2 Euler–Poisson solution of MVC

We add Lagrange multipliers to represent the constraint that the curve spans the endpoints. Adding the Lagrange multipliers and eliminating κ in favor of θ , we seek to minimize

$$E = \int_0^l (\theta'')^2 + \lambda_1 \sin \theta + \lambda_2 \cos \theta \, ds . \quad (7.6)$$

To apply this equation to the MVC, we substitute $x = s$ and $y(x) = \theta(s)$, so that

$$F(x, y, y', y'') = (y'')^2 + \lambda_1 \sin y + \lambda_2 \cos y . \quad (7.7)$$

The relevant partial derivatives are

$$\frac{\partial F}{\partial y} = \lambda_1 \cos y - \lambda_2 \sin y , \quad (7.8)$$

$$\frac{\partial F}{\partial y'} = 0 , \quad (7.9)$$

$$\frac{\partial F}{\partial y''} = 2y'' , \quad (7.10)$$

$$\frac{d^2}{dx^2} \frac{\partial F}{\partial y''} = 2y'''' . \quad (7.11)$$

Applying 7.5 and reversing the variable substitution, we derive an ordinary differential equation for the MVC,

$$\lambda_1 \cos \theta - \lambda_2 \sin \theta + 2\theta'''' = 0 . \quad (7.12)$$

Because at this point we are concerned only about the existence of the Lagrange multipliers, not their actual values, we can ignore the constant factor of $\frac{1}{2}$ and $-\frac{1}{2}$ for λ_1 and λ_2 , respectively.

$$\theta'''' + \lambda_1 \cos \theta + \lambda_2 \sin \theta = 0 . \quad (7.13)$$

Note the pleasing similarity to the elastica (MEC) equation,

$$\theta'' + \lambda_1 \cos \theta + \lambda_2 \sin \theta = 0 . \quad (7.14)$$

As a special case when λ_1 and λ_2 become zero, all second order polynomial spirals $\kappa(s) = k_0 + k_1s + k_2s^2$ are solutions to the MVC equation, as the the third derivative of κ (the fourth derivative of θ) vanishes. These solutions are analogous to circular arcs $\kappa(s) = k_0$ as solutions of the elastica when the corresponding Lagrange multipliers become zero.

Again as in the MEC case, we can reformulate this equation solely in terms of first and higher derivatives of θ , also eliminating the Lagrange multipliers.

First, we differentiate Equation 7.13, giving

$$\theta^{(5)} + \theta'(-\lambda_1 \sin \theta + \lambda_2 \cos \theta) . \quad (7.15)$$

To integrate Equation 7.13, we first multiply by θ' , giving

$$\theta'''\theta' + \theta'(\lambda_1 \cos \theta + \lambda_2 \sin \theta) = 0 . \quad (7.16)$$

This equation can readily be integrated, giving

$$\theta'''\theta' - \frac{(\theta'')^2}{2} + (\lambda_1 \sin \theta - \lambda_2 \cos \theta) + C = 0 . \quad (7.17)$$

Combining with Equation 7.15,

$$\theta^{(5)} + \theta''(\theta')^2 - \theta'(\theta'')^2/2 + C\theta' = 0 . \quad (7.18)$$

Rewriting in terms of κ ,

$$\kappa'''' + \kappa''\kappa'^2 - \frac{\kappa(\kappa')^2}{2} + C\kappa = 0 . \quad (7.19)$$

When the length is not constrained, the constant C becomes zero. This follows from transversality, a standard technique in the calculus of variations [19, p. 345], and is analogous to the way the corresponding constant becomes zero in the case where the MEC is not length-constrained. Then, the equation becomes

$$\kappa'''' + \kappa''\kappa'^2 - \frac{1}{2}\kappa(\kappa')^2 = 0 . \quad (7.20)$$

Equation 7.20 was also given by Brook et al [13]; their solution does not encompass additional length constraints (nonzero values of C). This equation is well suited to standard numerical techniques such as Runge–Kutte integration for very precise computation of the shape of the curve.

The MVC spline is defined as the curve that globally minimizes the MVC metric over the space of all curves that interpolate the input points. As in the MEC, this curve is made of segments of a parametrized primitive curve, joined together at the control points with continuity constraints. The MVC primitive curve (defined by Equation 7.20) is four parameter, and it is G^4 -continuous at the control points, as opposed to two parameters and G^2 for the MEC.

At the endpoints of the MVC, $\kappa' = 0$ and $\kappa'' = 0$ (again, analogous to the endpoint condition of the MEC, $\kappa = 0$). It's fairly easy to see why the latter conditions hold. At an endpoint, it is

possible to add at least points lying on a circular arc of matching tangent and curvature. The resulting MVC spline must assign exactly that circular arc to the additional segments, as the value of the MVC functional itself for the arc is zero, and if there is any other curve with a lower value for the functional, then it would have been a better solution for the original problem. And since the MVC is G^4 -continuous, then κ' and κ'' must match the values for a circular arc, which are zero. These endpoint conditions can also be derived straightforwardly (if abstractly) from the transversality condition, a principle of the calculus of variation which determines scalar parameters from the solution space to minimize the global functional (see a textbook such as Elsgolts [19] for more details).

There is also a simple equation for the MVC relating curvature and a Cartesian coordinate, analogous to $\kappa = \lambda y$ for the MEC case. To derive it, we first assume without loss of generality that $\lambda_1 = 0$; unlike the formulation in terms of curvature alone, the new equation will not be rotationally invariant. We then observe that $\frac{dy}{ds} = \sin \theta$. Substituting that into Equation 7.13, we get

$$\theta'''' + \lambda_2 y' = 0. \quad (7.21)$$

Integrating (the constant of integration can be set at zero, since that corresponds to a translation of the curve), substituting $\kappa = \theta'$, and adjusting the constants so $\lambda = -\lambda_2$, the resulting simple formulation is

$$\kappa'' = \lambda y. \quad (7.22)$$

When $\lambda = 0$, this equation is equivalent to that of the Euler spiral, in much the same way that $\lambda = 0$ is the value for which the elastica becomes a circle. Kimia et al [44] erroneously gave this as the general solution to the MVC equation, correctly observing that the Euler spiral is a special case of the MVC, but they apparently overlooked the possibility of nonzero values of λ .

7.3 Polynomial spiral spline

Another approach to four-parameter splines is to generalize the Euler spiral to an implicit equation where curvature is a cubic polynomial function of arc length. The general equation is:

$$\kappa(s) = a + bs + cs^2 + ds^3. \quad (7.23)$$

This curve family is a very good approximation to the MVC. Assuming small turning angles, y is nearly proportional to s , so substituting Equation 7.22, $\kappa'' \approx c_1 s + c_2$. Integrating twice (and bringing about two more constants of integration) yields Equation 7.23. This approximation is analogous to that of the Euler spiral approximating the MEC, but can be expected to be even closer, as the nonlinear terms affect only higher-order derivatives of curvature.

As in the MVC, the corresponding spline can be defined in terms of enforcing G^4 -continuity across control points. These are four constraints per control point. Since each segment requires four constraints to determine the four parameters of the polynomial, for an open curve there are four additional parameters, or two for each endpoint. One approach is to set tangent and curvature explicitly, but a more natural approach (in the interpolating spline framework) is to set $\kappa' = 0$ and

$\kappa'' = 0$ at the endpoints, as in the MVC. In this respect, the cubic polynomial spline is even more closely related to the MVC than the Euler spiral is to the MEC, because in the latter case the endpoint conditions differ; the natural end condition is constant curvature (a circular arc), while the natural endpoint condition for the MEC is zero curvature. In the four-parameter world, the MVC endpoint conditions make sense and there is no need to adjust them.

The cubic polynomial spiral spline was first proposed by Ohlin [67], by analogy of making the pentic (polynomial) spline nonlinear in the same way that the MEC is the nonlinear counterpart of the cubic spline. Although Ohlin did express the natural variational formulation of polynomial splines, and cited Mehlum’s work indicating that the Euler spiral spline was a consistent (extensional) approximation to the MEC, he did not explicitly mention the MVC or state that this new spline is a good approximation to it.

The idea of a curve defined as curvature being a cubic polynomial function of arc length goes back even further. In 1936, Bloss proposed a railroad transition spiral in which the curvature is a simple cubic polynomial with respect to arc length [45]. The Euler spiral (or clothoid, as it is nearly universally known in that application domain) is popular in railroad track design. The linear change in curvature is desirably smooth on the interior of the segment, but the literature has also long recognized the limitations of smoothness at the joins between clothoid segments. These higher derivatives of curvature become more significant as train speeds increase. In particular, high speed train tracks are almost invariably banked (or “superelevated”) so that the rail on the outside of the curve is higher than the inside, to a degree roughly proportional to the curvature. Thus, passengers of a train traversing the tracks experience a *roll acceleration* roughly equal to the *second* derivative of curvature. At the join of two Euler spiral segments, this roll acceleration is an impulse. Usually, the suspension of the train provides enough cushioning to avoid an unpleasant lurch, but curves with better continuity properties are clearly desirable for a smooth, quiet ride. The Bloss transition spiral is a very early example of a curve with higher continuity, to avoid such impulses.

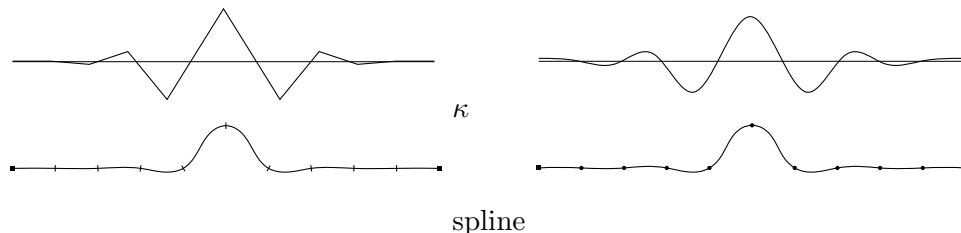


Figure 7.5. G^2 spline has better locality than G^4 .

If G^4 -continuity is smoother than G^2 , then why not use it everywhere? Unfortunately, the additional smoothness comes at the cost of locality. As shown in Figure 7.5, the effects of perturbing a point ripple out almost twice as far as for the Euler spiral spline. Since, for most applications, the difference in smoothness is theoretical, rather than visible to the eye, it’s better to enjoy the superior locality properties of the G^2 -continuous spline.

7.4 Generalized constraints

The motivation for four-parameter splines showed a number of specific examples of blended corners, straight-to-curve transitions, and so on. This section presents a highly general technique that subsumes all these examples, and provides a broad palette of curve transitions to the designer.

A single segment of cubic polynomial spline has four free parameters. In general, these can fit any tangent and curvature constraints at the endpoints. One such example is to fit the given tangent constraints of adjoining straight segments, as well as zero curvature, as in the blended corners of Figure 7.2.

The approach of this section will be to start from special cases, then progressively generalize. First, observe that the input to the solver for Figure 7.2 must distinguish which segments are straight and which are the curved corners. We propose the notation of annotating points with half-circles. A segment between the two straight sides of the half-circle is a straight line. A segment between the two rounded sides is a blended corner. For such segments, the solver fits a four-parameter spline segment to match the tangent and curvature of the adjoining straight segment.

The next generalization is to allow more than one segment in the curved section. In fact, any G^4 -continuous spline interpolating a sequence of control points has two free parameters at the endpoints. For the most natural end conditions, these are determined by constraining $\kappa' = 0$ and $\kappa'' = 0$, but it's also possible to determine them by constraining tangent and curvature constraints to specific values. In the case of the right half of Figure 7.4, there are two segments bracketed between the half-circle sides of the one-way constraints, and the tangents and curvatures are, as above, matched to the adjoining straight segments. One important observation is that perturbing the central point does not affect the curve beyond the one-way constraints.

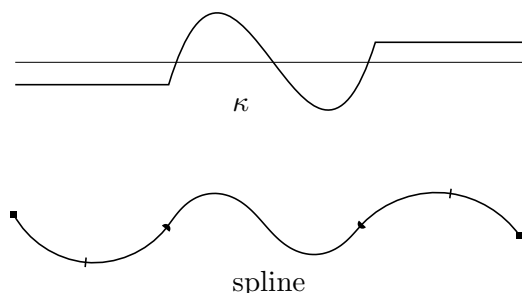


Figure 7.6. Matching nonzero curvatures.

The next generalization is to match adjoining curvatures other than zero. While straight-to-curve transitions are common, sometimes the adjoining section will have some curvature, rather than be straight. An example is shown in Figure 7.6. The general technique for the solver is to compute the adjoining curve segments first, treating the flat side of the one-way constraint the same as an endpoint, then use the tangent and curvature value at that point to constrain the spline segment on the curved side. Generalizing further, the one-way constraint is treated as a dependency, so that the spline on the curved side depends on the tangent and curvature of the other side, but not vice versa. In many cases, it's possible to solve the entire spline by doing a topological sort of the segments, and solving in order of dependency.

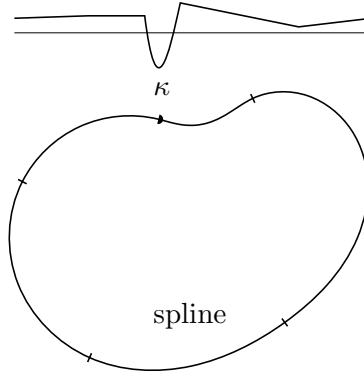


Figure 7.7. A one-way constraint creates a cyclic dependency.

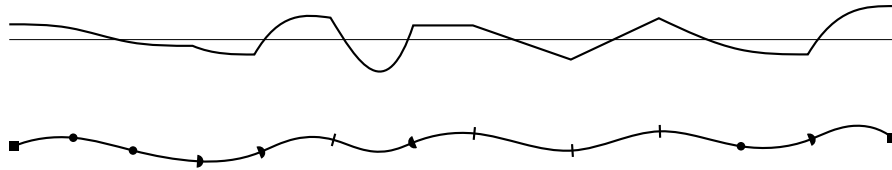


Figure 7.8. Examples of different kinds of constraint points.

However, even this approach is not maximally general. Allowing arbitrary annotations of control points can create a cyclic dependency, as in Figure 7.7, which even so has a meaningful solution. More importantly, the techniques above only allow G^4 -continuous joins for interior points, while, as we have seen, Euler spiral segments (G^2 -continuous joins) have better locality properties.

In the general approach, each segment between two adjacent control points is assigned a cubic polynomial spiral segment (i.e. as defined by Equation 7.23). Note that an Euler spiral segment fits into this framework; the parameters c and d are both zero. Equivalently, κ'' is constrained to be zero at both endpoints. For all types of internal points (other than corners, which are considered endpoints) are also constrained to have G^2 -continuity, in other words, both tangent angle θ and curvature κ are constrained to be equal on both sides of a control point. Different points have additional constraints, depending on the type of point.

The total number of degrees of freedom (and thus, number of constraints) is four times the number of segments. If each endpoint has two constraints and each internal point has four, then the total number of constraints is satisfied.

The complete list of point types, and associated mathematical constraints, is given in the table below.

For a given sequence of points with constraint annotations, the resulting spline is defined as the curve that is piecewise cubic polynomial spline, such that the tangent, curvature, and higher-order derivatives satisfy the constraints given in the table above for each point.

Figure 7.4 shows a sequence of somewhat arbitrary constraints, parallel with a curvature plot



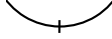

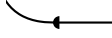
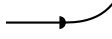
Type of constraint	Formula	Symbol
Left endpoint	$\kappa_r' = 0$	
	$\kappa_r'' = 0$	
Right endpoint	$\kappa_l' = 0$	
	$\kappa_l'' = 0$	
G^2 curve point	$\theta_l = \theta_r \quad \kappa_l'' = 0$	
	$\kappa_l = \kappa_r \quad \kappa_r'' = 0$	
G^4 curve point	$\theta_l = \theta_r \quad \kappa_l' = \kappa_r'$	
	$\kappa_l = \kappa_r \quad \kappa_l'' = \kappa_r''$	
One-way (curve to terminal)	$\theta_l = \theta_r \quad \kappa_r' = 0$	
	$\kappa_l = \kappa_r \quad \kappa_r'' = 0$	
One-way (terminal to curve)	$\theta_l = \theta_r \quad \kappa_l' = 0$	
	$\kappa_l = \kappa_r \quad \kappa_l'' = 0$	

Table 7.1. Constraints for curve points, and their corresponding formulae.

which shows the effect of the different types of constraints in the curvature domain. All segments between two G^2 curve points have linear curvature. All segments between a G^2 point and an endpoint (or between a G^2 point and the straight side of a one-way constraint) have constant curvature, i.e. the segment is a circular arc. All of the points have at least G^2 continuity, although if corner points were added, those would have only G^0 .

Section 8.3 presents numerical techniques for satisfying such a system of constraints, and Chapter 9 discusses techniques for efficiently drawing the polynomial spline curves that result. With all these components in place, the spline is both highly flexible for drawing the types of shapes likely to arise in font design, and also entirely practical for interactive editing.

Chapter 8

Numerical toolbox

One of the reasons for the enduring popularity of polynomial-based splines is the ease of computation, and straightforward numerical properties such as stability. By contrast, variational techniques are considered to be extremely slow and unwieldy. For example, Moreton’s MVC solver took about 2.5 billion CPU cycles (75 s on the hardware of the day) to compute the MVC solution to the 7-point data set posed by Woodford [90]. Even though more refined techniques have since been published, the prejudice continues. Kimia et al [44] state, “The major shortcomings of the elastica are that (i) they are not scale-invariant, (ii) they are computationally expensive to derive, and (iii) they do not lead to circular completion curves when the point-tangent pairs are, in fact, co-circular (i.e., both lie on the same circle).”

This chapter presents a toolbox of numerical techniques for computing sophisticated splines quickly, and with high precision. Section 8.1 presents a polynomial approximation to a fundamental integral that encompasses both the Euler spiral and the general polynomial spiral used as a primitive in Chapter 7. Section 8.2 presents a very efficient solver fitting an Euler spiral segment to arbitrary endpoint tangent constraints. Section 8.3 presents a fast Newton solver that globally satisfies the continuity constraints for the interpolating spline. Section 8.4 presents a fast 2-D lookup table approach for solving arbitrary two-parameter extensional splines using a single code framework.

The combination of these techniques, along with optimized drawing code, presented in Chapter 9, yields an interpolating spline that is extremely efficient to compute and easily provides interactive response speeds. The complete spline is implemented and released in the open source libspiro library, which has also been integrated into FontForge, a font editing tool, and Inkscape, a vector drawing editor.

The software release accompanying this thesis also includes a JavaScript implementation, which runs in standard Web browsers.

8.1 Numerical integration of polynomial spirals

Almost all calculations involving polynomial spiral curves can be reduced to an evaluation of a single definite integral,

$$spiro(k_0, k_1, k_2, k_3) = \int_{-0.5}^{0.5} e^{i(k_0 s + \frac{1}{2}k_1 s^2 + \frac{1}{6}k_2 s^3 + \frac{1}{24}k_3 s^4)} ds. \quad (8.1)$$

Here, $k_0 \dots k_3$ are the curvature and its first three derivatives at the center of a curve segment of arc length 1. More generally (for an arbitrary degree polynomial),

$$spiro(\mathbf{k}) = \int_{-0.5}^{0.5} e^{i \sum_{0 \leq j} \frac{1}{(j+1)!} k_j s^{j+1}} ds. \quad (8.2)$$

Fixing the limits to ± 0.5 simplifies the math and (as will be seen) creates symmetries which can be exploited to improve numerical accuracy, but does not lose generality. An integral over any limits can be expressed in terms of the spiro integral, as follows:

$$\begin{aligned} \int_{s_0}^{s_1} e^{i(k_0 s + \frac{1}{2}k_1 s^2 + \frac{1}{6}k_2 s^3 + \frac{1}{24}k_3 s^4)} ds = \\ (s_1 - s_0) e^{i(k_0 s + \frac{1}{2}k_1 s^2 + \frac{1}{6}k_2 s^3 + \frac{1}{24}k_3 s^4)} \\ spiro((s_1 - s_0)(k_0 + k_1 s + \frac{1}{2}k_2 s^2 + \frac{1}{6}k_3 s^3), \\ (s_1 - s_0)^2(k_1 + k_2 s + \frac{1}{2}k_3 s^2), \\ (s_1 - s_0)^3(k_2 + k_3 s), \\ (s_1 - s_0)^4 k_3), \end{aligned} \quad (8.3)$$

$$\text{where } s = \frac{s_0 + s_1}{2}.$$

Note that, as a special case of Equation 8.3, the classic Fresnel integrals also reduce readily to this integral.

$$\int_0^s e^{it^2} dt = s e^{i \frac{s^2}{8}} spiro(\frac{s}{2}, s^2, 0, 0). \quad (8.4)$$

The spiro integral can be thought of as analogous to the sine and cosine functions of the generalized polynomial spiral. This connection is clear when considering a circular arc.

$$\frac{\sin x}{x} = spiro(2s, 0, 0, 0). \quad (8.5)$$

The absolute value of the spiro integral is the ratio of chord length to arc length of the curve segment (note that expressing this quantity as a ratio is invariant to scaling, and it is obviously invariant to rotation and translation as well). For a given curve segment with chord length Δx , the curvature and its derivatives along the curve (with arc length s normalized to the range $[-0.5, 0.5]$) are given by a simple formula,

$$\kappa^{(j)}(s) = \left(\frac{|spiro(\mathbf{k})|}{\Delta x} \right)^{j+1} \sum_{0 \leq i} \frac{k_{i+j}}{(i+1)!} s^i. \quad (8.6)$$

8.1.1 Polynomial approximation of the integral

There are a number of approaches to numerically computing an integral such as Equation 8.1. However, a straightforward numerical integration technique such as Runge–Kutte would converge only relatively slowly, requiring significant computation time to achieve high accuracy. The approach of this section is to compute polynomial approximations of the integral, keeping all terms up to a certain order and discarding the rest.

A key insight is that the angle $\theta(s)$ is a straightforward polynomial function of s . Thus, to integrate $e^{i\theta(s)}$, take the Taylor’s series expansion of e^x , substitute the polynomial $i\theta(s)$ for x , and expand symbolically, keeping terms only up to the given order.

$$\theta(s) = k_0 s + \frac{k_1}{2} s^2 + \frac{k_2}{6} s^3 + \frac{k_3}{24} s^4. \quad (8.7)$$

Below is a worked example retaining terms for e^{ix} up to quadratic, then integrating. The approximation is

$$e^{ix} = 1 + ix - \frac{1}{2}x^2 + O(x^3). \quad (8.8)$$

Substituting $x = \theta(s)$, we get

$$e^{i\theta(s)} = 1 + ik_0 s - \frac{ik_1 - k_0^2}{2} s^2 + O(s^3). \quad (8.9)$$

Integrating, we get

$$\int e^{i\theta(s)} ds = f(s) = s + \frac{ik_0}{2} s^2 - \frac{ik_1 - k_0^2}{6} s^3 + O(s^4). \quad (8.10)$$

Evaluating $f(0.5) - f(-0.5)$, we find that all even polynomial terms (including s^4) drop out:

$$f(0.5) - f(-0.5) \approx 1 + \frac{ik_1 - k_0^2}{24}. \quad (8.11)$$

Thus, as promised, fixing the limits of integration to ± 0.5 provides a symmetry that both decreases the number of terms to compute and also increases the accuracy by one degree.

Retaining all terms up to quartic yields an approximation not much more complicated. Note that this approximation actually uses all four parameters.

$$\Delta x \approx 1 + \frac{ik_1 - k_0^2}{24} + \frac{ik_3 - 6ik_0^2 k_1 + k_0^4 - 4k_0 k_2 - 3k_1^2}{1920}. \quad (8.12)$$

8.1.2 Higher-order polynomial approximations

Higher-order polynomials offer a good way to increase the accuracy of the approximation at modest cost. However, manually working out the coefficients becomes quite tedious. This section describes a technique for automatically generating code for an arbitrary polynomial degree.

For functions of a single variable, the Taylor's series expansion gives the coefficients of the approximating polynomial simply and unambiguously (it is, of course, possible to use refinements such as Chebyshev polynomials to redistribute the errors more evenly, but the underlying concept remains). Here, the task is to approximate a function of four parameters. It is possible to generalize the Taylor's series expansion to functions of more than one parameter, but a simplistic approach would result in a blowup of coefficients; for a function of four parameters, the number of coefficients grows quadratically with the order of the polynomial.

A more sophisticated approach is represent each successive x^j as a vector of polynomial coefficients. Because of the symmetries of the equation (many odd terms will drop out), for $j \geq 3$, it is most efficient to compute x^j by combining the vectors for x^{j-2} and x^2 . Then, these polynomials are substituted into the Taylor's series expansion for e^x . A fully worked out example, for order 8, is given below.

$$\begin{aligned}
t_{11} &= k_0 \\
t_{12} &= \frac{k_1}{2} \\
t_{13} &= \frac{k_2}{6} \\
t_{14} &= \frac{k_3}{24} \\
t_{22} &= t_{11}^2 \\
t_{23} &= 2t_{11}t_{12} \\
t_{24} &= 2t_{11}t_{13} + t_{12}^2 \\
t_{25} &= 2(t_{11}t_{14} + t_{12}t_{13}) \\
t_{26} &= 2t_{12}t_{14} + t_{13}^2 \\
t_{27} &= 2t_{13}t_{14} \\
t_{28} &= t_{14}^2 \\
t_{34} &= t_{22}t_{12} + t_{23}t_{11} \\
t_{36} &= t_{22}t_{14} + t_{23}t_{13} + t_{24}t_{12} + t_{25}t_{11} \\
t_{38} &= t_{24}t_{14} + t_{25}t_{13} + t_{26}t_{12} + t_{27}t_{11} \\
t_{44} &= t_{22}^2 \\
t_{45} &= 2t_{22}t_{23} \\
t_{46} &= 2t_{22}t_{24} + t_{23}^2 \\
t_{47} &= 2(t_{22}t_{25} + t_{23}t_{24}) \\
t_{48} &= 2(t_{22}t_{26} + t_{23}t_{25}) + t_{24}^2 \\
t_{56} &= t_{44}t_{12} + t_{45}t_{11} \\
t_{58} &= t_{44}t_{14} + t_{45}t_{13} + t_{46}t_{12} + t_{47}t_{11} \\
t_{66} &= t_{44}t_{22} \\
t_{67} &= t_{44}t_{23} + t_{45}t_{22} \\
t_{68} &= t_{44}t_{24} + t_{45}t_{23} + t_{46}t_{22} \\
t_{78} &= t_{66}t_{12} + t_{67}t_{11} \\
t_{88} &= t_{66}t_{22} \\
u &= 1 - \frac{t_{22}}{24} - \frac{t_{24}}{160} - \frac{t_{26}}{896} - \frac{t_{28}}{4608} + \frac{t_{44}}{1920} + \frac{t_{46}}{10752} + \frac{t_{48}}{55296} - \frac{t_{66}}{322560} - \frac{t_{68}}{1658880} + \frac{t_{88}}{92897280} \\
v &= \frac{t_{12}}{12} + \frac{t_{14}}{80} - \frac{t_{34}}{480} - \frac{t_{36}}{2688} - \frac{t_{38}}{13824} + \frac{t_{56}}{53760} + \frac{t_{58}}{276480} - \frac{t_{78}}{11612160}
\end{aligned} \tag{8.13}$$

Obviously, for higher degrees, the formulas are too complex to be worked out by hand. The accompanying software distribution includes a utility, `numintsynth.py`, that synthesizes C code for an arbitrary degree. Asymptotically, the number of lines of code grows as $O(n^2)$, so for any reasonably low degree, the code is quite concise. Additionally, the structure of the dependency

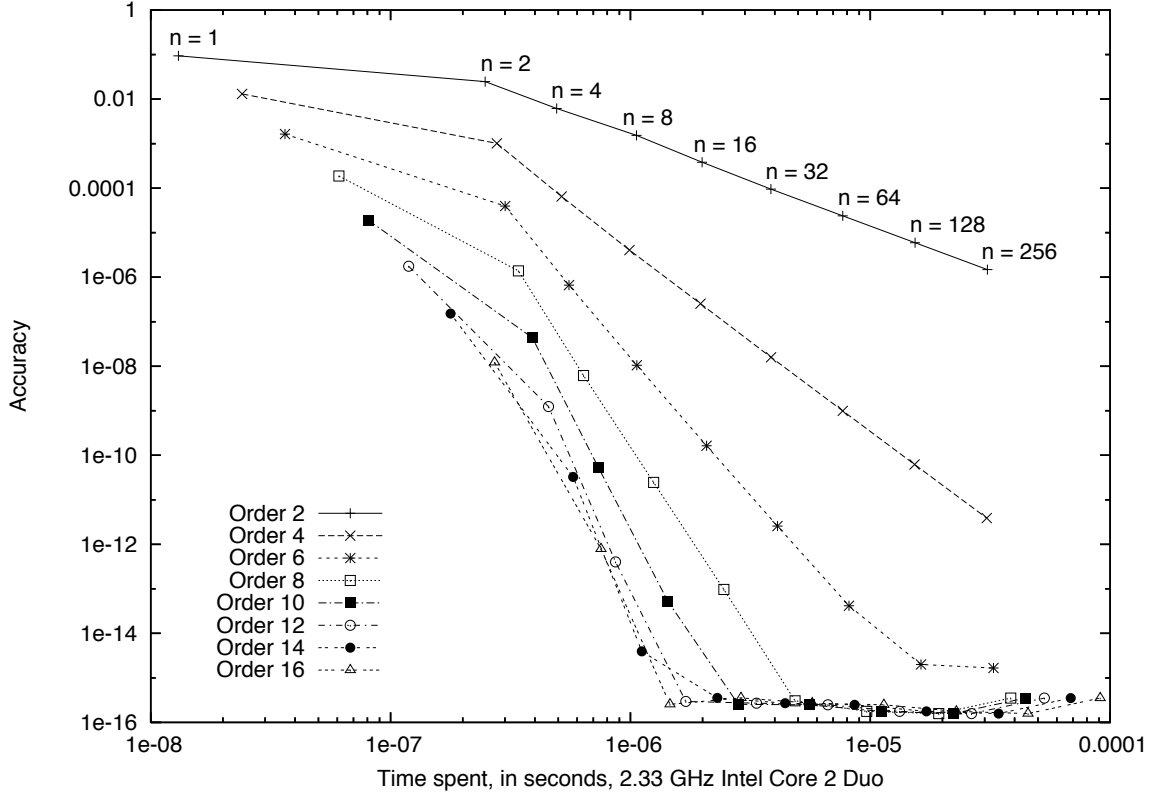


Figure 8.1. Precision vs. timing for different integration strategies.

graph permits a high level of instruction-level parallelism, making the computation very efficient on typical CPUs.

8.1.3 Hybrid approach

The higher order the polynomial, the more accurate the result. However, code complexity grows significantly, and for small angles, accurate results are possible with a low-order polynomial.

A hybrid approach is to implement a polynomial of fixed order, and use it for parameter values where the error is within tolerance. For parameter values outside this tolerance, subdivide. The range $[-0.5, 0.5]$ is split into n equal subdivisions, each is evaluated using Equation 8.3, and the result is the sum of all such intervals.

The extreme case of an order-2 polynomial is equivalent to Euler integration. The code is simple, but convergence is expected to be poor.

The optimum order of polynomial is determined empirically. All even orders of polynomial between 2 and 16 were measured, and the accuracy plotted against CPU time. The results are shown in Figure 8.1.3, measured on a MacBook Pro with a 2.33 GHz Core 2 Duo processor. Each successive datapoint doubles the total number of subdivisions. This particular plot is for evaluating $\text{spiro}(1, 2, 3, 4)$, but other combinations of arguments all give similar results.

The $n = 1$ case is relatively faster than $n > 1$, because no sine and cosine functions are needed for the subdivision. As can be seen, low order polynomials require a large number of subdivisions to provide accurate results. However, there is diminishing return as the order of the polynomial grows – while the number of intervals needed is decreased, the time needed to compute each one increases. As can be seen from the graph, orders greater than 12 don't improve overall performance significantly.

An empirically determined estimate for the error of the order 12 spiro approximation is:

$$|spiro_{12}(k_0, k_1, k_2, k_3) - spiro(k_0, k_1, k_2, k_3)| \approx (0.006|k_0|^2 + 0.03|k_1| + 0.03|k_2|^{\frac{2}{3}} + 0.025|k_3|^{\frac{1}{2}})^6 \quad (8.14)$$

With n subdivisions, the error scales approximately as n^{-12} . Thus, it should be straightforward to compute a value for any given precision.

8.2 Determining Euler spiral parameters

One important primitive is to compute the parameters for Euler spiral segments, given a fixed chord length and tangents relative to the chord. This primitive is needed in the global determination of angles to satisfy the C^2 curvature continuity constraints across knots, and is also independently of interest in vision applications, for computing a Euler spiral segment for use in shape completion.

There are quite a number of algorithms for this in the literature, the most detailed of which is given by Kimia et al [44]. Their solution represents the curve in terms of two parameters:

$$\kappa(s) = \gamma s + \kappa_0, \text{ with } 0 \leq s \leq L \quad (8.15)$$

Numerically, their solution is complicated. They first construct a biarc to approximate the desired spiral segment, then use a gradient descent optimization method for the two parameters, computing the Euler spiral at each iteration.

Our solution is similar in many respects, but more direct; the biarc as an initial approximation is no longer needed. There are two key insights. First, by specifying the curvature of the curve from the middle, as opposed to one of the endpoints, one of the parameters has a closed-form solution. Second, it is easier to compute Euler spiral segments of fixed arc length than fixed chord length. Once that's done, scaling the curve to fit the original chord length constraint does not affect the angle constraint.

We represent the curve to solve as

$$\kappa(s) = k_0 + k_1 s, \text{ with } -0.5 \leq s \leq 0.5. \quad (8.16)$$

Integrating, and fixing $\theta(0) = 0$,

$$\theta(s) = k_0 s + \frac{k_1}{2} s^2. \quad (8.17)$$

Again, the curve can be given in parametric form simply by integrating the angle:

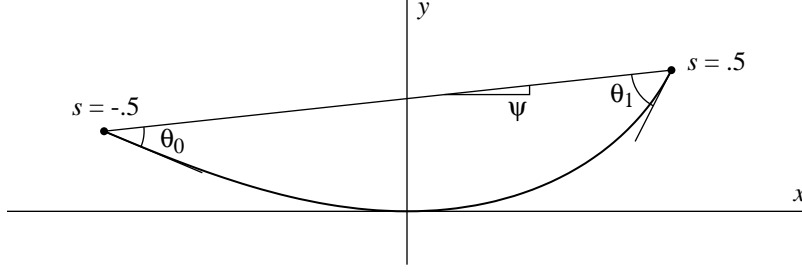


Figure 8.2. Determining Euler spiral parameters.

$$x(s) + iy(s) = \int_0^s e^{i\theta(t)} dt . \quad (8.18)$$

The endpoint angles with respect to the chord can be represented in terms of the angle of the chord in this coordinate frame.

$$\theta_0 = \psi - \theta(-0.5) , \quad (8.19)$$

$$\theta_1 = \theta(0.5) - \psi , \text{ where} \quad (8.20)$$

$$\psi = \arctan \frac{y(0.5) - y(-0.5)}{x(0.5) - x(-0.5)} . \quad (8.21)$$

But the formula for $\theta(s)$ is known (Equation 8.17), so

$$\theta_0 = \frac{k_0}{2} - \frac{k_1}{8} + \psi , \quad (8.22)$$

$$\theta_1 = \frac{k_0}{2} + \frac{k_1}{8} - \psi . \quad (8.23)$$

From Equations 8.22 and 8.23 we can immediately solve for k_0 :

$$k_0 = \theta_0 + \theta_1 . \quad (8.24)$$

We have now reduced the search space to a single dimension, k_1 . We will use a simple polynomial approximation to derive an initial guess, an approximate Newton–Raphson method to derive a second guess, and the secant method to further refine the approximation.

Using the small angle approximation $e^{i\theta} \approx 1 + i\theta$, Equation 8.18 becomes

$$\begin{aligned} x(s) + iy(s) &\approx \int_0^s 1 + i\theta(t) dt \\ &= s + i \left(\frac{k_0}{2} s^2 + \frac{k_1}{6} s^3 \right) . \end{aligned} \quad (8.25)$$

Combining Equations 8.21 and 8.25 (note the cancellation of even terms, just as for the integration in the previous subsection),

$$x(.5) - x(-.5) \approx 1, \quad (8.26)$$

$$y(.5) - y(-.5) \approx \frac{k_1}{24}, \quad (8.27)$$

$$\psi \approx \arctan \frac{k_1}{24}. \quad (8.28)$$

Applying the small angle assumption $\arctan \theta \approx \theta$ and substituting Equation 8.22,

$$\theta_0 \approx \frac{k_0}{2} - \frac{k_1}{12}, \quad (8.29)$$

$$\theta_1 \approx \frac{k_0}{2} + \frac{k_1}{12}. \quad (8.30)$$

Thus, we can derive a good approximation to k_1 :

$$k_1 \approx 6(\theta_1 - \theta_0). \quad (8.31)$$

This approximation may be sufficient for low-precision applications. A considerably more accurate technique is the secant method, which excels at finding the roots of nearly-linear one-dimensional functions. In this application, the function to be solved is angle error as a function of k_1 , with θ_0 and θ_1 given.

The secant method needs two initial guesses. For the first, we choose $k_1 = 0$, simply because the angle error is trivial to calculate at that value. For the second guess, the approximation of equation 8.31 works well for $\theta_0 + \theta_1 < 6$ or so, but as this value approaches 2π , it significantly overestimates k_1 and sometimes fails to converge. The guess $6(\theta_1 - \theta_0)(1 - \frac{\theta_0 + \theta_1}{2\pi})^3$ has been found to guarantee convergence for all $\theta_0 + \theta_1 < 2\pi$, and is also slightly more accurate, thus requiring fewer iterations to converge to a given precision.

Complete code for this method is appealingly concise; the Python version (shown in Figure 8.3) is only a dozen lines of code. Further, only a small handful of iterations is required to approach floating-point accuracy across the range, and a mere two or three for small angles. Given fast code for evaluating the spiro function, this approach is efficient indeed.

8.3 Global constraint solver

A common problem in computing interpolating splines is solving a global system of constraints. For example, in the framework of two-parameter extensional splines, the constraints are curvature continuity across control points. Other splines may demand a more complex set of constraints, such as the generalized four-parameter splines of Section 7.4.

Most generally, the goal of the solver is to produce a vector of parameters, such that all the stated constraints are met. Most often, the relationship between the parameters and the corresponding constraints is *nearly linear*. For example, as demonstrated in Section 8.2, in the small angle

```

def fit_euler(th0, th1):
    k1_old = 0
    e_old = th1 - th0
    k0 = th0 + th1
    k1 = 6 * (1 - ((.5 / pi) * k0) ** 3) * e_old

    for i in range(10):
        x, y = spiro(k0, k1, 0, 0)
        e = (th1 - th0) + 2 * atan2(y, x) - .25 * k1
        if abs(e) < 1e-9: break
        k1_old, e_old, k1 = k1, e, k1 + (k1_old - k1) * e / (e - e_old)

    return k0, k1

```

Figure 8.3. Python code for computing Euler spiral parameters.

approximation the relationship between endpoint tangent angles and curvature is given by linear Equations 8.29. Thus, one viable approach is to express the relationship between the parameters and constraints as a matrix, and solve it numerically. Hobby’s spline [37] uses this approach, and achieves approximate G^2 -continuity. The resulting matrix is tridiagonal, which admits a particularly fast $O(n)$ and simple solution.

However, this approach yields only an approximate solution, and as the bending angles increase, so do the curvature discontinuities in the resulting spline. A good general approach is to use Newton iteration on the entire set of vector parameters, using the Jacobian matrix of partial derivatives to refine each iteration. When a solution exists, convergence is quadratic, and experience shows that three or four iterations tend to suffice. This general technique resurfaces a number of times in the literature, including Malcolm [54], Stoer [82], and Edwards [18], the latter being an especially clear and general explication of using Newton-style iteration to solve nonlinear splines.

Newton iteration is one of the most efficient techniques, but far from the only workable approach. Since many of the splines are expressed in terms of minimizing an energy functional, a number of researchers use a gradient descent approach, such as Glass [26], and, more recently, Moreton [64]. However, gradient descent solvers tend to be much slower than Newton approaches. For example, Moreton’s conjugate gradient solver for MVC curves took 137 iterations, and 75 seconds, to compute the optimal solution to a 7 point spline [64, p. 99]. By contrast, the techniques in this chapter easily achieve interactive speeds even in a JavaScript implementation that runs in standard Web browsers.

8.3.1 Two-parameter solver

For two-parameter splines, the solution is particularly easy. The vector of parameters is the tangent angle at each control point, and the constraints represent G^2 -continuity across control points. Thus, the vector to be solved consists of the difference in curvature across each control point.

The vector of tangent angles is initialized to some reasonable values (in our implementation, the initial angles are those assigned by Séquin, Lee, and Yen’s circle spline [80], although likely a

simpler approach would suffice). At each iteration, the curvatures at the endpoints of each segment are computed (κ_i^l representing the curvature at the left endpoint of segment i , and κ_i^r the curvature at the right). The differences in curvature $\Delta\kappa_i = \kappa_i^r - \kappa_{i+1}^l$ form the error vector. At the same time, the Jacobian matrix represent the partial derivatives of the error vector with respect to the tangent angles,

$$J_{ij} = \frac{\partial \Delta\kappa_i}{\partial \theta_j} . \quad (8.32)$$

Then, a correction vector $\Delta\theta$ is obtained, by solving the Jacobian matrix for the given error vector. Since the Jacobian matrix is tridiagonal, the solution is very fast.

$$\Delta\theta = \mathbf{J}^{-1} \Delta\kappa . \quad (8.33)$$

Finally, the θ_i values are updated for the next iteration. In all cases but the most extreme, the new value is just $\theta_i + \Delta\theta_i$. When bending angles are very large, this iteration can fail, and a second pass is attempted, using only half $\Delta\theta_i$ as the update; convergence is then only linear rather than quadratic, but it is more robust.

While each segment is specified by two parameters, this solver is written in terms of one parameter per control point, the tangent angle. This approach guarantees G^1 -continuity by construction, as the same tangent is used on both sides of a control point.

One approach to determining the second parameter is to use an additional local solver, determining the two spiral parameters from the two tangent endpoints. For the Euler spiral spline, this local solver is described in Section 8.2. Another approach (described in more detail in Section 8.4) is to precompute a two-dimensional lookup table with the values.

In the software distribution, the clearest and most robust implementation of this solver is in `spiro.js`. The top-level Newton iteration is contained in `refine_euler`, the computation of the coefficients of the Jacobian matrix is in `get_jacobian_g2`, and the solution of the matrix uses the generic band diagonal solver `bandec` and `banbks`, adapted from Numerical Recipes [71].

8.3.2 Four-parameter solver

The same general approach is used for four-parameter splines, but the layout of parameters is a bit different. For one, the solver directly computes all parameters of each segment of the spline, rather than using a two-stage approach. This solver implements the general constraints as described in Section 7.4.

The inner loop for computing the error vector is essentially a spiro integral (see Section 8.1). For the Jacobian matrix, for simplicity we use central differencing to compute the partial derivatives, even though an analytical technique would also be feasible.

Each constraint gets one element in the error vector and one row in the Jacobian matrix. For a constraint of the form, say, $\kappa_2'' = \kappa_3''$, the entry in the error vector is the deviation from that equality, in this case $\kappa_2'' - \kappa_3''$. Each column in the Jacobian matrix represents one parameter of the polynomial spline, i.e. there are four times as many columns as curve segments. For the system to be well-determined, there must be an equal number of (linearly independent) constraints as

```

Set up structure of matrix:
    column = parameter of spline
    row = constraint

Initialize parameter vector to all zero

Repeat until done:
    Compute error vector of deviations from constraints
    Compute norm of error vector
    Below total error threshold?
        If so, done
    Compute partial derivative matrix of constraint wrt spline parameter
    Invert matrix (band-diagonal)
    Compute dot product of error vector and inverse of matrix
    Add to parameter vector

```

Figure 8.4. Pseudocode for computing four-parameter splines.

parameters, i.e. the Jacobian matrix must be square. Thus, the constraints described in 7.4 are carefully designed to make the count come out even. In particular, for a sequence of G^2 -continuous curve points, additional constraints zero out the higher-order derivatives of the polynomial spline, so that the resulting curve segments are segments of the Euler spiral.

The algorithm for the four-parameter solver is presented in pseudocode form in Figure 8.4. Complete working code for the solver is included in the open-source libspiro package. This package contains an efficient C implementation of this solver, the numerical methods being contained in `spiro.c`. The outer loop of the Newton iteration is contained in the `spiro_iter` function, with the computation of the Jacobian matrix in `compute_pderivs`, and the solution of the matrix in `bandec11` and `banbks11`, again adapted from Numerical Recipes, specialized to the size of the band.

An example of the four-parameter solver at work is shown in Figure 8.5. The first iteration (i.e. the linear approximation) is the thin red curve, and the second is in green. By the third iteration, the constraints are satisfied to within visual tolerance.

In practice, both solvers run efficiently and robustly. For the same problem, the two-parameter solver is a bit more robust than the four-parameter, but of course the latter is much more general and can solve a much wider class of spline problems.

8.4 2-D Lookup table

The family of two-parameter, extensional splines admits a particularly simple and efficient implementation based on 2-D lookup tables. Since the curve segments are defined by two parameters, for any given generator curve it is practical to precompute a lookup table indexed by the tangent angles at the endpoints. The lookup table stores the curvatures at the endpoints (used for enforcing curvature continuity across control points), as well as the position and tangent angle of a midpoint, used for subdivision while rendering the curve. With this precomputation done only once, drawing

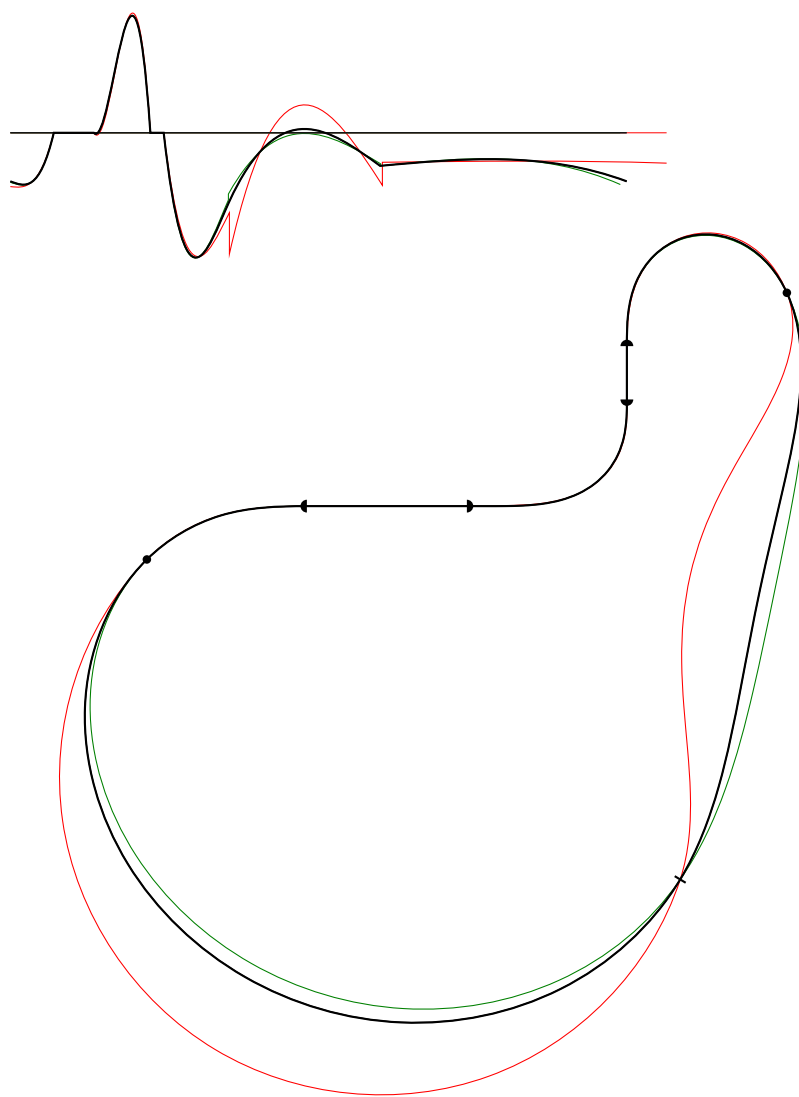


Figure 8.5. Example of four-parameter solver.

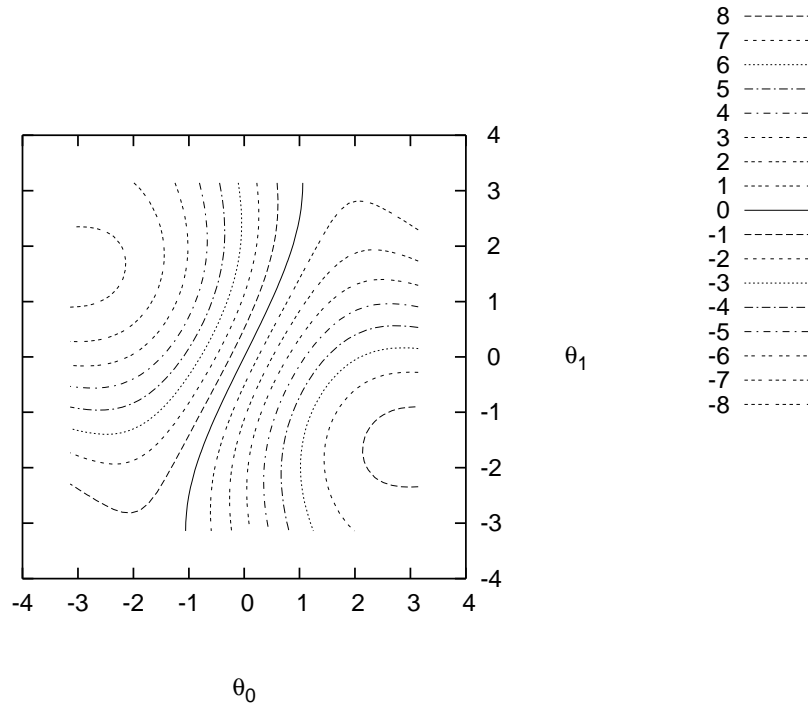


Figure 8.6. Contour plot of curvature lookup table for Euler spiral.

an arbitrary spline segment is only slightly more complex than the standard de Casteljau subdivision for rendering Bézier curves; the only difference is the use of a lookup table rather than an algebraic formula to determine the midpoint.

An example curvature lookup table, for the Euler spiral, is shown in Figure 8.4. The horizontal axis is the tangent angle, in radians, on the left side, the vertical axis is the tangent angle on the right, and the value is the resulting curvature on the left side; the other value follows by symmetry. Moreover, the table itself is anti-symmetric, which can further reduce storage requirements. Note that the relationship is nearly linear for small angles. The slope of that linear relationship is closely related to locality. Because these values are so smoothly behaved, a simple representation as a two-dimensional lookup table with 32×32 entries, and using bicubic interpolation for lookups, gives a maximum curvature error of 10^{-4} across the range of the function. This error scales with the fourth power of the linear size of the table; thus, for example, if a higher precision is desired, a 64×64 table could reduce the error to about 6×10^{-6} .

Chapter 9

Converting to Bézier curves

By far the most popular representation of curves is piecewise polynomial, most often with the coefficients represented in Bernstein–Bézier form. This popularity is well earned – the representation is simple, numerically stable, and the primitive is quite supple, capable of representing a wide range of curves with relatively few control points. Both cubic and quadratic Béziers are widely used. The PostScript language [70] established cubic Béziers as a de facto standard for 2D graphics, and the TrueType font format and Macromedia Flash are based on quadratics, presumably because the rasterizer is even simpler to implement than for cubics.

There are several reasons to convert piecewise clothoid curves, or piecewise polynomial spiral curves, into Bézier representation: interactive display using graphics API’s based on the PostScript imaging model, creation of font files in industry-standard formats, and export of curves to CAD and other design tools. All such tools are capable of accepting Bézier curves, but few, if any, can handle piecewise clothoids.

The conversion to Bézier form involves a tradeoff between the conciseness of the result and the amount of computation needed. For some applications, such as interactive display, the curve will be computed, displayed, and then discarded. A somewhat wasteful Bézier representation is perfectly acceptable as long as the overall computation time is small. For other applications, the curve will be computed, saved into a file such as a font, then embedded thousands of times in document files, each of which in turn may be rendered thousands of times. In such applications, a significant investment in computing time to reduce the size of the Bézier representation can pay off handsomely. This section will describe both fast and cheap methods for converting polynomial spirals into Béziers, as well as much slower techniques, suitable for creating optimally concise Bézier representations.

There are several techniques in the literature for creating polynomial approximations of clothoid curves. Sánchez-Reyes and Chacón [77] use a family of polynomials called *s-power series*, which is a fairly powerful and general technique. However, the error in their technique converges only as $O(n^4)$ in the number of subdivisions, while convergence of $O(n^6)$ is possible for cubic Béziers.

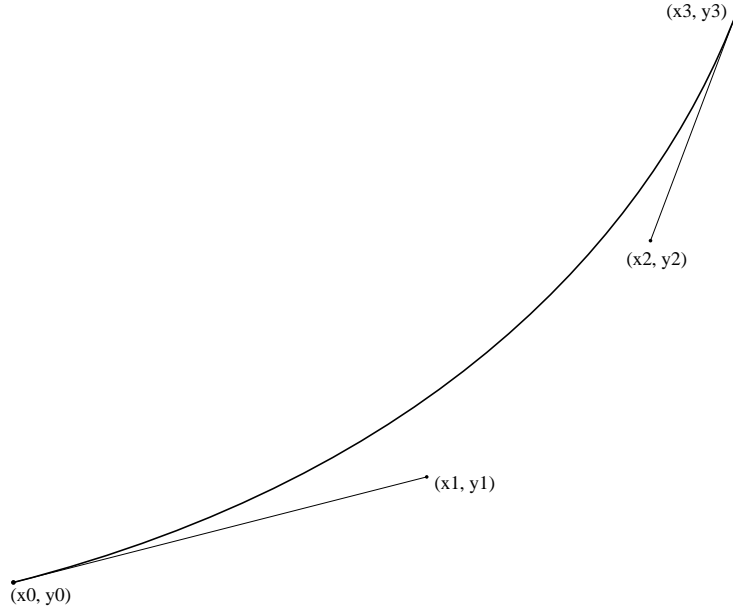


Figure 9.1. A cubic Bézier.

9.1 Parameters

Assuming fixed endpoints, quadratic Béziers have two free parameters. These are exactly sufficient to determine the tangent angles at the endpoints. Thus, the problem of generating a quadratic Bézier given the endpoints is fully determined; the only remaining problem is to determine the endpoints, or, in other words, how to subdivide the original curve to be fit.

Cubic Béziers have four parameters. Two of these determine the tangent angles. The other two have no direct geometric interpretation, and can be considered free parameters useful for making the curve fit the target more precisely.

Cubic Béziers are often described in terms of two *control arms*, or vectors from the endpoints to the control points. In the typical Bézier of Figure 9.1, these vectors are (x_0, y_0) to (x_1, y_1) and (x_3, y_3) to (x_2, y_2) , respectively. The tangent angles of these control arms are the tangent angles of the resulting curve at the endpoints, so the lengths of these control arms are the free parameters.

9.2 Error metrics

Fitting of a cubic Bézier to a target curve can be seen as a simple optimization problem. Over the parameter space, what set of parameters produces the closest fit to the target curve? However, for such a formulation to be meaningful, we must choose an error metric. What does it mean for one fit to be “closer” than another to the target curve?

In the literature, the maximum distance deviation, or “flatness” is the most commonly used

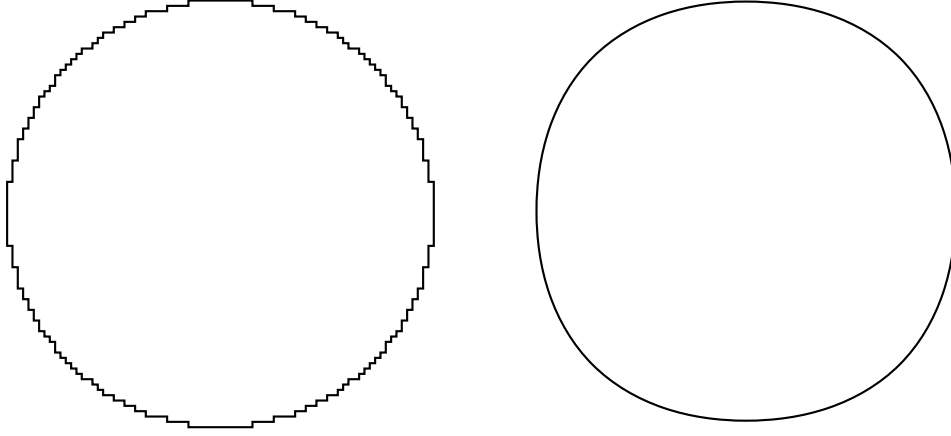


Figure 9.2. Two circle approximations with equal distance error.

metric. However, this metric has several shortcomings. To address these shortcomings, this section presents other error metrics based on angle, rather than distance.

Figure 9.2 shows two approximations to a circle with equal maximum distance error. The approximation on the left is quantized to an integer grid, while the approximation on the right is smooth but warped out of shape. Even though the actual maximum distance error from the true circle is the same, the one on the right looks considerably smoother. Thus, it is clear this metric does not accurately predict perceived smoothness, and optimizing to minimize it may not necessarily produce the best results.

Thus, other error metrics seem more promising, in particular those based on angle error rather than absolute distance error. Assume, for the sake of simplicity, that the Bézier approximation and the target curve have identical total arc length l . Then, we propose the *absolute angle metric*

$$E_{\text{abs}} = \int_0^l |\theta_{\text{approx}}(s) - \theta(s)| \, ds. \quad (9.1)$$

Note that this error metric is extremely closely related to the standard flatness metric. Like the flatness metric, it is expressed in units of distance, and indeed it represents a fairly tight upper bound on the flatness. Consider the distance between points at arc length $d \leq l$ from the curve start, here using complex coordinates for algebraic convenience:

$$z_{\text{approx}} - z = \int_0^d e^{i\theta_{\text{approx}}(s)} - e^{i\theta(s)} \, ds. \quad (9.2)$$

But note the inequality

$$|e^{i\theta_{\text{approx}}(s)} - e^{i\theta(s)}| \leq |\theta_{\text{approx}}(s) - \theta(s)|. \quad (9.3)$$

Note also that the integral 9.2 is bounded from above by E_{abs} because the absolute angle error is everywhere positive. Thus, for all points on the original curve, there exists a point on the approximate curve no more distant than E_{abs} from this point, and thus from the target curve.

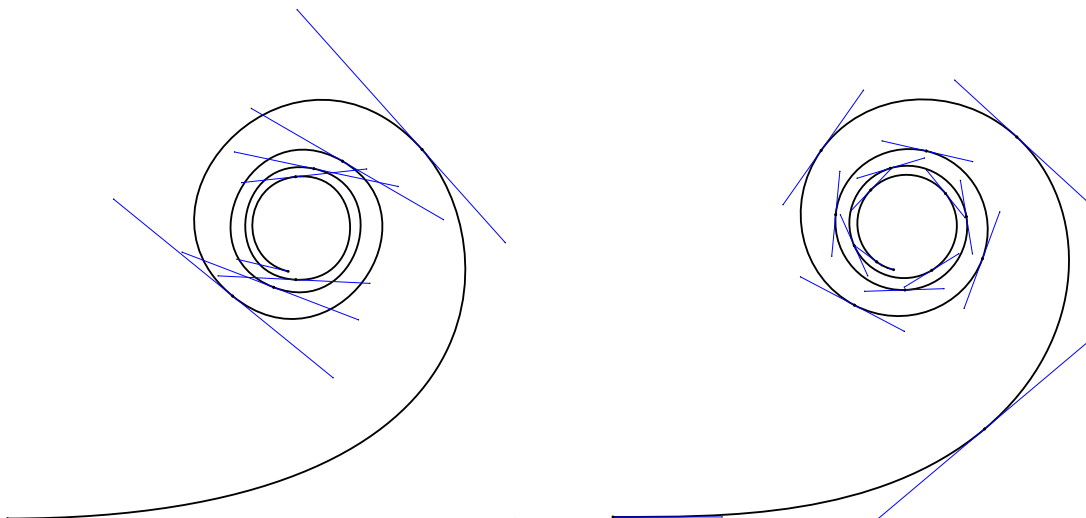


Figure 9.3. Euler spiral approximation drawn using .03 and .001 error threshold, respectively.

However, this metric still penalizes “bumps,” or short segments of relatively large angle error relatively lightly. Better visual results may thus be derived from minimizing the *square* of the angle error, integrated over the arc length as before.

$$E_{\text{sq}} = \left(\int_0^l (\theta_{\text{approx}}(s) - \theta(s))^2 ds \right)^{1/2}. \quad (9.4)$$

From standard properties of norms, $\sqrt{l}E_{\text{sq}} \geq E_{\text{abs}}$ holds, which means that E_{sq} also places a strong upper bound on the flatness error metric. This property implies that minimizing the E_{sq} metric will also yield a curve with a good flatness metric as well, although the converse is not always true.

Figure 9.3 shows examples of this error metric in use. Both approximations of the Euler spiral segment use a number of cubic Bézier curves, each with the approximately the same E_{sq} error metric. Each segment on the left has an error of approximately 0.03, and each segment on the right approximately 0.001 (relative to the total size of the figure being about 1). As can be seen, an error of 0.03 results in a somewhat distorted curve, while 0.001 is essentially indistinguishable from the original curve. This figure also provides evidence that converting to Bézier form by solving an optimization problem in terms of the E_{sq} metric gives good results, even for a relatively crude error threshold.

9.3 Simple conversion

A simple, fast conversion to cubic Bézier form is as follows. Like the s-power series, and like quadratic Béziers, the convergence is only $O(n^4)$, although the constant factor is considerably better than for quadratics, and this cubic solution can easily accommodate an inflection point.

In this solution, set the length of both control arms to one-third the total arc length of the



Figure 9.4. Three Bézier solutions: fast, equal arc length, and fully optimized. Curvature plot is the top graph, the Bézier is on bottom.

curve. An example, fitting the section of Euler spiral from parameter $t = 0.5$ to $t = 1.1$, is shown as the leftmost instance in Figure 9.4. Both the Bézier (bottom curve) and the plot of curvature against arc length (top curve) are shown. In the ideal Euler spiral, of course, curvature is a linear function of arc length.

9.4 Equal arc length, equal arm length solution

A somewhat more accurate solution can be obtained by solving the length of both control arms so that the total arc length of the Bézier matches that of the target curve. An example is shown as the top right instance of Figure 9.4.

There is no simple formula for the total arc length of a Bézier, so for this solution we rely on numerical integration. Since the relation between the control arm length and the total arc length is smooth and monotonic, a simple numerical solving technique such as the secant method gives robust, accurate results in a few iterations.

9.5 Fully optimized solution

In the fully general case, the lengths of the control arms need not be equal. To find the single best-fitting cubic Bézier is, thus, a minimization problem over a two-dimensional parameter space – the lengths of the control arms.

An example of the error metric over this parameter space is plotted in Figure 9.5. This plot is the root mean squared angle error for all combinations of left and right control arm length, with a target of the Euler spiral segment from $t = 0.5$ to $t = 1.0$. For clarity, only isolines are shown.

The first notable feature of this plot is that there are two local minima of the error metric. Thus, care must be taken to choose the best global minimum. The actual curves and curvature plots of these two local minima are shown in Figure 9.6. In this case, the one in which the left control arm is longer is clearly the better fit. In general, though, either local minimum may be globally optimal. Figure 9.7 shows a fit of a different section of the Euler spiral, from $t = 0.5$ to $t = 0.85$, and the two local minima are nearly indistinguishable, both in overall error metric and in curvature plot.

Also plotted in Figure 9.5 is the curved line representing all values of left and right control arm length so that the total arc length of the Bézier matches that of the target curve. Notably, error is locally minimized across this line, suggesting a numerical approach where the outer loop searches for the optimum ratio between left and right control arm lengths, and an inner loop determines the total length of the control arms, given their ratio fixed, such that the arc length of the Bézier matches that of the target curve.

As can be seen in the rightmost example in Figure 9.4, the fully optimized solution finds a Bézier that matches the original Euler spiral segment with impressive accuracy. Even in the curvature plot, the deviation from ideal is barely perceptible.

9.6 Global optimum

Generating an optimized Bézier outline of a shape requires not only computation of individual segments, but also deciding the subdivision points at which the original curve is to be divided.

In general, an optimized Bézier representation of a curve has three elements:

- The number of subdivisions is minimal with respect to the error metric threshold.
- The position of the subdivision points minimizes the error metric.
- Each individual segment minimizes the error metric given fixed endpoints.

Of course, the third element is the problem addressed in the previous section. This section describes techniques to determine the subdivision points, and essentially uses the generation of individual Bézier segments (with computation of error metric) as an inner loop.

As before, there is a tradeoff between the amount of time spent optimizing the Bézier representation and the degree to which the representation actually meets these goals. For applications such as quick interactive display, where the Bézier will be drawn once, it makes sense to compute it as quickly as possible, even if the number of segments is substantially above the minimum. For

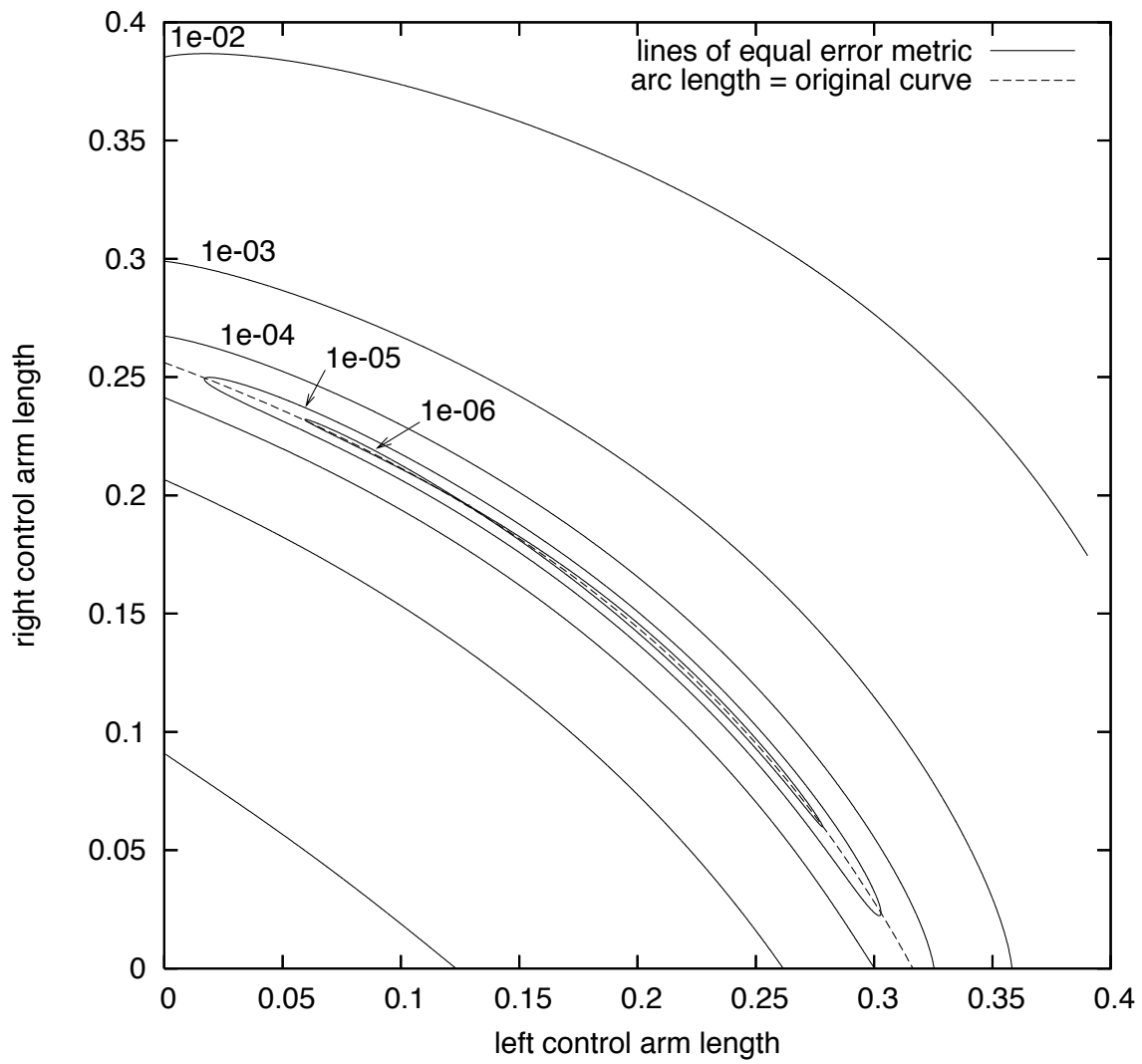


Figure 9.5. Approximation errors over parameter space for cubic Bézier fit.



Figure 9.6. Local minima of cubic Bézier fitting, $t=[0.5, 1.0]$. Curvature plots are the top two graphs, the Béziers are on the bottom.

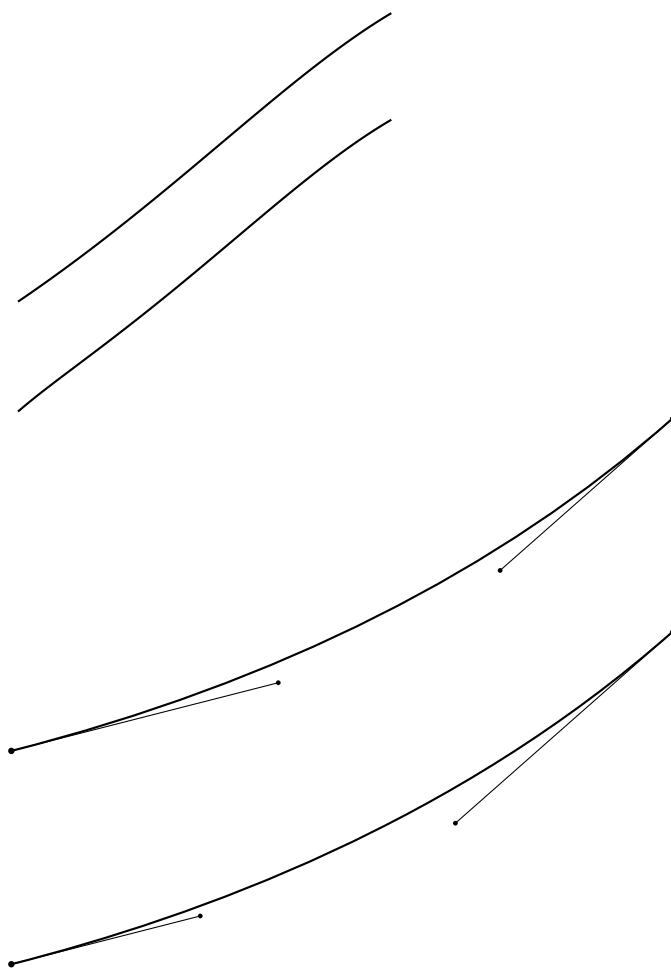


Figure 9.7. Local minima of cubic Bézier fitting, $t=[0.5, 0.85]$. Curvature plots are the top graphs, the Béziers are on the bottom.

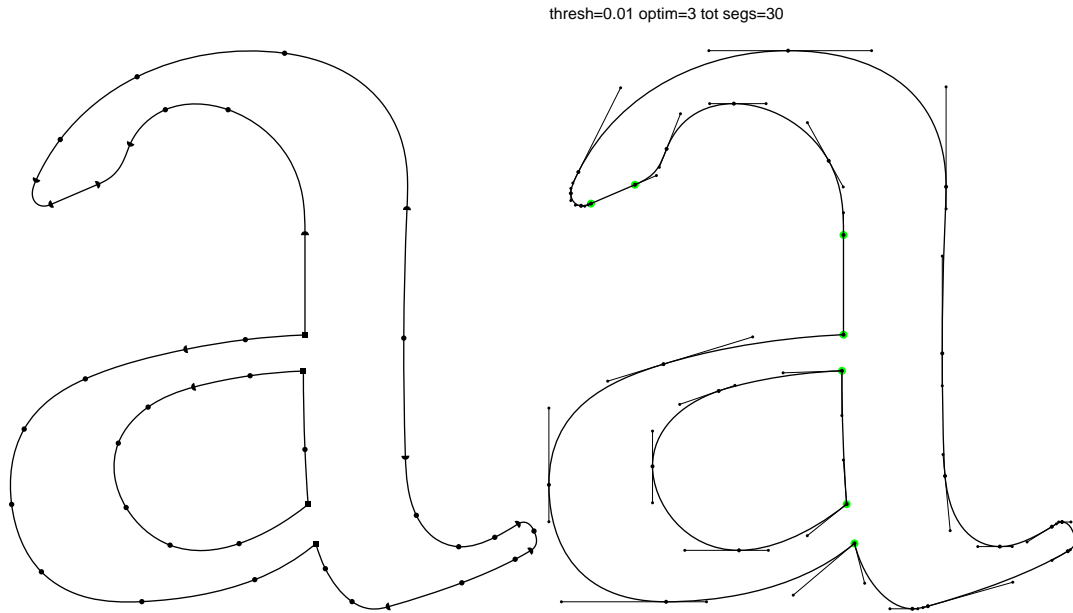


Figure 9.8. Conversion into optimized cubic Béziers.

other applications, such as generating font files to be embedded in documents served over the Web, it makes sense to spend a lot of computing time up front to optimize the representation.

Figure 9.8 shows an example of an optimized conversion into cubic Bézier representation, suitable for encoding into a PostScript Type 1 font format or any of its successors, such as CFF (Compact Font Format), which can be directly embedded into a PDF file, or included in an OpenType container. This font format has the additional requirement that subdivision points appear at all horizontal and vertical tangents. For conversion into simple vector drawings such as PostScript or SVG (Scalable Vector Graphics), this requirement could be relaxed.

9.6.1 Hybrid error metric

The purest approach to the error metric for the curve would be to use the same metric as for an individual segment. For a metric based on L^2 -norm, this is computed easily enough by summing the squares of the individual error metrics of the segments.

However, it is significantly easier to work with a hybrid metric in which an L^∞ -norm is used to combine the error metrics; in other words the error metric for the whole curve is the maximum metric for each individual segment. If the metric for individual segments is also an L^∞ -norm, this is equivalent to the purest approach. However, the L^2 -norm is easier to compute accurately using numerical integration (it is a considerably smoother function), so the result is a hybrid approach.

The error metric for an individual segment is clearly monotonic with respect to the subdivision points on either side. Otherwise, if making the segment shorter by moving a subdivision point inwards were to increase the error metric, the new curve must clearly not be an optimum Bézier fit. Given this property and the objective of minimizing an L^∞ -norm, all optimum Bézier repre-

representations of curves have the property that each individual Bézier segment has an identical error metric.

In theory, it would be possible to optimize for the pure L^2 -norm, but in practice it would be computationally much more expensive, and the results would be almost imperceptibly different. For reasoning about this approach, it is perhaps simplest to accept the L^2 -norm metric as a reasonably good approximation to the L^∞ -norm.

9.6.2 Simple subdivision

The simplest technique for meeting a particular error metric is to simply fit a single Bézier segment to the curve, accept it if the error metric threshold is met, and subdivide the curve in half if not.

Note that if the error metric is an L^∞ -norm, then the same error threshold can be applied to both subdivisions. If the error metric is an L^2 -norm, then the threshold for each subdivision is $\sqrt{0.5}$ of the original threshold.

One refinement is to subdivide at the midpoint of angle traversed, rather than at the midpoint of arc length. For sections not including an inflection point, this is simply the angle halfway between the tangent angles of the subdivision points, but for sections for which curvature changes sign, a more sophisticated approach is needed, using the integral of the absolute value of curvature as the quantity to subdivide.

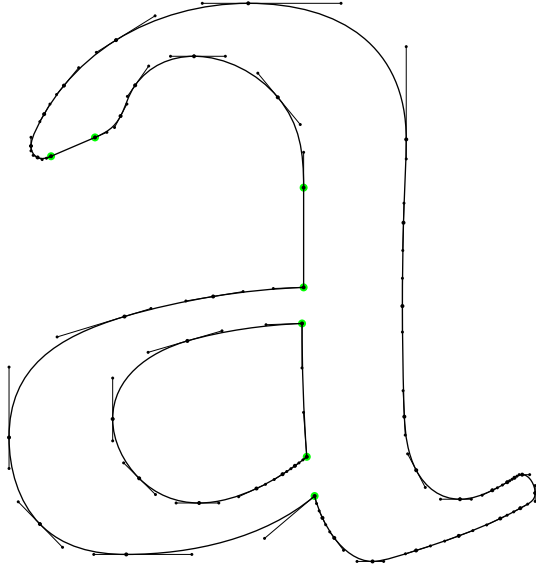
Even though simple subdivision doesn't produce fully optimum results, the number of segments is fairly good. Using the same 'a' from Cecco, using full optimization for the individual Bézier segments, as described in Section 9.5, results in 34 segments. Using the less computationally expensive equal arc length approach as described in Section 9.4 increases the count to 47. The conversions are shown in the upper right and upper left quadrants, respectively, of Figure 9.9.

9.6.3 Smarter subdivision

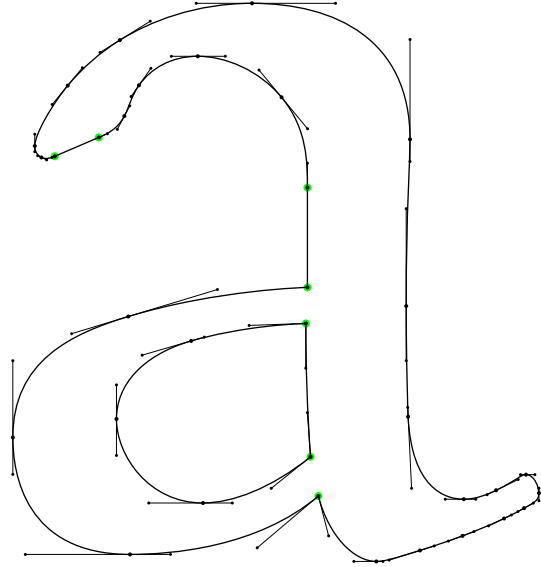
Simple subdivision often results in more segments than absolutely needed. In order to produce an actually optimum Bézier representation, we propose a smarter subdivision technique. It consists of four nested loops. In order of outer to inner:

- Choose the number of subdivisions needed (a simple approach is to count up from 1, and stop when the error threshold is met).
- Start with an initial dumb subdivision (equal arc length is fine), then iteratively refine by choosing the pair of adjacent segments with the greatest discrepancy of error, and finding the subdivision point joining those segments so that the error is equal.
- A slightly modified secant technique selects the new subdivision point so that the error metric of the two segments is balanced. Simple bisection would also work but its convergence isn't quite as fast.
- The inner loop computes the optimized Bézier segments (along with the error metric) on either side of the control point, using the algorithms of Section 9.5.

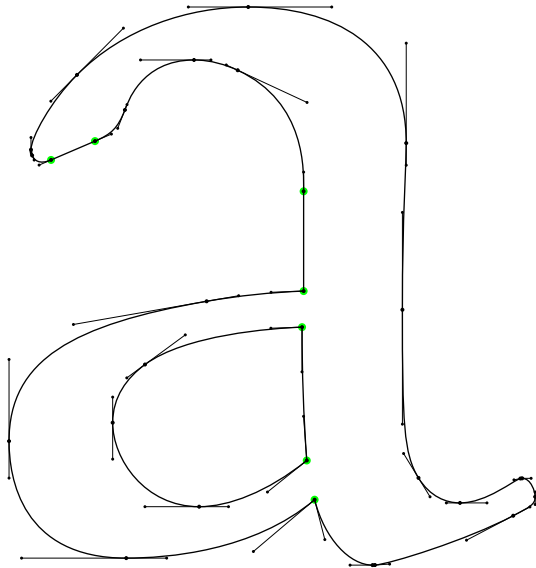
thresh=0.01 optim=0 tot segs=47



thresh=0.01 optim=1 tot segs=34



thresh=0.01 optim=2 tot segs=30



thresh=0.01 optim=3 tot segs=30

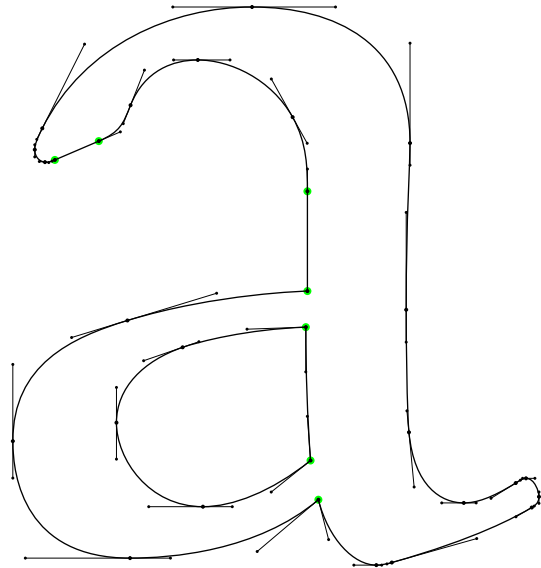


Figure 9.9. Four different levels of optimization: subdivision in half with equal arm length, subdivision in half with fully optimized Bézier segments, greedy subdivision with fully optimized segments, and fully optimized.

Figure 9.9 shows four different levels of optimization. In the top two examples, each segment is subdivided exactly in half (by arc length) until the error metric is below the threshold. In the top left, the Béziers are chosen to have equal arc length as the true curve, but with equal length control arms, as in Section 9.4. In the top right, each Bézier segment is fully optimized, as in Section 9.5.

In the global optimum (the lower right corner of Figure 9.9), the locations of the subdivision points are themselves chosen to be optimum. The lower left corner is another intermediate stage of optimization, to be discussed in the next section.

9.6.4 Alternative optimum algorithm

The algorithm presented in the previous section gives optimum results reasonably quickly. This section presents an alternative algorithm, which also produces optimum results, but using slightly different techniques. It is in some ways more systematic, in particular immediately yielding the number of segments needed, rather than relying on a search to find that number. Also, this technique may be faster in cases when only a minimal number of segments is needed (meeting the error threshold), rather than the additional criterion that the error metric be minimized given that number of segments.

Again, the central observation is that in an optimized representation, the error metric for each segment is the same. Given the error metric target, the algorithm walks from one endpoint of the curve along it, until the end is reached, one segment at a time. For each segment, the left endpoint is already known (again, starting at the endpoint of the curve). Bisection (or some other variant of the shooting method) determines the right endpoint such that the error metric of the segment is equal to the target. Keep in mind that the error metric of the fit is monotonic in the position of the right endpoint, so bisection is guaranteed to work.

Simply setting the error target to the threshold yields a minimal number of segments, but the last segment is likely to have a much lower error metric than the target (and, thus, be shorter than needed). An example is shown in the lower left corner of Figure 9.9. Note that the total number of segments (30) is equal to the fully optimized case in the lower right corner. Yet, some of the subdivision points seem rather lopsided. For example, the subdivision point on the top curve of the bowl is too far to the right; the error on the curve to its left is exactly the target, and to its right, much smaller. There are other similar examples.

Even so, this technique can be adapted (at the expense of more computation) to producing a fully optimized solution. The key is to determine the error target such that the last segment has the same error metric as the rest of the segments. If the target is too small, then the algorithm will use more segments than needed. If too large, then the error metric of the last segment will be significantly less than all the others. Thus, an outer bisection loop can determine this error target.

An additional small optimization can speed the convergence of this bisection: in trials where the error target is too large, the error metric of the last segment yields a lower bound on the final error target for the curve; if it is larger than the lower bound for the bisection, it can replace it. This is most helpful especially when the guess is close to the actual value.

These results are essentially identical to those of the algorithm of the previous section, and often slower. This algorithm requires 23.3 seconds (in a slow Python implementation) to produce the lower left corner of Figure 9.9 and 129 seconds for the fully optimized version. The algorithm of the previous section gives equivalent results in 48.5 seconds, and for other examples the discrepancy

is even greater. In settings where the absolute minimum number of Bézier segments is not required, simple subdivision with fully optimized segments produces the upper right corner in 8.6 seconds, with 34 segments, and simple subdivision with equal arm length, equal arc length produces the upper left corner, 47 segments in 0.9 seconds.

Such times are still undesirably large for interactive use. Clearly, reimplementation in a faster language such as C would improve the compute time by at least an order of magnitude, and other numerical improvements may also be possible. But as the last phase in a font designing process that lasts many hours, even these speeds are acceptable. The next chapter gives discusses applications of interpolating splines for the production of fonts in more detail.

Chapter 10

Applications in font design

A major motivation for this work is to develop better tools for producing fonts.

Existing font formats in the literature use many different representations of the curves. Usually the user interface for designing the curves corresponds closely to the representation. Today, by far the most common representation is Bézier curves, both cubic in the case of PostScript Type 1 fonts [2], and quadratic in the case of TrueType fonts [3]. (The more recent OpenType specification [69] can serve as a container for either of these two underlying formats).

However, a number of alternate formats, popular in their day, use other techniques. Metafont [37] and Ikarus [42], in particular, both used their own flavor of interpolating splines. Perhaps not coincidentally, both these formats supported *variation*, or representing a parametrized family of shapes rather than a single static outline.

Font shapes have a number of requirements beyond simple interpolating splines. The shapes of outlines have corners and straight line sections in addition to smooth curves. In general, the outlines forming a letter shape are closed, although in general not entirely made from smooth points like the inner and outer contours of a typical ‘o’.

Further, letter-shapes often contain transitions between straight and curved sections. These shape patterns are easily expressed using the constraints described in Section 7.4, and the results preserve G^2 -continuity, yielding visually smooth transitions. Table 10 summarizes these constraints, as well as the visual symbol used to represent them, used consistently in the following font outlines.

Inconsolata is an original design by the author, drawn entirely using the splines described in this thesis. Most of the lowercase alphabet is shown in Figures 10.1 and 10.2. Most of the curves were drawn using G^4 -continuity, but using some G^2 points when curvature varies more dramatically, for example in the inner curve of the tail of the ‘g’. Of course, straight-to-curve transitions (which have G^2 -continuity) are used extensively as well, for example in the ‘u’.

The design intent is classical, and draws inspiration from such metal types as the Franklin Gothic series by Morris Fuller Benton (see [58, p. 142] for showings of this and many other metal fonts), as opposed to more recent designs such as TheSans Mono, which show their Bézier underlying format rather clearly.

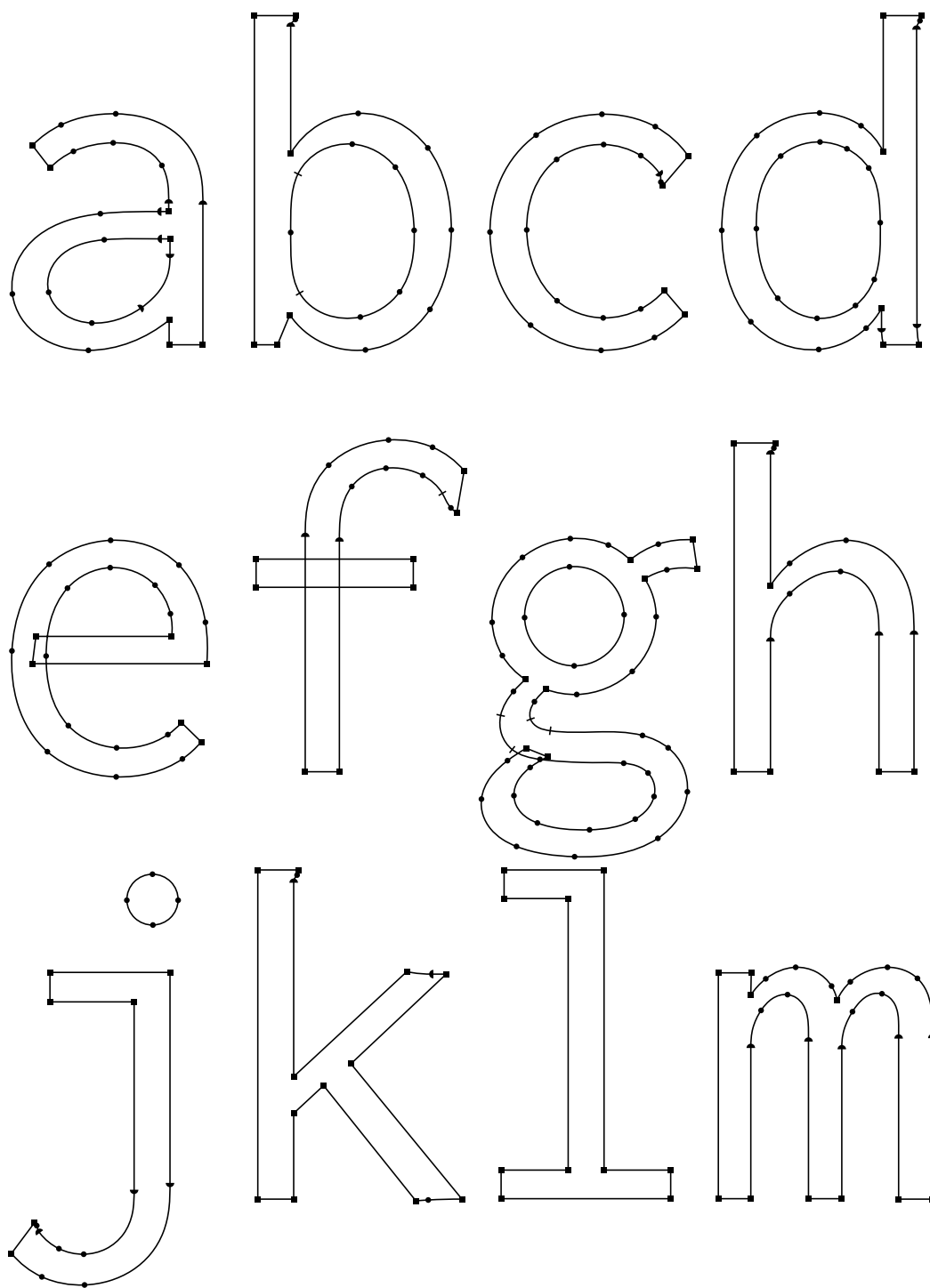


Figure 10.1. Selected glyphs from Inconsolata.

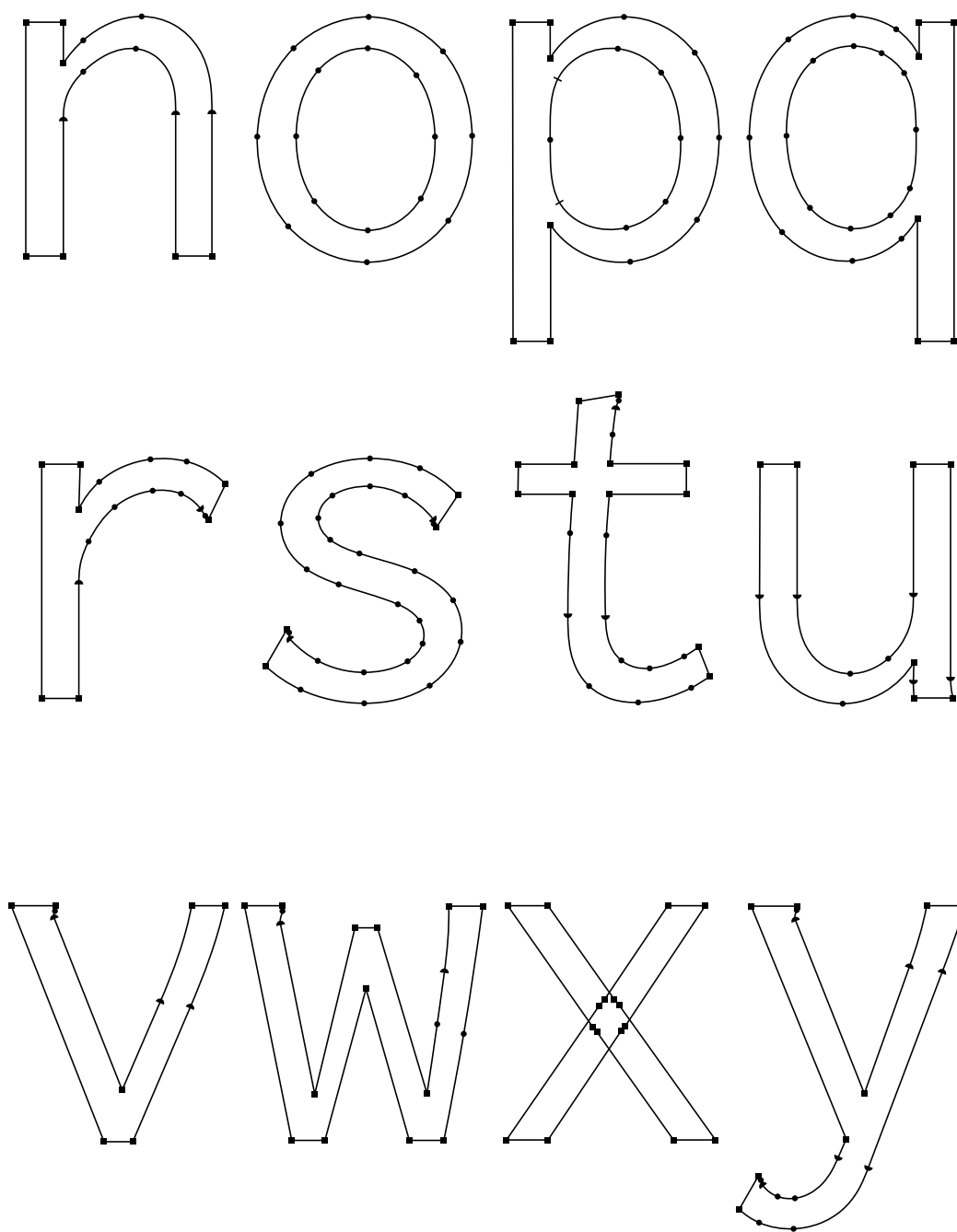


Figure 10.2. Selected glyphs from Inconsolata.





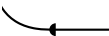
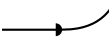
Type of constraint	Symbol
Left endpoint	
Right endpoint	
G^2 curve point	
G^4 curve point	
One-way (curve to terminal)	
One-way (terminal to curve)	

Table 10.1. Constraints used to draw font outlines.

Inconsolata is also used for the `typewriter` style in the formatting of this thesis.

Figures 10.3 and 10.4 show a lowercase alphabet carefully traced from ATF Baskerville Italic (again, see [58, p. 26] for a sample). These outlines generally use G^4 -continuous constraints, and often use one-way constraints to isolate sections of high curvature. Very special care was taken to trace the outlines faithfully. Figure 10.5 shows one scan (enlarged $25\times$ from the original) with the outline superposed.

Figures 10.6 and 10.7 are a draft of the capital letters from Cecco, a more recent original design by the author. As opposed to Inconsolata, the curves were drawn primarily with G^2 -continuous curves. These curves are visually just as smooth as the G^4 , and fewer points were needed, in general, to describe them. These shapes are also classical, inspired in large part by Pacioli by Giovanni Mardersteig [55].

Many interpolating splines in the literature are used for generating *variations* in fonts, rather than a single static shape. Since fonts are desired in a wide range of weights, one of the more common applications is to generate these continuously. One of the most straightforward techniques, pioneered by Ikarus [42], is to start with two masters with similar structure, representing thin and thick extremes, and linearly interpolate the positions of the control points between the two. Then, an interpolating spline is fit through these control points. Note that there are *two* senses of interpolation here – one for the control points, another for the spline.

The splines described in this thesis work especially well for such interpolation. An example interpolation is shown in Figure 10.10, derived from the masters in Figure 10.9. Figure 10.11 more clearly shows the nature of this interpolation, with the linear motion of the control points with respect to the interpolation parameter represented by the equal spacing of the resulting contour lines.

One advantage of the splines of this thesis over the Ikarus spline is that fewer control points are needed to achieve good smoothness, as shown in Figure 10.8. As can be seen, approximately half as many control points are needed, and the smoothness and quality of the result is favorable. Second, the one-way constraints are more robust under interpolation than ordinary interpolating

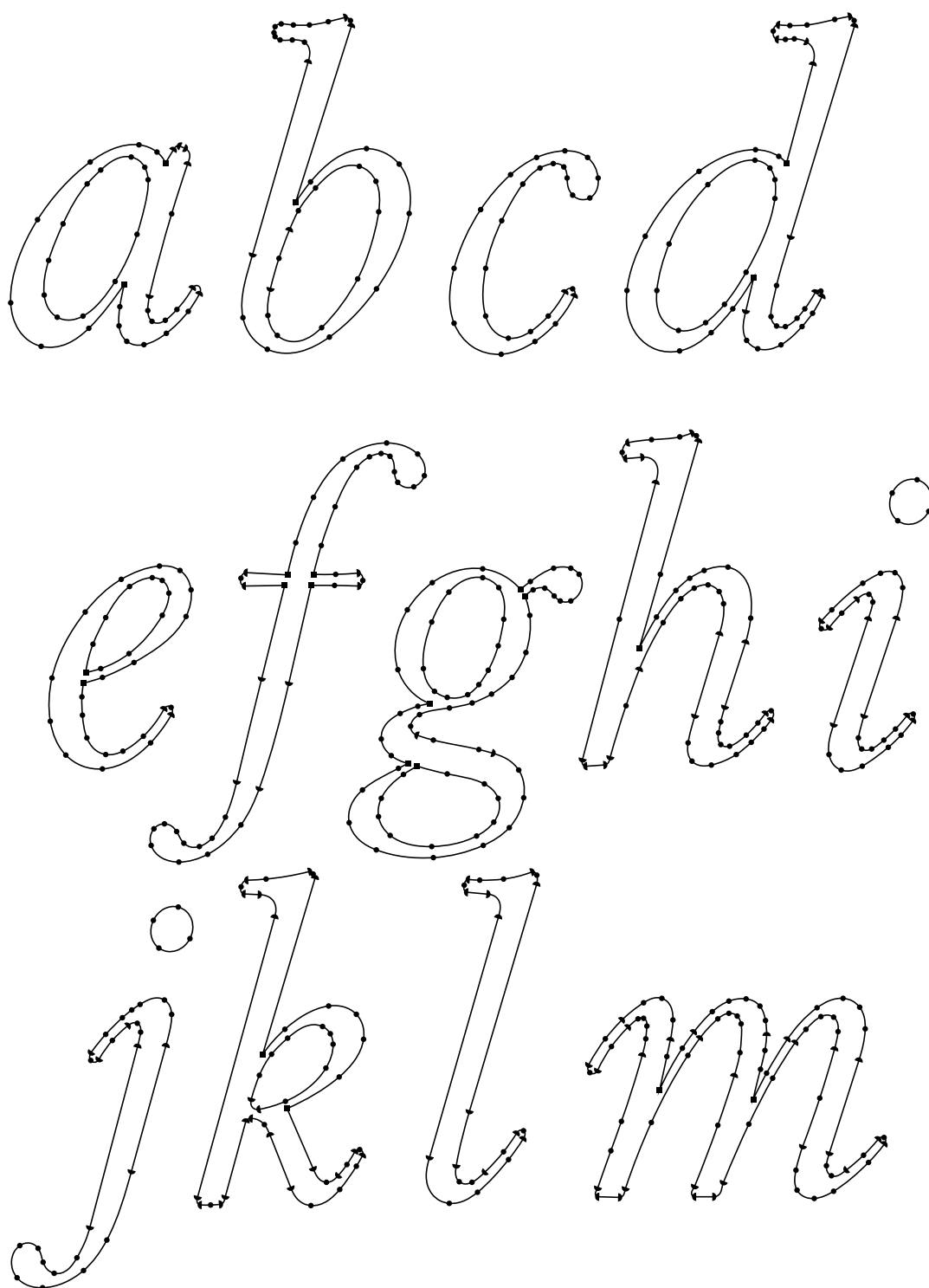


Figure 10.3. Selected glyphs from Baskerville Italic.

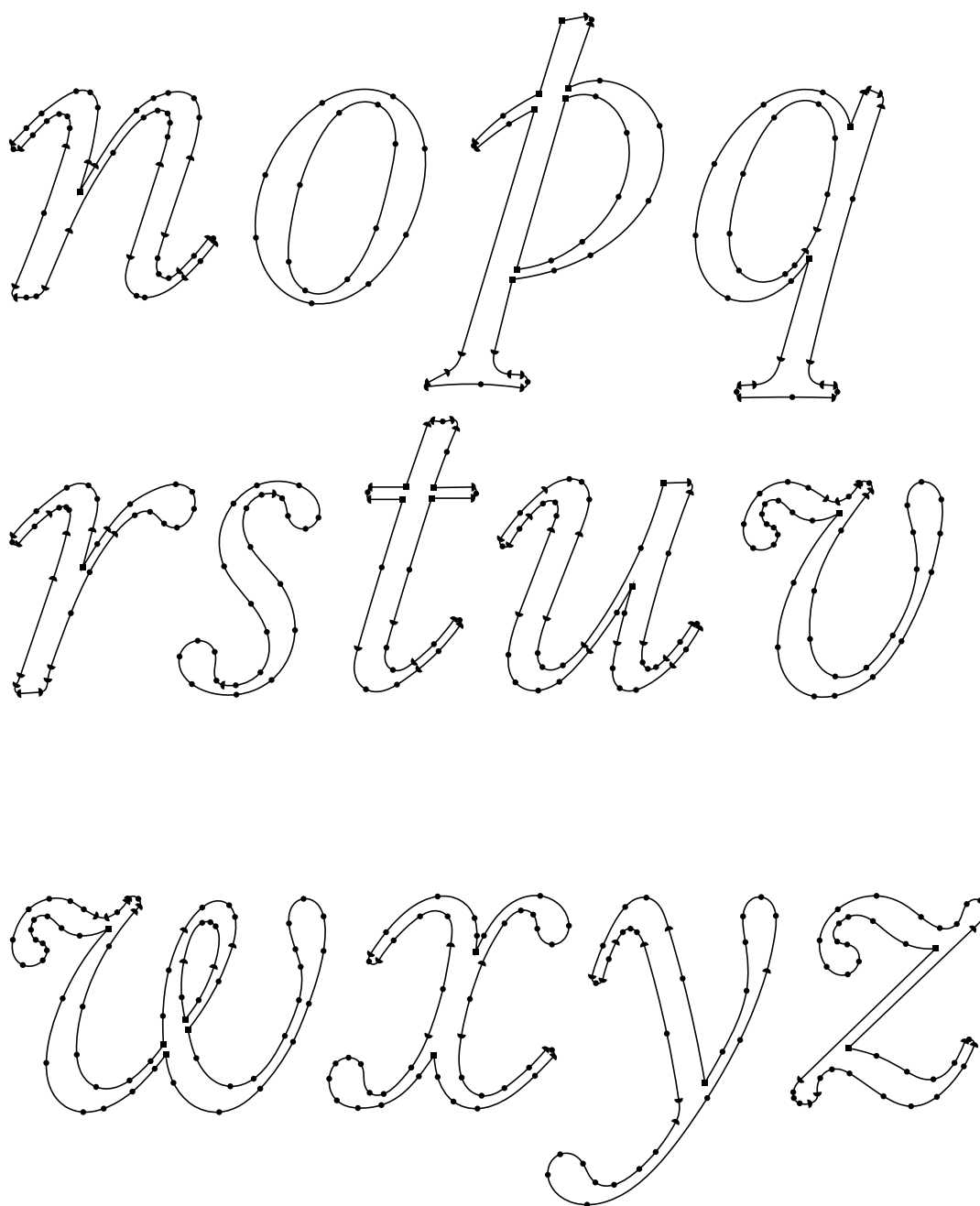


Figure 10.4. Selected glyphs from Baskerville Italic.



Figure 10.5. Lowercase ‘w’ from Baskerville Italic, scan with outline.

splines, for such things as straight-to-curve joins, as in the upper right corner curve of this “a”. In particular, G^2 -continuity is always preserved.

By contrast, as shown in Figure 10.12, interpolation between two curves defined by multiple cubic Bézier segments doesn’t preserve even G^1 -continuity, and a “kink” can appear, even when the source curves have a high degree of continuity (in this case, G^∞ across the join, because both source curves have been subdivided from a single Bézier curve). Even though the figure shows a fairly extreme example, it is likely that unless special care is taken, small discontinuities can arise in ordinary use (perhaps a degree or less, so barely visible). Such cases particularly arise when interpolating between different widths, as in this example; it requires special care to place the subdivisions so that the interpolations are continuous.

In this particular example, the interpolation master curves are subdivided from a single Bézier segment. If the subdivision point had been chosen at the same parameter for both, then the result would have been at least G^1 -continuous. Alternatively, if the curves had been subdivided at the point of the same tangent angle, the result also would have been G^1 -continuous. It is certainly possible for designers to carefully choose subdivision points to achieve smooth results, as witnessed by a relatively small number of Multiple Master releases from designers such as Robert Slimbach. However, this requirement leads to a steeper learning curve and entails ongoing difficulties in use. Using a spline that is G^2 -continuous by construction promises to be better in both respects, and make the design of fonts with variation more accessible and productive than with previous tools.

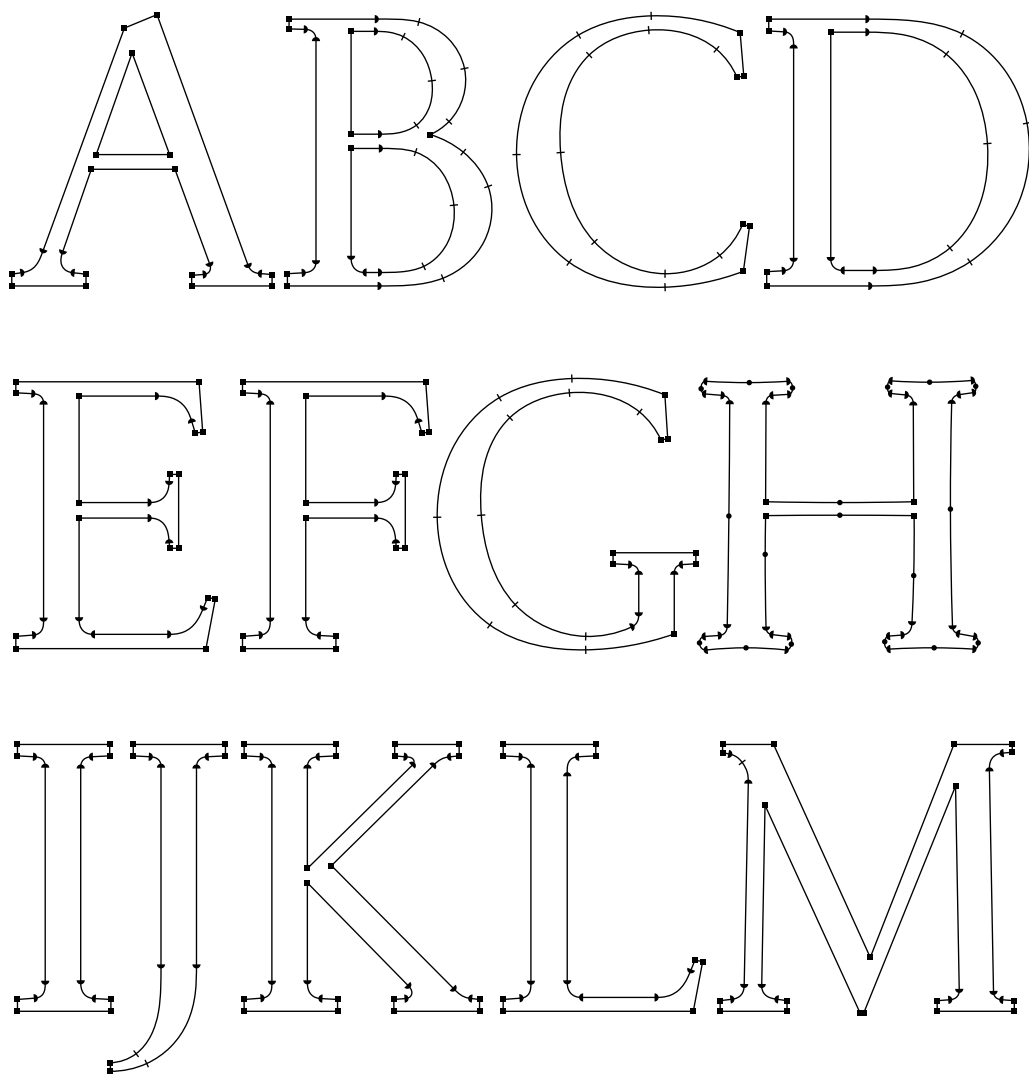


Figure 10.6. Selected glyphs from Cecco.

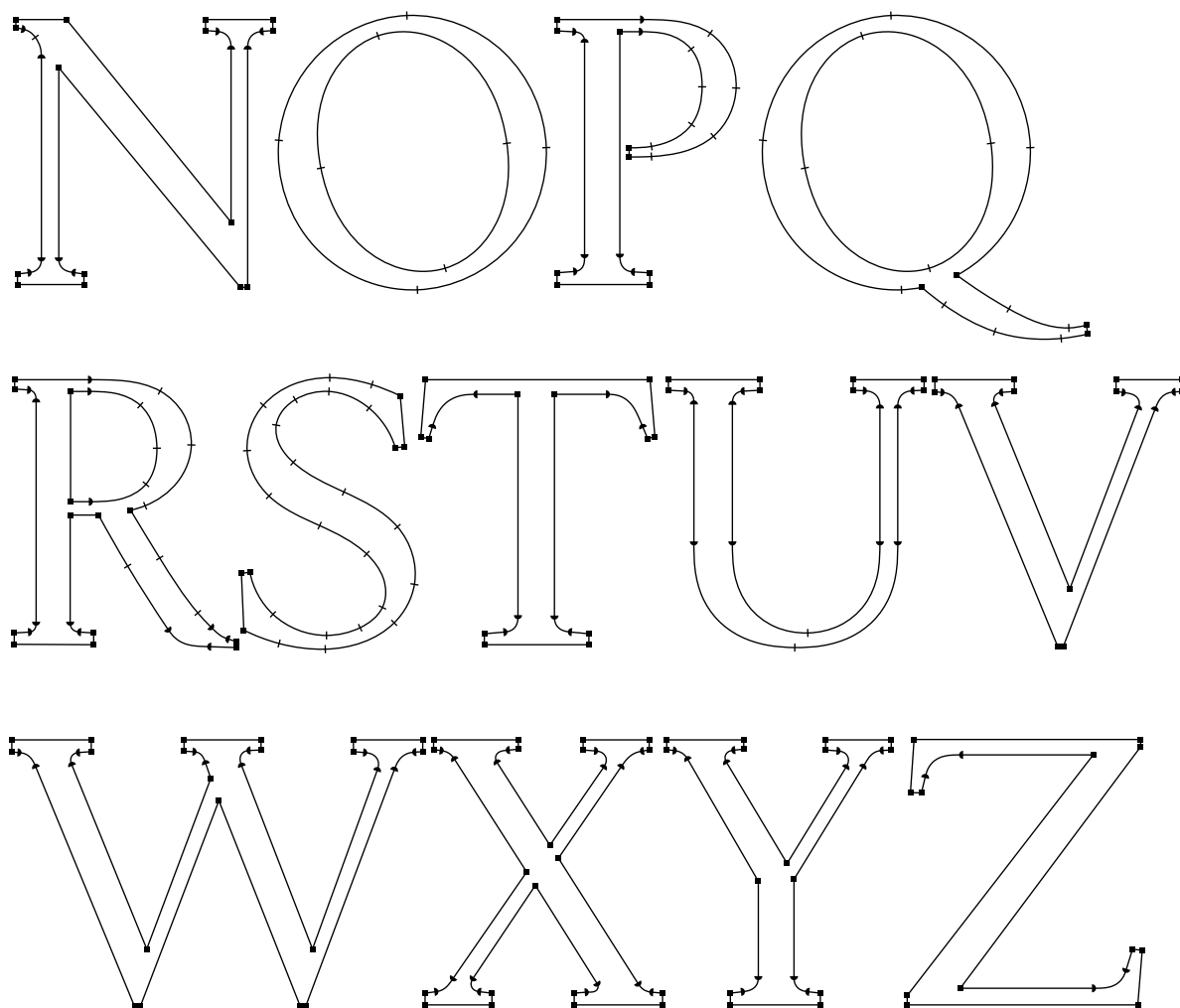


Figure 10.7. Selected glyphs from Cecco.

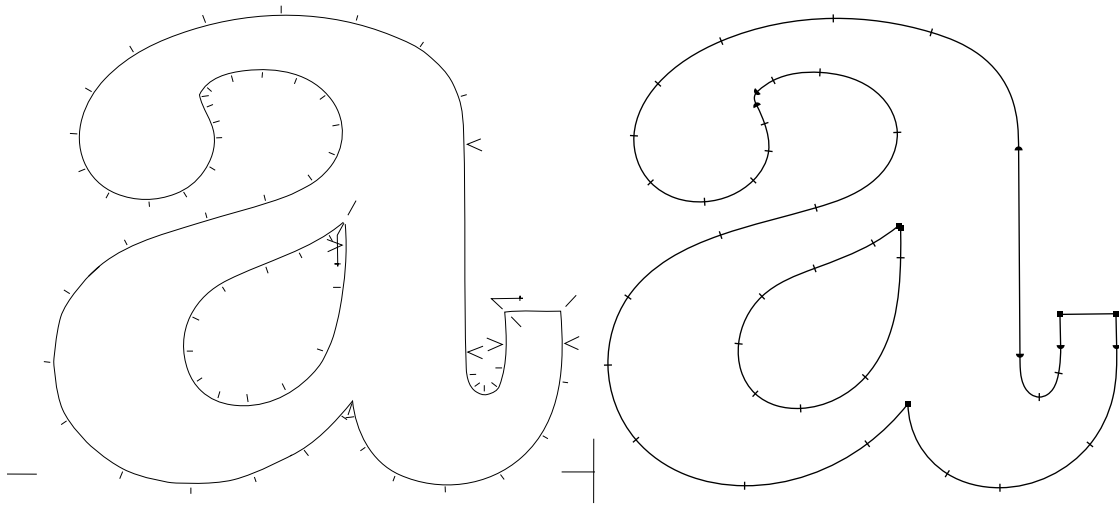


Figure 10.8. Typical letter-shape using IKARUS, with Spiro comparison.

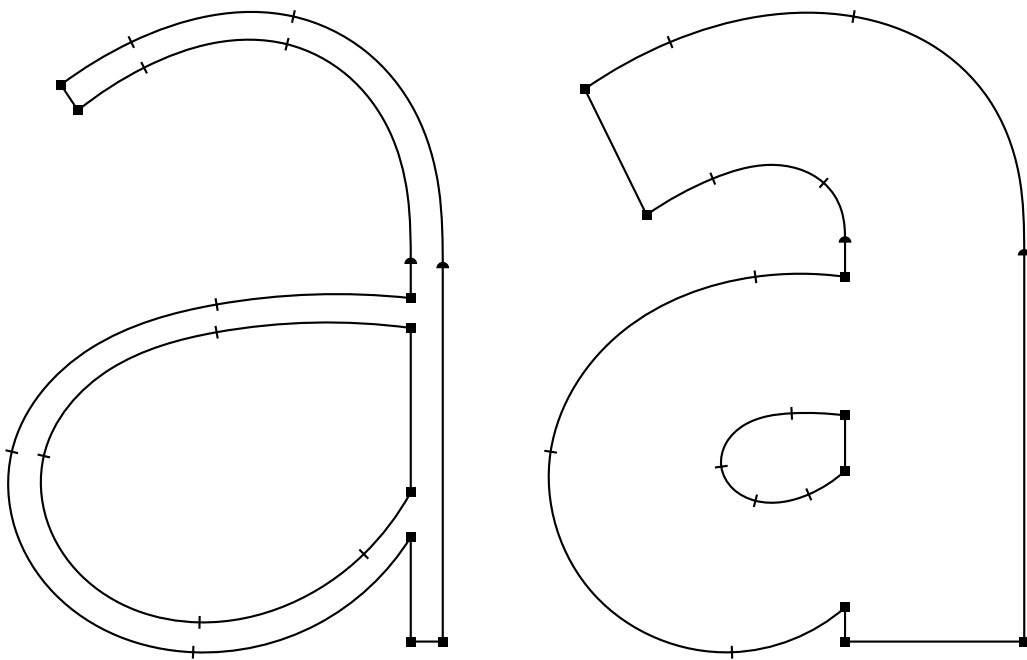


Figure 10.9. Interpolation masters.



Figure 10.10. Interpolation between extremes.

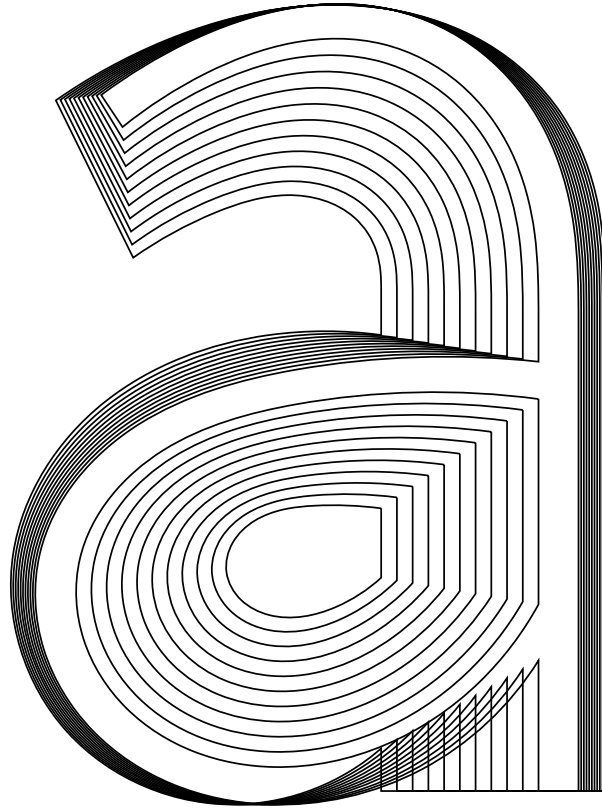


Figure 10.11. Interpolation between extremes, superimposed.

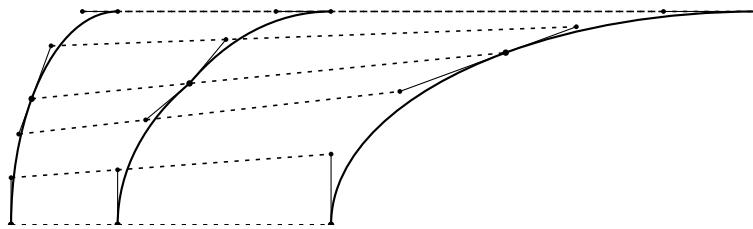


Figure 10.12. Interpolation between cubic Béziers fails to preserve G^1 -continuity.

Chapter 11

Space curves

The focus of this thesis has been on curves in the plane, but much of the literature on splines is also concerned with the generalization to space curves – a continuous mapping from arc length to points in space, $\mathbb{R} \rightarrow \mathbb{R}^3$. For example, Moreton’s investigation of the MEC and MVC [64] is equally concerned with space curves as with planar curves.



Figure 11.1. Space curves rendered into ironwork.

Space curves have many applications, including architectural ironwork. The interpolating spline techniques of this thesis have already shown much promise for this application. Figure 11.1 shows

ironwork space curves which have been constructed starting from planar interpolating curves. These pieces are by Terry Ross, as is the Curve Maker software (implemented in the Ruby language as a plugin for the Google SketchUp 3D CAD tool), adapted from the algorithms described in this thesis [74].

Many of the techniques presented here are also applicable to space curves. There are a number of viable paths.

First, variationally defined splines are also valid in three-space. The MEC, in particular, generalizes straightforwardly. Its definition in three-space is the same as on the plane; it is the curve that minimizes bending energy (L2 norm of curvature) that also interpolates the points.

Second, space curve splines may be defined as curve segments drawn from a parametrized space, with additional curvature constraints.

This chapter begins with a brief review of the properties of space curves (Section 11.1). Section 11.2 discusses the generalization of the MEC to space curves, including the property that this family of curves has three parameters, in contrast to two for the planar MEC, as well as the continuity properties of the space MEC used as an interpolating spline. Section 11.3 presents Mehlum’s generalization of the Euler spiral to space splines, a particularly appealing approach. Section 11.4 discusses a generalization of log-aesthetic curves to space. Finally, Section 11.5 considers the possibility of applying similar techniques to the problem of interpolating surfaces, a much more challenging problem.

11.1 Curves in space

In Cartesian coordinates, the representation of a space curve simply adds a z coordinate to the x and y of plane curves. The intrinsic representation also adds an additional dimension. While a plane curve is characterized by its curvature, a space curve is characterized by curvature and torsion. The plane curve with constant curvature is a circle, and the corresponding space curve with constant curvature and torsion is the helix.

An excellent introduction to the differential geometry of space curves is Chapter 2 of Kreyszig’s *Differential Geometry* [48]. This section simply restates some of the elementary properties of space curves, for convenience.

The *moving trihedron* (also known as the Frenet frame) consists of three orthogonal vectors. For a curve defined as a vector \mathbf{x} as a function of arc length s , the *tangent vector* is $\mathbf{t} = d\mathbf{x}/ds$, the *principal normal* is defined as

$$\mathbf{p} = \frac{d\mathbf{t}/ds}{|d\mathbf{t}/ds|} = \frac{1}{\kappa} \frac{d^2\mathbf{x}}{ds^2}. \quad (11.1)$$

The tangent and principal normal are orthogonal, and both lie in the plane of the osculating circle. The third component of the moving trihedron is the *binormal*, and is orthogonal to the osculating circle.

$$\mathbf{b} = \mathbf{t} \times \mathbf{p}. \quad (11.2)$$

Torsion is the rate of change of the direction of the binormal vector.

$$\tau = -\mathbf{p} \cdot \frac{d\mathbf{b}}{ds} . \quad (11.3)$$

An important result is that a curve is planar if and only if torsion is everywhere zero.

The concept of geometric continuity is similar in space as in the plane. In particular, a space curve has G^2 -continuity if its osculating circle is continuous along its arc length. Torsion need not be continuous in a G^2 -continuous space curve.

The bending energy of a wire is, as in the planar case, proportional to the integral of the square of curvature. Note that the bending energy does not directly depend on torsion. A physical wire may also have *twist*, but this is not considered here; for spline applications, it may be assumed that the wire can untwist to find its minimum energy configuration.

11.2 Minimum energy curve in space

Likely the first actual solution to the shape of the MEC in space was by Kirchoff, who found that the kinetic analogy for the planar elastica (see Section 5.11) also generalizes to three-space fairly straightforwardly. While the planar elastica corresponds to a pendulum rotating around a fixed point, constrained to a plane, the minimum energy curve in space corresponds to a spinning top, also gyrating around a fixed point. Goss, in his Ph.D. thesis, traces out the history of this development [29, p. 37]. The result is also presented in Greenhill's 1892 book on the applications of elliptic curves [33]. Max Born's Ph.D. thesis also contains a derivation of the differential equations of the curve from variational principles [11], but his choice of spherical polar coordinates makes the equations fairly awkward.

A more modern derivation of the shape of the MEC in space is contained in Mehlum's work on nonlinear splines, published in 1974 [59] and later refined [61, 60]. Most importantly, Mehlum discovered a simple relationship between curvature and torsion for all solutions of the elastica [59, Eq. 2.21].

$$\kappa^2 \tau = C \quad (11.4)$$

Mehlum also derived the following equation for the curvature as a function of arc length. Since a space curve is completely characterized by its curvature and torsion, given the above relation, the shape is determined by a differential equation in one variable.

$$[(\kappa^2)']^2 + \kappa^2 [(\kappa^2 - 2D)^2 - 4\psi^2] + 4C^2 = 0, \quad (11.5)$$

where C , D , and ψ are arbitrary constants [59, Eq. 2.24]. Compare with Equation 3.10, the equation for the elastica in the plane.

11.2.1 MEC as spline in 3-space

When used as a spline, the MEC has the following properties:

- Each segment is an energy-minimizing space curve, constrained by the *tangent vectors* relative to the chord at both endpoints.
- G^2 -continuity is preserved across control points, as in the planar case. However, torsion may be discontinuous.
- Each curve segment is controlled by three parameters. The additional parameter corresponds to the deviation of the two tangent vectors from being planar.

Each tangent vector has two components (equivalent, for example, to the spherical polar coordinates θ and ϕ). This would seem to suggest that there are *four* parameters for each curve segment. However, counting in this way would include one parameter for rotations around the chord, which, as a rigid body transformation, is not actually a true parameter of the underlying curve family. Thus, the actual number of parameters is three. The *relative* rotation around the chord is related to the overall torsion on the curve; when this rotation is zero, the torsion vanishes and the curve becomes the planar MEC.

Of course, the MEC in three-space has the same drawbacks as the planar MEC, including the lack of roundness and existence. Thus, a generalization of the Euler spiral is also worth seeking. As of now, there are two viable proposals for such a generalization: Mehlmum's spiral, discussed in Section 11.3, and a generalization of log-aesthetic curves to space, discussed in Section 11.4.

Except in the planar case, a space MEC never has an actual inflection point, i.e. a point in which curvature is zero. In the nearly-planar case, the curvature drops very low, and torsion peaks in this region, so that the plane of the osculating circle swings through nearly 180° .

As for higher degrees of continuity, Moreton [64] described a reasonable (but computationally intensive) techniques for computing a good approximation to the MVC, providing G^4 continuity. Likely, either an exact solution based on variational techniques, or a good approximation, would yield similar results with much less computational cost.

11.3 Mehlmum's spiral

One beautiful proposal for generalizing the Euler spiral into space is by Even Mehlmum [59]. It can be understood both as an approximation to the space MEC (entirely analogous to the way in which the Euler spiral is an approximation to the planar elastica), and as a geometric construction in which the Euler spiral is rolled up onto the surface of a sphere.

To derive the curve from the space MEC, assume that $\kappa^2 \ll 2D$. Then, Equation 11.5 simplifies to

$$[(\kappa^2)']^2 - 4\alpha^2\kappa^2 + 4C^2 = 0, \quad (11.6)$$

where α is a constant (it is equal to $\psi^2 - D^2$, combining these two constants into one). This equation, however, has a simple analytic solution,

$$\kappa^2 = \alpha^2 s^2 + \frac{C^2}{\alpha^2}. \quad (11.7)$$

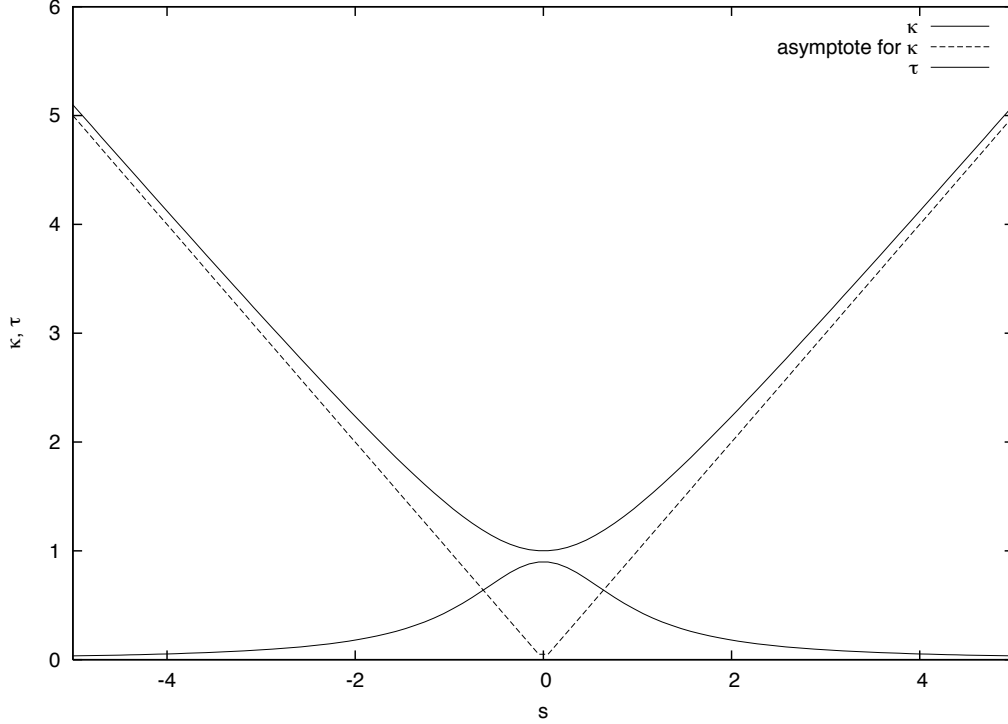


Figure 11.2. Curvature as a function of arc length in the Mehlum spiral.

When $C = 0$, this equation simplifies to $\kappa = \alpha s$, the equation of the Euler spiral. In general, it is the equation of a hyperbola, as shown in Figure 11.2.

A space curve requires both curvature and torsion. For the latter, simply retain Equation 11.4 from the planar MEC. Thus, these equations fully define a family of curves suitable as the basis for an interpolating spline. This family was used as the basis of the KURGLA 2 interpolating spline algorithm in the Autokon system [59].

Further analysis by Mehlum and Wimp [61] revealed a surprising geometric property of this curve – it lies entirely on the surface of a sphere. That analysis is based on a well-known intrinsic equation relating curvature and torsion, as a necessary and sufficient condition for a space curve to lie on a sphere [61, Eq. 2.8].

$$\frac{\tau}{\kappa} - \left(\frac{\kappa'}{\tau \kappa^2} \right)' = 0 \quad (11.8)$$

For values of curvature following the function described in Equation 11.7, this relation (after a bit of algebraic manipulation) is shown to be equivalent to Equation 11.4.

Even more recently, Mehlum discovered a purely geometric construction of this space curve (credited jointly with Professor Ronald Resch), based on rolling up an Euler spiral onto the surface of a sphere [60]. He writes,

Draw a Cornu spiral with wet ink in the xz plane. The curvature of the spiral is specified by $\kappa = \alpha s$. Place a sphere with radius $R = \frac{\alpha}{C}$ at the xz plane with the tangent point at the origin. Roll the sphere along the Cornu spiral such that the axis of rotation always goes through the center of the sphere and that point on the evolute of the spiral that corresponds to the touching point of the sphere at the spiral. The sphere is always tangent to the xz plane during the rotation.

Mehlum's 1994 paper [60] also contains, in an impressive feat of analysis, closed-form formulas for the Cartesian coordinates of the space curve as a function of arc length, using confluent hypergeometric functions.

This curve seems a very good choice for an interpolating space curve, both because of its excellent practical properties (generalized from those of the Euler spiral) and for its mathematical beauty. I know of no other implementations than the original Autokon system.

11.4 Log-aesthetic space curves

Another fruitful avenue for exploration is the generalization of log-aesthetic curves to space [91]. Recall that the planar log-aesthetic curve is defined as a straight-line plot of curvature vs (scaled) derivative of curvature when both are plotted on a logarithmic scale. The space log-aesthetic curve applies the analogous constraint to the torsion as well. Just as the circle is a special case of the planar log-aesthetic curve, the helix is a special case of the corresponding space curve. There are a number of additional parameters of this curve family, so determining these parameters remains an interesting open problem.

The parameter space admits many different types of curves. In particular, some have both curvature and torsion increasing as a function of arc length. These curves generally resemble a corkscrew. One of Mehlum's main results is that, in a space curve resulting from the physical bending of a wire, the curvature and torsion vary oppositely (see Equation 11.4). In regions of high curvature, torsion is low, and vice versa. Perhaps this heuristic could be applied to curves from the log-aesthetic family to reduce the dimensionality of the parameter space while preserving desirable solutions.

11.5 Surfaces

While well beyond the scope of this thesis, some of the concepts described here also generalize to surfaces. Moreton's work [64] is one of the earliest presentations of an interpolating surface defined in terms of minimizing an energy functional. These techniques are known to be extremely computationally intensive. Furthermore, the choice of energy functional to minimize is not obvious (just as in the planar case). Among the proposals are MES (Minimum Energy Surface), analogous to the planar MEC, and MVS (Minimum Variation Surface), analogous to the planar MVC.

One proposal for such an energy functional to minimize is the MVS_{cross} functional of Joshi and Séquin [41]. For the input configurations tested, the surface minimizing this functional is more balanced and aesthetically pleasing than the pure MES and MVS functionals.

A very intriguing direction to explore is whether surface segments may be drawn from a parametrized family in the same way that curve segments are drawn in the planar and space-curve cases. Such an approach holds the promise of much faster computation than iteratively evolving a surface to reduce its energy functional. Candidates for such surfaces include a family of patches that minimize an energy functional (analogous to the MEC), generalizations of the log-aesthetic curve family to surfaces, and quasi-static liquids [83], which are defined as energy minimizing in the presence of certain additional constraints.

Chapter 12

Conclusion

Interpolating splines are a useful and important tool for designing curves. However, choosing one from the multiplicity in the literature is not easy. While fairness is obviously a crucially important property, there has been no universal consensus on what makes an ideal interpolating spline, or even how fairness should be measured. In this thesis, I have argued that an “ideal” spline must share two properties with the Minimum Energy Curve, the mathematical idealization of the mechanical spline based on a thin strip of elastic material. First, it must be extensional, meaning that adding an additional control point on the curve does not change the shape of the curve. Second, the spline should be based on a two-parameter family of primitive curves, joined together at control points with G^2 -continuity. Higher degrees of continuity (also corresponding to more parameters for the primitive curve) are possible, but come at the cost of poorer locality.

Based on evidence from existing splines in the literature, two parameter splines have better locality than splines defined from a larger parameter space. Two-parameter splines can easily achieve G^2 -continuity, and no higher degree of continuity is required for a planar curve to appear smooth to the human visual system.

In fact, there is in general a tradeoff between fairness and locality. A degenerate spline consisting of straight lines has perfect locality but terrible fairness. More usefully, the interpolating spline based on the log-aesthetic curve covers a range of this tradeoff, based on a scalar parameter (the exponent of the relation between curvature and arc length).

A significant result of this thesis is that all two-parameter, extensible splines can be defined in terms of a generator curve, presented in Section 4.8. As a result, for this important subclass of G^2 -continuous splines, the design space is reduced to choosing the best curve, both in terms of the aesthetic beauty of the curve itself and also the properties of the corresponding spline.

The Euler spiral is an excellent choice for this generator curve. In fact, the Euler spiral is better in nearly every respect than the MEC, the only exception being minimization of the MEC energy functional itself. However, as demonstrated by human experiment (see Section 4.8.1), MEC energy is not a good predictor of perceived fairness. The Euler spiral spline has identical locality properties, and, unlike the MEC, is round (producing a circular arc when the control points are co-circular), and a solution nearly always exists.

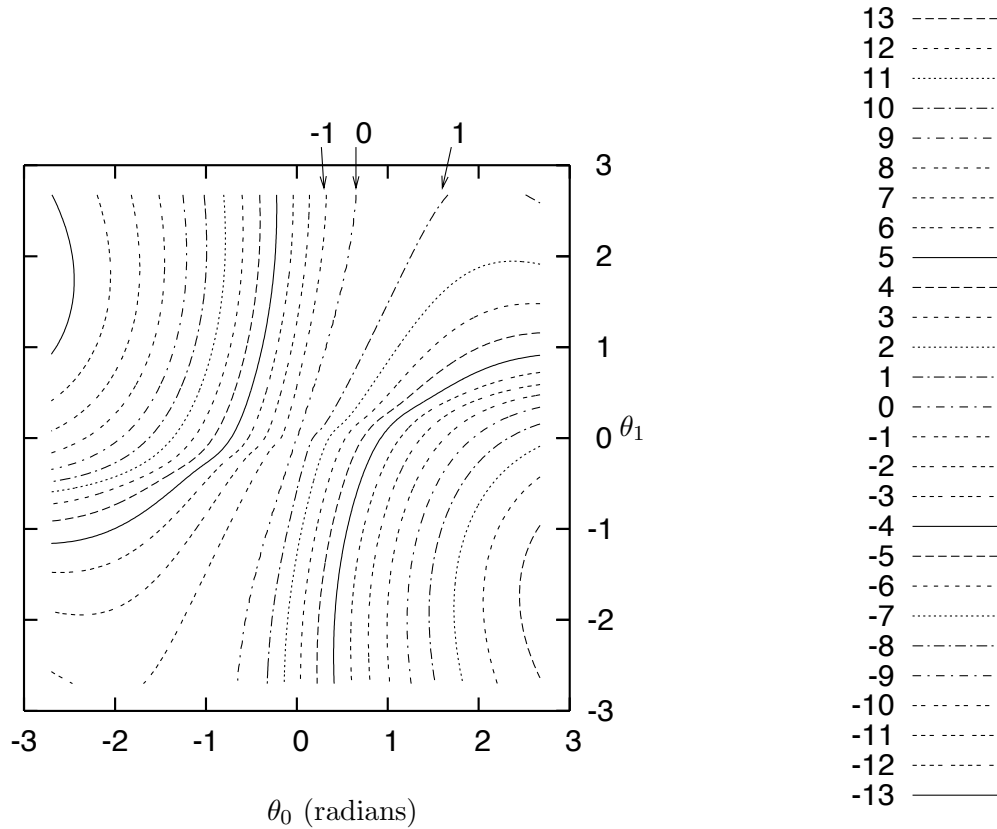


Figure 12.1. Contour plot of curvature lookup table for $\alpha = 2$ log-aesthetic curve.

One promising candidate for the generating curve is the family of log-aesthetic curves, parameterized by the exponent α . At $\alpha = 1$, the Euler spiral is one instance of this family. Higher values of α exhibit better locality, but overly large values have poor perceived fairness. However, as measured by empirical study, values of α between 1.5 and 2 have both better locality and better fairness than the Euler spiral. Based on these properties, it would seem like the log-aesthetic spline with a higher exponent would be an even better interpolating spline than that based on the Euler spiral. Indeed, for solving a spline based on a static sequence of control points, the log-aesthetic spline tends to produce good results.

However, in interactive use, a weakness of the log-aesthetic spline emerges. For many configurations of input points, a small perturbation to the input causes a fairly significant change to the curve, even when the angles are small. One way to analyze this phenomenon is to look at the relation between tangent angles and curvature. While it is very nearly linear for the Euler spiral (see Figure 8.4), the log-aesthetic plot has significant nonlinearities, even for very small angles, as shown in Figure 12.1.

An excellent way to achieve better fairness (albeit at the expense of locality) is to require higher degrees of geometric continuity. In the variational framework, the Minimum Variation Curve (MVC) is a four-parameter spline with G^4 -continuity. A strong motivation was to fix the roundness problem

of the MEC. While previous researchers have used numerical techniques to minimize the functional (especially Moreton [64]), one contribution of this thesis is to develop the exact general equation for the MVC. Happily, the MVC has a very simple formula relating curvature with Cartesian coordinates: $\kappa'' = \lambda y$, which also has an appealing parallelism with the corresponding equation for the MEC, $\kappa = \lambda y$.

The MVC has a simple approximation, writing curvature as a cubic polynomial in terms of arc length. It is entirely analogous to the Euler spiral spline, which can be seen as a simpler, more robust approximation to the MEC. This curve admits very efficient numerical techniques (presented in Section 8.1), for fast and robust computation. Interestingly, the cubic polynomial spiral spline was first described by Ohlin in 1987 [67], but with little impact at the time, probably because Ohlin did not adequately characterize its properties.

The G^4 -continuous polynomial spiral spline (or “spiro” for short) is quite practical, and has been used to draw many font outlines. Because of its poorer locality properties relative to the Euler spiral spline, it requires more points to adequately describe letter-shapes without exhibiting the ripple effect, but only about a factor of two, based on experience drawing multiple fonts (described in detail in Chapter 10). Because G^2 -continuity is sufficient for the perception of fair curves, I have taken to using primarily the Euler spiral spline for my later font work.

For pure interpolating splines, the tradeoff between fairness and locality seems somewhat un-compromising. In practice, the biggest problem is that regions of high curvature causes ripples, analogous to “ringing” in digital filters. Fortunately, adding explicit one-way constraints to the control points isolates these regions of high curvature and prevents ripples entirely. The implementation of these constraints uses the same four-parameter curve primitive as used to generate splines with G^4 -continuity. Instead of enforcing continuity of higher derivatives of curvature, these “one-way” (asymmetrical) constraints enforce zero derivatives of curvature on one side only of the control point, while still preserving G^2 -continuity everywhere. In practice, these constraints are especially useful for making smooth transitions from straight to curved sections, a common configuration in letter-shapes.

A major motivation of this thesis was to explore the idea of the best possible interpolating spline, regardless of computational complexity, based on the idea that Moore’s Law would ultimately make even computationally extremely expensive splines practical. However, work on efficient implementation proved very fruitful, and the numerical techniques developed in this thesis are competitive with cruder, polynomial-based splines.

Especially for font design, it is important to represent the resulting shapes in a format that can be widely used. Because the industry-standard font formats are all based on Bézier curves, I developed algorithms for converting curves into this form, producing curves that are visually nearly identical to the original spline curves (based on a carefully designed error metric), while using an optimally small number of segments, to keep file size for the resulting fonts small.

A significant focus of this work was to quantify the properties of the various interpolating splines explored in this thesis, to facilitate choosing the best one for any given application. In using these splines for drawing a number of fonts, I also found they brought an aesthetic quality to the work. Designing on the screen (rather than starting on paper), I found that I could readily achieve the smooth, classical shapes I desired. Knowing that the curves have a strong mathematical foundation added to this aesthetic sense, as did the fact that the numerical techniques were based on rapidly computing nearly exact analytical solutions, rather than merely crude approximations. It is my

hope that in publishing this work and releasing the code under an open-source license, a wider audience of designers will experience the pleasure of working with these superior tools.

Bibliography

- [1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth Dover printing, tenth GPO printing edition, 1964.
- [2] Adobe Systems Inc. *Adobe Type 1 Font Format*, 1990.
- [3] Apple Computer Inc. *TrueType Reference Manual*, October 1996.
<http://developer.apple.com/textfonts/ttrefman/index.html>.
- [4] Raymond Clare Archibald. Euler integrals and Euler’s spiral. *American Mathematical Monthly*, 25(6):276–282, June 1917.
- [5] Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.
- [6] Daniel Bernoulli. The 26th letter to Euler. In *Correspondence Mathématique et Physique*, volume 2. P. H. Fuss, October 1742.
- [7] James Bernoulli. Quadratura curvae, e cujus evolutione describitur inflexae laminae curvatura. In *Die Werke von Jakob Bernoulli*, pages 223–227. Birkhäuser, 1692. Med. CLXX; Ref. UB: L Ia 3, p 211–212.
- [8] James Bernoulli. *Jacobi Bernoulli, Basiliensis, Opera*, volume 1. Cramer & Philibert, Geneva, 1744.
- [9] James Bernoulli. *Jacobi Bernoulli, Basiliensis, Opera*, volume 2. Cramer & Philibert, Geneva, 1744.
- [10] Garrett Birkhoff and Carl R. de Boor. Piecewise polynomial interpolation and approximation. *Proc. General Motors Symp. of 1964*, pages 164–190, 1965.
- [11] Max Born. *Untersuchungen über die Stabilität der elastischen Linie in Ebene und Raum, under verschiedenen Grenzbedingungen*. PhD thesis, University of Göttingen, 1906.
- [12] Max Born. *My Life and Views*. Scribner, 1968.
- [13] Alexander Brook, Alfred M. Bruckstein, and Ron Kimmel. On similarity-invariant fairness measures. In *Scale Space and PDE Methods in Computer Vision*, number 3459 in LNCS, pages 456–467. Springer, 2005.
- [14] Ian D. Coope. Curve interpolation with nonlinear spiral splines. *IMA Journal of Numerical Analysis*, 13(3):327–341, 1993.

- [15] Marie Alfred Cornu. Méthode nouvelle pour la discussion des problèmes de diffraction dans le cas d'une onde cylindrique. *Journal de Physique théorique et appliquée*, pages 5–15, 1874.
- [16] Charles Lee Crandall. *The Transition Curve*. Wiley, 1893.
- [17] Lawrence D'Antonio. The fabric of the universe is most perfect: Euler's research on elastic curves. In *Euler at 300: an appreciation*, pages 239–260. Mathematical Association of America, 2007.
- [18] John A. Edwards. Exact equations of the nonlinear spline. *Trans. Mathematical Software*, 18(2):174–192, June 1992.
- [19] L. Elsgolts. *Differential Equations and the Calculus of Variations*. Mir Publishers, Moscow, 1970.
- [20] Leonhard Euler. *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici lattissimo sensu accepti*, chapter Additamentum 1. eulerarchive.org E065, 1744.
- [21] Leonhard Euler. On the values of integrals extended from the variable term $x = 0$ up to $x = \infty$, 2007. Translation into English by Jordan Bell.
- [22] G. Forsythe. Notes taken at the GM Research Laboratories Symposium on approximation of functions. Special collection SC98 2–6, Stanford University, 1964.
- [23] A. H. Fowler and C. W. Wilson. Cubic spline, a curve fitting routine. Technical Report Y-1400, Oak Ridge Natl. Laboratory, September 1962.
- [24] Craig G. Fraser. Mathematical technique and physical conception in Euler's investigation of the elastica. *Centaurus*, 34(3):211–246, 1991.
- [25] Augustin Jean Fresnel. Memoir on the diffraction of light. In Henry Crew, editor, *The Wave Theory of Light*, pages 79–144. American Book Company, 1900.
- [26] J. M. Glass. Smooth curve interpolation: A generalized spline-fit procedure. *BIT*, 6(4):277–293, 1966.
- [27] James Glover. Transition curves for railways. In *Minutes of Proceedings of the Institution of Civil Engineers*, volume 140, pages 161–179, 1900.
- [28] Lawrence E. Goodman and William H. Warner. *Statics*. Wadsworth, 1963.
- [29] Victor Geoffrey Alan Goss. *Snap buckling, writhing and loop formation in twisted rods*. PhD thesis, University of London, London, 2003.
- [30] Victor Geoffrey Alan Goss. The history of the planar elastica: Insights into mechanics and scientific method. *Science & Education*, 2008.
- [31] I. Grattan-Guinness. *Convolutions in French Mathematics, 1800–1840*. Birkhäuser, 1990.
- [32] Alfred Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica*, chapter 3.5, Clothoids, pages 64–66. CRC Press, Boca Raton, FL, 2nd edition, 1997.
- [33] Alfred George Greenhill. *The Applications of Elliptic Functions*. Macmillan, London, 1892.

- [34] Stefan Gumhold. Designing optimal curves in 2D. In *Proc. CEIG*, pages 61–76, Sevilla/Spain, July 2004.
- [35] Peter M. Harman and Alan E. Shapiro, editors. *The critical role of curvature in Newton’s developing dynamics*. Cambridge University Press, 2002.
- [36] Arthur Lovat Higgins. *The Transition Spiral and Its Introduction to Railway Curves*. Van Nostrand, 1922.
- [37] John D. Hobby. Smooth, easy to compute interpolating splines. Technical Report STAN-CS-85-1047, Stanford University, 1985.
- [38] B. K. P. Horn. The curve of least energy. Technical Report A.I. Memo. No. 612, MIT AI Lab, January 1981.
- [39] Josef Hoschek, Dieter Lasser, and Larry L. Schumaker. *Fundamentals of Computer-aided Geometric Design*. AK Peters, 1993.
- [40] Robert Alexander Houstoun. *A Treatise on Light*. Longmans, Green and Co., 1915.
- [41] Pushkar Joshi and Carlo Séquin. Energy minimizers for curvature-based surface functionals. *CAD And Applications*, 4(5):607–617, 2007.
- [42] Peter Karow. *Digital Formats for Typefaces*. URW Verlag, April 1987.
- [43] Peter Karow. Digital punch cutting. *Electronic Publishing*, 4(3):151–170, September 1991.
- [44] Benjamin B. Kimia, Ilana Frankel, and Ana-Maria Popescu. Euler spiral for shape completion. *Int. J. Comput. Vision*, 54(1-3):157–180, 2003.
- [45] Louis T. Klauder. A better way to design railroad transition spirals. *ASCE Journal of Transportation Engineering*, May 2001. Date is date submitted.
- [46] Donald E. Knuth. *Computer Modern Typefaces*, volume E of *Computers and Typesetting*. Addison-Wesley, Reading, MA, USA, 1986.
- [47] Donald Ervin Knuth. *Digital typography*, volume 78 of *CSLI lecture notes*. CSLI Publications, Stanford, Calif., 1999.
- [48] Erwin Kreyszig. *Differential Geometry*. Dover Publications, 1991.
- [49] Boris I. Kvasov. *Methods of shape-preserving spline approximation*. World Scientific, 2000.
- [50] Pierre Simon Laplace. *Œuvres complètes de Laplace*, volume 4. Gauthier-Villars, 1880.
- [51] Erastus H. Lee and George E. Forsythe. Variational study of nonlinear spline curves. *SIAM Rev.*, 15(1):120–133, 1973.
- [52] A. E. H. Love. *A Treatise on the Mathematical Theory of Elasticity*. Dover Publications, fourth edition, 1944.
- [53] D. H. MacLaren. Formulas for fitting a spline curve through a set of points. Technical Report 2, Boeing Appl. Math. Report, 1958.

- [54] Michael A. Malcolm. On the computation of nonlinear spline functions. *SIAM J. Numer. Anal.*, 14(2):254–282, April 1977.
- [55] Giovanni Mardersteig. *The Officina Bodoni: An Account of the Work of a Hand Press, 1923-1977*. Edizioni Valdonega, 1980.
- [56] Dan Margalit. History of curvature.
http://www3.villanova.edu/maple/misc/history_of_curvature/k.htm, 2003.
- [57] James Clerk Maxwell. Capillary action. In *Encyclopædia Britannica*, pages 256–275. Henry G. Allen, 9th edition, 1890.
- [58] Mac McGrew. *American Metal Typefaces of the Twentieth Century*. Oak Knoll Books, 2nd, revised edition, 1993.
- [59] Even Mehlum. Nonlinear splines. *Computer Aided Geometric Design*, pages 173–205, 1974.
- [60] Even Mehlum. Appell and the apple (nonlinear splines in space). In M. Dæhlen, T. Lyche, and L. L. Schumaker, editors, *Mathematical Methods for Curves and Surfaces: Ulvik, Norway*. Vanderbilt University Press, 1994.
- [61] Even Mehlum and Jet Wimp. Spherical curves and quadratic relationships for special functions. *J. Austral. Math Soc. Ser. B*, 27(1):111–124, 1985.
- [62] William Hallows Miller. *The Elements of Hydrostatics and Hydrodynamics*. J. & J. J. Deighton, 1831.
- [63] Kenjiro T Miura, J Sone, A Yamashita, and Toru Kaneko. Derivation of a general formula of aesthetic curves. In *In: Proceedings of the Eighth International Conference on Humans and Computers*, pages 166–171, 2005.
- [64] Henry Moreton. *Minimum Curvature Variation Curves, Networks and Surfaces for Fair Free-Form Shape Design*. PhD thesis, University of California, Berkeley, Berkeley, California, 1992.
- [65] Henry P. Moreton and Carlo H. Séquin. Scale-invariant minimum-cost curves: Fair and robust design implements. *Computer Graphics Forum*, 12:473–484, 1993.
- [66] Stephen L. Moshier. *Methods and Programs for Mathematical Functions*. Prentice Hall, 1989.
- [67] S. C. Ohlin. Splines for engineers. In Guy Marechal, editor, *Eurographics '87*, pages 555–565, 1987.
- [68] W. A. Oldfather, C. A. Ellis, and Donald M. Brown. Leonhard Euler’s elastic curves. *Isis*, 20(1):72–160, November 1933.
- [69] OpenType specification, version 1.6 draft.
<http://www.microsoft.com/typography/otspec/>, April 2009.
- [70] Adobe Press. *PostScript Language Reference Manual*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1985.
- [71] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

- [72] Thomas Preston and Charles Jasper Joly. *The Theory of Light*. Macmillan, 1901.
- [73] William John Macquorn Rankine. *A Manual of Civil Engineering*. Charles Griffin, 17th edition, 1889.
- [74] Terry Ross. More about curve maker.
<http://www.drawmetal.com/moreaboutcurvemaker>.
- [75] John Roulier and Thomas Rando. Measures of fairness for curves and surfaces. In Nickolas S. Sapidis, editor, *Designing Fair Curves and Surfaces: Shape Quality in Geometric Modeling and Computer-Aided Design*, chapter 5. SIAM, 1994.
- [76] Louis Saalschütz. *Der belastete Stab unter Einwirkung einer seitlichen Kraft*. B. G. Teubner, Leipzig, 1880.
- [77] Javier Sánchez-Reyes and Jesús Miguel Chacón. Polynomial approximation to clothoids via s-power series. *Computer-Aided Design*, 35(14):1305–1313, 2003.
- [78] I. J. Schoenberg. *Selected Papers*, volume I. Birkhäuser, Boston, 1988. edited by Carl de Boor.
- [79] Peter Selinger. Potrace: a polygon-based tracing algorithm.
<http://potrace.sourceforge.net/potrace.pdf>, September 2003.
- [80] Carlo H. Séquin, Kiha Lee, and Jane Yen. Fair, G2- and C2-continuous circle splines for the interpolation of sparse data points. *Computer-Aided Design*, 37(2):201–211, 2005.
- [81] R Sridharan. Physics to mathematics: from lintearia to lemniscate – I. *Resonance*, pages 21–29, April 2004.
- [82] Josef Stoer. Curve fitting with clothoidal splines. *J. Res. Nat. Bur. Standards*, 87(4):317–346, 1982.
- [83] Kurt W. Swanson, Kenneth A. Brakke, and David E. Breen. Physics-based surface modeling using quasi-static liquids. *CAD And Applications*, 6(6):759–779, 2007.
- [84] Arthur N. Talbot. *The Railway Transition Spiral*. U. Illinois, 1899. Reprinted from The Technograph No. 13.
- [85] Isaac Todhunter. *A History of the Theory of Elasticity and of the Strength of Materials*, volume 1. Cambridge University Press, 1886.
- [86] C. Truesdell. *The Rational Mechanics of Flexible or Elastic Bodies: 1638–1788*. Birkhäuser, 1960. Introduction to Leonhardi Euleri Opera Omnia Vol X et XI, Series 2.
- [87] C. Truesdell. The influence of elasticity on analysis: The classic heritage. *Bull. AMS*, 9(3):293–310, November 1983.
- [88] C. Truesdell. *Der Briefwechsel von Jacob Bernoulli*, chapter Mechanics, especially Elasticity, in the Correspondence of Jacob Bernoulli with Leibniz. Birkhäuser, 1987.
- [89] William Whewell. *Analytical Statics: A supplement to the fourth edition of an elementary treatise on Mechanics*. J. & J. J. Deighton, Cambridge, England, 1833.
- [90] C. H. Woodford. Smooth curve interpolation. *BIT*, 9:69–77, 1969.
- [91] Norimasa Yoshida, Ryo Fukuda, and Takafumi Saito. Log-aesthetic space curve segments. In *SIAM/ACM Joint Conference on Geometric and Physical Modeling*, 2009.